Distributional effects of payment for performance in the health sector

Examining effects on structural quality, performance outcomes and service utilisation in Tanzania

Peter John Binyaruka

Thesis for the Degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2018

UNIVERSITY OF BERGEN

Distributional effects of payment for performance in the health sector

Examining effects on structural quality, performance outcomes and service utilisation in Tanzania

Peter John Binyaruka

Thesis for the Degree of Philosophiae Doctor (PhD)
at the University of Bergen

2018

Date of defence: 07.11.2018

Year: 2018

Title: Distributional effects of payment for performance in the health sector

Name: Peter John Binyaruka

Print: Skipnes Kommunikasjon / University of Bergen

Distributional effects of payment for performance in the health sector

Examining effects on structural quality, performance outcomes and service utilisation in Tanzania

Peter John Binyaruka

Thesis for the degree philosophiae doctor (PhD)

at the University of Bergen

2018

Date of defence: 7th November 2018

Scientific environment

This PhD research is a product of collaboration between the University of Bergen (UiB) and other partner institutions, which includes the Ifakara Health Institute (IHI), Chr. Michelsen Institute (CMI) and the London School of Hygiene and Tropical Medicine (LSHTM). At UiB, I was admitted as a Masters' degree student in International Health in August 2014; and as a doctoral candidate in August 2015 at the Department of Global Public Health and Primary Care, Faculty of Medicine and Dentistry. I was a member of the Research Group for Global Health Priorities at UiB, and Research Group for Global Health at CMI. I also utilised a conducive research environment at CMI, where I was given an office space to successfully write my PhD thesis. The moments I was in Tanzania, all my research activities and write-ups were done at the IHI office.

The research was conducted at IHI with the impact evaluation thematic group, and at UiB with the CIH. The field work for data collection was done by the IHI under my supervision and coordination. The data collection was part of the large project, which conducted an impact evaluation of the payment for performance (P4P) programme. The P4P programme was implemented by the Government of Tanzania and funded by the Government of Norway. It was through the P4P evaluation project, I developed the research questions for my PhD research.

My PhD training classes were obtained from various institutions including the University of Bergen at CIH in Norway; the Norwegian School of Economics (NHH) in Norway; University of Lausanne (Switzerland); University College London (UCL) in the United Kingdom; and University of Heidelberg (Germany). The dissemination of my PhD outputs has been done in three international conference: at the 11th International Health Economics Association (IHEA) congress in Milan-Italy in 2015; at the 12th IHEA congress in Boston-United States in 2017; and at the 10th GLOBVAC conference in Trondheim-Norway in 2017.

Professor Gaute Torsvik from the CMI and University of Oslo, was my main supervisor. My co-supervisors were Professor Bjarne Robberstad from the CIH at UiB, and Dr Josephine Borghi from the LSHTM in the United Kingdom.

My PhD studies at the University of Bergen was funded by the Norwegian State Education Loan Fund (Statens Lånekassen) through the Quota programme.

Dedication

To my parents: John Binyaruka and Elizabeth Nyango.
I also dedicate this work to my brother Phillip and my sister Regina.

Acknowledgements

First and foremost, I thank God for blessing me with this opportunity and for giving me strength, motivation, endurance, commitment and drive to pursue and complete my PhD studies. In the journey of my PhD studies several people have been involved in many ways. I would like to express my heartfelt thanks to you all. As I cannot mention each and every one, those who may not be acknowledged please accept my sincere apologies.

I extend my heartfelt thanks to my wonderful supervisors: Professors Gaute Torsvik (main supervisor), Bjarne Robberstad (co-supervisor), and Josephine Borghi (co-supervisor). Your tireless support, guidance, contributions, constructive comments, inspiration and insightful feedback have given me confidence and strength to write and present all components of my PhD research. I wish to express my appreciation to Gaute Torsvik, my main supervisor, for your constructive critiques especially in the theoretical perspectives and logical reasoning. Bjarne Robberstad, I am very grateful for your close support and immediate advice on training, research and administrative issues at the University of Bergen, *asante sana!* I sincerely acknowledge the support, guidance, courage, and motivation provided by Josephine Borghi. You recruiting me as a young researcher after my Masters' degree, and you have ever continued to inspire and mentor my career development in research. Your mentorship in doing researches, analyses, write-ups, presentation, will remain valuable in my career. I also recognise the support and guidance given by Dr Ottar Mæstad, who leads our Tanzanian P4P evaluation project, and through this project my PhD sub-studies were embedded. I further thank Josephine Borghi, Ottar Mæstad and Gemini Mtei for your sincere recommendation for me to pursue my PhD in Bergen, Norway.

I sincerely appreciate the Norwegian government through the Quota scholarship programme for the financial support to pursue my PhD studies in Norway (2014–2018). I appreciate the opportunity to enrol and study at the University of Bergen, through the Centre for International Health (CIH). I acknowledge the valuable financial support on research write-ups and dissemination that were given by the UK-DFID through the RESYST consortium (www.resyst.lshtm.ac.uk) under the leadership of Professor Kara Hanson. My thanks also go to the Chr. Michelsen Institute (CMI), under the leadership of Dr Ottar Mæstad, for providing an office space that was suitable for doing my PhD research. I thank my home institution, Ifakara Health Institute (IHI), for allowing me to pursue my PhD studies and remain partly employed on various projects.

I would like to thank all staff at CIH for providing an excellent environment for my PhD studies. Many thanks to CIH director, Professor Bente Moen, and all other Professors/ lecturers at CIH. I similarly thank all staff at CMI for being friendly during my stay. *Tusen takk!* I thank the administrative staff at CIH and CMI for their prompt responses and support during my study period: Linda Forshaw, Ingvild Hope, Gunhild Koldal, Borgny Lavik, Solfrid Vikøren, Guri Stegali and Merete Leby. You are a great team!

Special thanks also go to my friends and classmates in Bergen, this PhD journey would have been challenging without your company. My fellow Tanzanians “Uni-Bergen”: Amani, George, Frida, Peter Nestory, Israel, Lutango, Patience, Doreen, Angela Mwakimonga, and Abraham Laizer. Thanks to Victor Chimhuntu and your family for hosting me when I first arrived in Bergen, and you continued to support me while in Bergen. I thank my fellow CIH doctoral candidates: Jobiba, Alemayehu, Peter Hangoma; and my Masters’ classmates at CIH: Simon Kibira, Angela Akol, Anselme Sanou, Anna Hammer, Maija Rankinen, Robin Cole, Mohamed Ali, and the late Jane Macha.

I am thankful to the whole P4P evaluation research team, field workers, respondents, and government officials for their valuable cooperation to accomplish the research.

My parents, brother Phillip and sister Regina, thanks for your moral support and prayers. Mulakoze chane!

Abstract

Introduction: Payment for performance (P4P) involves the allocation of financial incentives to health workers and/or facilities for reaching pre-defined performance targets or measures. P4P has been used in high-income countries (HICs) to improve healthcare quality, and recently has been applied in low-and middle-income countries (LMICs) to improve the coverage and quality of health services and strengthen health systems. The available evidence on the effect of P4P is mixed, but with some promising results of improvements in the incentivised indicators. However, most evaluations of P4P have focused on average programme effects on the incentivised services, paying little attention to distributional effects of P4P. Specifically, little is known about the effects of P4P on structural quality of care (e.g. availability of medical commodities), and similarly on the understanding of the heterogeneity of the P4P effects among subgroups of providers and populations. This PhD work aims to fill that knowledge gap. It estimates the effect of P4P on the availability and stock-out of medical commodities, and examine the differential effects of P4P across subgroups of health facilities and populations in Tanzania.

Data sources: The study collected data in intervention and control areas through facility and household surveys, and facility payments data from administrative records. Baseline data were collected in January 2012 with a follow-up 13 months later. Facility survey across 150 facilities (75 facilities from each study arm) included data on the availability and stock-out of medical commodities (drugs, supplies and equipment), and facility characteristics. Household survey across 3000 women who delivered within 12 months prior to the survey (20 women per facility catchment area), and a similar sample size in the follow-up survey, captured information about individual and household characteristics and maternal and child health service utilisation.

Analyses: A difference-in-differences (DID) regression model was used to estimate the average effects of P4P on the availability and stock-out of medical commodities (Paper I). The DID model was further extended by including a three-way interaction term (i.e.

average effect and subgroup indicator) to capture the differential effects of P4P across facilities' subgroups (Paper I and II), and across populations subgroups (Paper III). Assessment of differential effects were based on outcomes which improved significantly due to P4P (i.e. availability of drugs and supplies, institutional delivery rates and uptake/ provision of antimalarial drugs during antenatal care (ANC)). Descriptive measures of inequality were also used to assess the distribution of facility payouts across facilities' subgroups (Paper II).

Results: Paper I reports that P4P improved the availability of drugs and supplies and reduced their stock-out rates, but had no effect on the availability of medical equipment. The improved effects were generally similar across facilities, but relatively higher among facilities serving a poor population and located in rural areas. Paper II finds that facility payments were initially higher among higher level facilities (hospitals and health centres than dispensaries), the better resourced than worse resourced facilities, and facilities serving wealthier than poorer populations, but these inequalities in payouts declined over time. The effect of P4P on institutional delivery rates was greater among facilities with low baseline performance, serving middle wealth populations, located in rural areas, than among their counterparts; whereas the effect on provision of antimalarial drugs was similar across facilities subgroups. Paper III finds that the effect of P4P on institutional deliveries was greater among women in the poorest households, who lived in rural areas and who did not have health insurance than among their counterparts. P4P effect on the uptake of antimalarial drugs was equally distributed across population subgroups.

Conclusion: The study findings suggest that the P4P programme can improve structural quality of care in terms of the availability of medical commodities. It can further enhance more equitable performance among facilities as the worse-off providers improved most in this study. Similarly, P4P can enhance equitable service utilisation since the service use increased mostly among the worse-off populations.

List of abbreviations

ANC	Antenatal Care
CCT	Conditional Cash Transfer
CHF	Community Health Fund
DED	District Executive Directors
DHS	Demographics and Health Survey
DID	Difference-in-Differences
DMO	District Medical Officers
GDP	Gross Domestic Product
HICs	High-Income Countries
HMIS	Health Management Information System
IMR	Infant Mortality Rate
IPT	Intermittent Preventive Treatment
LMICs	Low- and Middle-income Countries
MCH	Maternal and Child Health
MoHSW	Ministry of Health and Social Welfare
MMR	Maternal Mortality Ratio
MSD	Medical Stores Department
PBF	Performance-Based Financing
PHQID	Premier Hospital Quality Incentive Demonstration
P4P	Payment for Performance
QOF	Quality and Outcome Framework
RBF	Results-Based Financing
RMNCH	Reproductive, Maternal, Newborn and Child Health
UHC	Universal Health Coverage
USD	US Dollar
U5MR	Under-5 Mortality Rate
WHO	World Health Organisation

List of papers

1. Binyaruka P, Borghi J: **Improving quality of care through payment for performance: examining effects on the availability and stock-out of essential medical commodities in Tanzania.** *Trop Med Int Health* 2017, 22(1):92-102.

2. Binyaruka P, Robberstad B, Torsvik G, Borghi J: **Does payment for performance increase performance inequalities across health providers? A case study of Tanzania** (Submitted and under review)

3. Binyaruka P, Robberstad B, Torsvik G, Borghi J: **Who benefits from increased service utilisation? Examining the distributional effects of payment for performance in Tanzania.** *Int J Equity Health* 2018, **17**(1):14.

Table of contents

Table of Contents

List of Tables

List of Figures

1.0 Introduction

1.1 Health system challenges

A health system consists of organisations, institutions, resources, people and actions whose primary purpose is to promote, restore or maintain people's health [1]. A health system is made up of six building blocks: service delivery, health workforce, information system, medical supplies and drugs, financing, and governance [2]. A functioning health system is fundamental to people's health. However, most health systems especially in developing countries suffer from various challenges including insufficient resources, limited government accountability, inadequate service delivery, poor information systems, inadequate supply of medicine, and limited technology. The failures within health systems lead to poor health outcomes and persistent inequities in health status [3]. For example, inadequate financial and human resources leads to poor health service quality and quantity and ultimately poor health outcomes [4-6].

In terms of service delivery, having health systems that can deliver quality health services efficiently and equitably are critical for achieving improved health outcomes and financial protection [2]. However, poor service quality and insufficient coverage of life-saving interventions exist in most settings. In low-and middle-income country (LMIC) settings, for example, the performance of health systems is critical as resources for health are much more constrained in these settings. In fact, improving the performance of healthcare delivery systems is an important objective globally [7, 8]. Thus, it is necessary to allocate the available human and financial resources efficiently and equitably while improving the health system performance [9-11].

Most health systems face the problem of a shortage and retention of health workers, especially in poor settings [12-16]. However, it has been shown that it is possible to increase health care supply with the given stock of health workers, since they perform below their capacity on service delivery [17, 18]. Indeed research has shown there is typically a gap between what health workers know how to do and what they actually do for their patients [8, 17, 19, 20]. The gap between knowledge and practice is largely due to low motivation and absenteeism among health care providers, as well as limited resources for health in low-income settings [21]. To avoid wastage of resources or inefficacies in providers' performance, there is an urgent need to improve health worker productivity. The approach of paying providers based on their performance or results has been suggested to redress the concerns of absenteeism, low motivation and poor performance of health workers [17, 19, 22].

1.2 Payment for performance (P4P) strategy

P4P is a financing strategy which involves financial incentives being rewarded to health workers and/ or facilities for reaching pre-specified performance measures or targets related to quality and quantity of health services. P4P involves purchasing of identified set of health services and quality. Purchasing refers to the process by which funds are allocated to healthcare providers to obtain services on behalf of identified groups or the entire population [1, 10, 23]. Purchasing can be done passively or strategically. Passive purchasing implies following a predetermined budget/ simply paying bill when presented, while strategic purchasing involves a continuous search for the best ways to maximise health system performance by deciding which interventions or services should be purchased, how and from whom [1].

In most countries, health care systems have traditionally been financed by paying for inputs (e.g. human resources, drugs, supplies, infrastructure, etc.) and this approach is considered as passive purchasing of health care services [10, 23]. There is, however, an increasing trend in applying an approach of paying for results (e.g. P4P approach) –that pays based on results on various aspects such as service delivery, consultations, service quality and coverage. Thus, P4P is considered as active and strategic purchasing [9, 10, 23]. P4P approach is largely implemented as an additional to input-based financing, and is taken forward due to slow and insufficient progress on health-related *Millennium Development Goals* and *Sustainable Development Goals* especially in LMICs [22, 24-26]. It is also one of the strategies developed to improve performance of healthcare providers [22, 27]. P4P is based on the notion that providing financial incentives to health providers based on performance will motivate them to exert more effort to achieve better outcomes [27-29].

The P4P approach started in high-income countries (HICs), especially in the United States and in the United Kingdom, with the aim of improving healthcare quality [30-32]. To date, P4P has increasingly been used in LMICs to improve coverage and quality of health services, as well as to reform and strengthen health systems [22, 28, 33]. In this way, P4P facilitates the progress to achieve health-related development goals [22, 28, 33].

There are several terminologies referring to paying for results, which are commonly used and sometimes used interchangeably in literature [28, 33, 34]. These include results-based financing (RBF), performance-based financing (PBF), payment for performance (P4P), conditional cash transfer (CCT), cash on delivery and others. The RBF is a broad term which involves a cash payment or non-monetary transfer made to a national or sub-national government, manager, provider, payer or consumer of health services after predefined results have been reached and verified [34]. The RBF strategy incorporates both the demand-side and supply-side incentive programmes. Specifically, the demand-side programmes such as the CCT and voucher schemes, rewards patients for improved behaviour and health outcomes. On the other hand, the supply-side programmes such as PBF and P4P, rewards health workers/ facilities based on their performance. The focus of this thesis is on P4P which encompasses the entire range of incentive approaches on the supply-side.

1.3 P4P theoretical perspective

The economic justification of provider P4P programmes is based on agency theory, i.e. the principal-agent theory [35-37]. This theory describes a principal-agent relationship, which involves the principal (employer or payer) and the agent (employee or service provider). In the health sector, there are multiple principals and multiple agents, including the provider being the agent of the patient/ payer [21]. The principal-agent relationship is directed at the agency relationship, such that the principal delegates a task and authority to the agent who receives a compensation for doing that task. However, this relationship is faced with a problem, the principal-agent problem, that is based on two sources [35, 38]. First, the interests/ preferences of the principal and agent on the goals of the organisation are not perfectly aligned and independent. For instance, service provider (agent) may not perfectly act on behalf of the payer/ patient (principal) [39]. Second, there is an information asymmetry between the principal and agent. Specifically, the principal faces imperfect information about the effort exerted by the provider (agent), principal cannot observe and reward the effort of the agent, and the agent/ provider is risk averse [35-37, 40, 41].

The principal-agent theory therefore recommends linking financial incentives with some performance measures, as for P4P approach, in order to align the agent's objective function with the principal's (i.e. increasing outputs or outcome) [38, 42]. Financial incentives are applied in the principal-agent model based on two assumptions [38]: financial incentives triggers greater motivation to produce the output that the organisation cares for, and greater motivation leads to better performance. In particular, P4P approach addresses the principal-agent problem through incentive contracting, that is, by adding a conditional incentive to the principal-agent contract [43]. The principal attempts to structure the contractual relationship for an agent to perform the desired work by the principal [44], and in a way that the objectives of both principal and agent are fully aligned [40]. As a result, P4P relies on the assumption that the incentives or payments conditional on performance will enhance desired behaviours with few unwanted effects [45]. Despite the focus on financial incentives driven from the principal-agent theory to address the agency problem, there is evidence that other forms of motivation apart from financial incentive (e.g. good leadership and supportive management) also matters [46-48].

1.4 P4P design and structure

The design and structure of P4P varies in many ways across settings. There are several varieties of design features and structures to consider when designing an incentive scheme as described elsewhere [19, 27, 49-52]. These includes the frequency of payment, size of payment, target setting for rewards, incentive based on loss or gain, and individual or group incentives. For the interest of this thesis, one of the design elements, i.e. the target setting or rewarding system, is discussed below along with potential implications for providers' response. P4P schemes can reward using, for example, fee-for-service, relative performance, single absolute threshold targets, multiple thresholds targets, or geographical/ equity targeting [27, 33, 49, 50, 53]. Each approach is described further below.

A ***fee-for-service approach*** involves purchasing from the first service provided/ consultation (e.g. outpatient visit), and at the same value for each subsequent service provided [33, 54]. As this approach leads to more services provision and increases coverage, most P4P schemes in LMICs use a fee-for-service approach that is conditional on quality performance scores [33]. A ***relative performance target***, or tournament approach, involves ranking participants based on their performance and rewarding a share of top performers (e.g. top 10% of providers) [27, 50]. This approach sometimes includes penalties for lower performers, and eventually encourages competition, and has been applied in the United States [55]. An ***absolute target*** uses a single threshold target, e.g. >75% of immunised children, meaning that only providers who can immunise more than

75% of children are rewarded. This absolute or linear target can enhance divergence in performance if some providers are far above or below the target [19, 49, 50, 56, 57]. Improvement is most likely for providers/facilities closest to the threshold target. Top performers have no incentive to improve, and those far below the target are likely to perceive it as unattainable, a phenomenon referred to as “goal-gradient” theory [56]. A further approach is the use of ***multiple thresholds targets*** which rewards improvements and features all providers in the performance. Some evidence suggests that multiple thresholds targets can enhance convergence in performance [27, 49, 53]. This is because they account for baseline performance and provide incentives for lower performers to catch up. Lastly, ***geographical or equity target*** aims to improve equity by providing high incentive bonus to providers serving the disadvantaged clients or remote populations [33, 50].

1.5 P4P and health systems

Health systems in HICs are performing relatively better than in LMICs [5]. Thus, this sub-section focuses on P4P and health systems in LMICs, since P4P schemes aim to strengthen and reform health systems in these settings. P4P as implemented in LMICs is a reform package with a range of potential attributes to strengthen the health systems [22, 28]. P4P as a reform package ensures the relationship between organisational units within a health system is based on contractual terms with clearly defined performance targets or measures, and gives organisation units substantial decision rights (autonomy) over their resources [21, 22]. P4P schemes incorporate not only financial incentives but also other health systems’ elements (e.g. verification, supervision, health management information systems, financial management through bank accounts, accountability, etc.). Although health systems in LMICs are characterised by complexity, the effects of P4P on health systems are increasingly been studied [58].

P4P through contractual performance incentives can impact the health system through additional financing, improved availability of medical commodities, improved governance and accountability, and improved human resource's productivity. The effects on health system financing can be through the bonus payments among health workers and additional resourcing earmarked to the health facility for facility improvement. Health system financing can also be affected through P4P when providers' attempt to increase user fees to boost facility revenue, encourage enrolment in health insurance schemes, or reduce user fees to attract more patients for performance improvement. The effect on medical commodities can be realised, for example, through incentivising provision of drugs to patients; through facility-level bonus payments which can be used to procure commodities which are commonly out of stock; and by incentivising regional and district health managers to reduce drug stock-out rates. The health system effects through reallocation of resources at the facility or district level are possible, because P4P gives organisational units substantial decision rights or autonomy over their resources [22, 33].

P4P can affect health system governance through increased supervision, verification of performance data, transparency and accountability [22, 28]. Accountability can also be strengthened through providers' responsiveness to users [22], and through community involvement by enhancing health facility governing committees [59]. P4P may also change the organisational culture with improved team work and working sprit among key health system stakeholders. P4P is further expected to affect human resources for health in many other ways. The financial incentives through P4P can increase the staff motivation in order to improve the quality of service delivery and overall productivity [22, 28, 60]. Financial incentives can even reduce the brain drain and encourage providers to work in remote areas [22]. Financial incentives in P4P scheme, while enhancing extrinsic forms of motivation, might also undermine or "crowd-out" intrinsic motivation especially when health workers have a sense of professionalism [21, 38, 61-63]. However, this effect on intrinsic motivation depends on how health workers perceive the financial bonuses within P4P schemes (as fair/ unfair, as a form of recognition and supportive or as form of control) [38, 61]. P4P is also more than external financial incentives, as it encompasses many attributes (e.g. performance feedback, autonomy, and supervision) that can improve instead of crowding-out intrinsic motivation [33]. The health management information systems may also improve through P4P as remuneration is based on proper reporting systems [22].

1.6 P4P and heterogeneous effects

P4P programmes aim to incentivise providers to change their behaviour to improve health service delivery, and obtain financial rewards [64]. Based on this assumption, P4P can improve service delivery on the supply-side which in turn triggers a demand-side response and improves service utilisation. However, while service delivery on the supply-side and service utilisation on the demand-side can improve on average due to P4P, such improvements are rarely uniform across providers and/ or service users, respectively. Thus, P4P can lead to heterogeneous effects among providers on the supply-side, and among service users on the demand-side [65]. However, these heterogeneous effects may arise due to either varied responses to incentives among providers (supply-side) or varied responses to improved health services among populations (demand-side). These two pathways are briefly discussed below.

Heterogeneous P4P effects among providers: Health providers/ facilities are not uniform and may respond differently to incentives. The initial/ baseline performance, for example, may differ across providers (i.e. lower vs. higher baseline performers), and affects subsequent performance. Thus, setting performance targets or measures based on baseline performance is critical for heterogeneous performance among providers [19, 27, 49, 50, 53, 56, 64]. It is whether performance targets give an incentive to improve performance among lower baseline performers, higher baseline performers, or both. Further, health providers/ facilities may differ on structural factors which may favour some facilities to better perform than others. For example, structural factors based on facility-characteristics (e.g. availability of medical commodities, ownership, level of care, staffing level, etc.) and area-based characteristics (e.g. catchment population wealth status, rural-urban location, etc.) may affect facility performance and lead to heterogeneous effects on performance [28, 33, 58, 66-70]. Specifically, facilities with wealthier catchment populations for example may respond better to incentives, as they can more readily increase service use, and user fees contributions [66, 67, 71, 72]. Moreover, facilities with greater availability of medical inputs, as a marker for quality of care [73-75] will be better able to increase patient demand than their counterparts.

***Heterogeneous P4P effects among populations*:** Heterogeneous P4P effects among population subgroups may arise due to either varied providers' responses to incentives or varied responses among populations themselves. Providers are likely to adopt several strategies in order to improve service quality and attract more patients to facilities [22, 33], but patients' responses to different strategies may differ and lead to demand-side heterogeneous P4P effects. One such strategy could be to make services more affordable [22]. This can be, for example, through reducing user fees or by reducing drug shortages (e.g. procure drugs that are stocked-out) to protect patients from incurring costs of purchasing drugs [76, 77]. To improve responsiveness to service users could be another strategy, for example, by being kinder during service delivery [77]. However, providers might also attempt to cherry-pick patients or focus on easy-to-reach populations (i.e. underserved but easily reached) in order to meet the performance targets for rewards [29, 78]. This approach of cherry-picking leaves the hard-to-reach (i.e. poorest with greatest need) underserved. However, to serve the hard-to-reach population needs providers to exert greater effort and time [79]. The efficiency gains can be reached in the case of cherry-picking patients but at the expenses of inequity [80].

Household and individual-based characteristics may also affect how they respond to improved health services in the supply-side. According to Andersen's behavioural model of healthcare utilisation, the use of health services is a function of patient's propensity to use services (predisposing factors), factors that facilitate or impede access and use (enabling factors), and perceived need for healthcare (need factors) [81, 82]. These factors by Andersen's model among others are also social determinants of health [83-85], and affects the demand-side responses to healthcare access and use. For example, reduced financial barriers to accessing care, resulting from provider response to incentives, may stimulate demand especially for poor and/or uninsured individuals, since they are more responsive to a change in healthcare costs consistent with demand theory [86, 87]. The improvement in quality of care supplied may also increase the demand for health services [88]; and likely the better-off populations (e.g. wealthier, educated, and urban residents) may benefit more from quality improvements simply because they use services more than their counterpart populations [84, 85, 89-93].

Although there are potential interactions between the demand-side and supply-side responses to P4P, the health care sector does not operate like a classic free market [41, 94]. There are some cases where the demand-side response may be weak, for example, when some demand-side barriers to accessing care (e.g. cultural and information barriers) are not affected by the supply-side response to incentives [7, 94-96].

2.0 An overview of the literature: Payment for Performance (P4P)

2.1 Introduction

In this section, I present an overview of the literature on P4P from high-income countries (HICs) and low-and middle-income countries (LMICs). I separate the P4P literature between HICs and LMICs since there are differences in context, scheme design, and objectives between settings. This overview focuses on the history of P4P across settings, and the effect of P4P in relation to service utilisation, quality of care, health outcomes, costs, and heterogeneity/ inequalities.

2.2 P4P in High–income countries (HICs)

2.2.1 Introduction

The introduction of P4P schemes in high-income countries (HICs) was pioneered by the United States [30] and the United Kingdom [32]. These schemes continued to be implemented in other developed countries including Canada, New Zealand, Taiwan, Israel, France, Australia and Germany [31, 97]. The focus of P4P in developed countries has been to improve the quality of care [31]. P4P for physicians, for example, has focused on process and outcome measures related to chronic diseases, as well as primary prevention (e.g. screening and immunisations) [44]. However, the hospital-based P4P has focused not only on process quality measures but also on health outcome measures [55, 98].

In the United States, a variety of P4P schemes were introduced [30]. These includes the California P4P (Quality Incentive Programme, QIP), which rewards physician groups based on five ambulatory care quality indicators and five patient-reported measures of service quality [53, 57]. Another P4P scheme in the United States is the Premier Hospital Quality Incentive Demonstration (PHQID). The PHQID rewards inpatient quality of care and outcome measures regarding five clinical conditions: acute myocardial infarction, heart failure, pneumonia, coronary artery bypass surgery, and hip and knee replacement [55, 99, 100]. The first phase of the PHQID started in 2003 to 2006, while its extension began in 2006 to 2009 [101]. In 2004, the United Kingdom government introduced one of the world's largest P4P programme, the Quality and Outcome Framework (QOF) [32]. The QOF targeted family practitioners as the main primary care physicians in the United Kingdom.

2.2.2 Evidence base of P4P in HICs

Despite the widespread implementation of P4P schemes in HICs, there are still mixed evidence of their effects on quality of care improvements, health outcomes, inequalities or whether these approaches are cost-effective [31, 94, 100, 102-112]. Some studies with evidence regarding quality of care, health outcomes and inequalities are discussed below.

Effects on quality of care: Most P4P studies in HICs show improved quality of care measures [31, 55, 94, 103-105]. For instance, the hospital-based P4P scheme in the United States, PHQID, improved most of the process measures of quality of care [55, 103], with limited incremental impact on processes of care for acute myocardial infarction [113]. However, after five years of the Premier HQID implementation, there was no significant difference in performance on process quality measures between Premier hospitals and matched hospitals for comparison [114]. In Hawaii, over a 4-year period of implementation, Chen et al [115] found that a P4P programme in a preferred provider organisation health plan improved quality of care measures for four conditions. In California, the P4P scheme for physician groups (Quality Incentive Programme) revealed that although quality improved for most conditions after P4P, only quality measures for cervical cancer screening improved significantly [53, 57]. Furthermore, P4P programme in the United Kingdom shows that family practitioners improved significantly in quality of care at the early stage of the programme, but such an improvement slowed once targets were reached and even declined for non-incentivised conditions [108, 116-118].

Effects on health outcomes: Available evidence shows that P4P does not seem to reduce mortality rates with few exceptions. For instance, two studies in the United States [113, 119] assessed the early effects of P4P as the hospital-based Premier HQID on mortality reduction for the four incentivised conditions –heart failure, pneumonia, acute myocardial infarction and coronary-artery bypass grafting (CABG). They found that the Premier HQID did not reduce risk-adjusted mortality for all conditions including acute myocardial infarction [113, 119]. A longer term assessment of the Premier HQID also revealed a lack of P4P effect on mortality reduction [120]. In the United Kingdom, however, P4P was associated with an overall reduction in mortality for three incentivised conditions combined, and specifically a significant mortality reduction for pneumonia [98]. The programme reduced mortality in the United Kingdom compared to the United States possibly because the United Kingdom programme had larger bonus size, no self-selection of hospitals to participate, and presence of good communication and feedback among participants [98]. However, the effects of the United Kingdom-based P4P on reduced mortality as initially reported were not maintained in a longer term [121]. Further, Fleetcroft et al [122] reported the evidence on mortality reduction in the United Kingdom across general practices. In terms of health gain, however, there was no clear relationship between the size of financial incentive and health gain for indicators included in QOF for an average general practice in the United Kingdom [123]. Similarly, Ryan et al [124] have recently compared the P4P effect on population mortality between the United Kingdom and other HICs not exposed to P4P (as a synthetic control group) and found no significant decrease in mortality in the United Kingdom after P4P.

Effects on inequalities: P4P effects on inequalities among service users and among providers have been reported in HICs. On the demand-side, P4P generally reduced inequalities in access to quality healthcare between population socioeconomic groups, but had no effect on inequalities with respect to age, sex and ethnicity [31, 105, 109, 110]. On the supply-side, P4P reduced performance inequalities across health providers, in such a way that low baseline performers improved most over time [53, 55, 65, 67, 95, 115, 125]. Also, providers serving lower socioeconomic populations underperformed initially but improved over time [53, 65, 67, 125]. In terms of payments, Ryan et al [68] in the United States found that hospitals treating wealthier populations initially received higher incentive payments than hospitals serving poorer populations, but these inequalities in payments declined over time. Other studies have shown unclear associations between performance and characteristics such as provider's type, size, urban/rural location and staffing level [65, 105].

2.3 P4P in low– and middle–income countries (LMICs)

2.3.1 Introduction

P4P is increasingly being implemented in LMICs with support from donors including the World Bank Health Results Innovation Trust Fund –HRITF [33]. In LMICs, this move is driven by the apparent failure of traditional input-oriented funding to achieve much progress on improving service coverage and quality especially for maternal and child health services [7, 33, 40]. Experiments with performance incentives are also being stimulated by the concern of providers' absenteeism and provision of insufficient service quality due to low productivity (i.e. large know-do gap) [8, 17, 19, 20, 126]. P4P in LMICs is therefore promoted to improve productivity as well as to reform and strengthen the health care system, and facilitate the progress towards health-related development goals [22, 28, 33].

Haiti and Cambodia were the first low-income countries to apply payment for results through performance contracts. P4P was applied to the public sector in Cambodia from 1999 [127, 128], while non-governmental organisations were contracted in Haiti from 1995 and the approach was termed as performance-based contracting [129]. This payment approach was not rolled out nationally in Haiti nor in Cambodia, despite some promising results. In Africa, Rwanda pioneered the implementation of P4P with several pilots from 2002. In 2005, Rwanda decided to scale up P4P nationally [130-132]. The experience from Rwanda inspired and attracted a lot of attention to many other African countries, such as Burundi that rolled out the P4P scheme nationally by 2010 [133]. To date, more than 30 African countries including Tanzania are currently implementing and scaling up P4P (World Bank Health Results Innovation Trust Fund, 2013)[1].

2.3.2 Evidence base of P4P in LMICs

[1] With support from the governments of Norway and the United Kingdom through the *Health Results Innovation Trust Fund* –HRITF, the Bank has helped more than 30 countries implement large-scale pilot efforts in RBF.

Despite growing implementation of P4P programmes in LMICs, the evidence base on the effects of P4P is limited with inconclusive findings [7, 40, 58, 134-136]. The available evidence can be summarised across countries and by themes such as effects on service utilisation and costs, quality of care, and effects on inequalities.

Effects on service use and costs: In Rwanda, P4P led to an increase in utilisation of institutional delivery and child preventive care [132] and further led to improved health worker productivity [60]. Further evidence in Rwanda shows that P4P increased the probability of HIV testing among individuals and even strongly among married individuals [137]. In Burundi, a pilot study from selected provinces found that P4P increased the rate of institutional deliveries, antenatal care (ANC) utilisation, and use of modern family planning services [138]. When evaluating the national programme in Burundi, P4P was associated with an increase in the probability of received full vaccinations for children, while the effect on institutional deliveries was only borderline significant [133]. In Tanzania, Binyaruka et al [77] revealed a couple of positive effects of a P4P pilot scheme in Pwani region. They found that P4P was associated with an increase in institutional deliveries, and provision of antimalarial drugs during ANC, and both were among the incentivised services. The Tanzanian P4P was also associated with a reduction in probability of paying out-of-pocket for delivery care, although the average amount paid did not change. P4P in Malawi, that was combined with conditional cash transfer for pregnant women, did not affect the household costs associated with seeking obstetric care, while reduced time to seek such care [139]. By using facility-level administrative data in Burkina Faso, Steenland et al [140] found that P4P increased ANC visits, institutional deliveries, and postnatal care visits.

In South Kivu Province of the Democratic Republic of Congo (DRC), P4P was associated with an increase in the annual per capita revenues from patient user fees, and an increase in the per capita out-of-pocket health spending from the household survey [141]. These effects on revenues and spending was linked to the P4P design that allowed health providers in P4P areas to negotiate with the communities on the user fees increase. In Katanga Province of the DRC, health workers exerted more effort by reducing fees, absenteeism, increasing outreaches and improving staff motivation [142]. Despite such an increase in effort, there were no changes in the utilisation of health services by the population and even lowered staff revenues due to reduced user fees. A recent P4P pilot in the Republic of the Congo that focused in rural regions (Niari, Plateaux and Pool) found that the scheme significantly improved curative visits, patient referral, vitamin A uptakes, HIV testing and assisted deliveries as measured from facility surveys [143]. In two provinces of Mozambique, P4P was found to increase the provision of HIV testing and treatment, increase of at least four ANC visits, postnatal consultations, and facility-based deliveries [144]. In Cameroon, De Walque [24] found that P4P led to significant increases in utilisation of child and maternal vaccinations, use of modern family planning, and significantly reduced formal and informal user fees.

Moreover, P4P in Cambodia raised the rate of institutional deliveries in public facilities, but no effect on other incentivised services such as ANC and infant vaccinations [128]. In Afghanistan, P4P had no impact on improving service coverage for incentivised services [145]. Studies from Haiti showed that participating non-governmental organisations (NGOs) health facilities outperformed the rest in terms of complete immunisation coverage, prenatal care, assisted deliveries and postnatal care [21, 129]. Consistent with previous evaluations, a recent evaluation in Haiti using facility-level data showed that P4P improved health care delivery, especially on services for under 1 children and pregnant women [146]. However, this study used few NGO facilities with P4P compared to non-P4P facilities (i.e. 15 vs 202), and non-P4P facilities included NGO and public facilities. A P4P scheme in China, for village doctors in rural areas, reduced health care spending for services, and reduced unnecessary care and prescriptions [147, 148].

Effects on quality of care: Most P4P schemes in LMICs do not explicitly incentivise quality of care, but rather these schemes purchase quantity of services and adjust the quantity-based payouts with quality scores [33, 149]. P4P schemes also incentivise service indicators as content of care that link to process quality of care. Quality of care is a multidimensional concept but typically considered in three components (i.e. structural, process, and outcome) [73, 150]. The quality scores for P4P that used to adjust the quantity indicators payout relies on structural quality and resource availability indicators [149]. The effect of P4P with respect to quality is currently skewed towards structural and process quality [134].

The effects of P4P on *structural quality* are generally mixed. In South Kivu of DRC, P4P was associated with an increase in staff availability and improved patient perceptions of drug availability [141]. A study in Katanga province of DRC, however, found negative effects on a structural quality index [142]. The Tanzanian P4P scheme was associated with an increase in availability of drugs and supplies, with no effect on the availability of equipment [76]. In Burundi, no effect of P4P was found on drug availability as perceived by patients [138]; while in Malawi, P4P improved the availability of both functioning equipment and essential drugs [151]. P4P in Cameroon also significantly improved the availability of essential equipment, and qualified health workers [24]. In Rwanda, P4P scheme improved the presence of maternity-related staff, the presence of covered waiting areas and facility management [152]. In Afghanistan, however, the availability of drugs and equipment were not affected by a P4P programme [145].

Regarding *process quality*, there is an evidence that P4P improved the quality of ANC in terms of adherence to clinical guidelines/ contents of care in Rwanda [60, 132] and in Burundi [133]. P4P also improved providers' practices on most attributes during ANC in Egypt [153]. In Tanzania, although there was no effect on quality of ANC for overall adherence to guideline except for some contents of care such as IPT2, P4P increased providers' kindness as reported by patients during delivery care [77]. In Malawi, however, P4P scheme had no effect on birth assistants' adherence to clinical protocols [151]. P4P in Cameroon increased satisfaction of care among patients and providers [24]. Overall perceived quality of care from the household surveys increased due to a P4P pilot in the Republic of the Congo [143]. A randomised study by Peabody [154] in the Philippines found that P4P improved process quality scores among physicians as measured by clinical knowledge performance vignette.

Further, there is limited evidence on the effect of P4P on *health outcomes* in LMICs [54], with the exception of Rwanda, the Philippines and Cambodia. In Rwanda, P4P was associated with improved health outcomes for child nutritional outcomes [60]. P4P in the Philippines improved child health outcomes with respect to wasting and reported health status [155]. Other studies in the literature, however, found P4P did not have any effect on health outcomes. For example, it did not reduce neonatal mortality in Cambodia [128], nor morbidity from diarrhoea, fever or acute respiratory infections in Rwanda [156].

Effects on inequalities: P4P heterogeneous effects have the potential to affect inequalities on service utilisation on the demand-side (among population subgroups/ service users) and inequalities on facility performance on the supply-side (among providers). On the demand-side, available evidence of P4P on utilisation inequalities among population subgroups is limited and varies across service types in LMICs [58]. For example, the effect of P4P on institutional delivery rates was greater among wealthier populations (pro-rich) in most settings [128, 133, 157] but there was an indication that it was greater among poorer groups (pro-poor) and among rural populations (pro-rural) in Tanzania [77, 158]. The effect of P4P on institutional deliveries was greater among women with health insurance in Rwanda [157] or a maternity care voucher in Cambodia [128] than their counterparts, but a greater effect among uninsured women was reported in Tanzania [158]. The effect of P4P on family planning coverage was pro-rich in Rwanda [157], and the effect on immunisation coverage was pro-poor in Burundi [133]. However, studies based on Rwanda Demographic Health Survey (DHS) data reported no differential effect by socioeconomic groups on the use of maternal care [159] and on child curative care seeking [156].

Despite increasing evidence of P4P on inequalities among population subgroups (demand-side) in LMICs, there is only one published study from a LMIC, Rwanda [70], that examined performance inequalities (or heterogeneity of P4P effect) across facilities (supply-side) and only by baseline levels of facility quality. Sherry et al [70] found that facilities in the middle of the baseline quality distribution generally improved most across a broader range of rewarded services. A forthcoming study from Tanzania (Binyaruka et al.), which is one of the articles of this thesis, will supplement the evidence on the supply-side heterogeneity of P4P effect across health facilities.

2.4 Research gaps

Health systems face considerable challenges in providing good quality services for better health outcomes, especially in LMICs. Several initiatives such as P4P have been applied to address some of the challenges. Although there are some promising results of P4P in improving the incentivised indicators or services in LMICs [7, 40, 58, 134-136], there is still little and mixed evidence on structural quality of care and on the heterogeneity of the P4P effects.

Evidence on the effect of P4P on structural quality of care through improved availability of medical commodities (drugs, supplies and equipment) remain scant and mixed [134], despite being a precondition for service delivery [73, 150]. Also, the shortage of medical commodities associates with low levels of patient satisfaction [75], and leads to out-of-pocket payments among patients [74, 160]. From the literature, some studies report on the effect of P4P on the availability of medical comodities that was measured subjectively through patients' perceptions [138, 141], rather than objectively through facility register checklists/ direct observations [142, 145, 151]; and only one study reports on stock-out rates [142]. Further, neither of the previous studies on the P4P effect on medical commodities explains on the potential pathways through which the programme effect occurred, nor examined the potential heterogeneity of the P4P effect on medical commodities across facilities of different characteristics.

Moreover, most evaluations of P4P have focused on average programme effects on incentivised services, with little attention to distributional effects across health providers (supply-side) and across population subgroups (demand-side) especially in LMICs [7, 58, 65, 70]. Evidence on P4P heterogeneous effects is crucial since there is a growing awareness that average effects may mask important heterogeneous programme effects [65, 161-166]. It is therefore important not only to understanding average P4P effects but also heterogeneous P4P effects in order to inform programme design and scale-up.

Limited studies have examined the heterogeneity of P4P effect on service use and quality across population subgroups in LMICs [58], but mainly focused on population socioeconomic groups rather than a broader range of subgroups of social determinants. The use of subgroups based on a variety of social determinants help to better understand the exisitence and potential drivers of heterogeneity of programme effect across populations of different characteristics in a broader perspective. Furthermore, the heterogeneous effects of P4P on performance across health providers (supply-side) are limited in LMICs, despite great variation in health facility readiness to deliver services [167]. Only a study from Rwanda [70] reports on the supply-side heterogeneous effect of P4P on service provision by baseline levels of facility quality. However, this study neither assessed the heterogeneity of P4P effect on facility performance based on baseline levels of performance outcomes (service use), nor based on baseline area-based and other facility-based characteristics. There is also no study in LMICs that assessed the distribution of P4P payouts based on area-based and facility-based characteristics.

3.0 Study objectives

3.1 General objective

The aim of the study is to examine the effect of P4P on the availability and stock-out of medical commodities, and to assess the distribution of the effects of P4P on medical commodities, performance outcomes, and utilisation outcomes in Tanzania.

3.2 Specific objectives

1. To examine the effects of P4P on the availability and stock-out of medical commodities, and assess the distributional effects across health facilities in Tanzania **(Paper I)**
2. To assess the distributional effects of P4P on facility performance outcomes across subgroups of health facilities in Tanzania **(Paper II)**
3. To assess the distributional effects of P4P on utilisation outcomes across population subgroups in Tanzania **(Paper III)**

4.0 Methods

4.1 Study setting

4.1.1 Country profile

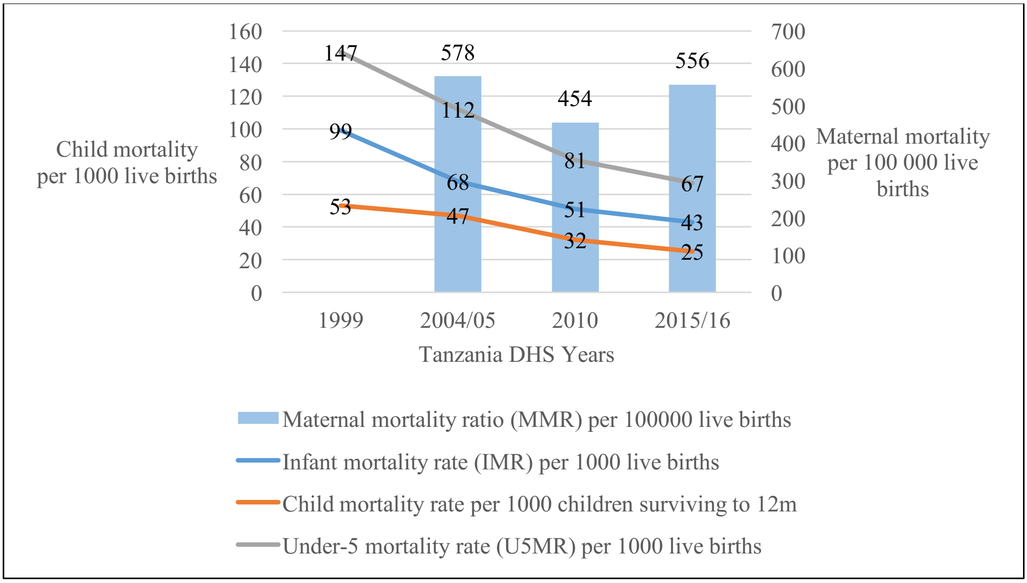
Tanzania is a country in Eastern Africa along the coast of the Indian Ocean. Tanzania has a total area of 945,087 square kilometres, and the largest country in East Africa.

According to the 2012 census survey, its population was 45 million people (and estimated to be nearly 56 million in 2016), with an average annual growth rate of 2.7% (3.1% in 2016) and total fertility rate of 5.5 live births per woman (5.1 in 2016) [168-170]. The population growth rate is higher than the average rate of 2.6% per year for sub-Saharan Africa, and the fertility rate is also higher than that of sub-Saharan Africa of 4.7 births per woman in 2010–15 [171]. About 70% of the population in Tanzania lives in rural areas, and about 46% of children are below 15 years of age [168, 169]. Administratively, Tanzania is divided into 31 regions, and each region is subdivided into several districts, wards, and villages. Tanzania is a low-income country according to the World Bank classification. In 2016, the gross domestic product (GDP) of Tanzania was around USD 47.4 billion, and GDP per capita around USD 879.2 [170]. Tanzania has the annual economic growth rate around 7% which is higher than the average rate of around 3% for sub-Saharan Africa [170, 172]. A number of sectors such as agriculture, tourism, service and mining contribute significantly to the economy, and particularly in terms of employment and GDP growth.

4.1.2 Health status in Tanzania

According to the Tanzanian 2012 population census, the life expectancy at birth was 62 years [169], which is slightly higher than average life expectancy in Africa of 60.2 years in 2010-15 [171]. Tanzania has made progress on the reduction of child mortality over time [168, 173, 174]. According to the recent Tanzanian DHS, the under 5 mortality rate (U5MR) has dropped from 147 deaths per 1000 live births in 1999 to 67 deaths per 1000 live births in 2015 [168]. The infant mortality rate (IMR) has dropped from 99 deaths per 1000 live births in 1999 to 43 deaths per 1000 live births in 2015; while the child mortality rate dropped from 53 deaths per 1000 live births in 1999 to 25 deaths per 1000 live births in 2015. Such a declining trend in child mortality has been associated with an increase in coverage of key child survival interventions such as integrated management of childhood illness, insecticide-treated nets, vitamin A supplementation, immunisation and exclusive breastfeeding practices [174]. A further reason for improved child survival might be an increase in external financing for child health more than three-fold from 2002 [173].

Figure 1: Trends in health indicators from the Tanzanian DHS (1999 –2016)



Source: Author's own contribution from the TDHS data

Notes: IMR–rate of dying before the first birthday; Child mortality rate –rate of dying between the first and the fifth birthday; U5MR –rate of dying between birth and the fifth birthday; MMR –annual rate of female deaths per 100 000 live births from any cause related to pregnancy or childbirth.

However, maternal mortality ratio (MMR) in Tanzania has shown little improvement over the last 11 years, as it stands at 556 maternal deaths per 100 000 live births [168, 173]. The unfavourable progress on reducing maternal deaths is partly due to unskilled home delivery (almost 37% of births still occurring at home [168]), and those who deliver in facilities are faced with poor quality of maternal health services [173]. The current MMR is slightly lower than that of 2005 (i.e. MMR=578), but higher than the ratio reported in

2010 (i.e. MMR=454). In general, both child and maternal mortality rates in Tanzania are far from the Sustainable Development Goals of reducing MMR to less than 70 per 100 000 live births and U5MR to less than 25 per 1000 live births by 2030 [175].

In terms of nutritional status for children under 5 years of age, about 34% of children are stunted, 5% are wasted, and 14% are underweight in Tanzania [168]. Although the prevalence of wasting has remained almost unchanged since 1999, the prevalence of stunting and underweight has been declining steadily since 1996 (as it was 50% for stunting and 24% for underweight).

The use of maternal and child health (MCH) services has increased over time in Tanzania, but with marked imbalance along the continuum of MCH care [168, 173]. According to the TDHS [168], the coverage of at least one ANC visit was almost universal, 51% of pregnant women went for at least 4 ANC visits, 63% of women delivered in health facilities, and about two-thirds received first postnatal care in seven days after delivery.

4.1.3 The health system in Tanzania

The decentralised health system

Tanzania, like other developing countries, has recognised the role of both central and local government to foster economic growth. The process of decentralisation of government functions with several sectoral reforms, including health services, began around 1990s in Tanzania. The aim of these reforms was rooted in improving efficiency, equity, and resource mobilization, through leadership, accountability and partnership at all levels [176]. A typical policy change has been decentralisation, which involves the transfer of power and authority from the central government to local authorities [177, 178]. The decentralisation process in Tanzania took place mainly in three domains: fiscal, political and administrative. Under decentralisation, the local governments should identify priorities and set plans for the allocation and use of resources in order to address local needs, while the central government provides technical support, verification of the relevance of priorities and assists with resource mobilisation [179]. Furthermore, the central government provides grants to local government and then provides autonomy to local government to generate their own resources and allocate these accordingly to prioritised developmental activities.

In the health sector specific, the district-level managers are responsible for preparing annual health sector plans to implement health programmes in their facilities, and they are responsible for generating and managing resources for the district. District managers (Council Health Management Teams) are supported by a Regional Health Management Team, while the health facility governing committees oversee the implementation of plans and the management of resources at the facility level. The decentralised health system gives great autonomy to the district council and uses a needs-based resource allocation formula which can potentially reduce inequalities in resource allocation between rural and urban districts [180]. However, there is evidence that most local governments in Tanzania face inadequate and unreliable financing sources for public service provision [181]. This makes district councils dependent on central government grants, although they further face delays in the disbursement of these funds from the central government [180, 181]. To deal with these delays district councils borrow money from projects in the council and they use money generated from their own source like cost sharing [177].

Organisation and structure

The public sector is the largest sector of the Tanzanian health system, with private for profit and the faith-based organisation/ voluntary sector as important supplements [182]. More than 60% of facilities are publicly owned. The public health system has a hierarchical administrative structure, and is organised in a referral structure with dispensaries and health centres providing primary health care services, followed by district hospitals, regional hospitals, and national referral hospitals. However, the referral structure is hardly followed due to typical bypassing scenarios [183, 184]. Some of the faith-based organisation hospitals have a service agreement with government to offer services as Designated District Hospitals in districts that lack a district hospital. The central government through the Ministry of Health and Social Welfare (MoHSW) oversees most hospitals, and the local government authorities oversee the primary care facilities. As previously discussed, the health sector in Tanzania is also decentralised with great autonomy been given to local governments in terms of budgeting and planning for health service delivery.

Human resources for health in Tanzania

In most settings, especially in LMICs, human resources for health are in shortage and poorly distributed [6]. In Tanzania, the health workforce density has recently been estimated at around 5.5 of doctors, nurses, and midwives per 10 000 population, which is far below the WHO minimum density threshold of 23 per 10 000 population [6, 173, 185]. A further shortage of health workforce is noted with respect to specialised cadres. The staffing level in Tanzania when compared to MoHSW's staffing guidelines is generally low. For example, Manzi et al [186] found only 20% of the recommended number of clinical staff and 14% of the recommended number of nurses had been employed in Southern Tanzania. The distribution of health workforce in Tanzania is also marked with geographical imbalances, and specifically in favour of urban settings [173, 187]. It was estimated that only 31% of health professionals were found in rural facilities in Tanzania [188], despite the fact that most people are residing in rural areas (i.e. 70%). Primary health care facilities are mostly located in rural areas and serve the poor with greatest need, but they face a huge staff shortage problem. This pattern is typical in LMICs, and it reflects an inverse care law since the staffing level is inversely related to poverty and level of need [189, 190].

Medical commodities in Tanzania

In 1993, Tanzania established the Medical Stores Department (MSD) as an autonomous department of the MoHSW. The MSD is responsible for the procurement, storage, and distribution of medical drugs and supplies in the country. However, the MSD supply chain suffers from a shortage of commodities, inadequate budget allocations, inadequate tracking mechanisms and late delivery of required commodities [182, 191-193]. As a result, facilities experience regular shortages of essential drugs and supplies especially in the public sector [182, 188, 192, 193]. For example, out of 1297 facilities surveyed in 2012, only 41% stocked the 14 essential tracer medicines at the time of the survey [188]. An assessment in 2010 found that the MSD fulfilled 68% of hospital orders and 67% of orders from health centres and dispensaries [194].

In terms of ordering, public health facilities order medical commodities on a quarterly basis, based on an estimate of quantity needs. They submit requests to the district who review and send them to the MSD and distribute medical commodities to facilities (the 'pull' system) [192, 195, 196]. Districts and health facilities can also use their own funds (e.g. insurance contributions, user fees and P4P bonus payments) to procure commodities in case of stock-outs [182, 192, 197]. Non-public hospitals that are contracted by districts to deliver services on behalf of the MoHSW also receive medical commodities from the MSD. All other non-public facilities either procure commodities from the MSD, foreign or local manufacturers, privately owned accredited drug dispensing outlets and pharmacies [198-200]. Some commodities (vaccines, antiretrovirals, vitamin A and family planning) are managed through disease-specific vertical programmes, which are financed externally, and distributed via the MSD or directly to facilities [182, 201, 202].

Health financing in Tanzania

The health financing system in any country has three main functions: revenue collection, pooling and purchasing [9]. Revenue collection involves raising or mobilising funds to pay for health services; the pooling function involves pooling together resources across individuals to share the risks associated with ill health; and the purchasing function involves transfer of pooled resources by service purchaser to the service provider on behalf of the beneficiaries who contributed into the pool [9, 10]. In Tanzania, the health financing system is highly fragmented with various sources and modes of financing (Table 1). Health care is largely financed internally through domestic sources, that is 64%, while 36% is through external sources [203]. The domestic sources include general taxation, out-of-pocket payments, and health insurance schemes. According to the National Health Accounts, about 6% of the GDP is invested in health care, and 12% of government expenditure or total budget is spent on health, which is below the Abuja Target of 15% [197, 203]. Out-of-pocket payments account for about 23% of the total health expenditure, while the contribution of prepayment health insurance schemes in total financing is insignificant [203, 204].

Table 1: Health financing system and functions in Tanzania

Resource collection	Pooling of funds	Purchasing
• General taxation • External resources (donors)	• Pooled by government from tax-based and from donors –Central government –Local government	• Government as a purchaser: –Direct budget allocation –Salary • Government via basket funding • Direct donor projects/ vertical programmes • Payment for performance (pilot & roll-out)
• Health insurance o Social Health Insurance (e.g.,	• NHIF, NSSF pools risk for formal sector	• NHIF & NSSF purchases services from their network

NHIF, NSSF-SHIB)	workers	of health facilities (public mainly) in terms of fee-for-services
○ Community based health insurance (e.g. CHF/TIKA)	• CHF (under NHIF) pools risk for informal workers	• CHF purchases services in terms of capitation/ fee-for-services
○ Private insurance	• Private pools risk in each scheme	• Private pool purchases services from their facilities
• Out-of-pocket (OOP) payment		• Households purchases services directly (OOP)

Source: Author's own contribution from the literature.

Notes: NHIF=National health insurance fund for public formal workers; NSSF=National social security fund for private formal workers; SHIB=Social health insurance benefit, which is funded from general NSSF contributions; CHF=Community health fund; TIKA=Tiba kwa kadi; CHF/TIKA is a voluntary scheme for informal workers; OOP= Out-of-pocket.

In 1999, the Government of Tanzania introduced the National Health Insurance Fund (NHIF) for public formal sector employees, followed by the Community Health Fund (CHF) in 2001 for the population in the informal sector in rural areas [204, 205]. In 2009, "*Tiba kwa Kadi*" (TIKA) was introduced which operates like CHF but it focuses on urban settings. The CHF/TIKA membership is based on household enrolment, and allows up to 6 household members. An annual contribution per household varies across district councils, but it ranges between Tanzanian shillings 5000 –15000 (i.e. between 2 –7 USD/ year) with no co-payments [206]. A waiver is granted to households which are unable to pay an annual fee [205]. The government of Tanzania through the NHIF provides a matching grant to the CHF/TIKA contributions at the district level [182]. The National Social Security Fund (NSSF) initiated the Social Health Insurance Benefit (SHIB) program in 2006 for private formal sector workers. SHIB is financed from general NSSF contributions. Both the NHIF and NSSF-SHIB are funded through payroll deductions. The NHIF is mandatory and its benefit package covers about 11 services (www.nhif.or.tz) as per standard treatment guidelines issued by the MoHSW. The CHF/TIKA is voluntary and covers mostly public primary health care.

As an effort to move towards Universal Health Coverage (UHC), the Government of Tanzania aims to improve the health care financing system through various reforms [204, 207, 208]. Such an effort involves strengthening the insurance schemes and expanding their coverage; also to ensure services are affordable, equitable, accessible, and of good quality. In the last decade up to now, the following health care financing reforms have been considered in Tanzania:

- Harmonising management and administration of CHF with the NHIF in 2008
- Introducing TIKA which is similar to the CHF but for the urban informal sector.
- CHF to engage non-government providers through service agreements to improve service availability; which is part of the public-private partnership policy.
- Making CHF/ TIKA uniform across the country in terms of benefits package, contribution rates, and provider payment mechanisms.
- Developing a national health financing strategy which proposes a national health insurance scheme to reduce fragmentation of health insurance schemes.
- Recently, Tanzania is introducing a *direct health facility financing* mechanisms to improve efficiency and effectiveness of resources use and management by direct allocation/ transfer of health basket fund to all health facilities' bank accounts.

Despite the introduction of various health financing reforms, the coverage of health insurance is gradually increasing, but remains low and variant in the country. A recent estimate shows the health insurance coverage ranges between 10–15% in Tanzania [168, 204]. A number of challenges to coverage expansion has been documented such as inability to pay, poor quality of care provided, poor staff attitudes, large population in the informal sector, lack of awareness on risk pooling, lack of provider choice, and limited benefit packages [173, 204, 205, 209, 210]. Tanzania also has exemption and waiver policies for some population groups. It aims to protect the poor and vulnerable groups, which include pregnant women, under five children, elders above 60 years, and patients suffering from TB and HIV/AIDS [204, 211]. However, the enforcement of an exemption policy is generally weak in Tanzania, as a result eligible patients are paying out of pocket [77, 212, 213].

4.2 P4P in Tanzania

The government of Tanzania through the MoHSW, with support from the Government of Norway, introduced a P4P pilot scheme in Pwani region (2011 –2014). The objectives of the pilot were to inform the national P4P roll out programme, and to accelerate the reduction of maternal, neonatal and child morbidity and mortality through improving reproductive, maternal, newborn and child health (RMNCH) services and quality. Pwani is one of 31 regions in the country and is comprised of seven districts with more than 209 health facilities. It has a population of just over a million [169]. The scheme was implemented in all facilities providing RMNCH services in the region irrespective of ownership status. The Tanzanian P4P rewarded health providers based on performance in relation to utilisation of specific services (e.g. institutional delivery) or for care provided during a service (e.g. provision of antimalarial drugs during ANC) as described elsewhere [77, 214].

The Tanzanian P4P scheme rewarded the performance of health workers and their health facilities based on two methods of target setting (Table 2): A single threshold for all facilities (absolute coverage target which is fixed), and multiple thresholds based on performance in the previous cycle (relative change). With multiple thresholds, a facility could fall into one of five groups based on their performance in the previous cycle: Group 1 (0-20% coverage of said indicator), group 2 (21-40%), group 3 (41-70%), group 4 (71-85%) and group 5 (>85%). Each of these five groups has its own absolute threshold target, with group 5 being required to maintain coverage due to a limited scope for improvement (a ceiling effect) (Table 2). For a single threshold target, all facilities have a single absolute target irrespective of the previous/ baseline performance. The strategies to reach facility-level performance targets were left to the discretion of the health workers at the individual facilities. However, district and regional managers provided supportive supervision to ensure performance. Health managers at district and regional levels were also rewarded depending on the performance of facilities in their district and region, and had additional performance targets linked to management, timely deaths audit, and reduction of stock-outs of essential drugs (e.g. antimalarials, antibiotics) in their districts/ region, respectively.

Table 2: Service indicators and performance targets for P4P implementing facilities in Tanzania

P4P service indicators	**Method**	**Baseline coverage (previous cycle)**				
		0–20%	**21–40%**	**41–70%**	**71–85%**	**85%+**
Coverage indicators						
% of institutional deliveries	Percentage point increase	15%	10%	5%	5%	Maintain
% of mothers attending a facility within 7 days of delivery.	Percentage point increase	15%	10%	5%	5%	Maintain
% of women using long term contraceptives	Percentage point increase	20%	15%	10%	Maintain above 71%	Maintain
% children under 1 year received measles vaccine	Overall result	50%	65%	75%	80%+	Maintain
% children under 1 year received Penta 3	Overall result	50%	65%	75%	80%+	Maintain
% of complete partographs	Overall result	80%	80%	80%	80%+	Maintain above 80%

HMIS reports submitted to district managers on time and complete	Overall result	100%	100%	100%	100%	100%
Content of care indicators						
% ANC clients receiving two doses of IPT	Overall result	80%	80%	80%	80%+	Maintain above 80%
% HIV+ ANC clients on ART	Overall result	40%	60%	75%	75%+	Maintain
% of children receiving polio vaccine (OPV0) at birth	Overall result	60%	75%	80%	80%+	Maintain

Notes: 85%+ = 85% or more; 80%+ = 80% or more; HMIS=Health Management Information System; ANC=Antenatal care; IPT=Intermittent preventive treatment.

Source: Ministry of Health and Social Welfare (2011): The Pwani Pay for Performance (P4P) Pilot in Pwani region, Tanzania: The Design Document.

The implementation of the scheme was overseen by the Pilot Management Team, comprised of MoHSW and Clinton Health Access Initiative officials. Performance data were compiled by facilities and verified by the Pilot Management Team every six months (one cycle) before distributing payouts. Data reporting followed the existing Health Management Information System (HMIS). All facilities with a P4P scheme must have a bank account to receive performance payouts. The scheme rewarded facilities either full or partial payments depending on their achievement level. Full payment was made if 100% of a given target was achieved as pre-specified, 50% of payment was made for 75%<100% achievement, and no payment was made for a target achievement below 75%. The maximum potential payout a facility could earn was USD 820 per cycle for dispensaries, USD 3220 per cycle for health centres and USD 6790 per cycle for hospitals. P4P payouts were additional to funding for operational costs and health worker salaries. It included staff bonuses (approximately 10% of their monthly salary if all targets were fully attained) and facility funds earmarked to support improvement or demand creation initiatives (10% of the total in hospitals and 25% in lower level facilities). The maximum potential payout for district and regional managers was USD 3000 per cycle, if all the targets were fully achieved.

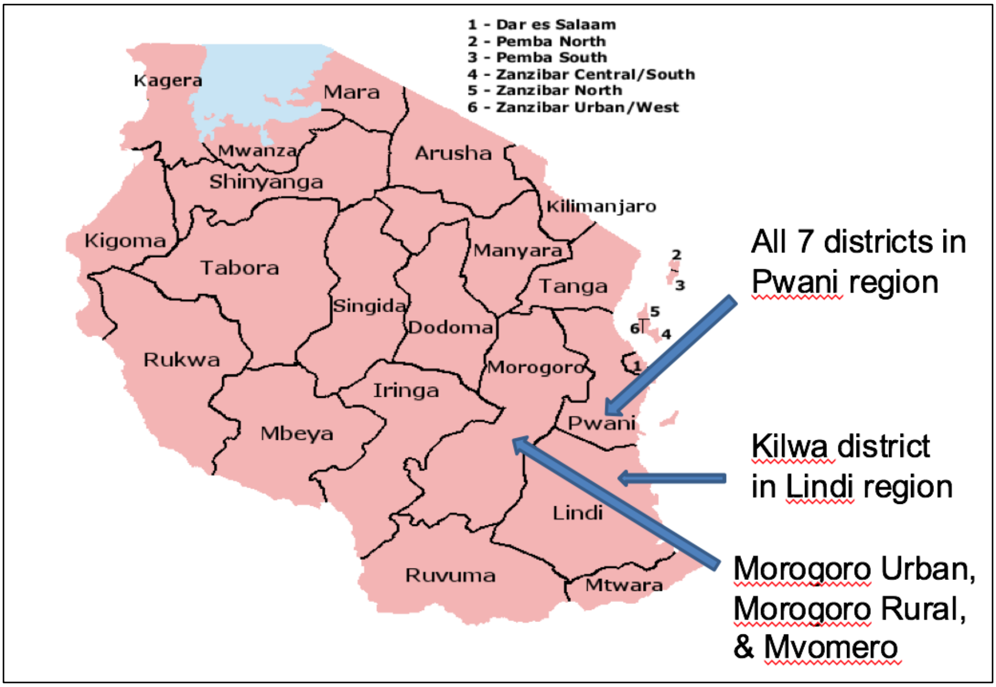
The P4P programme in Tanzania was the subject of a process and impact evaluation. I was part of the team that focused on the impact evaluation component. I oversaw the fieldwork preparation, trained the fieldworkers, oversaw all rounds of data collection, lead the data analysis of the household and facility survey data, and participated in the write-up of publications and be involved in the dissemination of results. The impact evaluation found a significant positive effect of P4P on two out of eight incentivised service indicators: institutional delivery rate and provision of antimalarial drugs during ANC [77]. P4P was also associated with a number of process changes such as increased availability of drugs and supplies, increased supportive supervision, a reduced chance of paying user fees, and greater provider kindness during delivery care [59, 76, 77, 215]. This PhD work is based on further analyses of facility and household survey data to determine whether there were heterogeneous effects of the P4P among subgroups of populations and facilities in Tanzania.

Based on experience and lesson learned from the P4P pilot, the Government of Tanzania with financial support from the World Bank decided to roll out the programme. The initial phase of roll out started in 2016 with 8 regions (i.e. Shinyanga, Mwanza, Pwani, Tabora, Simiyu, Geita, Kagera and Kigoma). These initial regions were selected due to poor health outcomes and high poverty index. The P4P programme was slightly re-designed prior to roll out, and was rebranded as a Results-based financing (RBF) scheme. The changes in the design included an increased number of incentivised indicators covering the outpatient department care and quality of care indicators; indicators for community health workers; indicators for MSD offices; paying per service (fee-for-service) conditional on quality scores instead of paying for performance targets on service coverage; payment cycle (from bi-annual to quarterly); and a higher proportion of bonus payments earmarked for facility improvement (from 10-25% to 75% in the RBF roll out).

4.3 Study sites

The evaluation study of the P4P pilot was conducted in three regions (i.e. Pwani, Morogoro and Lindi) out of 31 regions in Tanzania. The P4P pilot was implemented in all seven districts in Pwani region, whereas three districts from Morogoro and one district from Lindi were selected for comparison purposes. Pwani and Morogoro regions are in the eastern zone, while Lindi region is in the Southern zone of Tanzania. The population estimates in Pwani region were just above a million, whereas in Morogoro region the population estimates were just above two million, and less than a million in Lindi region [169]. Pwani region is next to Dar es Salaam city, while Morogoro and Lindi are neighbouring regions to Pwani (Figure 2).

Figure 2: Map of Tanzania with location of the study sites



4.4 Study design

An impact evaluation study attempts to measure the causal impact of a programme or policy on an outcome of interest [161, 216]. It seeks to answer cause-and-effect questions. The programme's impact is identified by comparing the observed outcomes of participants with an estimate of what would have been the outcome of participants in the absence of a programme (unobserved as counterfactual outcome) [216]. The main challenge in designing an impact evaluation study relies on how to deal with the evaluation problem. The evaluation problem exists because only one outcome at any point in time can be observed per unit of observation, but not both outcomes for the same unit of observation with and without a programme/ intervention [161, 216]. This problem leads to the challenge of finding a good counterfactual group due to missing data, that is to find or create a convincing and reasonable comparison group for programme participants [161, 216]. Failure to find a reasonable comparison/ control group may lead to biased estimates of programme impact because of selection bias.

The randomisation process addresses the problem of selection bias at the level of randomisation, that is, both groups should be similar in observed and unobserved factors [161, 216, 217]. Randomised experiments are considered *gold standard* for causal inference. However, randomisation is not always feasible because of ethical issues, lack of compliance, being expensive, poor external validity, contamination, spill-over effects, politically unacceptable within a targeted area, and selective attrition [218]. Thus, quasi-experimental designs are preferred to attribute casual inference in the absence of randomisation as described elsewhere [161, 216, 219].

The P4P evaluation study in Tanzania used a quasi-experimental design, which was a controlled before and after study design. It was due to the fact that the Government of Tanzania introduced a P4P programme in one region in the absence of randomisation. This was partly due to political reason of not accepting provision of financial incentives to some facilities/ districts within a region and not to others. With a controlled before and after study design, surveys were done in two-time period (before and after the introduction of P4P) and from two study arms (intervention and comparison districts) as previously described. Comparison districts were selected such that they were similar as possible to intervention districts in terms of poverty, literacy rates, rates of institutional deliveries, infant mortality, population per health facility, and the number of children under one year of age per capita [214].

4.5 Sampling and data sources

The health facility was the primary sampling unit in the survey. This study included all 6 hospitals and 16 health centres that were eligible for the P4P scheme, and a random sample of 53 eligible dispensaries in the intervention arm. A similar number of facilities were included in the comparison arm. To assess RMNCH service utilisation in the population, 20 households from the catchment area of each health facility were randomly sampled. A household to be eligible had to have a woman aged (15–49 years) who had delivered in the 12 months prior to the survey. To sample eligible households, the study identified first the village(s) from a facility's catchment area, and randomly sampled four hamlets (sub-villages/ streets) from each village. Then, five eligible households were randomly sampled from each hamlet to make a total of 20 households per facility's catchment area. In total, 3000 households with eligible women in both arms at baseline were surveyed, and a similar number in the follow-up survey. Furthermore, the P4P Pilot Management Team provided data on facility total payouts which reflect performance on all incentivised indicators for the 75 facilities in the intervention area over seven payment cycles (2011 –2014). The payout data were used to assess the inequality in payout distribution as a proxy for performance inequality across health facilities.

4.6 Data collection

The baseline survey for data collection was carried out in January 2012[2] in seven intervention districts and in four comparison districts, with a follow up survey 13 month later. The facility and household surveys were used to capture data from the supply-side and the demand-side, respectively. The facility survey questionnaire was administered to the facility in-charge or other experienced health worker. The facility survey collected information on facility ownership, level of care, availability and stock-out of medical commodities (drugs, supplies and equipment), availability of infrastructures and utilities (electricity and clean water), facility distance from district headquarters, and rural/urban district location.

The household survey questionnaire was in two components –household head and woman survey. The household head survey was administered to the household head, and captured

[2] Note that the programme started in 2011 and first payouts tied to performance was made in September 2011. Therefore, to get around the risk of early P4P effects in the intervention areas, we sampled women aged 16–49 years who delivered between October 2010 and October 2011 during the baseline survey in early 2012. (See Borghi et al [214] for more details).

information on household background characteristics (e.g. household size, health insurance status, and ownership of assets and housing particulars for assessing the household socioeconomic status). The woman's questionnaire administered to an eligible woman with a child of less than 12 months of age, and captured data on background characteristics (e.g. age, marital status, education occupation, religion, and number of births), and service utilisation for RMNCH services. For the case where the eligible woman was also the household head, such a woman was administered with both sets of questionnaires.

The survey of data collection was done by 48 data collectors with three coordinators on each round of data collection. These data collectors were grouped into 8 teams of 6 people each, including a supervisor per team. All data collectors were trained for one week before the pilot of tools and the actual survey of data collection. The pilot of survey tools aimed to pre-test the tools before the actual survey to ensure all questions were clear and relevant, and possible revisions were done. The survey of actual data collection in all 11 districts took almost two months.

Ethical approval for the evaluation study was obtained from the Institutional Review Board of the Ifakara Health Institute (approval number: 1BI1IRB/38) and the Ethics Review Board of the London School of Hygiene & Tropical Medicine. The P4P Pilot Management Team which included members of the Ministry of Health approved the study design and protocol. Introduction letters were sent to respective District Executive Directors (DED) copied to District Medical Officers (DMO) informing them about the evaluation study and its objectives prior to fieldworks of data collection. The research team provided an information sheet at the district level (DMO's and DED's offices), and district officers (DMO and DED) provided introduction letters for the team to all facility in-charges and community leaders. All study participants were given the information sheets and consent forms prior to conducting the interviews. Moreover, this study utilises aggregate data on health services utilisation. It does neither utilise sensitive health information attributable to individuals, nor does it concerns the conduct of research to generate new knowledge on health and disease. Consequently, we considered the study to fall outside the scope of the Norwegian Health Research Act, and that submission to the Regional Ethics Committee was not required.

4.7 Variable measurement

A number of variables of interest were considered in this study. The variable types and measurements are shown below for each of the papers of this thesis. Some of the variables overlap but there are also some differences across papers.

Paper I

The main outcomes for this paper included the availability of RMNCH medicines, medical supplies and functioning equipment on the day of the survey, and whether there was a stock-out of medicines and supplies at the facility in the 90 days preceding the survey. In terms of availability measurement, if a commodity was available on the day of the survey, the outcome was coded 1 and 0 otherwise. For stock-out measurement, if a commodity was out of stock for at least one day in 90 days prior to the survey the outcome was coded 1 and 0 otherwise. Medical commodities were classified in terms of their therapeutic use as: antibiotics, antimalarials, antihypertensives, antidiarrheal, antiretrovirals, oxytocics, vaccines, family planning, vitamin A, medical supplies and medical equipment (Appendix S1a, Paper I). There were 37 items of essential drugs, 11 medical supplies and 16 functioning equipment. Commodities were further differentiated between items which relate directly to P4P targets and those which do not, to examine eventual spill-over effects. Items were also classified according to their beneficiary/ recipient group along the RMNCH continuum of care based on the World Health Organisation (WHO) classification of priority medicines [220, 221]. Composite scores were generated for each classified subgroup based on an un-weighted mean score across items in the group. The composite score can be interpreted as the mean percentage availability/ stock-out rate within the grouping across facilities. The proportion of facilities with availability/ stock-out of the respective commodity groups were captured. When generating indices each commodity item was given equal weight for ease of interpretation, although some of the items may be more important than others in enhancing better health outcomes.

Paper II

This paper used two sets of performance outcomes to assess performance inequalities across health facilities. First, a "payout score" for each facility in the intervention arm, defined as the percentage of bonus payout received relative to the total potential amount if all targets had been fully achieved. Payout scores were generated for each of the seven payment cycles (2011 –2014) per facility, and aggregated into an average score for all cycles. Note that, a payout for each cycle was in aggregate form to reflect facility performance on all incentivised indicators within a cycle. Second, this study considered the two incentivised services which improved significantly as a result of P4P [77]. These two services also had different incentive designs for target setting: the coverage of institutional deliveries (multiple thresholds target) and provision of two doses of intermittent preventive treatment (IPT2) for malaria during ANC (single threshold target). The average service coverage rates for these two services were estimated at the facility level based on outcomes measured from households in the facility catchment area, which was used as a proxy for facility performance.

Paper III

This paper also used the two outcome variables which improved significantly as a result of P4P: institutional deliveries and uptake of two doses of intermittent preventive treatment (IPT2) for malaria during ANC [77]. These were measured at the individual level as binary outcomes for whether a woman gave birth in a health facility and whether she received IPT2 during ANC, respectively. These outcome variables overlap in paper II and III, but they differ in terms of the level of measurement (facility-level in paper II versus individual-level in paper III).

4.8 Generating subgroups for distributional analyses

Subgroups of facilities (Paper I & II)

To assess the supply-side distributional effects of P4P, health facilities were classified into subgroups. The types of subgroups used in each paper and the justification for these is provided below.

Paper 1 assessed the distributional effects of P4P on the availability and stock-out of medical commodities across facilities. This analysis examined whether the effects of P4P differed with the wealth status of the facility catchment population (wealth subgroups), facility ownership (public vs. non-public), facility level of care (dispensary vs. health centre or hospital) and facility location (urban vs. rural). The choice of wealth subgroups was necessary to examine if benefits were pro-poor, given the greater burden of out-of-pocket payments on poorer groups due to stock-outs of medical commodities [74, 160, 193, 209]. The out-of-pocket payment for drugs also limits the affordability of and access to care, especially among the worse-off population. The analysis by facility ownership (public vs. non-public) was because of differing procurement and supply systems in public and non-public sectors as described earlier; while the analysis by level of care (dispensary vs. health centre or hospital) was due to the fact that dispensaries are typically worse-off in resources availability including drug availability [188, 222, 223]; and the analysis by location (rural vs. urban district) was done because facilities in urban districts are better connected by roads and easily accessible facilitating the distribution of commodities relative to those in rural districts. The wealth status of the facility catchment population was measured as the mean wealth index score across households in the catchment area at baseline. The wealth scores were derived using principal component analysis based on 42 items relating to household characteristics and asset ownership (Appendix S1c, Paper I) [224, 225]. Then, the average wealth score of the 20 households sampled within the facility catchment area was calculated. Facilities were further ranked by scores from poorest (low score) to least poor, and classified into three equal-sized groups (terciles): poorest, middle and least poor.

Paper II examined whether facility performance outcomes differed across facility subgroups. The first facility subgroups were based on baseline facility performance (above or below the median level for the two outcomes –rate of institutional deliveries and IPT2 coverage). The use of facility subgroups based on baseline performance was considered to test an *incentive design effect*, i.e. whether target setting design affects facility performance differently between lower and higher baseline performers. The second set of facility subgroups were considered to test the *structural effects*, i.e. whether the baseline facility- and area-based characteristics affects facility performance. Facility-based characteristics included: facility ownership (public owned vs. non-public); facility level of care (dispensary vs. health centre and hospital); baseline availability of utilities (electricity and water supply); and baseline availability of essential drugs (above/below the median in an un-weighted index based on the availability share of all 37 essential drugs (Appendix S1a, Table I)). Area-based characteristics included: facility location (rural vs. urban district) and the wealth status of the facility catchment population (poorest, middle and least poor) as previous described.

Subgroups of population (Paper III)

To assess the demand-side distributional effects of P4P, households were classified into subgroups based on individual and household-level characteristics. According to Andersen's behavioural model of healthcare utilisation [81, 82], the use of health services is a function of patient's propensity to use services (predisposing factors), factors that facilitate or impede access and use (enabling factors), as well as perceived need for healthcare (need factors). These factors among others are also social determinants of health [83-85]. Only predisposing and enabling factors were considered in this study since data on perceived illness was not available. Further, "perceived illness" could be argued to be of less relevance for maternal service utilisation outcomes, since care seeking is preventive to ill health and related to pregnancy status and not a function of ill health.

Predisposing and enabling factors were then used to generate population subgroups to assess the differential effects of P4P on service utilisation. The categorisation followed previous categorisation in the literature as well as based on context specific and frequencies across categories. Subgroups of predisposing factors included: marital status (married vs. none), maternal age (15–49) years (below vs. above the median age of 25), education (no education vs. primary level/above), occupation (farmer vs. non-farmer), religion (Muslim vs. non-Muslim), number of births/parity (parity 1 vs. parity 2/above), and household size (below vs. above the median size of 5 members). Subgroups of enabling factors included: health insurance status (any insurance vs. none), place of residence (rural vs. urban district), and household wealth status subgroups. The wealth subgroups were generated from wealth scores derived using principal component analysis based on 42 items of household characteristics and asset ownership (Appendix 1: Table 5, Paper III) [224, 225]. The household wealth scores were generated separately for baseline and follow-up samples, since participants differed over time. Households were ranked by wealth scores from poorest (low score) to least poor and classified into three-equal sized groups (terciles): poorest, middle and least poor. Subgrouping based on five-equal sized groups (quintiles) were also generated to examine the sensitivity of the findings to different wealth subgroupings.

4.9 Data analyses

The data analyses in all papers proceeded in two parts: descriptive analyses and difference-in-differences (DID) linear regression analyses.

4.9.1 Descriptive analyses

These included sample means comparison across study arms at baseline and equity analysis. The sample means at baseline for all the characteristics of facilities (**paper I and II**) and characteristics of population (**paper III**) were compared between intervention and comparison arms. The baseline assessment also examined the distribution of facility outcomes (medical commodities, performance outcomes) across facility subgroups for **paper I and II** respectively, and the distribution of population outcomes (service utilisation outcomes) across population subgroups for **paper III**. The baseline comparison of outcomes across subgroups of facilities and population generated the differences/ gaps between subgroups which indicates inequalities at baseline [226, 227]. T-tests were used to assess whether the gaps were significantly different from zero.

An equity analysis was further conducted for the distribution of P4P payouts as a proxy for facility performance in **paper II**. The equity analysis on the payout distribution used three measures of inequality: an absolute measure (the equity gap) and two relative measures (the equity ratio and the concentration index) [226, 227]. Equity measures identifies the unfair or unnecessary differences in outcomes across facilities subgroups [228]. The equity gaps and equity ratios were computed across all stratifying variables used, while the concentration indices were computed on a ranking variable of area-based wealth status. These three measures of inequality are further described below. The equity gaps and equity ratios were generated by comparing the performance payouts (payout scores) across facility subgroups. Specifically, the equity gap was measured as the difference in payout scores between facility subgroups, while the equity ratio was measured as the ratio of payout scores between facility subgroups. A positive (negative) equity gap and an equity ratio greater (less) than one in relation to wealth defines a pro-rich (pro-poor) distribution, respectively. An equity gap of zero and an equity ratio of one defines an equal distribution. T-tests were used to assess whether the equity gaps in payout distribution were significantly different from zero. Since seven payout cycles of 6-

months each were considered, the equity measures were applied to each payout cycle and to all cycles combined.

Concentration indices (CI) were used to measure wealth-related inequality in the distribution of performance payouts using an area-based wealth status [227, 229]. The CI is a relative measure of inequality that shows the gradient of an outcome of interest across multiple subgroups with natural ranking [226, 227, 229]. It indicates the concentration of an outcome of interest across ranked subgroups of interest. The CI ranges between [-1 and +1], with zero indicating equality between multiple subgroups. A positive value indicates that an outcome of interest is more prevalent in the highest ranked subgroup (e.g. the richest), while the negative values indicate that an outcome of interest is more prevalent in the lowest ranked subgroup (e.g. the poorest). Equation 1 shows the formula to estimate a CI of an outcome of interest [227].

$$\mathrm{CI} = \frac{2}{\mu}\mathrm{cov}\,(\mathrm{y_i}, \mathrm{R_i}), \qquad (1)$$

where $\mathrm{y_i}$ is the outcome of interest of the $\mathrm{i^{th}}$ individual/ facility; $\mathrm{R_i}$ is the fractional rank of the $\mathrm{i^{th}}$ individual/ facility (in terms of wealth status); μ is the mean of the outcome and cov denotes the covariance. The estimated CIs were also tested whether they were significantly different from zero and the p-values were estimated.

4.9.2 Difference-in-differences linear regression analyses

First, the DID analysis was used to identify the effect of P4P on the availability and stock-out of medical commodities (paper I). Second, the DID analysis was extended to identify the differential effects of P4P on medical commodities across facilities subgroups (paper I); the differential effects of P4P on facility performance outcomes across facilities subgroups (paper II); and differential effects of P4P on increased service utilisation across population subgroups (paper III).

Equation 2 estimates the average effects of P4P at the facility-level for **paper I**.

$$\mathrm{Y_{it}} = \beta_0 + \beta_1(\mathrm{P4P_i} \times \delta_t) + \beta_2\delta_t + \gamma_i + \varepsilon_{it} \qquad (2)$$

where $\mathrm{Y_{it}}$ is the outcome (availability/ stock-out of commodities) of facility i at time t. $\mathrm{P4P_i}$ is a dummy variable, taking the value 1 if a facility is exposed to P4P and 0 if not. This analysis controlled for time-invariant facility level determinants γ_i through facility fixed-effects estimation, and controlled for year-specific characteristics through δ_t year fixed-effects. The error term is denoted by ε_{it}. The average effect of P4P on the outcome is given by β_1 in equation 2.

An extension of the DID regression model with three-way interaction terms was used to identify differential effects of P4P across facilities' subgroups (paper I and II) and across population subgroups (paper III). The three-way interaction term was between the average P4P effect ($\mathrm{P4P_i} \times \delta_t$) and subgrouping variable (facility subgroup $\mathrm{G_i}$ / population subgroup $\mathrm{G_{ijt}}$). The associated two-order interaction terms were also included in the model, though the time-invariant interaction terms were dropped through fixed-effects estimation. The coefficient of interest is β_4 which indicates the differential effect of P4P across facility and population subgroups as shown in equation 3 and 4, respectively.

Equation 3 estimates the differential effects of P4P at the facility-level for **paper I** and **II**.

$$\mathrm{Y_{it}} = \beta_0 + \beta_1(\mathrm{P4P_i} \times \delta_t) + \beta_2\delta_t + \beta_3\mathrm{Z_{it}} + \beta_4(\mathrm{P4P_i} \times \delta_t \times \mathrm{G_i}) + \beta_5(\mathrm{P4P_i} \times \mathrm{G_i})$$
$$+\beta_6(\mathrm{G_i} \times \delta_t) + \gamma_i + \varepsilon_{it} \qquad (3)$$

where Y_{it} is the service coverage outcome (facility performance) of facility i at time t. $P4P_i$ is a dummy variable, taking the value 1 if a facility is exposed to P4P and zero otherwise. This estimation controlled for time-invariant facility-level characteristics γ_i through facility fixed-effects estimation, and included δ_t for year fixed-effects. Potential confounding factors such as time-varying facility-level covariates Z_{it} (availability of electricity and water supply, and the mean wealth index for households sampled in the catchment area of the facility) were controlled for. Note that the estimation of differential effects involved a series of regressions such that each regression has an indicator of subgrouping variable G_i. The G_i are time-invariant and only baseline values were taken for time-varying facility variables to capture the pre-existing structural effects. The error term is denoted by ε_{it}. The confidence interval was reported based on standard errors clustered at the facility level to account for serial correlation of ε_{it} at the facility level.

Equation 4 estimates the differential effects of P4P at the individual/ household-level for **paper III**.

$$Y_{ijt} = \beta_0 + \beta_1(P4P_j \times \delta_t) + \beta_2\delta_t + \beta_3 X_{ijt} + \beta_4(P4P_j \times \delta_t \times G_{ijt}) + \beta_5(P4P_j \times G_{ijt}) + \beta_6(G_{ijt} \times \delta_t) + \gamma_j + \varepsilon_{ijt} \qquad (4)$$

where Y_{ijt} is the utilisation outcome (institutional deliveries or uptake of IPT2) of individual i in facility j's catchment area and at time t. The intervention dummy variable $P4P_j$ takes the value 1 if a facility is in the intervention arm and 0 if it is in the comparison arm. The time invariant facility characteristics γ_j were controlled for through facility fixed-effects estimation; and included δ_t for year fixed-effects. Also, potential confounding factors that are time-varying such as individual and household-level covariates X_{ijt} (age, education, occupation, religion, marital status, parity, insurance status, household size, and household wealth status) were controlled for in the model. Note that the estimation of differential effects involved a series of regressions whereby each regression has an indicator of subgrouping variable G_{ijt} taken among covariates X_{ijt}, and the G_{ijt} were time-varying since sample participants differed over time. The error term is donated by ε_{ijt} in the model. The standard errors were clustered at the facility level/ facility catchment area to account for serial correlation of ε_{ijt} at the facility level.

4.10 Sensitivity analyses

Paper I, II and III

The first robustness check focused on clustering the standard errors[3]. Instead of clustering at the facility level to account for serial correlation of error terms at the facility level, this study also clustered the standard errors at the district level to correct for correlation of error terms across facilities within districts. To calculate robust standard errors clustered at the district level, the study used the bootstrapping method to adjust for the small number of clusters[4] [230].

Paper II

In this paper, an initial sensitivity analysis was used to re-estimate the model for institutional deliveries by excluding hospitals (8% of facilities per arm). This is because the performance indicator of institutional deliveries was aimed at primary care facilities (health centres and dispensaries), as opposed to hospitals that have a less clearly defined catchment population. Then, the mean wealth scores were reclassified into two quantiles (below or above the median) to check whether the wealth effect was sensitive to the classification of the wealth groupings. Lastly, apart from using a conventional parametric test (a t-test) to assess whether differences in payouts between subgroups were significant,

[3] The default ordinary least squares (OLS) standard errors which ignores data clustering commonly underestimate the true OLS standard errors. The clustering is essential as it accounts for any within-group dependence in estimating standard errors of regression parameter estimates (Cameron & Miller [230]).

[4] This study used 11 districts (7 intervention and 4 comparison districts) which were quite few for data clustering without bootstrapping approach.

a non-parametric test (Wilcoxon rank-sum test)[5] was also used considering that the distributions of payouts in all cycles were not normally distributed [231].

Paper III

The robustness checks in this paper included re-estimating the P4P differential effects by using wealth quintiles instead of wealth terciles to examine whether the results were sensitive to wealth group classification. Wealth status subgroups for each study arm were also generated and used to re-estimated the P4P differential effects by arm-based wealth subgroups to avoid the baseline imbalance in wealth status between study arms. This is contrary to initial subgrouping of wealth status which was first generated separately at baseline and then in the follow-up survey regardless of study arms. The regression model was also re-estimated by including three-way interactions with categorical variables which give multiple subgroups (e.g. categories of education levels (no education, primary, secondary and college/above)), instead of interactions with binary variables only (e.g. no education vs. some education). Lastly, a non-linear logit model was also applied in paper III, instead of linear model because paper III used binary outcome variables. All the analyses in this study were performed using STATA software (version 13).

5.0 Summary of results

5.1 Paper I

The facility characteristics at baseline were generally balanced across study arms, but facilities in the intervention arm were serving poorer populations than those in the comparison arm (Table 2, Paper I).

Average effects of P4P

[5] The use of parametric test depends on the assumption that data are normally distributed. This implies that samples from different groups are independent and that the variances between the groups are equal (Kitchen [231]).

The introduction of P4P was associated with an 8.4 percentage point increase in the availability of all 37 medicines combined (13.8% increase from baseline, p=0.002), and an 8.3 percentage point increase in the availability of medical supplies, though this was only borderline significant (12.9% increase from baseline, p=0.050) (Table 3, Paper I). There was no effect of P4P on the availability of functioning equipment. The effects of P4P were further identified for some medicines related with P4P targets (i.e. antimalarials, antihypertensives and oxytocics used for deliveries) and supplies (i.e. partograph), but not on vaccines, family planning and antiretrovirals. Effects were also observed for drug items that were not clearly linked to service targets, but were incentivised for district managers (i.e. antibiotics).

In terms of stock-outs, P4P was associated with a reduction in stock-outs of medicines and medical supplies (Table 4, Paper I). Particularly, most of the items which were found to increase significantly in availability, were also less likely to be out of stock. However, while there was no effect on the availability of vaccines and family planning medicines, we found a borderline significant reduction of stock-outs of these items (Table 4, Paper I). Similarly, while the effect on the availability of IPT and partograph was significant, there was no effect in terms of their stock-outs in the 90 days prior to the survey. P4P reduced the stock-out of medicines across the RMNCH continuum of care, and that of medical supplies benefiting mothers and newborns (Appendix 1b, Paper I). The effects of P4P on the availability of commodities were most pronounced for maternal, newborn and child medicines and reproductive health supplies.

Differential effects across facility subgroups

The overall effect of P4P on reducing the stock-out of medicines was pro-poor, with the reduction in facilities serving the poorest population being 24.5 percentage points greater than that in facilities serving the least poor tercile (p=0.019). Specifically, the effects on the stock-outs of antimalarials, antibiotics and oxytocics were pro-poor; effects on antimalarial availability were also marginally pro-poor (Table 5, Paper I). Further, P4P had a greater effect on the availability of medicines and medical supplies in facilities in rural districts than in urban districts. Similarly, the effect of P4P on the availability and stock-outs of antimalarials was greater in facilities in rural than in urban districts. The effect of P4P on the availability and stock-out of antihypertensives was greater in health centres and hospitals than in dispensaries. Further, there were no differential effects by facility ownership.

5.2 Paper II

Distribution of facility payouts

Generally, there was an increase in average payout scores between payment cycle 1 (50.1% of total potential payout) and cycle 7 (77.7%) (Table 3, Paper II). Facility payouts were pro-rich, because payout scores for facilities with least poor catchment populations were higher than for those with the poorest catchment populations. The pro-rich effect was supported by the positive equity gaps and concentration indices, as well as an equity ratio that is greater than one across all payment cycles (Table 3, column 5 –7). However, these pro-rich inequalities were generally stronger in early compared to later cycles (Table 3, Paper II). Apart from wealth subgroups, payout scores for facilities with higher drug availability at baseline were significantly higher than for those facilities with lower drug availability; payout scores for hospitals and health centres were higher than those for dispensaries (Table 4, Paper II). However, these payout gaps declined over time. The equity ratios between subgroups of other characteristics apart from wealth status were approximately one, which reflects a near equal distribution across subgroups.

Distribution of service coverage outcomes

At baseline, the institutional delivery rates and coverage of IPT2 during ANC were similar between most subgroups of facilities (Table 5, Paper II). However, baseline institutional delivery rates were higher among facilities that served least poor catchment populations than the poorest; while coverage of IPT2 was higher among facilities that served the poorest catchment populations than the least poor. Further, the coverage of IPT2 in the catchment area of dispensaries in the intervention arm was higher than the coverage around health centres and hospitals in the intervention arm; while dispensaries in comparison arm had lower levels of coverage in both outcomes (IPT2 and deliveries) at baseline than health centres and hospitals (Table 5, Paper II).

In terms of *differential effects*, there was a greater increase in institutional deliveries among facilities which started with lower baseline coverage than those with higher baseline coverage (by 13.0 percentage points, p=0.006) (Table 6, Paper II). There was also an evidence of a greater increase in institutional deliveries among facilities serving the middle wealth population than those serving the least poor wealth population (by 14.3 percentage points, p=0.004). In terms of the place of residence, there was a greater increase in institutional deliveries among facilities in rural than in urban districts (by 10.0 percentage points, p=0.030). There were no significant differential effects on the IPT2 coverage outcome across facility subgroups.

5.3 Paper III

The majority of individual and household characteristics were similar across intervention and comparison arms at baseline (Table 2, Paper III). The differences were noted for women in the intervention arm who were more likely to be married, non-farmers, and Muslim; and their households were more likely to be poor than their counterparts in the comparison arm.

Distribution of service utilisation at baseline

The institutional delivery rates in both arms were significantly lower for women in poorest and middle wealth households, and for women who were illiterate, farmers, with parity greater than one than for their counterpart women (Table 3, Paper III). By study arm specific, the rate of institutional deliveries was higher among intervention women with health insurance and from smaller households than among their counterpart women in intervention arm. In comparison arm, the rate of institutional deliveries was higher among urban women than rural women. However, the baseline uptake of IPT2 was generally similar across arms and population subgroups, except married women in the comparison arm, who were more likely to receive IPT2 than unmarried women (Table 3, Paper III).

Differential effects across population subgroups

P4P was associated with a significant increase in institutional delivery rates among women in the poorest and in the middle wealth status households, but not among women in the least poor households (Table 4, Paper III). However, when compared with the least poor subgroup, the effect of P4P was only marginally greater among women in the middle wealth status households only (p=0.094 for differential effect) (Table 4, Paper III). The effect of P4P on institutional deliveries was also significantly higher among women in rural districts compared to women in urban districts (p=0.028 for differential effect), and among uninsured than insured women (p=0.001 for differential effect). There were no differential effects of P4P on institutional deliveries among other subgroups, and no differential effects of P4P on the IPT2 outcome across any population subgroups (Table 4, Paper III).

6.0 Discussion

An approach of paying for results such as P4P has the potential to improve health system performance in LMICs as it rewards health providers based on their performance [22, 28]. P4P strategies are gaining popularity in many LMIC settings, but the evidence in terms of their effectiveness, cost and equity remain limited and mixed [7, 40, 58, 134-136, 232, 233]. This thesis, based on the P4P programme in Tanzania, contributes to a growing evidence base of P4P especially on the distributional effects of P4P in three aspects, as presented in three article papers of the thesis. The research approach in this thesis is quantitative and based on the programmes' impact evaluation. Since the robustness and validity of the results depends on how the research or evaluation was done, it is worth discussing the methodology of the study in more detail.

6.1 Methodological considerations

The methods used in this thesis rely on descriptive and regression analyses, which are both quantitative in nature. In quantitative research, particularly for impact evaluations, the validity of the findings is crucial to various stakeholders such as governments, policy makers and donors, as these findings have important implications for policy. Thus, it is worth to discuss the methodologies used in this thesis, and specifically based on two types of validity: internal and external validity.

6.1.1 Internal validity

Internal validity assesses the extent that a research study measured what it set out to measure in the study population [234]. In relation to impact evaluation, internal validity refers to an estimation of the true impact of a programme, that is net of all other potential bias and confounding factors, or that the comparison group represents the true counterfactual [216]. The internal validity can be undermined with the endogeneity problem for example, in many ways through biases (e.g. selection bias, information bias and/or confounding/ omitted variables). By definition, bias is a deviation of results or inferences from the truth or can be the processes leading to such deviation [234]. I then briefly discuss the type of bias in relation to this study.

Sample selection bias: This may refer to an absence of comparability between two study groups [234], especially in the context of impact evaluation. The lack of comparability may happen if the study participants/ sites are not randomly selected. Therefore, the question remains on how best to establish a robust counterfactual group in order to rule out selection bias and improve internal validity. In establishing unbiased causality, randomised experiments are often considered as credible approach to use [161, 216, 217].

For a successful randomisation, participants and non-participants exhibits similar characteristics before the programme, showing comparability. However, randomised experiments are rarely applied for policy evaluation [235], and in some cases they are argued to be unnecessary, inappropriate, impossible or inadequate (See Black [218] for a detailed discussion). Non-randomised experiments (e.g. quasi-experimental study designs) are commonly used in policy evaluation, especially where randomisation is not possible [161, 216, 219, 235, 236]. Although quasi-experimental study designs can mimic randomisation for causal inference, they rarely rule out completely the risk of encountering selection bias.

The Tanzanian P4P evaluation may potentially suffer from selection bias given the nature of the study design (i.e. controlled before and after design). The randomisation was not possible in Tanzania because the government selected an entire region to start implementing the P4P programme, and partly due to political reason as previously explained (See section 4.4). However, in order to establish a reasonable comparison group, we identified neighbouring districts which were comparable to the intervention districts on key variables such as: poverty and literacy rates, the rate of institutional deliveries, infant mortality, population per health facility, and the number of children under one year of age per capita [214]. Additionally, the DID regression-based approach was used in this thesis as a preferred method in a controlled before and after study design in order to remove the selection bias. This method compares the changes in outcomes over time between study arms, and thus accounts for any differences between study arms that are constant over time. As it uses regression-based methods, the *observed covariates* that are different between study arms are easily controlled for. The DID method also assumes *unobserved factors* are constant over time (time-invariant) because they are differenced out, and assumes there is no any *unobserved time-varying differences* exist between study arms [161, 216].

So, the identification of impacts through DID estimation relies on the key identifying assumption that the trends in outcomes would be parallel across study arms in the absence of the intervention, *i.e. parallel trends assumption* [161, 216]. That is to say, without an intervention, outcomes would need to increase or decrease at the same rate or trend in both study groups. This assumption can never be formally tested. However, this study supported this assumption by verifying that the pre-intervention trends in selected facility and household level outcomes were parallel [77, 215]. When verifying the assumption at household level, we used data of women surveyed in households. We first created a time trend variable based on the time of birth (single event per time) in the baseline survey data; then we ran a regression on outcomes (with longitudinal nature of the data) against time trend, intervention dummy, and their interaction to test for a divergence in pre-trends. A significant coefficient on the interaction term between intervention dummy and monthly time trend shows the difference/ divergence in the pre-trend. Four longitudinal outcomes during childbirth were used from household data (i.e. share of institutional deliveries, share of caesarean section deliveries, share of women who breastfeed within one hour of birth, and share of women who paid for delivery care). At the facility level, we similarly verified the *pre-intervention parallel trends assumption* based on monthly utilisation outcomes from patient register books prior to the start of the programme, that is from 2010 to 2012. The monthly utilisation outcomes include normal deliveries, vaccination data (polio, Measles and DPT), family planning visits, ANC visits, and outpatient visits. However, we were unable to verify the *pre-intervention parallel trends* based on other facility outcomes (e.g. the availability and stock-out of medical commodities) for which we had no data prior to the baseline survey.

Omitted variable bias/ confounding: This type of bias may happen when some variables are omitted from the analysis while they correlate with the outcome variable or other covariates in the model [237, 238]. In this thesis, we were fortunate that the data sources used for analysis captured a rich set of potential covariates/ confounders as observed factors. Through the DID regression model, all observed covariates were controlled for.

Yet, we cannot rule out the possibility of unobserved factors/ confounders due to measurement difficulties. The unobserved factors as omitted factors can either be fixed or varying over time. As previously mentioned, with a DID regression analysis, unobservable factors are assumed to be fixed (time-invariant) following the parallel trend assumption, and can be differenced out [161, 216]. This assumption seems to appeal with panel data [238]. This thesis used panel data at the facility level, but not at individual and household level. This is because the household survey was not necessarily performed on the same set of households over time, and rather their selection was based on whether the household had a woman who delivered in the 12 months prior to the survey at each round. Following the nature of panel data at facility level, we therefore applied facility-fixed effects estimation in order to control for observed and unobserved *time-invariant factors* that are heterogeneous across facilities; and similarly applied a year-fixed effects estimation to control the year-specific characteristics common to all observations, respectively.

Information bias: This refers to a systematic (non-random) measurement error which reflects inaccurate data reported or measured from respondents or participants [239, 240]. This type of bias is often observed during data collection [241], for example, when conducting surveys either at facility or household levels. Information bias may occur, for example, when the respondents do not know the exact answer to the survey questions, but they still provide answers [239]. It can also happen when respondents decide to either over-report or under-report the information relative to actual information. Typical potential sources of information bias include *recall bias* and *social desirability bias*. The sources of bias can range from how the interviews were conducted to what information was given and recorded. In this thesis, the information biases might have happened when measuring the outcomes and/or covariates through facility and household surveys. Respondents can inaccurately recall the past experience or information when asked (recall bias), but this depends on the length of the reference period. Thus, this bias is less likely in this study as a reasonable reference period was used that minimised the problem, i.e. a year to recall the experience on maternal service utilisation. In addition, social desirability bias refers to inaccuracy in reporting self-reported events with respect to social desirability [242]. If self-reported event is socially undesirable individuals are more likely to under-report, and potentially over-report the event that perceived socially desirable. Similarly, sensitive or personally threating events or behaviours are often under-reported. In this study, social desirability bias may seem less of an issue, since both facility and household surveys asked questions which were not sensitive, and facility survey involved mostly direct observations. Additionally, data collectors were fully trained in many aspects that ensures minimal measurement error; for example on data entry approach, on ensuring privacy and confidentiality, and on establishing an adequate rapport with respondents as these makes respondents comfortable to reveal valid responses [243].

Other methodological limitations

The findings of convergence in facility performance for institutional deliveries between worse-off and better-off performers, and convergence in utilisation of institutional deliveries between worse and better-off populations should be interpreted with caution. An initial interpretation might show that P4P improves performance and service utilisation among the worse-off providers and service users, respectively. However, these results might also reflect a regression to the mean principle[6] (a random fluctuation rather than a true causal effect). To disentangle the two hypotheses, the analyses may need longer term observations of data [244] or randomised experiment data [245, 246].

Paper II used proxy measures from a household survey based on a random sample of 20 households per facility to model service coverage outcomes as facility's performance outcomes and to measure the wealth status of the facility catchment population. To proxy facility-level outcomes based on household data raises two concerns for discussion. First,

[6] Regression to the mean (RTM) is a statistical phenomenon that occurs when repeated measurements are made on the same subject or unit of observation. It happens because values are observed with random error (Barnett et al [244]).

the proxy was from a sample of 20 households which may seem as not representative of the entire facility catchment population of women who delivered in the previous 12 months prior to the survey. However, a random sample of 20 households was reasonable and practical in this study because of two reasons: (i) the 20 households were reached based on a sample size calculation [214], and (ii) similar P4P evaluation studies have used sample sizes within that range to estimate the effects of P4P (e.g. 13 households in Rwanda [132] and 20 households in Cameroon [24]). A second concern was that sampled households were assumed to have used the health services from their nearest facility. This assumption might not always hold for services that are measured through household survey like institutional delivery, since has been associated with a high rate of client's bypassing the nearest facility [183, 184]. However, even the use of facility-based data from patient register books (though unreliable with several concerns like incompleteness) revealed an increase in the number of normal deliveries due to P4P in Tanzania, which is consistent with the finding from the household survey [77].

A further limitation was in *paper I* that several items of medical commodities were used when assessing the effect of P4P on the availability and stock-out rates of medical commodities. In terms of inference, assessing an impact of a programme on many items reflects multiple hypotheses testing which could potentially lead to a Type I error, i.e. of rejecting the null hypothesis when it is true [247]. However, this study reduced the risk of this error by generating composite scores for subgroups of commodities.

Furthermore, in *paper III*, the study may have been underpowered to detect the effect of P4P in some groups, for example among insured women and urban residents, possibly due to the smaller sample sizes for subgroups [248, 249]. The results of differential effects on deliveries by wealth status, health insurance and place of residence, were also slightly not consistent across all analytical specifications used in robustness checks (i.e. non-linear model, and a model that reports standard errors clustered at the district level). However, all analytical specifications consistently showed the lack of significant differential effects on deliveries for other subgroups of social determinants (i.e. marital status, age, education, occupation, religion, parity, and household size), and the lack of significant differential effects on IPT2 overall.

6.1.2 External validity

External validity means that results observed in one population or setting can be generalised to others [241, 250]. For an impact evaluation, external validity means that the impact estimated in the evaluation sample can be generalised to the population of all eligible units [216]. External validity is of utmost importance especially when the research findings are used to inform policy in the wider population of interest [216].

The findings of the evaluation of P4P scheme in Pwani region, Tanzania, can easily be generalised across Pwani region. This is because all hospitals, health centres, and non-public dispensaries offering RMNCH services together with a sample of public dispensaries were included in the study sample (i.e. 46% of all facilities in Pwani were included in the study) [214]. However, the generalisability of findings to other regions within Tanzania is more questionable as the performance of facilities, and level of service utilisation varies across regions. For example, according to the DHS data, Pwani region performed above the national average in most of the RMNCH indicators [168]. Pwani region is also next to the capital city of Dar es Salaam, where the MoHSW and the MSD responsible for distributing drug and supplies are located. This might potentially enhance the performance in Pwani compared to other regions in Tanzania. However, other regions in Tanzania especially those with lower performance have more scope for improvement if exposed with performance incentives. In order to learn and incorporate diversity across regions within Tanzania, the roll out of the scheme (RBF) started with Pwani region and moved to regions in Lake zone, almost in the northern part of Tanzania. Further assessment on performance variation based on the RBF roll out is needed, and this can be done since I am involved in the evaluation of the RBF roll out in Tanzania.

Moreover, the effects of P4P in Tanzania are not easily generalizable to other countries, due to the context-specific differences across settings. Some contextual factors influence the introduction, design, implementation and effectiveness of P4P schemes [58, 251-257]. For example, the contextual variations in institutional, social, political, cultural, organisational set-up and policy environment, may lead to varied implementation progress, providers' responses to P4P incentives and eventual effectiveness of the scheme. The effectiveness of the scheme, for instance, can be enhanced more in settings that have demand-side policies to reduce the barriers to access care (e.g. fee exemptions, health insurance, or conditional cash transfers) [128, 157]. Similarly, positive effects of P4P are more likely in settings with favourable health system structure, which may include functioning information systems, adequate supply of medical commodities, and adequate financial and human resources.

6.2 Discussion of the main findings

This section discusses the findings of this thesis in relation to other theoretical and empirical literature on performance based payments.

6.2.1 Can P4P improve the structural quality of care?

P4P is increasingly being applied in many settings to improve both health service quantity and quality. The Tanzanian P4P pilot scheme considered in this thesis may seem focused much on improving service use, with limited attention to quality of care, because the latter was not explicitly incentivised. Note that quality of care has three attributes namely *structural*, *process* and *outcomes* [73]. In Tanzania, the evidence of P4P effects on process quality has been documented [77], but with mixed findings. The effects of P4P on structural quality are presented in this thesis, but the P4P effects on health outcomes remains limited in Tanzania like in most developing countries.

In terms of structural quality, this study found that P4P was associated with an increase in the availability of essential drugs and supplies, but there was no effect on the availability of functioning equipment. P4P was also associated with a significant reduction in the stock-out of essential drugs and supplies. However, this study provides a partial assessment of the effects of P4P on structural quality of care, because other aspects of structural quality (e.g. human resources, organisation structure, and physical infrastructures [73]) were not considered due to a lack of data.

How does these results compare to others?

A recent review by Das et al [134] concluded that P4P is not effective in improving structural quality of care in LMICs, which is contrary to our findings. This conclusion was based on three studies on structural quality that were published between 1990 and 2014, and specifically from a study in Burundi and two studies in the DRC. A study in Katanga province of the DRC found negative effects on structural quality index, which includes drugs, vaccines and equipment [142]. Other two studies considered have used subjective measures (i.e. patients' perceptions) on drug availability. It was revealed that patients in Burundi perceived there were no P4P effects on drug availability [138], while patients in South Kivu province of the DRC perceived an improvement in drug availability [141]. Apart from such studies in a review, there are other studies being published recently that can be compared with our findings. For example, a study from Afghanistan found that P4P had no effects on the availability of drugs and equipment [145]. The finding that P4P improved the availability of essential drugs in Tanzania is consistent with recent evidence from P4P programme in Malawi [151]. However, contrary to the evidence from Tanzania, P4P improved the availability of functioning equipment in Malawi [151] and Cameroon [24].

The variation in results across settings could partly be explained by differences in programme designs. While the Tanzanian P4P programme directly incentivised the district managers to ensure drugs availability for their facilities, this was not clearly the case in other settings. District managers in Tanzania were incentivised because of their role in the process of procurement and supply of medical commodities to facilities. In Malawi, however, there were incentives to district managers' that were tied to equipment maintenance and management of drug supply across facilities [151]. These incentives to district managers explains the similarities of P4P effect between Malawi and Tanzania. In the DRC, however, facilities could channel a percentage of their bonus to districts to support the functioning of the districts [258]. Further, while up to 25%, 30% and 50% of the bonus payment could be used to procure drugs or for facility improvement in Tanzania, Malawi and Burundi, respectively [77, 138, 151], this was not clearly the case in the other settings.

How P4P can improve the availability of medical commodities?

According to the United Nations commission on life-saving commodities, P4P is considered as a strategy to improve access to life-saving commodities for RMNCH [221]. The question remains on the mechanisms through which P4P programmes can affect the availability of drugs and supplies. In Tanzanian, we conceptualised the pathways in two ways: direct and indirect pathway. The *direct pathway* refers to when the P4P programme directly incentives the availability of commodities. P4P in Tanzania incentivised district managers to reduce essential drug stock-out rates among facilities in their district, and similarly the P4P in Malawi tied incentives to equipment maintenance and management of drug supply across facilities [151]. In Tanzania, district managers are in the chain of procurement and supply of commodities to the facilities, and therefore they can efficiently influence this process in order to limit stock-outs. District managers also frequently visit facilities for supportive supervision and data verification, which creates an avenue to discuss or report on and deal with stock-outs. Further, the *indirect pathway* can be in two parts –either by incentivising the provision of commodities/ drugs, or through additional financial resources as bonus payments that can be used to procure drugs and supplies. In the case of the former, for example, P4P incentivised the provision of IPT during ANC, and therefore indirectly incentivised the availability of IPT stocks at facilities. In addition, the extra resources provided to facilities by P4P could be used to procure essential commodities which are out of stock or equipment. This pathway will only work where P4P payments are in part paid to facilities, and providers have autonomy in how they use these funds [22, 33].

It is also important to note that the Tanzanian P4P scheme did not affect the availability of equipment and some of the drugs. One potential reason is that some of the items –e.g. vaccines, antiretrovirals and family planning items, are procured through donor funding or vertical programmes, meaning they are not within the direct control of providers [182, 201, 202]. Also the higher level of availability for vaccines and family planning at baseline may have limited the scope for further improvement. The lack of effect on equipment availability may be due to the lack of incentives attached to equipment availability at the facility or the district level. The cost of equipment is also higher than that of many drugs and supplies, which may have deterred facilities from such investments. Thus, it seems incentivising managers on equipment availability might have served to improve this outcome.

How are the effects of P4P on medical commodities differed across facilities?

This is the first study to examine the heterogeneity of the effect of P4P on medical commodities, despite the importance of assessing distributional effects within program evaluation [65, 162]. It was important, for example, to assess whether facilities serving the poorer and rural populations improved most, since these populations face a greater burden of out-of-pocket payments due to drug stock-outs [74, 160]. The finding shows that the effects of P4P on drugs were generally stronger among facilities serving poorer and rural catchment population. These pro-poor and pro-rural effects may reflect the potential scope for improvement among facilities in poor and rural settings, as they performed poorly in drug availability at baseline than their counterparts. They further suggest that facilities in rural and poor communities responded strongly to P4P incentives in a bid to earn performance payouts for investing in reducing stock-out rates. Since district managers were paid based on performance of their facilities, maybe they focused much on strengthening the poorly performing facilities (e.g. in poor and rural areas) by for example addressing the issue of their stock-out rates to enhance overall district performance. Generally, the pro-poor effects on drugs are encouraging as are the pro-rural effects and these are consistent with UHC goals in the Sustainable Development Goal three.

Further, despite the differing procurement and supply systems in public and non-public sectors in Tanzania, the effects did not differ by facility ownership status. This might be due to the fact that non-public facilities often rely on procurement and supply systems of the MSD that public facilities also use. Effects also did not differ across facility level of care, regardless of the fact that primary care facility such as dispensaries are often worse-off in terms of drug availability compared to health centres and hospitals [188, 222, 223].

6.2.2 Does P4P increase or reduce performance inequalities?

Most countries especially in LMICs are providing performance incentives to health providers to improve the health system's performance [22, 28, 259]. The intention is logical and can improve efficiency and possibly equity. However, the equity dimension across providers if not well monitored and get worse may also worsen the pre-existing system performance. Given that health providers are not similar, then heterogeneous responses to incentives are expected when paying based on providers' performance. The evidence of whether a P4P programme leads to heterogeneous performance which may increase or reduce performance inequalities among providers/ facilities is limited in LMICs[7], despite the substantial variation in health facility readiness to deliver services in this context [167]. However, there is only one study as an exception which is from Rwanda [70]. This study has recently assessed the heterogeneous facility performance by baseline levels of quality, but they neither used other facility- and area-based characteristics to assess the heterogeneous performance, nor assessed how payouts were distributed across facilities of differing characteristics. Hence, a sub-study in this thesis contributes in these lacking aspects on heterogeneous facility performance on service coverage and payouts across facilities of differing characteristics.

The study findings showed that there were inequalities in the distribution of facility payouts which favoured the better-off facilities, but these inequalities in payouts declined over time. Note that P4P payouts reflects the performance in all incentivised indicators, and therefore the better-off facilities were better able to perform at the beginning and earned larger payouts than the worse-off facilities who seem to improve over time as well (showing convergence). The performance on the coverage of institutional deliveries was greater among facilities with initially lower levels of coverage, with middle wealth catchment populations, and located in rural areas than their counterpart facilities. These greater improvements among the worse-off facilities is partly due to an initial large scope for improvement among the worse-off facilities compared to their counterparts. Further, the performance on the coverage of antimalarials provision was similar across facilities.

How does these results compare to others?

As there is only one study from LMICs on supply-side heterogeneous performance, the findings from Tanzania can largely be compared with findings from HICs. However, it is

[7] This evidence is useful because an increase in performance inequality reflects inequality in payments, that may widen the resource gap and eventually increase inequality in healthcare provision between providers (Chien et al [66], Blustein et al [125]).

important to note that the programme design, context, and types of incentivised indicators differs across settings. In Rwanda, Sherry et al [70] found that facilities in the middle of the baseline quality distribution generally improved most across a broader range of rewarded services. This finding is consistently supporting the role of baseline facility characteristics in influencing facility performance, as in Tanzania both baseline facility- and area-based characteristics were considered for assess heterogeneous performance.

Further, the convergence in performance payouts over time is partly consistent with the “inverse equity hypothesis”[8] [71]. A study from the United States also found hospitals treating wealthier populations initially received higher incentive payments than hospitals serving poorer populations, but with a declining trend in payout inequalities over time [68]. In Tanzania, lower performers at baseline improved most in terms of institutional deliveries which enhanced convergence in performance, and this is consistent with P4P studies on quality improvement in the United Kingdom [67], in Canada [95] and in the United States [53, 55, 115, 120, 125]. The finding that facilities serving middle wealth populations with initial low coverage on deliveries improved more over time than those serving the least poor populations, is different to that reported in the United States and United Kingdom in relation to quality improvements [66, 67, 109, 116, 125, 260-262]. These studies found that providers serving low-income populations performed initially less well on quality improvement but improved most over time than those serving high-income populations. The pro-rural performance on institutional deliveries observed in Tanzania, differs with a finding of no association between performance and rural/urban location in the United States [263], and the findings from the United Kingdom showing less effect in rural than in urban areas [261, 262].

What are the potential mechanisms to affect performance inequality?

[8] The hypothesis suggests that better-off groups will initially benefit from a new intervention and widen inequalities, but over time the worse-off can eventually catch up.

Despite growing interest in performance incentives especially in LMICs, a lot remains unknown particularly regarding the exact mechanisms through which such schemes bring about change and how programme design affects this [58, 215, 264]. Establishing and testing a theory of change for P4P programmes remains crucial in understanding programme's impacts. This study hypothesised that the effects on performance inequalities will depend on existing structural factors, and how the incentives are designed. Performance inequalities can either be enhanced, reduced or remain unchanged between worse-off and better-off providers.

Incentive design pathway**:** The performance inequality may arise depending on how payouts are offered with respect to target setting. Despite several ways of target setting, the Tanzanian P4P programme used: (i) multiple thresholds targets based on baseline performance (e.g. for institutional deliveries), and (ii) single threshold targets (e.g. IPT2 coverage) irrespective of baseline performance. Based on the two designs in Tanzania, evidence shows that lower baseline performers had greater improvements in performance on institutional deliveries (with multiple thresholds), but the performance on IPT2 coverage (single threshold) was similar across all facilities. A greater improvement in institutional deliveries among lower baseline performers is possibly because of the design that used multiple thresholds targets. Some literature suggest that multiple thresholds targets can enhance convergence in performance as they account for baseline performance and provide incentives for lower performers to catch up [27, 49, 53].

From a theoretical perspective, a single threshold target, as used for IPT2 in Tanzania, can enhance divergence in performance as it fails to account for baseline performance [27, 49, 50, 53, 56, 57]. However, the finding in Tanzania is different from a theoretical prediction, as it shows similar improvement on IPT2 coverage due to P4P across facilities. Other factors beyond the incentive design for IPT2 may possibly explain such a finding of similar performance. Contextual factors, for example, of almost universal coverage of one ANC visit in Tanzania (i.e. more than 98%) [77, 168] may have led to minimal effort needed for most facilities to achieve the target for IPT2. Also the nature of the IPT2 indicator which is a content of care (within the control of the provider) as opposed to service use (which requires a change in household behaviour), possibly facilitated a similar response among providers.

In other settings, some studies have shown how programme design/target affects inequalities in performance outcomes. In the United States, for example, a reduction in payout inequalities was attributed to a change in the design of the scheme from rewarding top performers to rewarding for improvement where all providers were likely to receive an incentive payment [68]. Also the convergence in performance on quality improvement was partly linked to a design with multiple thresholds targets in the United Kingdom [67] and in Canada [95], and to a system that rewards the highest performers and penalised the lowest performers in the United States [53, 55]. Theoretically, for a design with multiple or single threshold targets, both convergence and divergence are possible outcomes but also depending on the structural context.

Structural effect pathway: Structural factors provide another pathway through which P4P may affect performance inequalities across providers [65, 69, 70]. This study considered structural factors such as facility characteristics (ownership status, level of care, and the availability of medical inputs) and area-based characteristics (wealth status of the catchment population and rural/urban location). The variation of these factors across facilities may explain performance inequalities at baseline and over time. In Tanzania, some structural factors were significantly associated with performance inequality. For example, a greater performance on the coverage of institutional deliveries was shown among facilities in middle wealth catchment population and in rural districts with initially low coverage than their counterpart facilities. It seems the P4P incentives were stronger among the worse-off providers as they improved more on delivery care coverage than their counterparts. In contrast, the payouts distribution favoured the better-off facilities initially (higher level facilities, in wealthier catchment areas, and with more medical commodities), but these inequalities/ payout gaps declined over time.

The convergence pattern on the coverage of institutional deliveries and on bonus payouts stands out as an encouraging finding, because the worse-off providers possibly responded positively to incentives and enhanced performance. Note that the two findings (on service coverage and payouts) cannot be compared directly because facility improvements were assessed *only on two incentivised services* (delivery care and IPT2 provision), but payout outcome reflects a total performance on *all the incentivised services*.

To this end, the hypothesised pathways to reduce or enhance performance inequalities cannot be conclusively confirmed by this study and remains open for discussion and future research. This is because multiple thresholds enhanced convergence on deliveries while single threshold did not lead to divergence in performance as hypothesised, and not all structural factors hypothesised associated with performance inequality. However, there is an indication that P4P can reduce performance inequalities by enhancing convergence in performance. Therefore, this study suggests that both the incentive design on target setting and structural factors matters for performance inequality.

6.2.3 Do the benefits of P4P reach the worse-off populations?

From a demand-side perspective, it is clear that more research is needed to monitor and evaluate how the benefits of P4P are distributed across a wider range of population subgroups. Providers' responses to incentives may affect not only the average P4P effects but also heterogeneous P4P effects across populations. For example, in order to meet performance targets, providers may extend services to underserved groups and enhance equity [22, 33], or may focus on easier-to-reach population and enhance inequity [29]. Therefore, the assessment of heterogeneous P4P effects across populations is crucial, and must look beyond economic status subgroups as commonly reported, but rather incorporate a wider range of social determinants subgroups in order to capture a broader range of population subgroups that may drive heterogeneous effects [85, 265]. This type of evidence is key to inform universal access policies [265-267].

The heterogeneous results from this study show that P4P increased institutional deliveries more among women in the middle wealth status households, among the uninsured, and among women living in rural areas than among wealthier, insured, and urban residing women. However, there were no any heterogeneous effects of P4P on institutional deliveries across other population subgroups of social determinants (e.g. education, occupation, age, parity). Thus, population wealth status, health insurance status and place of residence were the main drivers of demand-side heterogeneous P4P effects on institutional deliveries. Moreover, the effect of P4P on the uptake of antimalarial drugs was equally distributed across population subgroups.

How does these results compare to others?

While most studies on demand-side heterogeneous effects of P4P have disaggregated the effect across population economic status particularly in LMICs, this study used a broader range of social determinants subgroups.

In terms of wealth status, this study found that institutional deliveries increased more among middle wealth women than least poor women. This pro-middle wealth effect of P4P on institutional deliveries, as an indication of being pro-poor, is contrary to the pro-rich effect on deliveries reported in Burundi [133], Rwanda [157] and Cambodia [128]. The pro-rich effect in Cambodia was attributed to the lack of effective demand among the poorest women due to user fees [128]; whereas in Burundi it was attributed to other costs like transport because the user fees for deliveries were removed prior to P4P [133, 268]. However, other study in Rwanda and Burundi revealed a different pattern of results. For example, a pilot study in Burundi [138] and a study using DHS data in Rwanda [159] found similar P4P effect on deliveries across household's socioeconomic groups; and the results from Rwanda were attributed to low and uniform coverage of services at baseline. In other settings such as the DRC, providers implementing P4P negotiated user fees with communities and raised revenues without hurting the poorest [141]. However, the equity effects of this approach in the DRC were not assessed empirically. Additional evidence of a pro-poor effect of P4P has been shown on immunisation services in Burundi [133], and on quality of care improvement in the United Kingdom [31, 67, 105, 109, 110]. Generally, from the above studies, it seems the effect of P4P on socioeconomic equity remains mixed across settings and across targeted services.

Regarding the place of residence, this study found a greater increase in the institutional deliveries in rural than in urban populations. This finding differs from the P4P scheme in Rwanda that led to similar improvement in institutional deliveries between rural and urban populations [159]. However, the number of urban clusters in Rwanda (which were few compared to rural clusters) were thought to limit the power to detect the heterogeneity of P4P effect by place of residence [159], while our study had a slightly higher number of urban clusters compared to Rwanda (i.e. 28 versus 22 urban clusters). In the United Kingdom, the effect of P4P on quality of care was greater in urban areas than in rural areas [261, 262], while the effect of P4P on quality of care was similar between rural and urban areas in the United States [263]. Although the classification of rural–urban is context specific [269], the evidence on the heterogeneity of P4P effect between rural and urban populations remain limited and mixed across settings.

The Tanzanian P4P was associated with an increase in institutional deliveries among uninsured women, whereas a greater effect on institutional deliveries was found among women with health insurance in Rwanda [157] and a maternity care voucher in Cambodia [128]. The findings from Rwanda and Cambodia were attributed to reduced financial barriers to access care [128, 157], and this could be the case with a stronger enforcement of fee exemptions in Tanzania [77]. Efforts to reduce the demand-side barriers seem to enhance the equity effects of P4P in healthcare access and use. However, another study in Rwanda based on DHS, as nationally representative data, found a similar increase in deliveries due to P4P irrespective of women's health insurance status [159].

In Tanzania, there was also a similar distribution of institutional delivery rates and IPT2 uptakes across age groups prior to P4P, and the effect of P4P was equally distributed across age groups. In contrast, P4P studies in HICs found inequalities in quality of care across age groups that existed and persisted after the introduction of P4P [31, 105, 109, 110].

What are the potential mechanisms for P4P to preferentially benefit disadvantaged populations?

This study found that the use of institutional deliveries improved most among women in middle wealth households, uninsured, and in rural areas than their counterpart women. This was potentially due to the increased adherence to user fee exemption policy among public facilities, and also due to the improved availability of drugs, minimising the need to pay for drugs in private pharmacies [10, 74, 76, 77, 96, 212, 215, 270-272]. Greater improvements on institutional deliveries among the worse-off women suggest that these women were likely to have been more responsive to a change in healthcare costs [86, 87]. Such a responsiveness is consistent with demand theory, and previously explained conceptual framework that incentives stimulated a supply-side response to reduce financial barriers to access care which in turn stimulated the demand-side response on service utilisation mostly among the disadvantaged population subgroups.

The greater effect of P4P on institutional deliveries among uninsured women in Tanzania is also because baseline institutional delivery rates were higher among insured compared to uninsured women in the intervention arm, which gave a large scope for improvement among uninsured women. A further possibility would be that uninsured women were more responsive to reduced healthcare costs compared to insured women who were already covered. It is also likely that the statistical power to detect the effect among women with health insurance was limited because fewer women are insured in Tanzania[9] [204], compared to other countries like Rwanda [273, 274]. Further, the pro-rural effects of P4P on institutional deliveries in Tanzania, seem to reflects the fact that rural providers strongly responded to incentives (as found in paper II) and eventually triggered the demand of institutional delivery among women in those areas. This is likely because the effect of P4P on deliveries was greater among the worse-off women (e.g. poorer and uninsured)[158], and these women often reside in rural areas [275]. Despite an indication that P4P may benefit more the worse-off populations (poorer, uninsured and rural residents) especially on delivery care, further research is needed to better understand the demand-side heterogeneity of P4P effects.

6.3 Policy implications

This work contributes to reducing the knowledge gap in terms of P4P effects on medical commodities, and eventual heterogeneous effects of P4P on medical commodities, facility performance outcomes, and on service utilisation in LMICs. This study offers a number of policy implications. The finding that P4P improved the availability of drugs and supplies was linked to a design element of incentivising district managers to limit drug stock-outs in Tanzania. This study highlights the importance of incentivising district managers, given their role in doing supportive supervision, and in procurement and supply of medical commodities. A further P4P design element to be reinforced is financial autonomy through bank accounts and ensuring discretion in the use of funds at the facility level. Overall health facility autonomy is an important design element of P4P and is hypothesised to improve provider’s performance. These highlighted design elements can

[9] About 9% of women were insured in this study (Binyaruka et al [77]).

be strengthened in the P4P roll out design in Tanzania, and similarly be applied in P4P designs in other settings to maximise programme impact.

This study also shows that the design of incentives in setting targets can affect performance inequalities and eventually benefit facilities differently, but also structural factors and the nature of performance indicators themselves also matter for performance and eventual payouts. On that regard, it is crucial to carefully consider the incentive design structures such that they do not lead to inequalities in performance and payouts, and avoid reinforcing the gaps in resources and service provision across facilities. The inequalities in performance and payouts should not favour only facilities that are better-off at baseline (i.e. better able to perform and needs little improvement for higher payout) and rather the scheme should incentivise the worse-off facilities to improve, benefit and catch-up. For instance, it is possible that paying based on improvement/ per additional service delivered (e.g. fee-for-service) may incentive facilities that are worse-off at baseline, and those with large scope for improvement, to improve and receive payments; as opposed to paying based on target attainment (achieve/ maintain at a threshold) that favours the better-offs at the onset as they are better able to achieve targets/ maintain above a threshold. Similarly, this study found that payout distribution favoured facilities that were better-off at the onset (i.e. facilities with more medical commodities, serving wealthier population, and higher level facilities). Therefore, equity bonuses for disadvantaged providers/ facilities should be considered to redress such inequalities in performance and payouts resulting from pre-existing structural challenges (e.g. geographical inaccessibility, low catchment population size, and poverty). It is also important to harmonise the capacity to deliver services prior to P4P through a facility readiness assessment study and potential quality boosting investments (e.g. through initial start-up financial support).

The Tanzanian P4P scheme reduced the chances of paying informal user fees through providers' stronger enforcement of an exemption policy and also the scheme improved the availability of drugs and supplies. These effects of P4P seem to have reduced the demand-side financial barriers to access care especially among the worse-off populations (e.g. poorer, uninsured and rural residents) and enhanced equitable access and use of services particularly for institutional deliveries. Based on these findings, policy makers should consider to strengthen or introduce other complementary measures to reduce demand-side access barriers which seem to reinforce the P4P effect on service use. Examples of such demand-side initiatives may include pre-payment health insurance schemes, cash transfers, exemptions and voucher schemes. Although contexts may vary, both insurance and voucher schemes have increased the effects of P4P on the use of institutional delivery care in Rwanda and Cambodia, respectively. On that regard, P4P is likely to be most effective at reducing inequities in service use in settings where they offer free health services (or with an exemption policy) or in settings with other demand-side schemes to enhance access to care. Demand-side initiatives that complement supply-side interventions are of utmost importance to ensure universal access to care and reduce any pre-existing inequities in service utilisation. Policy makers in LMICs should therefore discuss and formulate mechanisms to ensure universal access, and stimulate both demand for and supply of healthcare services (e.g. combining demand-side and supply-side incentives) in order to facilitate the move towards UHC goal in Sustainable Development Goals three.

6.4 Research implications

This PhD work shows the potential of a P4P scheme to increase the availability of drugs and supplies, and it expands the understanding of the heterogeneity of the P4P effects across facilities and populations subgroups. However, further research is needed as some questions remain unanswered.

First, since other aspects of structural quality (e.g. human resources, organisation structure, and physical infrastructures) are not considered in this study due to a lack of data, future research in LMICs should aim to capture P4P impacts on the overall structural quality measures. Furthermore, because most P4P evaluations are increasingly examining process and structural quality of care in LMICs, more research is needed to identify programme effects on all attributes of quality of care including health outcomes. To improve population health, as an ultimate goal of any health system, various programmes like P4P should be assessed whether they improve not only service utilisation but also quality of care in multidimensional sense.

The understanding of P4P heterogeneous effects is useful to inform the design and scale-up, therefore, more comprehensive evidence that may shed light on P4P pathways and mechanisms (theory of change) is needed as may also facilitate to open up the "black box" of P4P. Further, process evaluation through qualitative research that may explain the sources and mechanisms in the context of heterogeneous P4P effects is of utmost importance to supplement quantitative evidence.

Moreover, this study used only two target setting methods that were used in Tanzania, single and multiple thresholds, to assess the relationship between incentive designs and facility performance inequalities. Since the findings on this aspect remain mixed in this thesis, further studies are needed across settings to assess the influence of incentive designs (e.g. paying providers based on fee-for-service or thresholds targets) on performance inequalities.

In addition, from a theoretical perspective, P4P as a supply-side intervention which affects providers' behaviour may trigger the demand-side responses and improve both service quality and use. Therefore, further insights on how supply- and demand-side interventions/ programmes interact and complement each other to affect outcomes are needed.

Lastly, the average and heterogeneous effects of P4P were assessed in short term over a 13-month period in this study. Since supply-side responses to performance incentives (among providers) and demand-side responses to improved services (among service users) are not linear as they change over time, there is a need to monitor and assess the longer term average and heterogeneous effects of P4P to better understand these schemes over time.

6.5 Conclusion

The study findings show the potential of P4P in improving structural quality of care, through improved availability of medical commodities and stock-outs reduction. The findings also demonstrate the existence of some drivers of heterogeneity of P4P effects on the supply- and demand-side in a low- income country.

- The first paper carried out in this thesis found that P4P was associated with improvements in terms of drugs and supplies availability, especially among the worse-off facilities which serve poorer populations, and are located in rural areas.

- In the second paper, P4P increased the coverage rates of institutional deliveries more among the worse-off facilities which serve middle wealth populations, located in rural areas, and those with low performance initially. The coverage of antimalarials provision (IPT2) increased equally across facilities. Facility bonus payouts for all incentivised indictors were initially higher among higher level facilities like hospitals, better resourced facilities, and facilities serving wealthier populations, but these inequalities in payouts declined over time.

- In the third paper, P4P increased the rate of institutional deliveries more among women in middle wealth status households, among the uninsured, and among women living in rural areas than their counterparts. The uptake of at least two doses of antimalarial drugs (IPT2) increased equality across population subgroups.

The findings about the heterogeneity of the P4P effects across facilities and population subgroups have important implications for inequalities in facility performance and inequalities in healthcare use, respectively. Therefore, these heterogeneous effects of P4P, and of other financing programmes, should be monitored over time and similarly be considered to inform the scale-up and designing of health financing schemes in Tanzania.

References

1. WHO: **The World Health Report 2000: Health systems: improving performance**. In. World Health Organization, Geneva, Switzerland; 2000.
2. WHO: **Monitoring the building blocks of health systems: A handbook of indicators and their measurement startegies**. In. World Health Organization, Geneva, Switzerland; 2010.
3. WHO: **Strengthening Health Systems for Health Outcomes: WHO's Framework for Action**. In. World Health Organization, Geneva, Switzerland; 2007.
4. Mills A, Hanson K: **Expanding access to health interventions in low and middle-income countries: constraints and opportunities for scaling-up**. *Special issue of the Journal of International Development* 2003, **15**(1):1-131.
5. Mills A: **Health Care Systems in Low- and Middle-Income Countries**. *New England Journal of Medicine* 2014, **370**(6):552-557.
6. WHO: **The World Health Report: Working together for health**. In. World Health Organization, Geneva, Switzerland; 2006.
7. Witter S, Fretheim A, Kessy FL, Lindahl AK: **Paying for performance to improve the delivery of health interventions in low- and middle-income countries**. *The Cochrane database of systematic reviews* 2012(2):Cd007899.
8. Rowe AK, de Savigny D, Lanata CF, Victora CG: **How can we achieve and maintain high-quality performance of health workers in low-resource settings?** *Lancet (London, England)* 2005, **366**(9490):1026-1035.
9. Kutzin J: **Health financing for universal coverage and health system performance: concepts and implications for policy**. *Bull World Health Organ* 2013, **91**(8):602-611.
10. WHO: **The World Health Report: Health Systems Financing: The path to universal coverage**. In. Edited by Geneva. World Health Organization, Geneva, Switzerland: World Health Organization; 2010.
11. Anand S, Barnighausen T: **Health workers at the core of the health system: framework and research issues**. *Health Policy* 2012, **105**(2-3):185-191.
12. Lehmann U, Dieleman M, Martineau T: **Staffing remote rural areas in middle- and low-income countries: a literature review of attraction and retention**. *BMC Health Serv Res* 2008, **8**:19.
13. Chen L, Evans T, Anand S, Boufford JI, Brown H, Chowdhury M, Cueto M, Dare L, Dussault G, Elzinga G: **Human resources for health: overcoming the crisis**. *The Lancet* 2004, **364**(9449):1984-1990.
14. WHO: **Increasing access to health workers in remote and rural areas through improved retention: global policy recommendations**: World Health Organization; 2010.
15. Bangdiwala SI, Fonn S, Okoye O, Tollman S: **Workforce resources for health in developing countries**. *Public Health Reviews* 2010, **32**(1):296.

16. Kwesigabo G, Mwangu MA, Kakoko DC, Warriner I, Mkony CA, Killewo J, Macfarlane SB, Kaaya EE, Freeman P: **Tanzania's health system and workforce crisis**. *Journal of public health policy* 2012, **33**(1):S35-S44.

17. Leonard KL, Masatu MC: **Professionalism and the know-do gap: exploring intrinsic motivation among health workers in Tanzania**. *Health Econ* 2010, **19**(12):1461-1477.

18. Maestad O, Torsvik G: **Improving the quality of health care when health workers are in short supply**. *CMI Working Paper* 2008.

19. Miller G, Babiarz KS: **Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs**. In: *NBER Working Paper No 18932.* 2013.

20. Das J, Gertler PJ: **Variations in practice quality in five low-income countries: a conceptual overview**. *Health Aff (Millwood)* 2007, **26**(3):w296-309.

21. Eichler R, Levine R: **Performance incentives for global health: potential and pitfalls**: CGD Books; 2009.

22. Meessen B, Soucat A, Sekabaraga C: **Performance-based financing: just a donor fad or a catalyst towards comprehensive health-care reform?** *Bull World Health Organ* 2011, **89**(2):153-156.

23. Kutzin J: **A descriptive framework for country-level analysis of health care financing arrangements**. *Health Policy* 2001, **56**(3):171-204.

24. De Walque DB, Robyn PJ, Saidou H, Sorgho G, Steenland MW: **Looking into the performance-based financing black box: evidence from an impact evaluation in the health sector in Cameroon**. In.: The World Bank; 2017.

25. Oxman AD, Fretheim A: **Can paying for results help to achieve the Millennium Development Goals? Overview of the effectiveness of results-based financing**. *Journal of Evidence-Based Medicine* 2009, **2**(2):70-83.

26. Montagu D, Yamey G: **Pay-for-performance and the Millennium Development Goals**. *Lancet (London, England)* 2011, **377**(9775):1383-1385.

27. Eijkenaar F: **Key issues in the design of pay for performance programs**. *The European journal of health economics : HEPAC : health economics in prevention and care* 2013, **14**(1):117-131.

28. Witter S, Toonen J, Meessen B, Kagubare J, Fritsche G, Vaughan K: **Performance-based financing as a health system reform: mapping the key dimensions for monitoring and evaluation**. *BMC Health Serv Res* 2013, **13**:367.

29. Ireland M, Paul E, Dujardin B: **Can performance-based financing be used to reform health systems in developing countries?** *Bull World Health Organ* 2011, **89**(9):695-698.

30. Rosenthal MB, Fernandopulle R, Song HR, Landon B: **Paying for quality: providers' incentives for quality improvement**. *Health Aff (Millwood)* 2004, **23**(2):127-141.

31. Eijkenaar F, Emmert M, Scheppach M, Schoffski O: **Effects of pay for performance in health care: a systematic review of systematic reviews**. *Health Policy* 2013, **110**(2-3):115-130.

32. Roland M: **Linking physicians' pay to the quality of care--a major experiment in the United kingdom**. *N Engl J Med* 2004, **351**(14):1448-1454.
33. Fritsche G, Soeters R, Meessen B: **Performance-Based Financing Toolkit**. In. Washington DC: The World Bank; 2014.
34. Musgrove P: **Financial and other rewards for good performance or results: a guided tour of concepts and terms and a short glossary**. *Washington, DC: World Bank* 2011:12.
35. Eisenhardt KM: **Agency theory: An assessment and review**. *Academy of management review* 1989, **14**(1):57-74.
36. Ross SA: **The economic theory of agency: The principal's problem**. *The American Economic Review* 1973, **63**(2):134-139.
37. Holmstrom B, Milgrom P: **Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design**. *Journal of Law, Economics, & Organization* 1991, **7**:24-52.
38. Lemière C, Torsvik G, Mæstad O, Herbst CH, Leonard KL: **Evaluating the Impact of Results-Based Financing on Health Worker Performance**. In. The World Bank, Washington DC: HNP Discussion Paper: The World Bank, Washington DC; 2013.
39. Grossman SJ, Hart OD: **An analysis of the principal-agent problem**. *Econometrica: Journal of the Econometric Society* 1983:7-45.
40. Eldridge C, Palmer N: **Performance-based payment: some reflections on the discourse, evidence and unanswered questions**. *Health Policy Plan* 2009, **24**(3):160-166.
41. Arrow KJ: **Uncertainty and the Welfare Economics of Medical Care**. *The American Economic Review* 1963, **53**(5):941-973.
42. Renmans D, Paul E, Dujardin B: **Analysing Performance-Based Financing through the Lenses of the Principal-Agent Theory**. In.: Universiteit Antwerpen, Institute of Development Policy (IOB); 2016.
43. Robinson JC: **Theory and practice in the design of physician payment incentives**. *Milbank Q* 2001, **79**(2):149-177, iii.
44. Christianson JB, Knutson DJ, Mazze RS: **Physician Pay-For-Performance**. *Journal of general internal medicine* 2006, **21**(S2).
45. Mannion R, Davies HT: **Payment for performance in health care**. *Bmj* 2008, **336**(7639):306-308.
46. Dolea C, Adams O: **Motivation of health care workers-review of theories and empirical evidence**. *Cahiers de sociologie et de demographie medicales* 2005, **45**(1):135-161.
47. Mathauer I, Imhoff I: **Health worker motivation in Africa: the role of non-financial incentives and human resource management tools**. *Human resources for health* 2006, **4**(1):24.
48. Franco LM, Bennett S, Kanfer R: **Health sector reform and public sector health worker motivation: a conceptual framework**. *Social science & medicine* 2002, **54**(8):1255-1266.

49. Mehrotra A, Sorbero ME, Damberg CL: **Using the lessons of behavioral economics to design more effective pay-for-performance programs**. *The American journal of managed care* 2010, **16**(7):497-503.
50. Rosenthal MB, Dudley RA: **Pay-for-performance: will the latest payment trend improve care?** *Jama* 2007, **297**(7):740-744.
51. Maynard A: **The powers and pitfalls of payment for performance**. *Health Econ* 2012, **21**(1):3-12.
52. Emanuel EJ, Ubel PA, Kessler JB, Meyer G, Muller RW, Navathe AS, Patel P, Pearl R, Rosenthal MB, Sacks L *et al*: **Using Behavioral Economics to Design Physician Incentives That Deliver High-Value Care**. *Annals of internal medicine* 2016, **164**(2):114-119.
53. Rosenthal MB, Frank RG, Li Z, Epstein AM: **Early experience with pay-for-performance: from concept to practice**. *Jama* 2005, **294**(14):1788-1793.
54. Lagarde M, Powell-Jackson T, Blaauw D: **Managing incentives for health providers and patients in the move towards universal coverage**. In*: 2010*: 'Global Symposium on Health Systems Research'16-19 November 2010. Montreux, Switzerland.; 2010.
55. Lindenauer PK, Remus D, Roman S, Rothberg MB, Benjamin EM, Ma A, Bratzler DW: **Public reporting and pay for performance in hospital quality improvement**. *N Engl J Med* 2007, **356**(5):486-496.
56. Heath C, Larrick RP, Wu G: **Goals as reference points**. *Cognitive psychology* 1999, **38**(1):79-109.
57. Mullen KJ, Frank RG, Rosenthal MB: **Can you get what you pay for? Pay-for-performance and the quality of healthcare providers**. *The Rand journal of economics* 2010, **41**(1):64-91.
58. Renmans D, Holvoet N, Orach CG, Criel B: **Opening the 'black box' of performance-based financing in low- and lower middle-income countries: a review of the literature**. *Health Policy Plan* 2016, **31**(9):1297-1309.
59. Mayumana I, Borghi J, Anselmi L, Mamdani M, Lange S: **Effects of Payment for Performance on accountability mechanisms: Evidence from Pwani, Tanzania**. *Soc Sci Med* 2017, **179**:61-73.
60. Gertler P, Vermeersch C: **Using Performance Incentives to Improve Health Outcomes.** In: *NBER Working Paper No 19046.* 2013.
61. Frey BS, Jegen R: **Motivation crowding theory**. *Journal of economic surveys* 2001, **15**(5):589-611.
62. Deci EL, Koestner R, Ryan RM: **A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation**. *Psychological bulletin* 1999, **125**(6):627.
63. Lohmann J, Houlfort N, De Allegri M: **Crowding out or no crowding out? A Self-Determination Theory approach to health worker motivation in performance-based financing**. *Soc Sci Med* 2016, **169**:1-8.
64. Rosenthal MB, Frank RG: **What is the empirical basis for paying for quality in health care?** *Med Care Res Rev* 2006, **63**(2):135-157.

65. Markovitz AA, Ryan AM: **Pay-for-Performance: Disappointing Results or Masked Heterogeneity?** *Med Care Res Rev* 2016.

66. Chien AT, Wroblewski K, Damberg C, Williams TR, Yanagihara D, Yakunina Y, Casalino LP: **Do physician organizations located in lower socioeconomic status areas score lower on pay-for-performance measures?** *J Gen Intern Med* 2012, **27**(5):548-554.

67. Doran T, Fullwood C, Kontopantelis E, Reeves D: **Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework**. *Lancet (London, England)* 2008, **372**(9640):728-736.

68. Ryan AM, Blustein J, Doran T, Michelow MD, Casalino LP: **The effect of Phase 2 of the Premier Hospital Quality Incentive Demonstration on incentive payments to hospitals caring for disadvantaged patients**. *Health Serv Res* 2012, **47**(4):1418-1436.

69. De Allegri M, Bertone MP, McMahon S, Mounpe Chare I, Robyn PJ: **Unraveling PBF effects beyond impact evaluation: results from a qualitative study in Cameroon**. *BMJ global health* 2018, **3**(2):e000693.

70. Sherry TB, Bauhoff S, Mohanan M: **Multitasking and Heterogeneous Treatment Effects in Pay-for-Performance in Health Care: Evidence from Rwanda**. *American Journal of Health Economics* 2017, **3**(2):192-226.

71. Victora CG, Vaughan JP, Barros FC, Silva AC, Tomasi E: **Explaining trends in inequities: evidence from Brazilian child health studies**. *Lancet (London, England)* 2000, **356**(9235):1093-1098.

72. Castro-Leal F, Dayton J, Demery L, Mehra K: **Public spending on health care in Africa: do the poor benefit?** *Bull World Health Organ* 2000, **78**(1):66-74.

73. Donabedian A: **The quality of care: how can it be assessed?** *Jama* 1988, **260**(12):1743-1748.

74. WHO: **Equitable access to essential medicines: a framework for collective action**. In. Geneva: World Health Organization; 2004.

75. Quick JD, Boohene N-A, Rankin J, Mbwasi RJ: **Medicines supply in Africa**. In.: British Medical Journal Publishing Group; 2005.

76. Binyaruka P, Borghi J: **Improving quality of care through payment for performance: examining effects on the availability and stock-out of essential medical commodities in Tanzania**. *Trop Med Int Health* 2017, **22**(1):92-102.

77. Binyaruka P, Patouillard E, Powell-Jackson T, Greco G, Maestad O, Borghi J: **Effect of Paying for Performance on Utilisation, Quality, and User Costs of Health Services in Tanzania: A Controlled Before and After Study**. *PLoS One* 2015, **10**(8):e0135013.

78. Ellis RP, McGuire TG: **Hospital response to prospective payment: moral hazard, selection, and practice-style effects**. *J Health Econ* 1996, **15**(3):257-277.

79. Gwatkin DR: **How much would poor people gain from faster progress towards the Millennium Development Goals for health?** *Lancet (London, England)* 2005, **365**(9461):813-817.

80. Le Grand J: **Equity versus efficiency: the elusive trade-off**. *Ethics* 1990, **100**(3):554-568.
81. Andersen R, Newman JF: **Societal and individual determinants of medical care utilization in the United States**. *The Milbank Memorial Fund quarterly Health and society* 1973, **51**(1):95-124.
82. Andersen R: **A behavioral model of families' use of health services**. *A behavioral model of families' use of health services* 1968(25).
83. Solar O, Irwin A: **A conceptual framework for action on the social determinants of health. Social Determinants of Health**. In: *Discussion Paper 2 (Policy and Practice).* 2010.
84. Marmot M, Friel S, Bell R, Houweling TA, Taylor S: **Closing the gap in a generation: health equity through action on the social determinants of health**. *Lancet (London, England)* 2008, **372**(9650):1661-1669.
85. CSDH: **Closing the gap in a generation: health equity through action on the social determinants of health**. In.; 2008.
86. Gertler P, Locay L, Sanderson W: **Are user fees regressive?: The welfare implications of health care financing proposals in Peru**. *Journal of econometrics* 1987, **36**(1-2):67-88.
87. Litvack JI, Bodart C: **User fees plus quality equals improved access to health care: results of a field experiment in Cameroon**. *Soc Sci Med* 1993, **37**(3):369-383.
88. Alderman H, Lavy V: **Household responses to public health services: cost and quality tradeoffs**. *The World Bank Research Observer* 1996, **11**(1):3-22.
89. Gabrysch S, Campbell OM: **Still too far to walk: literature review of the determinants of delivery service use**. *BMC Pregnancy Childbirth* 2009, **9**:34.
90. Say L, Raine R: **A systematic review of inequalities in the use of maternal health care in developing countries: examining the scale of the problem and the importance of context**. *Bulletin of the World Health Organization* 2007, **85**(10):812-819.
91. Barros AJ, Ronsmans C, Axelson H, Loaiza E, Bertoldi AD, Franca GV, Bryce J, Boerma JT, Victora CG: **Equity in maternal, newborn, and child health interventions in Countdown to 2015: a retrospective review of survey data from 54 countries**. *Lancet (London, England)* 2012, **379**(9822):1225-1233.
92. Boerma JT, Bryce J, Kinfu Y, Axelson H, Victora CG: **Mind the gap: equity and trends in coverage of maternal, newborn, and child health services in 54 Countdown countries**. *Lancet (London, England)* 2008, **371**(9620):1259-1267.
93. Victora CG, Barros AJ, Axelson H, Bhutta ZA, Chopra M, Franca GV, Kerber K, Kirkwood BR, Newby H, Ronsmans C *et al*: **How changes in coverage affect equity in maternal and child health interventions in 35 Countdown to 2015 countries: an analysis of national surveys**. *Lancet (London, England)* 2012, **380**(9848):1149-1156.

94. Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S: **Does pay-for-performance improve the quality of health care?** *Annals of internal medicine* 2006, **145**(4):265-272.

95. Li J, Hurley J, DeCicca P, Buckley G: **Physician response to pay-for-performance: evidence from a natural experiment**. *Health Econ* 2014, **23**(8):962-978.

96. Ensor T, Cooper S: **Overcoming barriers to health service access: influencing the demand side**. *Health policy and planning* 2004, **19**(2):69-79.

97. Eijkenaar F: **Pay for performance in health care: an international overview of initiatives**. *Med Care Res Rev* 2012, **69**(3):251-276.

98. Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M: **Reduced mortality with hospital pay for performance in England**. *N Engl J Med* 2012, **367**(19):1821-1828.

99. Das K, Anderson G, Payment R: **Premier Hospital Quality Incentive Demonstration**. *Health Policy Monitor* 2007.

100. Glickman SW, Peterson ED: **Innovative health reform models: pay-for-performance initiatives**. *The American journal of managed care* 2009, **15**(10 Suppl):S300-305.

101. Ryan A: **Hospital-based pay-for-performance in the United States**. *Health Econ* 2009, **18**(10):1109-1113.

102. Scott A, Sivey P, Ait Ouakrim D, Willenberg L, Naccarella L, Furler J, Young D: **The effect of financial incentives on the quality of health care provided by primary care physicians**. *The Cochrane database of systematic reviews* 2011(9):Cd008451.

103. Mehrotra A, Damberg CL, Sorbero ME, Teleki SS: **Pay for performance in the hospital setting: what is the state of the evidence?** *Am J Med Qual* 2009, **24**(1):19-28.

104. Flodgren G, Eccles MP, Shepperd S, Scott A, Parmelli E, Beyer FR: **An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes**. *The Cochrane database of systematic reviews* 2011(7):Cd009255.

105. Van Herck P, De Smedt D, Annemans L, Remmen R, Rosenthal MB, Sermeus W: **Systematic review: Effects, design choices, and context of pay-for-performance in health care**. *BMC Health Serv Res* 2010, **10**:247.

106. Langdown C, Peckham S: **The use of financial incentives to help improve health outcomes: is the quality and outcomes framework fit for purpose? A systematic review**. *J Public Health (Oxf)* 2014, **36**(2):251-258.

107. Christianson JB, Leatherman S, Sutherland K: **Lessons from evaluations of purchaser pay-for-performance programs: a review of the evidence**. *Med Care Res Rev* 2008, **65**(6 Suppl):5s-35s.

108. Gillam SJ, Siriwardena AN, Steel N: **Pay-for-performance in the United Kingdom: impact of the quality and outcomes framework: a systematic review**. *Ann Fam Med* 2012, **10**(5):461-468.

109. Alshamsan R, Majeed A, Ashworth M, Car J, Millett C: **Impact of pay for performance on inequalities in health care: systematic review**. *J Health Serv Res Policy* 2010, **15**(3):178-184.

110. Boeckxstaens P, Smedt DD, Maeseneer JD, Annemans L, Willems S: **The equity dimension in evaluations of the quality and outcomes framework: a systematic review**. *BMC Health Serv Res* 2011, **11**:209.

111. Emmert M, Eijkenaar F, Kemter H, Esslinger AS, Schoffski O: **Economic evaluation of pay-for-performance in health care: a systematic review**. *The European journal of health economics : HEPAC : health economics in prevention and care* 2012, **13**(6):755-767.

112. Mendelson A, Kondo K, Damberg C, Low A, Motuapuaka M, Freeman M, O'Neil M, Relevo R, Kansagara D: **The Effects of Pay-for-Performance Programs on Health, Health Care Use, and Processes of Care: A Systematic Review**. *Annals of internal medicine* 2017, **166**(5):341-353.

113. Glickman SW, Ou FS, DeLong ER, Roe MT, Lytle BL, Mulgund J, Rumsfeld JS, Gibler WB, Ohman EM, Schulman KA *et al*: **Pay for performance, quality of care, and outcomes in acute myocardial infarction**. *Jama* 2007, **297**(21):2373-2380.

114. Werner RM, Kolstad JT, Stuart EA, Polsky D: **The effect of pay-for-performance in hospitals: lessons for quality improvement**. *Health Aff (Millwood)* 2011, **30**(4):690-698.

115. Chen JY, Kang N, Juarez DT, Hodges KA, Chung RS: **Impact of a Pay-for-Performance Program on Low Performing Physicians**. *Journal for Healthcare Quality* 2010, **32**(1):13-22.

116. Doran T, Fullwood C, Gravelle H, Reeves D, Kontopantelis E, Hiroeh U, Roland M: **Pay-for-performance programs in family practices in the United Kingdom**. *N Engl J Med* 2006, **355**(4):375-384.

117. Doran T, Kontopantelis E, Valderas JM, Campbell S, Roland M, Salisbury C, Reeves D: **Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework**. *Bmj* 2011, **342**:d3590.

118. Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M: **Effects of pay for performance on the quality of primary care in England**. *N Engl J Med* 2009, **361**(4):368-378.

119. Ryan AM: **Effects of the Premier Hospital Quality Incentive Demonstration on Medicare patient mortality and cost**. *Health Serv Res* 2009, **44**(3):821-842.

120. Jha AK, Joynt KE, Orav EJ, Epstein AM: **The long-term effect of premier pay for performance on patient outcomes**. *N Engl J Med* 2012, **366**(17):1606-1615.

121. Kristensen SR, Meacock R, Turner AJ, Boaden R, McDonald R, Roland M, Sutton M: **Long-term effect of hospital pay for performance on mortality in England**. *N Engl J Med* 2014, **371**(6):540-548.

122. Fleetcroft R, Parekh-Bhurke S, Howe A, Cookson R, Swift L, Steel N: **The UK pay-for-performance programme in primary care: estimation of population**

mortality reduction. *The British journal of general practice : the journal of the Royal College of General Practitioners* 2010, **60**(578):e345-e352.

123. Fleetcroft R, Steel N, Cookson R, Walker S, Howe A: **Incentive payments are not related to expected health gain in the pay for performance scheme for UK primary care: cross-sectional analysis**. *BMC Health Serv Res* 2012, **12**:94.
124. Ryan AM, Krinsky S, Kontopantelis E, Doran T: **Long-term evidence for the effect of pay-for-performance in primary care on mortality in the UK: a population study**. *Lancet (London, England)* 2016, **388**(10041):268-274.
125. Blustein J, Borden WB, Valentine M: **Hospital performance, the local economy, and the local workforce: findings from a US National Longitudinal Study**. *PLoS Med* 2010, **7**(6):e1000297.
126. Chaudhury N, Hammer J, Kremer M, Muralidharan K, Rogers FH: **Missing in action: teacher and health worker absence in developing countries**. *The Journal of Economic Perspectives* 2006, **20**(1):91-116.
127. Soeters R, Griffiths F: **Improving government health services through contract management: a case from Cambodia**. *Health policy and planning* 2003, **18**(1):74-83.
128. Van de Poel E, Flores G, Ir P, O'Donnell O: **Impact of Performance-Based Financing in a Low-Resource Setting: A Decade of Experience in Cambodia**. *Health Econ* 2016, **25**(6):688-705.
129. Eichler R, Auxila P, Antoine U, Desmangles B: **Performance-based incentives for health: six years of results from supply-side programs in Haiti**. 2007.
130. Meessen B, Musango L, Kashala JP, Lemlin J: **Reviewing institutions of rural health centres: the Performance Initiative in Butare, Rwanda**. *Trop Med Int Health* 2006, **11**(8):1303-1317.
131. Soeters R, Habineza C, Peerenboom PB: **Performance-based financing and changing the district health system: experience from Rwanda**. *Bull World Health Organ* 2006, **84**(11):884-889.
132. Basinga P, Gertler PJ, Binagwaho A, Soucat AL, Sturdy J, Vermeersch CM: **Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation**. *Lancet (London, England)* 2011, **377**(9775):1421-1428.
133. Bonfrer I, Van de Poel E, Van Doorslaer E: **The effects of performance incentives on the utilization and quality of maternal and child care in Burundi**. *Soc Sci Med* 2014, **123**:96-104.
134. Das A, Gopalan SS, Chandramohan D: **Effect of pay for performance to improve quality of maternal and child care in low- and middle-income countries: a systematic review**. *BMC Public Health* 2016, **16**:321.
135. Oxman AD, Fretheim A: **Can paying for results help to achieve the Millennium Development Goals? A critical review of selected evaluations of results-based financing**. *Journal of Evidence-Based Medicine* 2009, **2**(3):184-195.

136. Fretheim A, Witter S, Lindahl AK, Olsen IT: **Performance-based financing in low- and middle-income countries: still more questions than answers**. *Bull World Health Organ* 2012, **90**(8):559-559A.

137. de Walque D, Gertler PJ, Bautista-Arredondo S, Kwan A, Vermeersch C, de Dieu Bizimana J, Binagwaho A, Condo J: **Using provider performance incentives to increase HIV testing and counseling services in Rwanda**. *J Health Econ* 2015, **40**:1-9.

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

139. Chinkhumba J, De Allegri M, Mazalale J, Brenner S, Mathanga D, Muula AS, Robberstad B: **Household costs and time to seek care for pregnancy related complications: The role of results-based financing**. *PLoS One* 2017, **12**(9):e0182326.

140. Steenland M, Robyn PJ, Compaore P, Kabore M, Tapsoba B, Zongo A, Haidara OD, Fink G: **Performance-based financing to increase utilization of maternal health services: Evidence from Burkina Faso**. *SSM - population health* 2017, **3**:179-184.

141. Soeters R, Peerenboom PB, Mushagalusa P, Kimanuka C: **Performance-based financing experiment improved health care in the Democratic Republic of Congo**. *Health Aff (Millwood)* 2011, **30**(8):1518-1527.

142. Huillery E, Seban J: **Pay-for-Performance, motivation and final output in the health sector: Experimental evidence from the Democratic Republic of Congo**. *Sciences Po Economics Discussion Papers* 2014.

143. Zeng W, Shepard DS, Rusatira JD, Blaakman AP, Nsitou BM: **Evaluation of results-based financing in the Republic of the Congo: a comparison group pre-post study**. *Health Policy Plan* 2018, **33**(3):392-400.

144. Rajkotia Y, Zang O, Nguimkeu P, Gergen J, Djurovic I, Vaz P, Mbofana F, Jobarteh K: **The effect of a performance-based financing program on HIV and maternal/child health services in Mozambique—an impact evaluation**. *Health policy and planning* 2017, **32**(10):1386-1396.

145. Engineer CY, Dale E, Agarwal A, Agarwal A, Alonge O, Edward A, Gupta S, Schuh HB, Burnham G, Peters DH: **Effectiveness of a pay-for-performance intervention to improve maternal and child health services in Afghanistan: a cluster-randomized trial**. *Int J Epidemiol* 2016, **45**(2):451-459.

146. Zeng W, Cros M, Wright KD, Shepard DS: **Impact of performance-based financing on primary health care services in Haiti**. *Health Policy Plan* 2013, **28**(6):596-605.

147. Wang H, Zhang L, Yip W, Hsiao W: **An experiment in payment reform for doctors in rural China reduced some unnecessary care but did not lower total costs**. *Health Affairs* 2011, **30**(12):2427-2436.

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

149. Josephson E, Gergen J, Coe M, Ski S, Madhavan S, Bauhoff S: **How do performance-based financing programmes measure quality of care? A descriptive analysis of 68 quality checklists from 28 low- and middle-income countries**. *Health Policy Plan* 2017.

150. Donabedian A: **Evaluating the quality of medical care. 1966**. *Milbank Q* 2005, **83**(4):691-729.

151. Brenner S, Wilhelm D, Lohmann J, Kambala C, Chinkhumba J, Muula AS, De Allegri M: **Implementation research to improve quality of maternal and newborn health care, Malawi**. *Bull World Health Organ* 2017, **95**(7):491-502.

152. Ngo DK, Sherry TB, Bauhoff S: **Health system changes under pay-for-performance: the effects of Rwanda's national programme on facility inputs**. *Health Policy Plan* 2017, **32**(1):11-20.

153. Huntington D, Zaky HH, Shawky S, Fattah FA, El-Hadary E: **Impact of a service provider incentive payment scheme on quality of reproductive and child-health services in Egypt**. *Journal of health, population, and nutrition* 2010, **28**(3):273.

154. Peabody J, Shimkhada R, Quimbo S, Florentino J, Bacate M, McCulloch CE, Solon O: **Financial incentives and measurement improved physicians' quality of care in the Philippines**. *Health Affairs* 2011, **30**(4):773-781.

155. Peabody JW, Shimkhada R, Quimbo S, Solon O, Javier X, McCulloch C: **The impact of performance incentives on child health outcomes: results from a cluster randomized controlled trial in the Philippines**. *Health Policy and Planning* 2013, **29**(5):615-621.

156. Skiles MP, Curtis SL, Basinga P, Angeles G, Thirumurthy H: **The effect of performance-based financing on illness, care-seeking and treatment among children: an impact evaluation in Rwanda**. *BMC Health Serv Res* 2015, **15**:375.

157. Lannes L, Meessen B, Soucat A, Basinga P: **Can performance-based financing help reaching the poor with maternal and child health services? The experience of rural Rwanda**. *Int J Health Plann Manage* 2016, **31**(3):309-348.

158. Binyaruka P, Robberstad B, Torsvik G, Borghi J: **Who benefits from increased service utilisation? Examining the distributional effects of payment for performance in Tanzania**. *Int J Equity Health* 2018, **17**(1):14.

159. Priedeman Skiles M, Curtis SL, Basinga P, Angeles G: **An equity analysis of performance-based financing in Rwanda: are services reaching the poorest women?** *Health Policy Plan* 2013, **28**(8):825-837.

160. Cameron A, Ewen M, Ross-Degnan D, Ball D, Laing R: **Medicine prices, availability, and affordability in 36 developing and middle-income countries: a secondary analysis**. *Lancet (London, England)* 2009, **373**(9659):240-249.

161. Khandker SR, Koolwal GB, Samad HA: **Handbook on Impact Evaluation: Quantitative Methods and Practices**. In. Washington DC: The World Bank; 2010.

162. Djebbari H, Smith J: **Heterogeneous impacts in PROGRESA**. *Journal of Econometrics* 2008, **145**(1):64-80.

163. Bitler MP, Gelbach JB, Hoynes HW: **What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments**. *American Economic Review* 2006, **96**(4):988-1012.

164. Bitler MP, Gelbach JB, Hoynes HW: **Distributional impacts of the self-sufficiency project**. *Journal of Public Economics* 2008, **92**(3):748-765.

165. Brand JE, Xie Y: **Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education**. *American sociological review* 2010, **75**(2):273-302.

166. Heckman JJ, Smith J, Clements N: **Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts**. *The Review of Economic Studies* 1997, **64**(4):487-535.

167. O'Neill K, Takane M, Sheffel A, Abou-Zahr C, Boerma T: **Monitoring service delivery for universal health coverage: the Service Availability and Readiness Assessment**. *Bull World Health Organ* 2013, **91**(12):923-931.

168. TDHS: **Tanzania Demographic and Health Survey and Malaria Indicator Survey 2015-16**. In. National Bureau of Statistics (NBS): Dar es Salaam; 2016.

169. NBS: **Tanzania Population and Housing Census: Population Distribution by Administrative Areas 2012.** In. National Bureau of Statistics (NBS): Dar es Salaam; 2013.

170. **TANZANIA Country profile -World Bank: [https://data.worldbank.org/country/tanzania]** [https://data.worldbank.org/country/tanzania]

171. United Nations: **World Population Prospects: The 2017 Revision, Key Findings and Advance Tables**. In., vol. Working Paper No. ESA/P/WP/248: United Nations, Department of Economics and Social Affairs 2017.

172. International Monetary Fund: **Regional Economic Outlook: Sub-Saharan Africa Restarting the Growth Engine**. In.: International Monetary Fund (IMF), Washington, D.C. ; 2017.

173. Afnan-Holmes H, Magoma M, John T, Levira F, Msemo G, Armstrong CE, Martinez-Alvarez M, Kerber K, Kihinga C, Makuwani A *et al*: **Tanzania's countdown to 2015: an analysis of two decades of progress and gaps for reproductive, maternal, newborn, and child health, to inform priorities for post-2015**. *The Lancet Global health* 2015, **3**(7):e396-409.

174. Masanja H, de Savigny D, Smithson P, Schellenberg J, John T, Mbuya C, Upunda G, Boerma T, Victora C, Smith T *et al*: **Child survival gains in Tanzania: analysis of data from demographic and health surveys**. *Lancet (London, England)* 2008, **371**(9620):1276-1283.

175. United Nations: **The Sustainable Development Goals Report 2016**. In. United Nations, New York; 2016.
176. Maluka SO, Hurtig AK, Sebastian MS, Shayo E, Byskov J, Kamuzora P: **Decentralization and health care prioritization process in Tanzania: from national rhetoric to local reality**. *Int J Health Plann Manage* 2011, **26**(2):e102-120.
177. Frumence G, Nyamhanga T, Mwangu M, Hurtig AK: **The dependency on central government funding of decentralised health systems: experiences of the challenges and coping strategies in the Kongwa District, Tanzania**. *BMC Health Serv Res* 2014, **14**:39.
178. Crook RC: **Decentralisation and poverty reduction in Africa: the politics of local–central relations**. *Public administration and development* 2003, **23**(1):77-88.
179. Makundi E, Nyoni J, Nanda P: **The implications of health sector reforms on reproductive health services: the case of Bukoba District, Kagera region, Tanzania study**. *Centre for Health and Gender Equity Working Papers* 2005.
180. Nyamhanga T, Fruemnce G, Mwangu M, Hurtig AK: **Achievements and challenges of resource allocation for health in a decentralized system in Tanzania: perspectives of national and district level officers**. *East African journal of public health* 2013, **10**(2):416-427.
181. Frumence G, Nyamhanga T, Mwangu M, Hurtig AK: **Challenges to the implementation of health sector decentralization in Tanzania: experiences from Kongwa district council**. *Global health action* 2013, **6**:20983.
182. MoHSW: **Tanzania Health Sector Strategic Plan (HSSP IV) 2015-2020**. In., vol. IV: Ministry of Health and Social Welfare (MoHSW): Dar es Salaam; 2015.
183. Kruk ME, Mbaruku G, McCord CW, Moran M, Rockers PC, Galea S: **Bypassing primary care facilities for childbirth: a population-based study in rural Tanzania**. *Health Policy Plan* 2009, **24**(4):279-288.
184. Kahabuka C, Kvale G, Moland KM, Hinderaker SG: **Why caretakers bypass Primary Health Care facilities for child care - a case from rural Tanzania**. *BMC Health Serv Res* 2011, **11**:315.
185. Countdown to 2015: **Fulfilling the health agenda for women and children—the 2014 Report.** In.: Geneva: World Health Organization.
186. Manzi F, Schellenberg JA, Hutton G, Wyss K, Mbuya C, Shirima K, Mshinda H, Tanner M, Schellenberg D: **Human resources for health care delivery in Tanzania: a multifaceted problem**. *Human resources for health* 2012, **10**:3.
187. Munga MA, Maestad O: **Measuring inequalities in the distribution of health workers: the case of Tanzania**. *Human resources for health* 2009, **7**:4.
188. MoHSW: **Tanzania Service Svailability and Seadiness Sssessment (SARA) 2012**. In. Ministry of Health and Social Welfare and Ifakara Health Institute: Dar es Salaam; 2013.
189. Hart JT: **The inverse care law**. *Lancet (London, England)* 1971, **1**(7696):405-412.

190. Willcox ML, Peersman W, Daou P, Diakite C, Bajunirwe F, Mubangizi V, Mahmoud EH, Moosa S, Phaladze N, Nkomazana O *et al*: **Human resources for primary health care in sub-Saharan Africa: progress or stagnation?** *Human resources for health* 2015, **13**:76.

191. Mkoka DA, Goicolea I, Kiwara A, Mwangu M, Hurtig AK: **Availability of drugs and medical supplies for emergency obstetric care: experience of health facility managers in a rural District of Tanzania**. *BMC Pregnancy Childbirth* 2014, **14**:108.

192. SIKIKA: **Medicines and Medical Supplies Availability Report. Using Absorbent Gauze Availability Survey as an Entry Point. A Case of 71 Districts and 30 Health Facilities across Mainland Tanzania.** In.: SIKIKA, Dar es Salaam, Tanzania; 2011.

193. Wales J, Tobias J, Malangalila E, Swai G, Wild L: **Stock-outs of essential medicines in Tanzania: A Political Economy Approach to Analysing Problems and Identifying Solutions.** In.: Twaweza ni sisi; 2014.

194. USAID: **Tanzania Health System Assessment 2010. Health Systems 20/20 project.** In.: Abt Associates Inc.: Bethesda, MD; 2011.

195. MoHSW: **Mapping of the Medicines Procurement and Supply Management System in Tanzania.** In.: Ministry of Health and Social Welfare (MoHSW): Dar es Salaam; 2008.

196. Euro Health Group: **The United Republic of Tanzania Drug Tracking Study.** In.: Euro Health Group: Denmark; 2007.

197. MoHSW: **Mid Term Review of the Health Sector Strategic Plan III 2009-2015: Health Care Financing. Technical Report, Ministry of Health and Social Welfare (MoHSW), United Republic of Tanzania**. In.: MoHSW, Dar es Salaam; 2013.

198. Zomboko FE, Tripathi SK: **Challenges in procurement and use of Donated medical-equipments: study of a selected referral hospital in Tanzania**. *Researchers World* 2012, **3**(4):41.

199. Rutta E, Shekalaghe E, Sillo H, Liana J, Johnson K, Embrey M, Lieber R, Valimba R, Kimatta S: **Accrediting retail drug shops to strengthen Tanzania's public health system: an ADDO case study**. *Journal of pharmaceutical policy and practice* 2015, **8**(1):23.

200. Centre for Pharmaceutical Management: **Accredited Drug Dispensing Outlets in Tanzania: Strategies for Enhancing Access to Medicines Program.** In.: Management Sciences for
Health: Arlington, VA; 2008.

201. MoHSW: **Tanzania Mainland Expanded Programme on Immunization (EPI) Review.** In.: Dar es Salaam, Ministry of Health and Social Welfare (MoHSW); 2010.

202. Yadav P, Tata HL, WHO G, Babaley M: **Storage and Supply Chain Management. The World Medicines Situation 2011**. In.: Geneva: World Health Organization; 2011.

203. **National Health Accounts: Tanzania** [http://apps.who.int/nha/database/ViewData/Indicators/en]
204. Mtei G, Makawia S, Masanja H: **Monitoring and evaluating progress towards Universal Health Coverage in Tanzania**. *PLoS Med* 2014, **11**(9):e1001698.
205. Mtei G, Mulligan J: **Community health funds in Tanzania: A literature review**. *Ifakara Health Research and Development Centre, Ifakara* 2007.
206. Macha J, Harris B, Garshong B, Ataguba JE, Akazili J, Kuwawenaruwa A, Borghi J: **Factors influencing the burden of health care financing and the distribution of health care benefits in Ghana, Tanzania and South Africa**. *Health Policy Plan* 2012, **27 Suppl 1**:i46-54.
207. Mtei G, Borghi J, Macha J, Kuwawenaruwa A, Makawia S: **Monitoring the implementation of Universal Health Coverage reforms in districts in Tanzania**. In: *Universal coverage in Tanzania and South Africa (UNITAS) Policy Brief.* Ifakara Health Institute; 2012.
208. Mills A, Ally M, Goudge J, Gyapong J, Mtei G: **Progress towards universal coverage: the health systems of Ghana, South Africa and Tanzania**. *Health Policy Plan* 2012, **27 Suppl 1**:i4-12.
209. Kamuzora P, Gilson L: **Factors influencing implementation of the Community Health Fund in Tanzania**. *Health Policy Plan* 2007, **22**(2):95-102.
210. Macha J, Kuwawenaruwa A, Makawia S, Mtei G, Borghi J: **Determinants of community health fund membership in Tanzania: a mixed methods analysis**. *BMC Health Serv Res* 2014, **14**:538.
211. Maluka SO: **Why are pro-poor exemption policies in Tanzania better implemented in some districts than in others?** *Int J Equity Health* 2013, **12**:80.
212. Kruk ME, Mbaruku G, Rockers PC, Galea S: **User fee exemptions are not enough: out-of-pocket payments for 'free' delivery services in rural Tanzania**. *Trop Med Int Health* 2008, **13**(12):1442-1451.
213. Manzi F, Schellenberg JA, Adam T, Mshinda H, Victora CG, Bryce J: **Out-of-pocket payments for under-five health care in rural southern Tanzania**. *Health Policy Plan* 2005, **20 Suppl 1**:i85-i93.
214. Borghi J, Mayumana I, Mashasi I, Binyaruka P, Patouillard E, Njau I, Maestad O, Abdulla S, Mamdani M: **Protocol for the evaluation of a pay for performance programme in Pwani region in Tanzania: a controlled before and after study**. *Implement Sci* 2013, **8**:80.
215. Anselmi L, Binyaruka P, Borghi J: **Understanding causal pathways within health systems policy evaluation through mediation analysis: an application to payment for performance (P4P) in Tanzania**. *Implement Sci* 2017, **12**(1):10.
216. Gertler P, Martinez S, Premand P, Rawlings L, Vermeersch C: **Impact evaluation in practice. The World Bank. Washington**. In.; 2011.
217. Duflo E, Glennerster R, Kremer M: **Using randomization in development economics research: A toolkit**. *Handbook of development economics* 2007, **4**:3895-3962.

218. Black N: **Why we need observational studies to evaluate the effectiveness of health care**. *Bmj* 1996, **312**(7040):1215-1218.

219. Blundell R, Dias MC: **Alternative Approaches to Evaluation in Empirical Microeconomics**. *The Journal of Human Resources* 2009, **44**(3):565-640.

220. Pronyk PM, Nemser B, Maliqi B, Springstubb N, Sera D, Karimov R, Katwan E, Walter B, Bijleveld P: **The UN Commission on Life Saving Commodities 3 years on: global progress update and results of a multicountry assessment**. *The Lancet Global health* 2016, **4**(4):e276-286.

221. United Nations: **UN Commission on Life-Saving Commodities for Women and Children: Commissioner's Report**. In. New York: United Nations; 2012.

222. Choi Y, Ametepi P: **Comparison of medicine availability measurements at health facilities: evidence from Service Provision Assessment surveys in five sub-Saharan African countries**. *BMC Health Serv Res* 2013, **13**:266.

223. Penfold S, Shamba D, Hanson C, Jaribu J, Manzi F, Marchant T, Tanner M, Ramsey K, Schellenberg D, Schellenberg JA: **Staff experiences of providing maternity services in rural southern Tanzania - a focus on equipment, drug and supply issues**. *BMC Health Serv Res* 2013, **13**:61.

224. Filmer D, Pritchett LH: **Estimating wealth effects without expenditure data--or tears: an application to educational enrollments in states of India**. *Demography* 2001, **38**(1):115-132.

225. Vyas S, Kumaranayake L: **Constructing socio-economic status indices: how to use principal components analysis**. *Health Policy Plan* 2006, **21**(6):459-468.

226. WHO: **Handbook on health inequality monitoring: with a special focus on low- and middle-income countries.** In. World Health Organization, Geneva, Switzerland; 2013.

227. O'Donnell O, Van Doorsslaer E, Wagstaff A, Lindelöw M: **Analyzing Health Equity Using Household Survey Data: A Guide to Techniques and Their Implementation**, vol. 434: World Bank Publications; 2008.

228. Whitehead M: **The concepts and principles of equity and health**. *Health promotion international* 1991, **6**(3):217-228.

229. Kakwani N, Wagstaff A, Van Doorslaer E: **Socioeconomic inequalities in health: measurement, computation, and statistical inference**. *Journal of econometrics* 1997, **77**(1):87-103.

230. Cameron AC, Miller DL: **A practitioner's guide to cluster-robust inference**. *Journal of Human Resources* 2015, **50**(2):317-372.

231. Kitchen CM: **Nonparametric vs parametric tests of location in biomedical research**. *Am J Ophthalmol* 2009, **147**(4):571-572.

232. Blacklock C, MacPepple E, Kunutsor S, Witter S: **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* 2016, **47**(4):309-324.

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

234. Grimes DA, Schulz KF: **Bias and causal associations in observational research**. *The Lancet* 2002, **359**(9302):248-252.

235. Imbens GW, Wooldridge JM: **Recent Developments in the Econometrics of Program Evaluation**. *Journal of Economic Literature* 2009, **47**(1):5-86.

236. Deeks J, Dinnes J, D'amico R, Sowden A, Sakarovitch C, Song F, Petticrew M, Altman D: **Evaluating non-randomised intervention studies**. *Health technology assessment (Winchester, England)* 2003, **7**(27):iii-x, 1-173.

237. Leightner JE, Inoue T: **Tackling the omitted variables problem without the strong assumptions of proxies**. *European Journal of Operational Research* 2007, **178**(3):819-840.

238. Angrist JD, Pischke J-S: **Mostly harmless econometrics: An empiricist's companion**: Princeton university press; 2008.

239. Eisele TP, Rhoda DA, Cutts FT, Keating J, Ren R, Barros AJ, Arnold F: **Measuring coverage in MNCH: total survey error and the interpretation of intervention coverage estimates from household surveys**. *PLoS Med* 2013, **10**(5):e1001386.

240. Cutts FT, Izurieta HS, Rhoda DA: **Measuring coverage in MNCH: design, implementation, and interpretation challenges associated with tracking vaccination coverage using household surveys**. *PLoS Med* 2013, **10**(5):e1001404.

241. Delgado-Rodríguez M, Llorca J: **Bias**. *Journal of Epidemiology & Community Health* 2004, **58**(8):635-641.

242. Coughlin SS: **Recall bias in epidemiologic studies**. *J Clin Epidemiol* 1990, **43**(1):87-91.

243. Nederhof AJ: **Methods of coping with social desirability bias: A review**. *European journal of social psychology* 1985, **15**(3):263-280.

244. Barnett AG, van der Pols JC, Dobson AJ: **Regression to the mean: what it is and how to deal with it**. *Int J Epidemiol* 2005, **34**(1):215-220.

245. Bland JM, Altman DG: **Statistics notes: some examples of regression towards the mean**. *Bmj* 1994, **309**(6957):780.

246. Davis C: **The effect of regression to the mean in epidemiologic and clinical studies**. *American journal of epidemiology* 1976, **104**(5):493-498.

247. Shaffer JP: **Multiple hypothesis testing**. *Annual review of psychology* 1995, **46**(1):561-584.

248. Hayes RJ, Bennett S: **Simple sample size calculation for cluster-randomized trials**. *Int J Epidemiol* 1999, **28**(2):319-326.

249. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafo MR: **Power failure: why small sample size undermines the reliability of neuroscience**. *Nat Rev Neurosci* 2013, **14**(5):365-376.

250. Steckler A, Mcleroy KR: **The importance of external validity**. *American journal of public health* 2008, **98**(1):9-10.

251. Bertone MP, Falisse JB, Russo G, Witter S: **Context matters (but how and why?) A hypothesis-led literature review of performance based financing in fragile and conflict-affected health systems**. *PLoS One* 2018, **13**(4):e0195301.

252. Olafsdottir AE, Mayumana I, Mashasi I, Njau I, Mamdani M, Patouillard E, Binyaruka P, Abdulla S, Borghi J: **Pay for performance: an analysis of the context of implementation in a pilot project in Tanzania**. *BMC Health Serv Res* 2014, **14**:392.

253. Paul E, Renmans D: **Performance-based financing in the heath sector in low- and middle-income countries: Is there anything whereof it may be said, see, this is new?** *Int J Health Plann Manage* 2018, **33**(1):51-66.

254. Chimhutu V, Lindkvist I, Lange S: **When incentives work too well: locally implemented pay for performance (P4P) and adverse sanctions towards home birth in Tanzania - a qualitative study**. *BMC Health Serv Res* 2014, **14**:23.

255. Chimhutu V, Tjomsland M, Songstad NG, Mrisho M, Moland KM: **Introducing payment for performance in the health sector of Tanzania- the policy process**. *Global Health* 2015, **11**:38.

256. Paul E, Albert L, Bisala BN, Bodson O, Bonnet E, Bossyns P, Colombo S, De Brouwere V, Dumont A, Eclou DS *et al*: **Performance-based financing in low-income and middle-income countries: isn't it time for a rethink?** *BMJ global health* 2018, **3**(1):e000664.

257. Kondo KK, Damberg CL, Mendelson A, Motu'apuaka M, Freeman M, O'Neil M, Relevo R, Low A, Kansagara D: **Implementation Processes and Pay for Performance in Healthcare: A Systematic Review**. *J Gen Intern Med* 2016, **31 Suppl 1**:61-69.

258. Fox S, Witter S, Wylde E, Mafuta E, Lievens T: **Paying health workers for performance in a fragmented, fragile state: reflections from Katanga Province, Democratic Republic of Congo**. *Health Policy Plan* 2014, **29**(1):96-105.

259. Shroff ZC, Tran N, Meessen B, Bigdeli M, Ghaffar A: **Taking results-based financing from scheme to system**. *Health Systems & Reform* 2017, **3**(2):69-73.

260. Ashworth M, Seed P, Armstrong D, Durbaba S, Jones R: **The relationship between social deprivation and the quality of primary care: a national survey using indicators from the UK Quality and Outcomes Framework**. *The British journal of general practice : the journal of the Royal College of General Practitioners* 2007, **57**(539):441-448.

261. Gravelle H, Sutton M, Ma A: **Doctor behaviour under a pay for performance contract: further evidence from the Quality and Outcomes Framework**. *CHE Research Paper 34* 2008.

262. Kontopantelis E, Buchan I, Reeves D, Checkland K, Doran T: **Relationship between quality of care and choice of clinical computing system: retrospective**

analysis of family practice performance under the UK's quality and outcomes framework. *BMJ Open* 2013, **3**(8):1-11.

263. Ryan AM, Blustein J: **The effect of the MassHealth hospital pay-for-performance program on quality**. *Health Serv Res* 2011, **46**(3):712-728.

264. Imai K, Keele L, Tingley D, Yamamoto T: **Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies**. *American Political Science Review* 2011, **105**(4):765-789.

265. O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, Evans T, Pardo Pardo J, Waters E, White H *et al*: **Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health**. *J Clin Epidemiol* 2014, **67**(1):56-64.

266. Evans DB, Hsu J, Boerma T: **Universal health coverage and universal access**. *Bull World Health Organ* 2013, **91**(8):546-546a.

267. Marmot M: **Universal health coverage and social determinants of health**. *Lancet (London, England)* 2013, **382**(9900):1227-1228.

268. Falisse J-B, Ndayishimiye J, Kamenyero V, Bossuyt M: **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* 2014, **30**(10):1251-1260.

269. Muula A: **How do we define'rurality'in the teaching on medical demography**. *Rural and remote health* 2007, **7**(1):653.

270. Yates R: **Universal health care and the removal of user fees**. *Lancet (London, England)* 2009, **373**(9680):2078-2081.

271. Lagarde M, Palmer N: **The impact of user fees on health service utilization in low- and middle-income countries: how strong is the evidence?** *Bull World Health Organ* 2008, **86**(11):839-848.

272. Idd A, Yohana O, Maluka SO: **Implementation of pro-poor exemption policy in Tanzania: policy versus reality**. *Int J Health Plann Manage* 2013, **28**(4):e298-309.

273. Sekabaraga C, Diop F, Soucat A: **Can innovative health financing policies increase access to MDG-related services? Evidence from Rwanda**. *Health Policy Plan* 2011, **26 Suppl 2**:ii52-62.

274. Lu C, Chin B, Lewandowski JL, Basinga P, Hirschhorn LR, Hill K, Murray M, Binagwaho A: **Towards universal health coverage: an evaluation of Rwanda Mutuelles in its first eight years**. *PLoS One* 2012, **7**(6):e39282.

275. Strasser R, Kam SM, Regalado SM: **Rural health care access and policy in developing countries**. *Annual review of public health* 2016, **37**:395-412.

Attachments: Scientific Papers

VOLUME 22 NO 1 PP 92–102 JANUARY 2017

Improving quality of care through payment for performance: examining effects on the availability and stock-out of essential medical commodities in Tanzania

Peter Binyaruka[1,2] and Josephine Borghi[3]

[1] *Ifakara Health Institute, Dar es Salaam, Tanzania*
[2] *Centre for International Health, University of Bergen, Bergen, Norway*
[3] *Department of Global Health and Development, London School of Hygiene & Tropical Medicine, London, UK*

Abstract

OBJECTIVE To evaluate the effects of payment for performance (P4P) on the availability and stock-out rate of reproductive, maternal, newborn and child health (RMNCH) medical commodities in Tanzania and assess the distributional effects.

METHODS The availability of RMNCH commodities (medicines, supplies and equipment) on the day of the survey, and stock-outs for at least one day in the 90 days prior to the survey, was measured in 75 intervention and 75 comparison facilities in January 2012 and 13 months later. Composite scores for each subgroup of commodities were generated. A difference-in-differences linear regression was used to estimate the effect of P4P on outcomes and differential effects by facility location, level of care, ownership and socio-economic status of the catchment population.

RESULTS We estimated a significant increase in the availability of medicines by 8.4 percentage points ($P = 0.002$) and an 8.3 percentage point increase ($P = 0.050$) in the availability of medical supplies. P4P had no effect on the availability of functioning equipment. Most items with a significant increase in availability also showed a significant reduction in stock-outs. Effects were generally equally distributed across facilities, with effects on stock-outs of many medicines being pro-poor, and greater effects in facilities in rural compared to urban districts.

CONCLUSION P4P can improve the availability of medicines and medical supplies, especially in poor, rural areas, when these commodities are incentivised at both facility and district levels, making services more acceptable, effective and affordable, enhancing progress towards universal health coverage.

keywords Policy evaluation, payment for performance, medical commodities, structural quality of care, RMNCH, health financing

Introduction

The availability of essential medical commodities (medicines, medical supplies and equipment) is a key component of effective service delivery required for maintaining population health [1]. Shortages of medical commodities are associated with poor structural quality of care, or poor quality relating to the attributes of the setting in which care delivery occurs [2, 3], low levels of patient satisfaction and preventable deaths [4–9]. Medicine and supply shortages in public facilities are also responsible for a large share of the out-of-pocket payments faced by households in low- and middle-income settings limiting the affordability of care [1, 10]. However, ensuring the availability of essential medical commodities remains a challenge for many low-income country health systems.

According to the United Nations Commission on Life-Saving Commodities, payment for performance (P4P) is a strategy to improve access to life-saving commodities for maternal and child health [11, 12]. P4P provides financial incentives to providers and/or healthcare managers based on the achievement of pre-defined performance targets and is currently being rolled out in many low-income countries [13, 14]. P4P could theoretically affect the availability of medical commodities by, for example, incentivising the provision of intermittent preventive treatment (IPT) for malaria during antenatal care (ANC), through facility-level bonus payments, which can be used to procure commodities, and by incentivising health care managers to reduce drug stock-out rates.

However, empirically, only four studies have reported on the effect of P4P on the availability of medical commodities in low-income countries. The effects are varied with no effects on the availability of drugs and equipment in Afghanistan [15]; no effects on patient perceptions of drug availability in Burundi [16]; an increase in patient perceptions of drug availability in the Democratic Republic of Congo (DRC) [17]; and a reduction in the availability of vaccines and equipment in another study from the DRC [18]. Only one study reports on stock-out rates [18] and none of the studies shed light on the pathways through which such changes occurred. Previous studies have not examined the potential heterogeneity of effects across facilities and effects on commodities related to non-incentivised services (spillover effects). This paper examines the effect of P4P on the availability and stock-outs of medicines, medical supplies and equipment for reproductive, maternal, newborn and child health in Tanzania and assesses whether these effects differed by facility location, level of care, facility ownership and socio-economic status of the facility's catchment population.

Methods

Study setting

Since the 1990s, Tanzania began a process of decentralisation of government functions including health services, involving the transfer of power from central to local government authorities [19]. As a result, district-level managers are responsible for preparing annual health sector plans and budgets to implement health programmes and renovations in facilities and are responsible for generating and managing resources for the district. District managers are supported by a regional health management team, while health facility governing committees oversee the implementation of plans and the management of resources at facility level. Public health facilities order medical commodities on a quarterly basis, based on an estimate of quantity needs; they submit requests to the district who review and send them on to the medical stores department (MSD) and distribute medical commodities to facilities (the 'pull' system) [20–22]. Districts and facilities can also use their own funds (e.g. insurance contributions, user fees and P4P bonus payments) to procure commodities in case of stock-outs [22–24]. Non-public hospitals that are contracted by districts to deliver services on behalf of the Ministry of Health and Social Welfare (MoHSW) also receive medical commodities from the MSD. All other non-public facilities either procure commodities from the MSD, foreign or local manufacturers, privately owned accredited drug dispensing outlets (ADDOs) and pharmacies [25–27]. Some commodities (vaccines, antiretrovirals (ARVs), vitamin A and family planning) are managed through disease-specific vertical programmes, which are financed externally, and distributed via the MSD or directly to facilities [24, 28, 29]. The MSD supply chain suffers from a shortage of commodities, inadequate budget allocations, inadequate tracking mechanisms and late delivery of required commodities [8, 22, 24, 30]. As a result, facilities experience regular shortages of essential drugs and supplies especially in the public sector [22, 24, 30, 31]. For example, out of 1297 facilities surveyed in 2012, only 41% stocked the 14 essential tracer medicines at the time of the survey [31]. An assessment in 2010 found that the MSD fulfilled 68% of hospital orders and 67% of orders from health centres and dispensaries [32].

P4P in Tanzania

In 2011, the MoHSW in Tanzania, with financial support from the Government of Norway, introduced a P4P scheme in Pwani region to improve reproductive, maternal, newborn and child health (RMNCH), which is ongoing. Pwani region is one of 30 regions in the country and has seven districts with more than 209 health facilities and a population of just over a million [33]. Financial incentives are given to health facilities, district and regional managers based on their performance on pre-defined service delivery targets (Table 1) [34, 35]. Most of the targets at facility level pertain to increases in service coverage, with four that involve the provision of medicines such as antiretroviral therapy (ART), IPT during ANC, vaccines and supplies such as partographs. District managers are rewarded for reducing the proportion of facilities in the district reporting stock-outs of essential medicines (Appendix S1a) for at least one week. Districts are required to verify facility performance reports, resulting in more frequent contact between district managers and providers which may also help reduce stock-outs. Facilities are required to open bank accounts to receive performance payments.

Facility and district performance data are verified every six months (one cycle). For dispensaries, the maximum payout, if all targets are fully attained, is USD 820 per cycle, while maximum payouts are USD 3220 and USD 6790 for health centres and hospitals, respectively. Incentive payouts at facility level include bonuses to staff (equivalent to 10% of monthly salary) and funds that can be used to procure drugs and supplies and for facility improvement (10% of the total in hospitals and 25% in lower level facilities). District and regional managers receive bonus payments of up to USD 3000 per cycle based on the performance of facilities in their district or region.

Table 1 Service indicators and performance targets for facilities

		Baseline coverage (previous cycle)				
Performance indicators	Method	0–20%	21–40%	41–70%	71–85%	85%+
Coverage indicators						
% of facility-based deliveries	Percentage point increase	15	10	5	5	Maintain
% of mothers attending a facility within 7 days of delivery	Percentage point increase	15	10	5	5	Maintain
% of women using long-term contraceptives	Percentage point increase	20	15	10	Maintain above 71	Maintain
% children under 1 year received measles vaccine	Overall result	50	65	75	80+	Maintain
% children under 1 year received Penta 3 vaccine	Overall result	50	65	75	80+	Maintain
% of complete partographs	Overall result	80	80	80	80+	Maintain above 80
HMIS reports submitted to district managers on time and complete	Overall result	100	100	100	100	100
Content of care indicators						
% ANC clients receiving two doses of IPT	Overall result	80	80	80	80+	Maintain above 80%
% HIV+ ANC clients on ART	Overall result	40	60	75	75+	Maintain
% children received polio vaccine (OPV0) at birth	Overall result	60	75	80	80+	Maintain

85%+ = 85% or more; 80%+ = 80% or more; HMIS, Health Management Information System Source: The United Republic of Tanzania, Ministry of Health and Social Welfare, 2011. The Coast Region Pay for Performance (P4P) Pilot: Design Document.

Study design

This study uses data from a controlled before and after study of the P4P scheme in Pwani region, Tanzania, conducted in all seven intervention districts and four comparison districts from Morogoro and Lindi regions [34, 35]. Baseline data were collected in January 2012 and 13 months later.

Data sources

The data on the availability and stock-outs of essential RMNCH commodities within the previous 90 days were collected through a survey of 75 facilities in each study arm. In the intervention arm, we included all 6 hospitals and 16 health centres that were eligible for the P4P scheme and a random sample of 53 eligible dispensaries. A corresponding number of facilities were surveyed in the comparison arm. The facility survey also documented facility characteristics and was administered to the facility incharge. To proxy the socio-economic status (SES) of the facility catchment population, we used data from a survey of 1500 households of women who had delivered in the previous 12 months prior to the baseline survey in each arm and a similar number in the follow-up survey (20 households sampled from the catchment area of each facility). More details on data sources and data collection are provided elsewhere [34, 35].

Outcome measures

Our main outcomes are the availability of RMNCH medicines, medical supplies and functioning equipment, and the stock-outs of medicines and supplies at the facility. If a commodity was available on the day of the survey, the outcome was coded 1 and 0 otherwise; if a commodity was out of stock for at least one day in the 90 days prior to the survey, the outcome was coded 1 and 0 otherwise (Appendix S1a).

Medical commodities were classified in terms of their therapeutic use as antibiotics, antimalarials, antihypertensives, antidiarrhoeal, anti-retrovirals (ARVs), oxytocics, vaccines, family planning, vitamin A, medical supplies and medical equipment (Appendix S1a). We differentiated between items that relate directly to a P4P target and those which do not, to examine eventual spillover effects. Items were also classified according to their beneficiary/recipient group along the RMNCH continuum of care based on the WHO classification of priority medicines [11, 12]. For each of these groupings, we generated composite scores based on an unweighted mean score across items in the group, which can be interpreted as the mean percentage availability/stock-out rate within the grouping across facilities. We measured the proportion of facilities with availability/stock-out of the respective commodity groups. In the generation of scores, we gave equal weight to each commodity item for ease of interpretation, but we acknowledge some of the items may be more effective than others in enhancing better health outcomes.

Subgroup effects

We examined whether the effects of P4P differed with the wealth of the facility catchment population to see whether benefits were pro-poor, given the greater burden of out-of-pocket payments from stock-outs on poorer groups [1, 10, 30, 36]. We also examined effects by facility ownership (public/non-public) given the differing procurement and supply systems in public and non-public sectors; level of care (dispensary/health centre or hospital) given that dispensaries are typically worse off in drug availability [7, 31, 37]; and whether the facility was in an urban or rural district as facilities in urban districts are better connected by roads facilitating the distribution of commodities relative to those in rural districts.

To generate a wealth score for each household in the catchment area of the facility based on their ownership of 42 household items and characteristics we used principal component analysis (PCA)[38, 39] (Appendix S1c). We then calculated the average wealth score of the 20 households sampled within the facility catchment area. We ranked facilities by these scores from poorest (low score) to least poor and split them into terciles (poorest, middle and least poor).

Statistical analysis

We compared facility characteristics and outcome scores across study arms by using t-tests adjusting for clustering at the facility level. We used a linear difference-in-differences regression model to identify the effects of P4P on the availability and stock-out of medical commodities (1):

$$Y_{it} = \beta_0 + \beta_1(P4P_i \times \delta_t) + \beta_2\delta_t + \gamma_i + \varepsilon_{it} \qquad (1)$$

where Y_{it} is the outcome of facility i at time t. $P4P_i$ is a dummy variable, taking the value 1 if a facility is exposed to P4P and 0 if not. We controlled for time-invariant determinants γ_i with facility fixed effects, and δ_t year fixed effects. The error term is ε_{it}. The effect of P4P on the outcome is given by β_1.

In order to examine subgroup effects, we included a triple interaction term between treatment effect ($P4P_i \times \delta_t$) and subgrouping variable $G_{i.}$ The associated two-order interaction terms were also included. The coefficient of interest for the differential effect is β_3 (2):

$$\begin{aligned} Y_{it} = \beta_0 + \beta_1(P4P_i \times \delta_t) + \beta_2\delta_t + \beta_3(P4P_i \times \delta_t \times G_i) \\ + \beta_4(P4P_i \times G_i) + \beta_5(G_i \times \delta_t) + \gamma_i + \varepsilon_{it} \end{aligned} \qquad (2)$$

For each of the effects, we report the confidence interval based on standard errors that are clustered at the facility level. As a robustness check, we clustered the standard errors at the district level and used the bootstrapping method to adjust for the small number of clusters [40]. We were unable to test whether the availability and stock-out outcomes were parallel between study arms prior to the intervention. However, we tested and confirmed that trends in facility-level utilisation for all incentivised services were parallel prior to the intervention [35, 41]. All analyses were performed using STATA version 13.

Ethical issues

The evaluation study received ethical approval from the Ifakara Health Institute Institutional Review Board and the Ethics Committee of the London School of Hygiene & Tropical Medicine. Study participants provided written consent to participate in this study, requiring them to sign a consent form that was read out to them by the interviewers. This consent form was reviewed and approved by the ethics committees prior to the start of the research.

Results

Baseline facility characteristics were fairly balanced across study arms (Table 2). However, facilities in the intervention arm were serving poorer populations than those in the comparison arm.

P4P was associated with an 8.4 percentage point increase in the availability of all 37 medicines combined (P = 0.002) and an 8.3 percentage point increase in the availability of medical supplies, although this was only borderline significant (P = 0.050) (Table 3). P4P had no effect on the availability of functioning equipment. Effects were noted for some medicines associated with P4P targets (antimalarials, antihypertensives and oxytocics used for deliveries) and supplies (partograph), although this effect was only borderline significant. There was no effect on vaccines, family planning and ARVs. Effects were observed for items that were not clearly linked to service targets, but were incentivised for district managers (antibiotics).

P4P was also associated with a reduction in stock-outs of medicines and medical supplies (Table 4). Most of those items where we found a significant increase in

Table 2 Baseline characteristics of health facilities

Facility characteristic	Intervention facilities (*n* = 75)	Control facilities (*n* = 75)	Difference (*P*–value)
Level of care			
Hospital (%)	8.0	8.0	0
Health centre (%)	21.3	21.3	0
Dispensary (%)	70.7	70.7	0
Ownership status			
Government/public facility (%)	84.0	82.7	1.3 (0.828)
Faith-based organisation (FBO) facility (%)	10.7	12.0	−1.3 (0.798)
Military/parastatal /private facility (%)	5.3	5.3	0 (0.652)
Infrastructure			
Electricity available (%)	68.0	66.7	1.3 (0.863)
Clean water available (%)	73.3	78.7	−5.3 (0.448)
Community/area features			
Facility in rural districts (%)	78.7	84.0	−5.3 (0.405)
Distance (km) from district headquarter, mean [SD]	56.9 [38.8]	62.9 [41.8]	−6.0 (0.367)
Poorest SES facilities (%)	40.0	26.7	13.3 (0.084)
Middle SES facilities (%)	34.7	32.0	2.7 (0.731)
Least poor SES facilities (%)	25.3	41.3	−16.0 (0.038)

SD is for standard deviation

availability were also less likely to be out of stock. In addition, there was a borderline significant 10.2 percentage point reduction in vaccine stock-outs (*P* = 0.073) and a 13.6 percentage point reduction in stock-outs of family planning medicines (*P* = 0.062) (Table 4). The effects of P4P on IPT and partograph stock-outs were not significant.

P4P reduced the stock-out of medicines across the RMNCH continuum of care and that of medical supplies benefiting mothers and newborns (Appendix S1b). Effects on availability were most pronounced for maternal, newborn and child medicines and reproductive health supplies.

The effect of P4P on the stock-outs of medicines overall was pro-poor, with reduction in facilities in the poorest tercile being 24.5 percentage points greater than that in the least poor tercile (*P* = 0.019); specifically, the effects on stock-outs of antimalarials, antibiotics and oxytocics were pro-poor; effects on antimalarial availability were also marginally pro-poor (Table 5). P4P had a greater effect on the availability of medicines and medical supplies in facilities in rural districts (by 10.4 percentage points, *P* = 0.051; and 22 percentage points, *P* = 0.003, respectively). Similarly, the effect of P4P on availability and stock-outs of antimalarials was greater in facilities in rural than urban districts (23.1 percentage points, *P* = 0.020; and 23.1 percentage points, *P* = 0.070, respectively). The effect of P4P on availability and stock-outs of antihypertensives was greater in health centres and hospitals than in dispensaries [by 19.9 percentage points (*P* = 0.020) and 26.1 percentage points (*P* = 0.064), respectively]. There were no differential effects by facility ownership.

When standard errors were clustered at the district level, the effects on the availability of antimalarials, oxytocics and delivery care drugs combined and, on stock-outs of oxytocics, vaccines and delivery care drugs combined were maintained (results not shown). However, the effects on composite indices for medicines combined and medical supplies were no longer significant.

Discussion

We examined the effects of P4P on the availability and stock-out rate of medical commodities for RMNCH. P4P was associated with significant improvements in availability and reductions in stock-outs of medicines and medical supplies, but had no effect on the availability of equipment. Among medicines, the main effects were for drugs associated with the delivery of some incentivised services: antimalarials, drugs to induce labour and manage bleeding (oxytocics) or manage hypertension during delivery (antihypertensives). However, there was little or no evidence of effects on medicines linked to other incentivised

Table 3 Effects of P4P on the availability of medical commodities mean score

	Baseline survey			Follow-up survey			Difference in differences, effect			
Category	P4P facilities	Control facilities	Difference	P4P facilities	Control facilities	Difference	*N*	Beta† [95% CI]	*P*–value	%*D**
Medicines combined (%)	60.8	65.7	−4.9**	63.9	60.7	3.2	295	8.4 [3.0 to 13.7]	0.002	13.8
Antimalarials – all (%)	60.3	69.9	−9.6**	69.7	59.3	−10.4**	295	20.5 [11.8 to 29.3]	0.000	33.9
Antimalarials – targeted (%)	74.6	93.2	−18.6***	96.0	90.7	5.3	295	25.2 [11.1 to 39.4]	0.001	33.8
Antibiotics (%)	36.3	39.9	−3.6	43.1	39.8	3.3	295	7.4 [0.8 to 14.1]	0.028	20.4
Antihypertensives (%)	36.2	37.1	−0.9	43.8	36.4	7.4*	295	8.7 [0.4 to 16.9]	0.040	24.0
Antidiarrhoeals (%)	60.6	63.5	−2.9	74.0	75.3	−1.3	295	1.9 [−12.5 to 16.3]	0.795	3.1
Oxytocics (%)	42.7	45.0	−2.3	45.8	32.9	−12.9***	295	15.0 [3.0 to 26.9]	0.014	35.1
Delivery care drugs – targeted (%)	39.5	41.1	−1.6	44.8	34.6	10.2***	295	11.8 [3.8 to 19.8]	0.004	29.9
ARVs – targeted (%)	55.4	50.3	5.1	57.4	60.4	−3.0	210	−7.9 [−20.3 to 4.7]	0.208	14.3
Vaccines – all (%)	94.8	92.9	1.9	96.9	92.9	4.0*	276	5.3 [−2.7 to 13.3]	0.193	5.6
Vaccines – targeted (%)	95.2	92.7	2.5	97.1	94.8	2.3	276	3.1 [−5.4 to 11.5]	0.475	3.3
Vitamin A (%)	91.9	91.8	0.1	92.9	92.9	0.0	276	3.2 [−8.7 to 15.0]	0.597	3.5
Family planning – targeted (%)	91.7	99.5	−7.8**	56.5	59.5	−3.0	255	7.3 [−4.6 to 19.3]	0.227	7.9
Medical supplies (%)	64.4	72.4	−8.0**	66.4	66.4	0.0	299	8.3 [0.01 to 16.5]	0.050	12.9
Partograph – targeted (%)	63.5	75.8	−12.3	77.0	76.0	1.0	274	16.1 [−3.0 to 35.3]	0.098	25.4
Medical equipment (%)	55.0	54.9	0.1	72.8	68.8	4.0	299	3.8 [−4.9 to 12.6]	0.391	6.9

Items included for medicines combined [37], medical supplies [11] and equipment [16]; 'targeted' are commodities linked to services targeted/incentivised by P4P; number of observations (*N*) is small for ARVs, family planning and vaccines because not all facilities stock these commodities; *the % *D* = (Beta / baseline mean) ×100, where the baseline mean of the dependent variable is for the intervention facilities; †the Beta is the estimated intervention effect controlling for a year dummy and facility fixed effects; *** denotes significance at 1%, ** at 5% and * at 10% level.

services such as vaccines, family planning, ARVs and supplies such as the partograph. P4P improved the availability/reduced stock-outs for some of the drugs that districts were incentivised for, including antibiotics (ampicillin, amoxicillin, gentamycin and flagyl). However, the scheme also reduced stock-outs of antibiotics that were not tied to any incentive (e.g. cotrimoxazole, chloramphenicol and crystapen injection). This suggests that P4P schemes have the potential to improve drug availability beyond those drugs that are directly linked to the delivery of incentivised services. Effects were generally equally distributed across facilities, with effects on medicine stock-outs being pro-poor in many cases, and greater in facilities in rural compared to urban districts. Greater improvements in the availability/stock-out reduction of antihypertensives in higher-level facilities are likely reflective of the greater number of obstetric referral cases at these facilities and associated need.

There are a variety of potential pathways to P4P effects on medicines and supplies in our study. The effect may in part be due to the provision of medicines being a pre-condition for meeting certain performance targets (e.g. IPT during ANC). The financial autonomy resulting from bank accounts enabled facilities to use bonus funds and cost sharing revenue (from user fees and community-based insurance) to procure drugs and supplies, consistent with findings from a process evaluation carried out alongside this study [42]. Incentives to district managers to limit drug stock-outs were also important, given the role of district managers in the procurement and supply process. By providing incentives to facilities and districts, the scheme ensured that stakeholders at all levels were working towards the same goals. The verification system under P4P also meant that district supervision was intensified, providing more opportunities for district managers to identify and address stock-outs of a wider range of drugs.

A number of medicines associated with incentivised services were not affected by P4P (vaccines, ARVs and family planning). The procurement of these items depended on donor funding [24, 28, 29]. The average availability of vaccines was above 94% at baseline (91% for family planning), so there was also little scope for improvement. Tanzania faced a problem with shortages of ARVs during the period of this study due to the introduction of a new treatment regimen, weak procurement

Table 4 Effects of P4P on the stock-out of medical commodities mean score

	Baseline survey			Follow-up survey			Difference in differences, effect			
Category	P4P facilities	Control facilities	Difference	P4P facilities	Control facilities	Difference	*N*	Beta† [95% CI]	*P*–value	%*D**
Medicines combined (%)	43.1	33.5	9.6***	26.7	30.4	−3.7	295	−13.6 [−22.1 to −5.1]	0.002	31.6
Antimalarials – all (%)	41.9	42.6	−0.7	29.8	40.4	−10.6**	295	−10.5 [−21.6 to 0.6]	0.064	25.1
Antimalarials – targeted (%)	27.1	18.9	8.2	6.7	10.7	−4.0	294	−13.3 [−29.8 to 3.2]	0.113	49.1
Antibiotics (%)	59.1	47.9	11.2**	41.0	45.2	−4.2	295	−16.6 [−29.5 to −3.8]	0.012	28.1
Antihypertensives (%)	57.0	46.0	11.0**	34.9	44.0	−9.1*	295	−21.0 [−35.1 to −6.9]	0.004	36.8
Antidiarrhoeals (%)	42.9	36.9	6.0	26.0	27.3	−1.3	294	−5.9 [−22.2 to 10.3]	0.472	13.8
Oxytocics (%)	55.2	39.3	15.9***	36.9	48.9	−12.0**	294	−27.2 [−43.7 to −10.7]	0.001	49.3
Delivery care drugs – targeted (%)	56.1	42.4	13.7***	35.9	46.4	−10.5**	295	−24.7 [−38.4 to −11.0]	0.000	44.0
ARVs – targeted (%)	40.6	32.5	8.1	25.0	25.0	0.0	210	−4.9 [−22.8 to 12.9]	0.585	12.1
Vaccines – all (%)	17.1	12.9	4.2	6.9	9.3	−2.4	276	−10.2 [−21.4 to 0.9]	0.073	59.6
Vaccines – targeted (%)	15.6	11.9	3.7	6.7	7.0	0.3	276	−7.4 [−18.8 to 4.1]	0.206	47.4
Vitamin A (%)	14.5	8.2	6.3	10.0	7.0	3.0	276	−6.6 [−20.3 to 7.0]	0.339	45.5
Family planning – targeted (%)	45.4	38.2	7.2	24.0	25.2	−1.2	255	−13.6 [−27.9 to 0.7]	0.062	29.9
Medical supplies (%)	39.7	29.4	10.3**	20.8	21.8	−1.0	286	−13.1 [−23.1 to −3.2]	0.010	32.9
Partograph – targeted (%)	33.9	18.6	15.3*	13.9	13.0	0.0	262	−12.3 [−31.9 to 7.3]	0.217	36.3

Items included for medicines combined [37], medical supplies [11] and equipment [16]; ‘targeted’ are commodities linked to services targeted/incentivised by P4P; number of observations (*N*) is small for ARVs, family planning and vaccines because not all facilities stock these commodities; *the % *D* = (Beta/baseline mean) ×100, where the baseline mean of the dependent variable is for the intervention facilities; †the Beta is the estimated intervention effect controlling for a year dummy and facility fixed effects; *** denotes significance at 1%, ** at 5% and * at 10% level.

mechanisms, and shortages of ARVs on the global market that were outside of facilities’ control [43]. The lack of effect on equipment availability may be due to the lack of incentives attached to equipment availability at the facility or the district level. The cost of equipment is also higher than that of many drugs and supplies, which may have deterred facilities from such investments.

Our study stands in contrast to a recent review from low- and middle-income countries concluding that P4P is not effective in improving structural quality of care [44]. However, our finding of increased availability of drugs is consistent with that reported from South Kivu Province in the DRC [17], but contrary to the findings from Afghanistan [15], Burundi [16, 45] and Katanga Province in the DRC [18] that showed no effects. The differences in context and variation in programme design likely explain the difference in effects. In Afghanistan, Burundi and the DRC drugs/supplies were incentivised through service targets, and providers had financial autonomy as in Tanzania [15–17], and in Burundi, up to 50% of the bonus could be used to procure drugs; however, this was not clearly the case in the other settings. Unlike the Pwani scheme, many schemes weight bonus payments with structural quality scores, which include the availability of drugs and supplies [15–17]. While facilities could channel a percentage of their bonus to districts in the DRC [46], districts were not directly incentivised, nor were they incentivised in other settings.

Despite the importance of assessing distributional effects within programme evaluation [47, 48], ours is the first study to examine the heterogeneity of the effect of P4P on medical commodities. The pro-poor effects on medicines are encouraging as are the pro-rural effects and these are consistent with universal health coverage (UHC) goals and efforts to meet sustainable development goal (SDG) 3.

There are several limitations to this study. First, we used household data from the facility catchment area to proxy the SES of the facility’s location based on a sample of 20 households that may not have accurately reflected the entire catchment population. Second, there was an imbalance in SES across study arms, although our results were reasonably robust when dividing facilities into SES groups in each arm separately. Third, we were unable to control for time-varying confounding factors due to a

Table 5 Heterogeneity of P4P effects on the availability and stock-out of medical commodities mean score

	Average effects			Differential effects by facility characteristics				
	Difference in differences, effect			Facility SES		Facility location (=1 if in rural district)	Ownership status (=1 if public facility)	Level of care (=1 if dispensary facility)
Category	*N*	Beta† [95% CI]	*P*-value	(=1 if poorest SES) *β* (*P*–value)	(=1 if middle SES) *β* (*P*–value)	*β* (*P*–value)	*β* (*P*–value)	*β* (*P*–value)
Panel A: Availability								
Medicines combined (%)	295	8.4 [3.0 to 13.7]	0.002	8.1 (0.173)	1.5 (0.817)	**10.4 (0.051)**	7.4 (0.180)	2.8 (0.578)
Antimalarials – all (%)	295	20.5 [11.8 to 29.3]	0.000	**18.9 (0.071)**	**25.9 (0.022)**	**23.1 (0.020)**	14.3 (0.119)	1.9 (0.844)
Antimalarials – targeted (%)	295	25.2 [11.1 to 39.4]	0.001	19.7 (0.319)	19.0 (0.333)	**24.2 (0.078)**	–23.1 (0.214)	–11.4 (0.522)
Antibiotics (%)	295	7.4 [0.8 to 14.1]	0.028	13.9 (0.113)	6.9 (0.409)	7.6 (0.421)	5.3 (0.617)	–5.6 (0.453)
Antihypertensives (%)	295	8.7 [0.4 to 16.9]	0.040	-9.3 (0.398)	–7.2 (0.485)	–2.6 (0.781)	12.8 (0.216)	**–19.9 (0.020)**
Oxytocics (%)	295	15.0 [3.0 to 26.9]	0.014	–10.8 (0.476)	–5.2 (0.731)	4.0 (0.762)	–3.4 (0.776)	5.4 (0.666)
Delivery care drugs – targeted (%)	295	11.8 [3.8 to 19.8]	0.004	–10.0 (0.322)	–6.2 (0.548)	0.7 (0.935)	4.7 (0.566)	–7.3 (0.365)
Medical supplies (%)	299	8.3 [0.01 to 16.5]	0.050	2.8 (0.815)	1.9 (0.849)	**22.0 (0.003)**	6.2 (0.522)	–4.9 (0.539)
Partograph – targeted (%)	274	16.1 [-3.0 to 35.3]	0.098	**41.5 (0.065)**	25.2 (0.307)	17.5 (0.510)	9.9 (0.652)	3.8 (0.850)
Panel B: Stock-out								
Medicines combined (%)	295	−13.6 [−22.1 to −5.1]	0.002	**−24.5 (0.019)**	−16.5 (0.110)	−8.9 (0.342)	2.5 (0.771)	7.4 (0.383)
Antimalarials – all (%)	295	−10.5 [−21.6 to 0.6]	0.064	**−23.6 (0.098)**	−23.6 (0.111)	**−23.1 (0.070)**	5.7 (0.624)	5.0 (0.696)
Antibiotics (%)	295	−16.6 [−29.5 to −3.8]	0.012	**−32.0 (0.032)**	−6.7 (0.668)	−15.3 (0.322)	−7.7 (0.624)	**23.9 (0.095)**
Antihypertensives (%)	295	−21.0 [−35.1 to −6.9]	0.004	−12.5 (0.465)	−5.9 (0.733)	−4.9 (0.743)	13.9 (0.263)	**26.1 (0.064)**
Oxytocics (%)	294	−27.2 [−43.7 to −10.7]	0.001	**−49.1 (0.017)**	−27.5 (0.171)	−14.4 (0.487)	−1.3 (0.946)	−17.7 (0.311)
Delivery care drugs – targeted (%)	295	−24.7 [−38.4 to −11.0]	0.000	**−30.8 (0.062)**	−18.4 (0.272)	−10.3 (0.512)	5.6 (0.680)	6.4 (0.649)
Vaccines – all (%)	276	−10.2 [−21.4 to 0.9]	0.073	1.0 (0.946)	−4.3 (0.749)	14.5 (0.160)	3.7 (0.749)	**17.9 (0.082)**
Family planning – targeted (%)	255	−13.6 [−27.9 to 0.7]	0.062	−29.6 (0.127)	−21.3 (0.128)	3.7 (0.842)	−11.9 (0.402)	8.2 (0.560)
Medical supplies (%)	286	−13.1 [−23.1 to −3.2]	0.010	−9.6 (0.467)	8.9 (0.471)	−1.9 (0.869)	−5.4 (0.543)	−0.9 (0.927)

Reference category in brackets: for poorest and middle SES (least poor SES), rural (urban), public (non-public) and dispensary (hospital & health centres); †the Beta is the estimated average intervention effect controlling for a year dummy and facility fixed effects; *β* is the estimated differential effects of P4P controlling for a year dummy and facility fixed effects; and statistically significant differential effects in bold (*P*-value < 0.10).

lack of data, but confounding bias due to time-invariant factors were adjusted through fixed effects estimation. Fourth, although we tested and confirmed the assumption of parallel trends in facility utilisation outcomes prior to the intervention, we failed to test that of drug availability and stock-out outcomes due to a lack of historical data on these outcomes. We were also unable to capture seasonal fluctuations in drug availability as this requires time series data which were not available. Finally, potential type I errors due to multiple hypotheses testing are a concern to inference; however, we used subgroups of items to minimise the risk of this error.

Conclusion

Our study has shown that P4P, when introduced with facility and district-level incentives and in a context where facilities and local government authorities have autonomy over the use of funds, can improve the availability of drugs and supplies and enhance good quality of care. This makes services more acceptable, effective and affordable, especially in facilities serving poor, rural populations, enhancing progress towards universal health coverage [1, 10].

Acknowledgements

We thank Kara Hanson, Masuma Mamdani, Gaute Torsvik, Bjarne Robberstad and Iddy Mayumana for reviewing the paper and for valuable comments and suggestions. We acknowledge the process evaluation findings provided by Masuma Mamdani and Iddy Mayumana that supplement this study. We also thank the whole P4P evaluation research team, including data collectors and field coordinators. The Government of Norway funded the data collection for the programme evaluation that was used in this paper. The Research Council of Norway and the UK Department for International Development as part of the Consortium for Research on Resilient and Responsive Health Systems supported the funding of the authors time undertaking data analysis and writing. This study is part of a PhD thesis at the University of Bergen for Peter Binyaruka, who is financially supported by the Norwegian State Education Loan Fund. The funding bodies had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. WHO. *Equitable Access to Essential Medicines: A Framework for Collective Action. Volume March*. World Health Organization: Geneva, 2004.
2. Donabedian A. The quality of care. How can it be assessed? *JAMA* 1988: **260**: 1743–1748.
3. Donabedian A. Evaluating the quality of medical care. *Milbank Mem Fund Q* 2005: **83**: 691–729.
4. Quick JD, Boohene N-A, Rankin J, Mbwasi RJ. Medicines supply in Africa. *BMJ* 2005: **331**: 709–710.
5. Uzochukwu BS, Onwujekwe OE, Akpala CO. Effect of the Bamako-Initiative drug revolving fund on availability and rational use of essential drugs in primary health care facilities in south-east Nigeria. *Heal Policy Plan* 2002: **17**: 378–383.
6. Macha J, Harris B, Garshong B *et al.* Factors influencing the burden of health care financing and the distribution of health care benefits in Ghana, Tanzania and South Africa. *Health Policy Plan* 2012: **27**(Suppl 1): 46–54.
7. Penfold S, Shamba D, Hanson C *et al.* Staff experiences of providing maternity services in rural southern Tanzania - a focus on equipment, drug and supply issues. *BMC Health Serv Res* 2013: **13**: 61.
8. Mkoka DA, Goicolea I, Kiwara A, Mwangu M, Hurtig A-K. Availability of drugs and medical supplies for emergency obstetric care: experience of health facility managers in a rural District of Tanzania. *BMC Pregnancy Childbirth* 2014: **14**: 108.
9. UNCoLSC. *Scaling Up Life Saving Commodities for Women, Children, and Newborns - An Advocacy Toolkit*. PATH: Washington, DC, 2015.
10. Cameron A, Ewen M, Ross-Degnan D, Ball D, Laing R. Medicine prices, availability, and affordability in 36 developing and middle-income countries: a secondary analysis. *Lancet* 2009: **373**: 240–249.
11. United Nations. *UN Commission on Life-Saving Commodities for Women and Children: Commissioner's Report*. United Nations: New York, 2012 (September).
12. Pronyk PM, Nemser B, Maliqi B *et al.* The UN Commission on Life Saving Commodities 3 years on: global progress update and results of a multicountry assessment. *Lancet Glob Heal* 2016: **4**: e276–e286.
13. Witter S, Fretheim A, Kessy FL, Lindahl AK. Paying for performance to improve the delivery of health interventions in low- and middle-income countries. *Cochrane Database Syst Rev* 2012, 2:CD007899.
14. Meessen B, Soucat A, Sekabaraga C. Performance-based financing: just a donor fad or a catalyst towards comprehensive health-care reform? *Bull World Health Organ* 2011: **89**: 153–156.
15. Engineer CY, Dale E, Agarwal A *et al.* Effectiveness of a pay-for-performance intervention to improve maternal and child health services in Afghanistan: a cluster-randomized trial. *Int J Epidemiol* 2016: **45**: 1–9.
16. Bonfrer I, Soeters R, Van de Poel E *et al.* Introduction of performance-based financing in burundi was associated with improvements in care and quality. *Health Aff (Millwood)* 2014: **33**: 2179–2187.
17. Soeters R, Peerenboom PB, Mushagalusa P, Kimanuka C. Performance-based financing experiment improved health

care in the Democratic Republic of Congo. *Health Aff (Millwood)* 2011: **30**: 1518–1527.

18. Huillery E, Seban J. *Pay-for-Performance, Motivation and Final Output in the Health Sector: Experimental Evidence from the Democratic Republic of Congo*. Sciences Po Economics Discussion Papers, 2014.
19. Frumence G, Nyamhanga T, Mwangu M, Hurtig AK. Challenges to the implementation of health sector decentralization in Tanzania: Experiences from kongwa district council. *Glob Health Action* 2013: **6**: 1–11.
20. MoHSW. *Mapping of the Medicines Procurement and Supply Management System in Tanzania*. Ministry of Health and Social Welfare (MoHSW): Dar es Salaam, 2008.
21. Euro Health Group. *The United Republic of Tanzania Drug Tracking Study*. Euro Health Group: Denmark, 2007 (August).
22. SIKIKA. *Medicines and Medical Supplies Availability Report. Using Absorbent Gauze Availability Survey as an Entry Point. A Case of 71 Districts and 30 Health Facilities across Mainland Tanzania*. SIKIKA: Dar es Salaam, Tanzania, 2011.
23. MoHSW. *Mid Term Review of the Health Sector Strategic Plan III 2009-2015: Health Care Financing*. Technical Report, Ministry of Health and Social Welfare (MoHSW), United Republic of Tanzania, 2013 (October 2013).
24. MoHSW. *Tanzania Health Sector Strategic Plan 2015 -2020 (HSSP IV)*. Ministry of Health and Social Welfare (MoHSW): Dar es Salaam, 2015 (July).
25. Centre for Pharmaceutical Management. *Accredited Drug Dispensing Outlets in Tanzania: Strategies for Enhancing Access to Medicines Program*. Management Sciences for Health: Arlington, VA, 2008.
26. Zomboko FE, Tripathi SK, Kamuzora FK. Challenges in procurement and use of donated medical-equipments: study of a Selected Referral Hospital in Tanzania. *J Arts Sci Commer* 2012: **4**: 41–48.
27. Rutta E, Liana J, Embrey M *et al*. Accrediting retail drug shops to strengthen Tanzania's public health system: an ADDO case study. *J Pharm policy Pract* 2015: **8**: 23.
28. MoHSW. *Tanzania Mainland Expanded Programme on Immunization (EPI) Review*. Ministry of Health and Social Welfare (MoHSW): Dar es Salaam, 2010 (July).
29. Yadav P, Lega Tata H, Bababley M. *Storage and Supply Chain Management*. The World Medicines Situation 2011, WHO: Geneva, 2011.
30. Wales J, Tobias J, Malangalila E, Swai G, Wild L. *Stock-Outs of Essential Medicines in Tanzania: A Political Economy Approach to Analysing Problems and Identifying Solutions*. TWAWEZA: Twaweza ni sisi, Dar es Salaam, Tanzania, 2014 (March).
31. MoHSW. *Tanzania Service Availability and Readiness Assessment (SARA) 2012*. Ministry of Health and Social Welfare (MoHSW) and Ifakara Health Institute: Dar es Salaam, 2013.
32. USAID. *Tanzania Health System Assessment 2010*. Health Systems 20/20 project. Abt Associates Inc.: Bethesda, MD, 2011 (January).
33. NBS. *2012 Population and Housing Census: Population Distribution by Administrative Areas*. National Bureau of Statistics (NBS): Dar es Salaam, 2013.
34. Borghi J, Mayumana I, Mashasi I *et al*. Protocol for the evaluation of a pay for performance programme in Pwani region in Tanzania: A controlled before and after study. *Implement Sci* 2013: **8**: 80.
35. Binyaruka P, Patouillard E, Powell-Jackson T, Greco G, Maestad O, Borghi J. Effect of paying for performance on utilisation, quality, and user costs of health services in Tanzania: a controlled before and after study. *PLoS One* 2015: **10**: e0135013.
36. Kamuzora P, Gilson L. Factors influencing implementation of the Community Health Fund in Tanzania. *Health Policy Plan* 2007: **22**: 95–102.
37. Choi Y, Ametepi P. Comparison of medicine availability measurements at health facilities: evidence from Service Provision Assessment surveys in five sub-Saharan African countries. *BMC Health Serv Res* 2013: **13**: 266.
38. Filmer D, Pritchett LH. Estimating wealth effects without expenditure data–or tears: an application to educational enrollments in states of India. *Demography* 2001: **38**: 115–132.
39. Vyas S, Kumaranayake L. Constructing socio-economic status indices: how to use principal components analysis. *Health Policy Plan* 2006: **21**: 459–468.
40. Cameron A, Miller D. A Practitioner's Guide to Cluster-Robust Inference. *J Hum Resour* 2015: **50**: 317–372.
41. Anselmi L, Binyaruka P, Borghi J. Understanding causal pathways within health systems policy evaluation through mediation analysis: an application to payment for performance (P4P) in Tanzania. (Forthcoming).
42. Mayumana I, Borghi J, Anselmi L, Mamdani M, Lange S. Effects of Payment for Performance on accountability mechanisms: evidence from Pwani, Tanzania. (Forthcoming).
43. SIKIKA. *Shortage of Antiretrovirals: What Went Wrong?* SIKIKA: Dar es Salaam, Tanzania, 2014.
44. Das A, Gopalan S, Chandramohan D. Effect of pay for performance to improve quality of maternal and child care in low- and middle-income countries: a systematic review. *BMC Public Health* 2016: **16**: 1–11.
45. Rudasingwa M, Soeters R, Bossuyt M. The effect of performance-based financial incentives on improving health care provision in Burundi: a controlled cohort study. *Glob J Health Sci* 2015: **7**: 15–29.
46. Fox S, Witter S, Wylde E, Mafuta E, Lievens T. Paying health workers for performance in a fragmented, fragile state: reflections from Katanga Province, Democratic Republic of Congo. *Health Policy Plan* 2014: **29**: 96–105.
47. Djebbari H, Smith J. Heterogeneous impacts in PROGRESA. *J Econom* 2008: **145**: 64–80.

48. Markovitz AA, Ryan AM. Pay-for-Performance: Disappointing Results or Masked Heterogeneity? *Med Care Res Rev* 2016: doi: 10.1177/1077558715619282.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 (a) List of medical commodities and classification. (b) Effects of P4P on mean score availability and stock-out of medical commodities [By RMNCH classification]. (c) Items used to construct household socioeconomic status score.

Corresponding Author Peter Binyaruka, Ifakara Health Institute, PO Box 78373, Dar es Salaam, Tanzania. E-mail: pbinyaruka@ihi.or.tz

II

Does payment for performance increase performance inequalities across health providers? A case study of Tanzania

Peter Binyaruka [a, b, c] *, Bjarne Robberstad [a], Gaute Torsvik [c, d], & Josephine Borghi [e]

[a] University of Bergen, PO Box 7804, N-5020, Bergen, Norway.
[b] Ifakara Health Institute, PO Box 78373, Dar es Salaam, Tanzania.
[c] Chr. Michelsen Institute, PO Box 6033, Bergen, Norway.
[d] University of Oslo, PO Box 1095, Oslo, Norway.
[e] London School of Hygiene & Tropical Medicine, 15-17 Tavistock Place, WC1H 9SH, London, UK.

* Corresponding author:
PO Box 78373, Dar es Salaam, Tanzania; E-mail address: pbinyaruka@ihi.or.tz

Highlights

- We examined the distribution of performance outcomes due to P4P across facilities.
- Inequality in payouts favoured better-off facilities, but declined over time.
- Lower baseline performers improved most on institutional deliveries coverage.
- Rural and middle wealth facilities improved most on deliveries coverage.
- Performance on antimalarial provision was similar across facilities.

Abstract

There is growing evidence evaluating the impact of payment-for-performance (P4P) schemes in the health sector, but there has been little attention to the distributional effects of P4P across health facilities, and whether P4P reduces or enhances performance inequalities across facilities. We examined the distribution of P4P bonus payouts and two service coverage outcomes: institutional deliveries and provision of antimalarials during antenatal care (ANC), which differed in terms of incentive design and across facility subgroups in Tanzania. We used data from 150 facilities from intervention and comparison areas in January 2012 and 13 months later. Service coverage outcomes and socioeconomic status of facility catchment populations were measured in a household survey, facility characteristics from facility survey, and data on performance payouts were obtained from the programme administrator. Descriptive inequality measures were used to examine the distribution of payouts across facility subgroups. Daifference-in-differences regression analyses were used to identify P4P differential effects on the two service coverage outcomes across facility subgroups. We found that performance payouts were initially higher among higher level facilities (hospitals and health centres than dispensaries), facilities with more medical commodities and among facilities serving wealthier populations, but these inequalities declined over time. P4P had greater effects on coverage of institutional deliveries among facilities with low baseline performance, serving middle wealth populations, and located in rural areas. P4P effects on antimalarials provision during ANC was similar across facilities. Performance inequalities were influenced by the design of incentives and a range of facility characteristics, however, the nature of the service being targeted is also likely to have affected provider response. Further research is needed to further examine the effects of incentive design on outcomes and researchers should be encouraged to report on design aspects in their evaluations of P4P and systematically monitor and report sub-group effects across providers.

Key words. Payment-for-performance, inequality, impact evaluation, incentive design, Tanzania

Introduction

Payment-for-Performance (P4P) programmes, involving financial incentives to health workers and/or health facilities for achievement of pre-defined performance outcomes, have been widely implemented. These programmes are generally aimed at improving quality of care especially in high-income countries (HICs) to improve quality of care (Eijkenaar et al. 2013), while in low- and middle-income countries (LMICs) the emphasise has also been on improving service coverage and to strengthen health systems (Meessen et al. 2011; Witter et al. 2013). However, mixed effects of P4P have been reported (Das et al. 2016; Eijkenaar et al. 2013; Gillam et al. 2012; Mendelson et al. 2017; Renmans et al. 2016; Witter et al. 2012).

Evaluations of P4P schemes have largely focused on average programme effects, with little attention to distributional effects (Markovitz & Ryan 2016). While the effects of P4P on service use inequalities among populations have been documented in the literature (Alshamsan et al. 2010; Binyaruka et al. 2018; Renmans et al. 2016; Van de Poel et al. 2016), there is little evidence of how P4P affects inequalities between health providers (Markovitz & Ryan 2016; Sherry et al. 2017) in relation to financing and service delivery outcomes. P4P is expected to reduce performance inequalities by motivating worse performing facilities to catch up (Fritsche et al. 2014; Meessen et al. 2011), but it is also possible that P4P increases performance inequalities by rewarding facilities that are better able to perform (Ireland et al. 2011). It is therefore important to assess how P4P affects different types of providers (Khandker et al. 2010; Markovitz & Ryan 2016), to ensure that P4P does not widen the resource gap and increase inequality in healthcare provision between providers (Blustein et al. 2010; Chien et al. 2012).

Some evidence from HICs shows that P4P can reduce performance inequalities by motivating lower performers to improve (Alshamsan et al. 2010; Markovitz & Ryan 2016). Despite substantial variation in health facility readiness to deliver services in LMICs (MoHSW 2013; O'Neill et al. 2013), there is limited evidence of the effect of P4P on performance inequalities across health facilities (Sherry et al. 2017). Therefore, we examined the distribution of P4P bonus payments and of the increased service coverage associated with P4P across health facilities in Tanzania.

Conceptual framework

To conceptualise the pathways to distributional effects of P4P, we adapted the theoretical framework by Rittenhouse et al. (2010) and Markovitz and Ryan (2016) to the Tanzanian context (Figure 1).

Suppose performance in period *t* (p_t) is given by facility-level effort (e_t), and a set of structural/ enabling factors (x_t): $p_t = p(e_t, x_t)$. Performance is also assumed differentiable and weakly increasing in both arguments: $\frac{\partial p}{\partial e} \geq 0$, $\frac{\partial p}{\partial x} \geq 0$. We then consider two types of facilities: those with high (p_0^H) and low baseline performance (p_0^L). At baseline we have the change: $\Delta^0 = p_0^H - p_0^L > 0$, and after P4P is introduced we have $\Delta^1 = p_1^H - p_1^L$. The incentive design structure and/or structural factors can modify the effects of P4P across facilities over time, resulting in convergence in performance/ positive distributional effects ($\Delta^0 > \Delta^1$); divergence in performance/ negative distributional effects

($\Delta^0 < \Delta^1$); or similar performance across facilities (i.e. zero distributional effects) ($\Delta^0 = \Delta^1$). We analysed the extent to which the incentive design and structural factors modify the effects of P4P across facilities.

Incentive design effect: P4P schemes can reward using fee-for-service, geographical targeting, relative performance, single absolute threshold targets, or multiple threshold targets (Eijkenaar 2013; Fritsche et al. 2014; Mehrotra et al. 2010; Rosenthal & Dudley 2007; Rosenthal et al. 2005). The distributional effects of P4P schemes will partly depend on how incentives, and especially targets, are designed.

Multiple threshold target designs can enhance convergence in performance (Eijkenaar 2013; Mehrotra et al. 2010; Rosenthal et al. 2005) because they account for baseline performance and provide incentives for lower performers to catch up. However, there is a study that reported the absence of systematic convergence in performance with this design in the UK (Sutton et al. 2012). Absolute single threshold/ linear targets can enhance divergence in performance if some providers are far above and below the target (Heath et al. 1999; Mehrotra et al. 2010; Miller & Babiarz 2013; Mullen et al. 2010; Rosenthal & Dudley 2007). Improvement is most likely for providers/facilities that are close to achieve the threshold target. Top performers have no incentive to improve, and those far below the target may perceive it as unattainable, a phenomenon referred to as "goal-gradient" theory (Heath et al. 1999). A single target design fails to account for any variation in baseline performance (Eijkenaar 2013; Mehrotra et al. 2010; Mullen et al. 2010; Rosenthal et al. 2005).

Structural effect: Variation in facility and area-based factors that are potentially responsible for inequalities in baseline performance, can also modify the effects of P4P programmes (Markovitz & Ryan 2016). This is given by $\frac{\partial p}{\partial x} \geq 0$. We further assume the change in effort devoted to affect performance $\frac{\partial p}{\partial e}$ is increasing in x, that is $\frac{\partial \frac{\partial p}{\partial x}}{\partial e} > 0$. If facilities invest initial bonus payments in enabling factors, this may improve their future performance, but general predictions of effects based on variation in structural factors are difficult to make (Markovitz & Ryan 2016). We hypothesise that public facilities in Tanzania are better able to respond to incentives than non-public providers, as they can offer free maternal and child health (MCH) services (under the fee exemption policy) and have more financial autonomy (Mayumana et al. 2017). However, it is also possible that P4P can level the playing field across providers of different ownership status (Meessen et al. 2011). We further hypothesise that facilities with greater resource availability (e.g. medical inputs) are better able to increase patient demand than their counterparts (Alderman & Lavy 1996; Donabedian 1988; WHO 2004) and that dispensaries are less able to respond to incentives compared to health centres and hospitals since they are more resource constrained (MoHSW 2013).

Regarding area-based factors, facilities with wealthier catchment populations may respond better to incentives, as they can more readily increase service use and revenue through user fees (Castro-Leal et al. 2000; Chien et al. 2012; Doran et al. 2008; Victora et al. 2000). Facilities in rural areas may be less able to respond to incentives than their urban counterparts, because of human resource shortages, poor road infrastructure, and more scattered and disadvantaged populations (Fritsche et al. 2014; Munga & Maestad 2009; Witter et al. 2013).

Apart from the above hypothesised pathways (incentive design and structural effect), provider response may also depend on the nature of the services targeted or incentivised. This is because performance improvement can be harder for some services compared to other services and this may confound the initial hypothesises of incentive design and structural effect. For instance, it is harder to change clients' behaviour for deliveries than for a content of care like IPT provision.

P4P in Tanzania

The public sector is the largest sector of the Tanzanian health system, with private for profit and the voluntary sector serving as important supplements (MoHSW 2015). The public health system has a hierarchal administrative structure, with dispensaries and health centres providing primary health care services, and district hospitals, regional hospitals, and national hospitals acting as referral facilities. The public health system in Tanzania is decentralised, with district-level managers being responsible for preparing annual health sector plans and generating and managing resources for the district.

In 2011, the Ministry of Health and Social Welfare (MoHSW) in Tanzania, with support from the Government of Norway, introduced a P4P scheme in all seven districts of Pwani region to improve MCH and inform the national P4P roll out. Pwani region has more than

300 health facilities covering a population of just over a million (NBS 2013). All facilities providing MCH services in Pwani were included in the scheme. P4P incentives were tied to coverage of services (e.g. institutional delivery) and content of care targets (e.g. provision of *intermittent preventive treatment* (IPT) for malaria during antenatal care (ANC)) (Binyaruka et al. 2015; Borghi et al. 2013). There were two methods of target setting (Table 1): a single threshold (absolute coverage target), and multiple thresholds based on performance in the previous cycle (relative change/ overall result) with five performance groups, each with their own absolute threshold: Group 1 (0-20% coverage of said indicator), group 2 (21-40%), group 3 (41-70%), group 4 (71-85%) and group 5 (>85%). Group 5 was required to maintain coverage. District and regional managers were rewarded for performance of facilities in their district or region.

Performance data were compiled by facilities and verified by the P4P implementing agency every six months (one cycle) before payments. The maximum payout per cycle differed by level of care: USD 820 per cycle for dispensaries; USD 3220 for health centres and USD 6790 for hospitals – the majority share of payout (90% in hospitals and 75% in lower level facilities) being staff bonuses, and the remainder for facility improvement/demand creation. Payments were additional to funding for operational costs and salaries which are unrelated to performance. Full payment was made if 100% of a given target was achieved, 50% of payment was made for 75-99% achievement, and no payment was made for lower levels of performance. Staff bonuses were equivalent to 10% of their monthly salary if all targets were fully attained. The maximum payout for district and regional managers was USD 3000 per cycle.

An impact evaluation of the P4P programme showed a significant positive effect on two out of eight incentivised service indicators: institutional delivery rate and provision of antimalarial during ANC (Binyaruka et al. 2015). The programme also increased the availability of drugs and supplies, increased supportive supervision, reduced payment of user fees, and resulted in greater provider kindness during delivery care (Anselmi et al. 2017; Binyaruka & Borghi 2017; Binyaruka et al. 2015; Mayumana et al. 2017).

Materials and methods

Study design and data sources

We used data from the impact evaluation of the P4P scheme in Pwani region, described elsewhere (Binyaruka et al. 2015; Borghi et al. 2013). The study surveyed all seven districts in Pwani region (intervention arm), and four districts from Morogoro and Lindi regions (comparison arm). Comparison districts were selected to be comparable to intervention districts in terms of poverty and literacy rates, the rate of institutional deliveries, infant mortality, population per health facility, and the number of children under one year of age per capita (Borghi et al. 2013).

Baseline data at facility and household-levels were collected in January 2012, with a follow-up round 13 months later. For each study arm, data on facility ownership, level of care, availability of medical inputs and rural/urban location was obtained from 75 sampled facilities providing MCH services (6 hospitals, 16 health centres and 53 dispensaries). Data on socioeconomic status of the facility catchment populations and service coverage rates were obtained from households with women who had delivered in the 12 months prior to the baseline and endline surveys. We randomly sampled 20 eligible households from each facility's catchment area, making a total of 1500 households in each arm per survey round. Facility payout data were obtained from the implementing agency for all incentivised indicators for the 75 intervention facilities in our sample over seven payment cycles (2011 –2014).

Performance outcomes

We considered two facility performance outcomes. First, we estimated for each payment cycle a "payout score" for each facility in the intervention arm, defined as the bonus payment received divided by the total potential payout if all targets had been met, multiplied by 100. Second, we estimated facility-level average service coverage rates for households in the facility catchment area from both study arms for the two incentivised services which improved significantly on average as a result of P4P (Binyaruka et al. 2015): the coverage of institutional deliveries (that used multiple-threshold target) and provision of two doses of *intermittent preventive treatment* (IPT2) for malaria during ANC (that used single threshold target).

Subgroups of facilities for distributional analyses

To examine whether incentive design and structural effects affected performance outcomes, we identified facility subgroups as shown in Table 2, pertaining to: their baseline performance for the two incentivised indicators (above or below the median);

facility characteristics (ownership, level of care, availability of utilities, rural-urban location); an un-weighted index of drug availability at baseline (Appendix: Table A1); and wealth of the catchment population, based on mean wealth index scores across households in the facility-catchment area generated by principal component analysis (Vyas & Kumaranayake 2006) (Appendix: Table A2).

Analysis

We first compared the sample means at baseline for each of the facility subgroups across study arms, and examined eventual differences between study arms using the t-test.

Distribution of bonus payouts

To assess how bonus payouts were distributed across intervention facilities, we used three measures of inequality: an absolute measure (the gap) and two relative measures (the ratio and the concentration index) (O'Donnell et al. 2008; WHO 2013). The gap was measured as the difference in payout scores between facility subgroups. The ratio was measured as the ratio of payout scores between subgroups. In relation to wealth subgroups, a positive (negative) gap and a ratio greater (less) than one defines a pro-rich (pro-poor) distribution, respectively. A gap of zero and a ratio of one defines an equal distribution. We tested whether the gaps were significantly different from zero by using t-tests.

The concentration index (CI) was computed on a ranking variable of area-based wealth status to examine wealth-related inequality in the distribution of payouts (Kakwani et al.

1997; O'Donnell et al. 2008). The CI ranges between [-1 and +1], with zero indicating equality between wealth subgroups, while negative and positive values indicating that payouts are pro-poor and pro-rich, respectively. We tested whether the CIs were significantly different from zero.

Distribution of service coverage outcomes

We measured the difference in mean baseline coverage of the two incentivised services between facility subgroups (the coverage gap (WHO 2013)) and tested for significant differences between subgroups.

We also assessed whether the effect of P4P on the coverage of these two incentivised services differed by facility subgroup. To this end, we used a linear difference-in-differences regression model with a three-way interaction term between the average treatment effect ($P4P_i \times \delta_t$) and facility subgrouping variable G_i. The associated two-order interaction terms were also included in the model as shown in Equation 1.

$$Y_{it} = \beta_0 + \beta_1(P4P_i \times \delta_t) + \beta_2\delta_t + \beta_3 Z_{it} + \beta_4(P4P_i \times \delta_t \times G_i) + \beta_5(P4P_i \times G_i)$$
$$+\beta_6(G_i \times \delta_t) + \gamma_i + \varepsilon_{it} \qquad (1)$$

where Y_{it} is the service coverage outcome of facility *i* at time *t*. $P4P_i$ is a dummy variable, taking the value 1 if a facility is exposed to P4P and zero otherwise. We controlled for unobserved time-invariant facility-level characteristics γ_i with facility fixed-effects estimation, and included δ_t for year fixed-effects. We also controlled for time-varying facility-level covariates Z_{it} (availability of electricity and water supply, and the mean wealth index for households sampled in the catchment area of the facility) as potential confounding factors. The error term is ε_{it}. We report the confidence interval based on standard errors clustered at the facility level to account for serial correlation of ε_{it} at the facility level. The coefficient of interest for the differential effect across facility subgroups is β_4.

Causal inference using the difference-in-differences approach relies on the key identifying assumption that the trends in outcomes would have been parallel across study arms in the absence of the intervention (Khandker et al. 2010). While this cannot be formally tested, we verified that the pre-intervention trends were parallel in women who had delivered in the past 12 months at baseline for the following outcomes for which we had monthly data: share of institutional deliveries, caesarean section deliveries, women who breastfeed within one hour of birth, and women who paid for delivery care (Anselmi et al. 2017; Binyaruka et al. 2015). We also verified that trends in facility service utilisation levels based on patient registers were parallel in the 2-year period preceding the introduction of P4P.

We performed some robustness checks. First, we re-estimated the model for institutional deliveries excluding hospitals (8% of facilities per arm), as hospitals have less clearly defined catchment populations. Second, we clustered the standard errors at the district level and used a bootstrapping method to adjust the small number of district–clusters (Cameron & Miller 2015). Third, we reclassified the mean wealth scores into two quantiles (below or above the median) to check whether the wealth effect was sensitive to classification of the wealth groupings. Lastly, apart from using a conventional parametric test (a t-test) to assess whether differences in payouts between subgroups were significant, a non-parametric test (Wilcoxon rank-sum test) was also used (Kitchen 2009). All the analyses were performed using STATA version 13.

Results

Facility and area-based characteristics were generally similar in the intervention and comparison arms at baseline (Table 2), although intervention facilities served poorer populations, and had marginally lower availability of drugs than comparison facilities.

Distribution of bonus payouts

There was an increase in average payout scores between payment cycle 1 (50.1% of total potential payout) and cycle 7 (77.7%) (Table 3), and the payouts were highest for facilities with least poor catchment populations. This pro-rich effect was supported by the positive equity gaps and concentration indices, and an equity ratio that is greater than one across all payment cycles (Table 3, column 5 –7). The inequalities were generally stronger in early compare to later cycles (Table 3).

Facilities with greater availability of drugs at baseline, and hospitals and health centres had significantly higher payout scores than facilities with more limited drug availability and dispensaries (Table 4, the gaps). The equity ratios were approximately one, near equality, between most subgroups (Table 4).

Distribution of service coverage outcomes

Baseline institutional delivery rates and coverage of IPT2 during ANC were similar between most facility subgroups (Table 5). However, baseline institutional delivery rates were highest among facilities with the least poor catchment populations, while coverage of IPT2 was highest among facilities with the poorest catchment populations. Coverage of IPT2 was higher among dispensaries that health centres and hospitals, but there were lower levels of coverage in both outcomes in the comparison arm at baseline (Table 5).

P4P resulted in a greater increase in institutional deliveries among facilities with lower baseline coverage levels than those with higher coverage levels (by 13.0 percentage points, p=0.006) (Table 6). P4P resulted in a greater increase in institutional deliveries among facilities serving middle wealth populations than those serving least poor populations (by 14.3 percentage points, p=0.004) (Table 6). P4P also resulted in a greater increase in institutional deliveries among facilities in rural compared to urban districts (by 10.0 percentage points, p=0.030). The effect of P4P on coverage of IPT2 was similar across all facility subgroups (Table 6).

The results on institutional deliveries were similar when we restricted the analysis to primary care facilities, except for the difference between rural/urban location which became insignificant (Table A3). The results were generally robust to clustering at the district level, except that there was no longer a differential effect on deliveries by wealth subgroups (Table A4). When two quantiles of wealth scores (lower and higher) were used, the differential effect for institutional deliveries became insignificant (Table A5). The use of non-parametric tests of differences between payouts across facilities revealed similar results to those using parametric tests (Table A6).

Discussion

We examined the distribution of P4P bonus payouts and programme effects on service coverage across facility subgroups in Tanzania. We specifically assessed whether performance was shaped by the design of the incentives and/or facility and area-based characteristics. This is the first study to examine the effects of P4P on bonus payout distribution and examine broadly whether there was supply-side heterogeneous P4P effects in a LMIC.

We found some evidence of an incentive design effect: lower baseline performers had greater improvements in the coverage of institutional deliveries (with multiple threshold targets); however, performance was similar across all providers in relation to IPT2 coverage (with a single threshold target). The characteristics of providers and their catchment populations were also found to matter, with hospitals and health centres, and facilities with wealthier catchment populations, and a better endowment of drugs, being better able to improve coverage of institutional deliveries and receive bonus payouts. However, the inequalities in payouts distribution declined over time. The effect of P4P on the coverage of institutional deliveries was also greater in rural facilities, with middle wealth catchment populations, however, effects on IPT2 coverage were similar across facility subgroups.

Our finding of convergence in performance payouts by wealth status over time is partly consistent with the "inverse equity hypothesis" (Victora et al. 2000). The hypothesis suggests that better-off groups will initially benefit from a new intervention, widening inequalities, but over time the worse-off will catch up. This is also consistent with US evidence that wealthier hospitals initially received higher payouts than their counterparts, but the distribution of payouts levelled over time due a change in the design of the scheme from only rewarding top performers only to rewarding any improvement where all providers were likely to receive a payout (Ryan et al. 2012).

The finding that P4P had greatest effect on institutional deliveries (with multiple threshold targets) among baseline lower performers is consistent with evidence on quality improvements from the UK (Doran et al. 2008), Canada (Li et al. 2014) and the US (Blustein et al. 2010; Chen et al. 2010; Jha et al. 2012; Lindenauer et al. 2007; Rosenthal et al. 2005). In Rwanda, however, most rewarded services based on fee-for-service improved most among facilities with middle baseline quality scores (Sherry et al. 2017). The convergence in performance in HICs was partly linked to a design with multiple threshold targets in the UK (Doran et al. 2008) and Canada (Li et al. 2014) and to a US design system that rewarded the highest performers and penalised the lowest performers (Lindenauer et al. 2007; Rosenthal et al. 2005). However, another study in the UK of a hospital incentive scheme with multiple thresholds found evidence of divergence in performance in relation to mortality outcomes linked to pneumonia but not for other conditions (Sutton et al. 2012).

Our finding that the effects of P4P on institutional deliveries differed according to the wealth status of facility catchment populations is somewhat different to that reported in the UK and US with respect to quality of care improvements (Alshamsan et al. 2010; Blustein et al. 2010; Chien et al. 2012; Doran et al. 2008; Gravelle et al. 2008; Kontopantelis et al. 2013). These studies found that providers serving low-income populations performed initially less well but improved most over time, whereas we found facilities serving middle wealth populations with initial low coverage improved more over time than those with least poor populations. Moreover, while we found that the effect of P4P on coverage of institutional deliveries was greater for rural facilities in Tanzania, a US study found no association between performance on quality and rural/urban location (Ryan & Blustein 2011); and studies in the UK showed that P4P had less effect in rural than in urban areas (Gravelle et al. 2008; Kontopantelis et al. 2013).

We found similar improvements on IPT2 coverage across facilities, which is in contrast to literature that suggests a design with a single threshold target, as used for IPT2, fails to account for baseline performance and can enhance divergence in performance (Eijkenaar 2013; Heath et al. 1999; Mehrotra et al. 2010; Mullen et al. 2010; Rosenthal & Dudley 2007; Rosenthal et al. 2005). Our finding might be explained by the almost universal coverage of one ANC visit in Tanzania (Binyaruka et al. 2015; TDHS 2016), and the nature of the targeted service (content of care, rather than service use) may have meant that minimal effort was needed for providers to achieve the target for IPT2.

Our results lend support to the notion that the incentive design, facility characteristics and the nature of services being targeted themselves, will determine how providers respond to P4P, their ability to achieve targets and receive bonus payouts, and the extent to which P4P leads to convergence or divergence in performance outcomes, or similar performance across providers. Although P4P is typically talked about as a single or uniform intervention, there is in fact substantial variation in incentive structures and scheme design across settings, and across the range of providers implementing P4P programmes (Eijkenaar 2013; Miller & Babiarz 2013). Our study supports the fact that these design details are crucial, particularly when it comes to determining the distributional effects of P4P across providers, and whether P4P will enhance or reduce existing performance inequalities (Rosenthal & Dudley 2007; Rosenthal & Frank 2006; Rosenthal et al. 2005; Ryan et al. 2012). Further research is needed to further examine the effects of incentive design on outcomes, and researchers should be encouraged to report on design aspects in their evaluations of P4P and systematically monitor and report subgroup effects across providers.

In addition to potential consideration to incentive design to increase the likelihood of reducing performance inequalities with the introduction of P4P, a number of policies could be introduced to tackle structural factors that contribute to inequalities in performance. For example, “equity bonuses” might be introduced to enhance performance among disadvantaged facilities so they benefit from payouts from the start (Fritsche et al. 2014; Meessen et al. 2011; Rosenthal & Dudley 2007). Facility readiness assessment studies and potential quality boosting investments are also important to harmonise the capacity to deliver services prior to P4P. These are standard practices for most P4P programmes funded by the World Bank in LMICs, and the national P4P rollout programme in Tanzania has similarly incorporated these practices.

This study has a number of limitations. First, the administrative data on payouts did not allow for a disaggregation of payouts by service indicator, and thus we used the total payout per cycle which reflects performance across all P4P indicators. Second, since information about payout distribution was limited to intervention facilities, our results represent associations rather than causal effects. Third, we used household data from a random sample of 20 households per facility to proxy service coverage at facility level and wealth status of the facility's catchment population and these may have not been representative of the entire catchment populations surrounding facilities. Furthermore, our analysis assumes that households in a facility's catchment population would have used the facility for care seeking, whereas it is possible that households bypassed their nearest provider to seek care at higher level or more distant facilities. Fourth, the finding of the convergence in coverage of institutional deliveries over 13 months may reflects a regression to the mean principle (a random fluctuation rather than a true causal effect) due to a 'shorter term' assessment (Barnett et al. 2005), although the distribution in terms of payouts over the 'longer term' of seven payment cycles showed a consistent pattern on convergence. Fifth, as our two service coverage outcomes differed both in terms of incentive design as well as the nature of the service being targeted, it was not possible to determine the extent to which the difference in provider performance response was due to the former or the latter. Finally, because of sample size constraints, we examined differential effects across facility subgroups using a three-way interaction term, and were unable to run separate models for each subgroup (subgroup effects) and compare their effects for better understanding of programme effect. We also classified baseline performance into two subgroups rather than five subgroups as used in the design, due to insufficient sample size. As a result, it was not possible to determine what effect the 'maintain coverage' target had on performance relative to the 'improve coverage' target.

Conclusion

In this study, P4P rewarded better-off facilities (hospitals, health centres, facilities with more medical commodities, and serving wealthier populations), more than worse of facilities in the short term; but these inequalities in the distribution of bonus payouts declined over time as worse of facilities caught up. The effect of P4P on coverage of institutional deliveries was greater among facilities with lower levels of baseline coverage, with middle wealth catchment populations, and located in rural areas; whereas the increase in IPT2 coverage was similar across facility subgroups. Performance inequalities were influenced by the design of incentives and a range of facility characteristics, however, the nature of the service being targeted is also likely to have affected provider response.

Acknowledgements: We would like to thank Kara Hanson and Ottar Mæstad for reviewing the paper and for valuable comments and suggestions. We also thank the whole P4P evaluation research team. We also acknowledge the support from all field respondents, and various government officials and P4P implementing agency.

Funding: This work was supported by the Government of Norway, which funded the data collection for the program evaluation that was used in this paper (grant numbers: TAN-3108 and TAN 13/0005. http://www.norad.no/en/). The UK Department for International Development (DFID) as part of the Consortium for Research on Resilient and Responsive Health Systems (RESYST) also supported the data analysis and writing of this paper. This study is part of a PhD thesis at the University of Bergen for Peter Binyaruka, who is financially supported by the Norwegian State Education Loan Fund. The funding bodies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Alderman H, Lavy V. 1996. Household responses to public health services: cost and quality tradeoffs. *The World Bank Research Observer*, **11**: 3-22.

Alshamsan R, Majeed A, Ashworth M, Car J, Millett C. 2010. Impact of pay for performance on inequalities in health care: systematic review. *J Health Serv Res Policy*, **15**: 178-84.

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. *Implement Sci*, **12**: 10.

Barnett AG, van der Pols JC, Dobson AJ. 2005. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol*, **34**: 215-20.

Binyaruka P, Borghi J. 2017. Improving quality of care through payment for performance: examining effects on the availability and stock-out of essential medical commodities in Tanzania. *Trop Med Int Health*, **22**: 92-102.

Binyaruka P, Patouillard E, Powell-Jackson T, et al. 2015. Effect of Paying for Performance on Utilisation, Quality, and User Costs of Health Services in Tanzania: A Controlled Before and After Study. *PLoS One*, **10**: e0135013.

Binyaruka P, Robberstad B, Torsvik G, Borghi J. 2018. Who benefits from increased service utilisation? Examining the distributional effects of payment for performance in Tanzania. *Int J Equity Health*, **17**: 14.

Blustein J, Borden WB, Valentine M. 2010. Hospital performance, the local economy, and the local workforce: findings from a US National Longitudinal Study. *PLoS Med*, **7**: e1000297.

Borghi J, Mayumana I, Mashasi I, et al. 2013. Protocol for the evaluation of a pay for performance programme in Pwani region in Tanzania: a controlled before and after study. *Implement Sci*, **8**: 80.

Cameron AC, Miller DL. 2015. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, **50**: 317-372.

Castro-Leal F, Dayton J, Demery L, Mehra K. 2000. Public spending on health care in Africa: do the poor benefit? *Bull World Health Organ*, **78**: 66-74.

Chen JY, Kang N, Juarez DT, Hodges KA, Chung RS. 2010. Impact of a Pay-for-Performance Program on Low Performing Physicians. *Journal for Healthcare Quality*, **32**: 13-22.

Chien AT, Wroblewski K, Damberg C, et al. 2012. Do physician organizations located in lower socioeconomic status areas score lower on pay-for-performance measures? *J Gen Intern Med*, **27**: 548-54.

Das A, Gopalan SS, Chandramohan D. 2016. Effect of pay for performance to improve quality of maternal and child care in low- and middle-income countries: a systematic review. *BMC Public Health*, **16**: 321.

Donabedian A. 1988. The quality of care: how can it be assessed? *Jama*, **260**: 1743-1748.

Doran T, Fullwood C, Kontopantelis E, Reeves D. 2008. Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework. *Lancet*, **372**: 728-36.

Eijkenaar F. 2013. Key issues in the design of pay for performance programs. *Eur J Health Econ*, **14**: 117-31.

Eijkenaar F, Emmert M, Scheppach M, Schoffski O. 2013. Effects of pay for performance in health care: a systematic review of systematic reviews. *Health Policy*, **110**: 115-30.

Fritsche G, Soeters R, Meessen B. 2014. Performance-Based Financing Toolkit. Washington DC: The World Bank.

Gillam SJ, Siriwardena AN, Steel N. 2012. Pay-for-performance in the United Kingdom: impact of the quality and outcomes framework: a systematic review. *Ann Fam Med*, **10**: 461-8.

Gravelle H, Sutton M, Ma A. 2008. Doctor behaviour under a pay for performance contract: further evidence from the Quality and Outcomes Framework. *CHE Research Paper 34*.

Heath C, Larrick RP, Wu G. 1999. Goals as reference points. *Cogn Psychol*, **38**: 79-109.

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

Jha AK, Joynt KE, Orav EJ, Epstein AM. 2012. The long-term effect of premier pay for performance on patient outcomes. *N Engl J Med*, **366**: 1606-15.

Kakwani N, Wagstaff A, Van Doorslaer E. 1997. Socioeconomic inequalities in health: measurement, computation, and statistical inference. *Journal of econometrics*, **77**: 87-103.

Khandker SR, Koolwal GB, Samad HA. 2010. Handbook on Impact Evaluation: Quantitative Methods and Practices. Washington DC: The World Bank.

Kitchen CM. 2009. Nonparametric vs parametric tests of location in biomedical research. *Am J Ophthalmol*, **147**: 571-2.

Kontopantelis E, Buchan I, Reeves D, Checkland K, Doran T. 2013. Relationship between quality of care and choice of clinical computing system: retrospective analysis of family practice performance under the UK's quality and outcomes framework. *BMJ Open*, **3**: 1-11.

Li J, Hurley J, DeCicca P, Buckley G. 2014. Physician response to pay-for-performance: evidence from a natural experiment. *Health Econ*, **23**: 962-78.

Lindenauer PK, Remus D, Roman S, et al. 2007. Public reporting and pay for performance in hospital quality improvement. *N Engl J Med*, **356**: 486-96.

Markovitz AA, Ryan AM. 2016. Pay-for-Performance: Disappointing Results or Masked Heterogeneity? *Med Care Res Rev*.

Mayumana I, Borghi J, Anselmi L, Mamdani M, Lange S. 2017. Effects of Payment for Performance on accountability mechanisms: Evidence from Pwani, Tanzania. *Soc Sci Med*, **179**: 61-73.

Meessen B, Soucat A, Sekabaraga C. 2011. Performance-based financing: just a donor fad or a catalyst towards comprehensive health-care reform? *Bull World Health Organ*, **89**: 153-6.

Mehrotra A, Sorbero ME, Damberg CL. 2010. Using the lessons of behavioral economics to design more effective pay-for-performance programs. *Am J Manag Care*, **16**: 497-503.

Mendelson A, Kondo K, Damberg C, et al. 2017. The Effects of Pay-for-Performance Programs on Health, Health Care Use, and Processes of Care: A Systematic Review. *Ann Intern Med*, **166**: 341-353.

Miller G, Babiarz KS. 2013. Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. *NBER Working Paper No. 18932.*

MoHSW. 2013. Tanzania Service Svailability and Seadiness Sssessment (SARA) 2012. Ministry of Health and Social Welfare and Ifakara Health Institute: Dar es Salaam.

MoHSW. 2015. Tanzania Health Sector Strategic Plan (HSSP IV) 2015-2020. Ministry of Health and Social Welfare (MoHSW): Dar es Salaam.

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

Munga MA, Maestad O. 2009. Measuring inequalities in the distribution of health workers: the case of Tanzania. *Hum Resour Health*, **7**: 4.

NBS. 2013. Tanzania Population and Housing Census: Population Distribution by Administrative Areas 2012. National Bureau of Statistics (NBS): Dar es Salaam.

O'Donnell O, Van Doorsslaer E, Wagstaff A, Lindelöw M. 2008. *Analyzing Health Equity Using Household Survey Data: A Guide to Techniques and Their Implementation*. World Bank Publications.

O'Neill K, Takane M, Sheffel A, Abou-Zahr C, Boerma T. 2013. Monitoring service delivery for universal health coverage: the Service Availability and Readiness Assessment. *Bull World Health Organ*, **91**: 923-31.

Renmans D, Holvoet N, Orach CG, 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 Plan*, **31**: 1297-309.

Rittenhouse DR, Shortell SM, Gillies RR, et al. 2010. Improving chronic illness care: findings from a national study of care management processes in large physician practices. *Med Care Res Rev*, **67**: 301-20.

Rosenthal MB, Dudley RA. 2007. Pay-for-performance: will the latest payment trend improve care? *Jama*, **297**: 740-4.

Rosenthal MB, Frank RG. 2006. What is the empirical basis for paying for quality in health care? *Med Care Res Rev*, **63**: 135-57.

Rosenthal MB, Frank RG, Li Z, Epstein AM. 2005. Early experience with pay-for-performance: from concept to practice. *Jama*, **294**: 1788-93.

Ryan AM, Blustein J. 2011. The effect of the MassHealth hospital pay-for-performance program on quality. *Health Serv Res*, **46**: 712-28.

Ryan AM, Blustein J, Doran T, Michelow MD, Casalino LP. 2012. The effect of Phase 2 of the Premier Hospital Quality Incentive Demonstration on incentive payments to hospitals caring for disadvantaged patients. *Health Serv Res*, **47**: 1418-36.

Sherry TB, Bauhoff S, Mohanan M. 2017. Multitasking and Heterogeneous Treatment Effects in Pay-for-Performance in Health Care: Evidence from Rwanda. *American Journal of Health Economics*, **3**: 192-226.

Sutton M, Nikolova S, Boaden R, et al. 2012. Reduced mortality with hospital pay for performance in England. *N Engl J Med*, **367**: 1821-8.

TDHS. 2016. Tanzania Demographic and Health Survey and Malaria Indicator Survey 2015-16. National Bureau of Statistics (NBS): Dar es Salaam.

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 Econ*, **25**: 688-705.

Victora CG, Vaughan JP, Barros FC, Silva AC, Tomasi E. 2000. Explaining trends in inequities: evidence from Brazilian child health studies. *Lancet*, **356**: 1093-8.

Vyas S, Kumaranayake L. 2006. Constructing socio-economic status indices: how to use principal components analysis. *Health Policy Plan*, **21**: 459-68.

WHO. 2004. Equitable access to essential medicines: a framework for collective action. Geneva: World Health Organization.

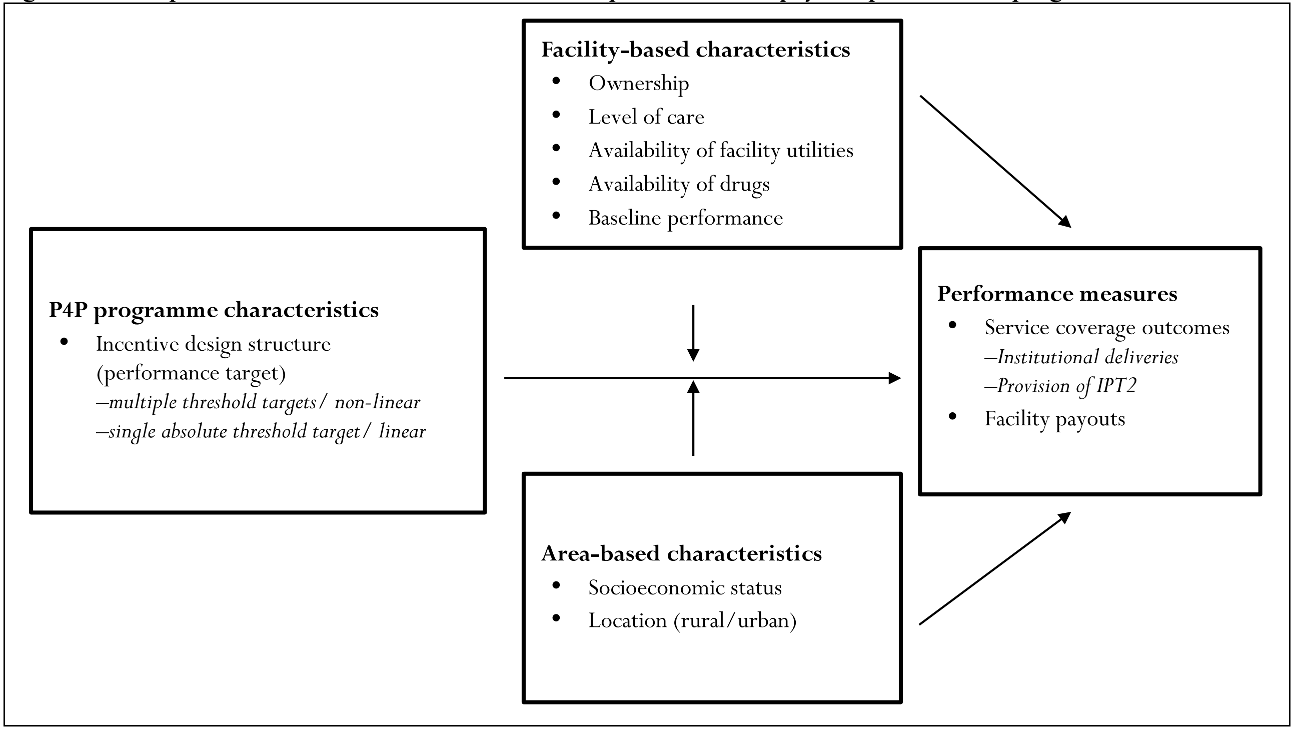
WHO. 2013. Handbook on health inequality monitoring: with a special focus on low- and middle-income countries. World Health Organization, Geneva, Switzerland.

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 Syst Rev*: Cd007899.

Witter S, Toonen J, Meessen B, et al. 2013. Performance-based financing as a health system reform: mapping the key dimensions for monitoring and evaluation. *BMC Health Serv Res*, **13**: 367.

Figure and Tables

Figure 1: Conceptual framework for the determinants of performance in pay-for-performance programmes.[1]


P4P programme characteristics
Incentive design structure (performance target)
–multiple threshold targets/ non-linear
–single absolute threshold target/ linear
Facility-based characteristics
Ownership
Level of care
Availability of facility utilities
Availability of drugs
Baseline performance
Area-based characteristics
Socioeconomic status
Location (rural/urban)
Performance measures
Service coverage outcomes
–Institutional deliveries
–Provision of IPT2
Facility payouts

[1] We modified a conceptual framework which was initially developed by (Rittenhouse et al. 2010) and (Markovitz & Ryan 2016).

Table 1: Service indicators and performance targets for facilities implementing P4P in Tanzania

P4P service indicators	Method	Baseline coverage (previous cycle)				
		0–20%	21–40%	41–70%	71–85%	85%+
Coverage indicators						
% of institutional deliveries	Percentage point increase	15%	10%	5%	5%	Maintain
% of mothers attending a facility within 7 days of delivery.	Percentage point increase	15%	10%	5%	5%	Maintain
% of women using long term contraceptives	Percentage point increase	20%	15%	10%	Maintain above 71%	Maintain
% children under 1 year received measles vaccine	Overall result	50%	65%	75%	80%+	Maintain
% children under 1 year received Penta 3	Overall result	50%	65%	75%	80%+	Maintain
% of complete partographs	Overall result	80%	80%	80%	80%+	Maintain above 80%
HMIS reports submitted to district managers on time and complete	Overall result	100%	100%	100%	100%	100%
Content of care indicators						
% ANC clients receiving two doses of IPT	Overall result	80%	80%	80%	80%+	Maintain above 80%
% HIV+ ANC clients on ART	Overall result	40%	60%	75%	75%+	Maintain
% of children receiving polio vaccine (OPV0) at birth	Overall result	60%	75%	80%	80%+	Maintain

Notes: 85%+ = 85% or more; 80%+ = 80% or more; HMIS=Health Management Information System; ANC=Antenatal care.

Source: The United Republic of Tanzania, Ministry of Health and Social Welfare. 2011. The Coast Region Pay for Performance (P4P) Pilot: Design Document. Health managers were rewarded based on the overall performance of facilities in their district/ region. Managers also had their own indicators which includes, maternal and newborn deaths audited properly and timely; reducing stock-out rates of essential drugs; timely reporting the facility data from district to regional level, and from regional to national level.

Table 2: Baseline facility and area-based characteristics by study arms

Characteristics	Description	Intervention (n=75)	Comparison (n=75)	Difference (p-value)
Panel A: Facility-based characteristics				
Facility ownership	=1 for public owned (%)	84.0	82.7	1.3 (0.828)
Facility level of care	=1 for dispensary (%)	70.7	70.7	0.0 (1.000)
Availability of facility utilities	=1 for electricity & water supply (%)	54.7	52.0	2.7 (0.745)
Availability of drugs –index	Mean index (0-1) of 37 drugs [SD]	0.61 [0.16]	0.66 [0.12]	**–0.05 (0.031)**
Availability of drugs –subgroup	=1 for availability below the median (%)	57.3	42.7	**14.6 (0.073)**
Baseline coverage level (deliveries)	=1 for facility below the median (%)	53.3	46.7	6.6 (0.418)
Baseline coverage level (IPT2)	=1 for facility below the median (%)	54.6	45.3	9.3 (0.256)
Panel B: Area-based characteristics				
Wealth status index	Mean wealth index [SD]	–0.43 [1.8]	0.32 [2.4]	**–0.75 (0.028)**
Wealth status –tercile 1	=1 for poorest population (%)	40.0	26.7	**13.3 (0.084)**
Wealth status –tercile 2	=1 for middle wealth population (%)	34.7	32.0	2.7 (0.731)
Wealth status –tercile 3	=1 for least poor population (%)	25.3	41.3	**–16.0 (0.038)**
Facility location	=1 for facility in rural district (%)	78.7	84.0	–5.3 (0.405)

Notes: Three quantiles (terciles) were used for wealth status of the facility's catchment population; Availability of drugs include 37 drugs and analysis used a dummy variable classified based on baseline availability distribution (=1 for availability below the median/bottom half and 0, otherwise), SD=Standard Deviation; Reference category in brackets: public (vs. non-public), dispensary (vs. health centre & hospital), with electricity and water supply at baseline (vs. none), baseline availability of drugs below the median/in bottom half (vs. top half), baseline lower performer/below the median (vs. higher performer), rural (vs. urban district); For distributional analyses, wealth index and drugs availability index were re-classified on each arm separately and equally to avoid the imbalance across arms at baseline.

Table 3: Distribution of facility payout scores by wealth status of the catchment populations

Payment Cycle	All	Area-based wealth status (Terciles)			Equity		Concentration Index (CI) (p-value)
	Mean [SD]	Least poor	Middle	Poorest	Gap (p-value)	Ratio	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CYCLE 1 (%)	50.1 [19.4]	54.7	52.3	43.1	**11.6 (0.027)**	1.27	**0.042 (0.099)**
CYCLE 2 (%)	50.3 [19.1]	58.4	49.7	42.4	**16.0 (0.002)**	1.38	**0.088 (0.000)**
CYCLE 3 (%)	64.6 [18.8]	69.2	65.1	59.6	**9.6 (0.062)**	1.16	**0.036 (0.054)**
CYCLE 4 (%)	67.5 [19.5]	67.8	69.6	65.1	2.7 (0.623)	1.04	0.007 (0.699)
CYCLE 5 (%)	74.5 [18.5]	75.3	74.9	73.4	1.9 (0.707)	1.03	0.007 (0.669)
CYCLE 6 (%)	69.6 [20.1]	72.0	75.3	61.3	**10.7 (0.046)**	1.17	**0.035 (0.058)**
CYCLE 7 (%)	77.7 [16.3]	79.2	76.9	76.9	2.3 (0.619)	1.03	0.006 (0.672)
Pooled–all cycles (1–7) (%)	64.7 [11.7]	68.1	66.3	60.5	**7.6 (0.015)**	1.13	**0.027 (0.022)**

Notes: Analysis restricted to intervention facilities only (n=75); P-values in column (5) are from t-test of the null hypothesis that the gap [column (2) – (4)] is equal to zero; P-values in column (7) are for testing the null hypothesis of zero concentration index; SD=Standard Deviation; Terciles for wealth status were generated with equal-size from intervention arm separately; Gap=Least poor–Poorest; Ratio=Least poor/Poorest; The results are generally similar in column (5) when non-parametric test (Wilcoxon rank-sum) is used (Table 6A).

Table 4: Distribution of facility payout scores by other subgroups of facilities

Facility subgroups	By payment cycle							Pooled average cycles
	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6	Cycle 7	Cycle 1-7
Facility location								
Rural (%)	52.2	48.5	66.3	69.5	76.4	71.3	80.0	66.4
Urban (%)	42.3	56.7	58.3	60.1	68.1	63.2	68.9	59.7
Gap (%)	9.9**	–8.2*	8.0	9.4*	8.3	8.1	11.1**	6.7*
Ratio	1.2	0.9	1.1	1.2	1.1	1.1	1.2	1.1
Ownership status								
Public owned (%)	49.9	49.5	66.0	68.8	75.8	70.0	78.4	65.6
Non-public (%)	50.9	54.4	56.9	60.6	66.7	67.1	73.6	61.5
Gap (%)	–1.0	–4.9	9.1	8.2	9.1	2.9	4.8	4.1
Ratio	1.0	0.9	1.2	1.1	1.1	1.0	1.1	1.1
Level of care								
Dispensary (%)	47.7	46.9	60.2	63.5	71.5	66.9	75.4	61.9
HC & hospital (%)	55.8	58.3	75.3	77.0	81.7	75.8	82.9	72.4
Gap (%)	–8.1*	–11.4**	–15.1***	–13.5***	–10.2***	–8.9**	–7.5**	–10.5***
Ratio	0.9	0.8	0.8	0.8	0.9	0.9	0.9	0.9
Electricity & water supply								
Available (%)	53.6	51.9	66.7	69.1	76.8	71.3	81.1	67.2
None (%)	45.9	48.3	62.1	65.5	71.7	67.5	73.5	62.2
Gap (%)	7.7*	3.6	4.6	3.6	5.1	3.8	7.6**	5.0*
Ratio	1.2	1.1	1.1	1.1	1.1	1.1	1.1	1.1
Availability of drugs								
Above the median (%)	50.6	58.6	68.3	72.2	76.0	74.6	79.3	68.5
Below the median (%)	49.7	41.8	61.0	62.9	73.2	64.6	76.0	61.5
Gap (%)	0.9	16.8***	7.3*	9.3**	2.8	10.0**	3.3	7.0***
Ratio	1.0	1.4	1.1	1.1	1.0	1.2	1.0	1.1

Notes: Analysis restricted to intervention facilities only (n=75); Gap is the difference in payout score between two subgroups of facilities; Ratio is the ratio of payout scores for two subgroups; The significance test was by t-test for the null hypothesis of gap equals zero; The results are generally similar when non-parametric test (Wilcoxon rank-sum) was used to test the significant of the gap (results not shown); *** denotes significance at 1%, ** at 5%, and * at 10% level.

Table 5: Baseline coverage levels by facility subgroups across study arms

Outcome variable/ subgrouping variable	Intervention arm (n=75)			Comparison arm (n=75)		
	Yes	No	Gap	Yes	No	Gap
	(1)	(2)	(3)	(4)	(5)	(6)
OUTCOME 1: Institutional deliveries						
Public facility (%)	84.6	84.7	–0.1	86.4	89.0	–2.6
Dispensary facility (%)	82.5	89.5	–7.0	85.3	90.7	–5.4**
Facility with utilities (electricity & water supply) (%)	86.9	81.7	5.2	88.3	85.4	2.9
Facility with drugs availability below the median (%)	83.9	85.2	–1.3	88.6	85.1	3.5
Facility with poorest catchment population (%)	84.6	89.7	–5.1*	81.5	92.7	–11.2***
Facility with middle wealth catchment population (%)	79.5	89.7	–10.2**	86.3	92.7	–6.4**
Facility in rural district (%)	83.9	87.1	–3.2	85.9	92.0	–6.1*
Lower performer (below the median) (%)	73.9	95.6	–21.7***	80.4	95.9	–15.5***
OUTCOME 2: Provision of IPT2						
Public facility (%)	50.2	50.6	–0.4	57.0	51.3	5.7
Dispensary facility (%)	53.8	41.7	12.1***	54.1	60.5	–6.4*
Facility with utilities (electricity & water supply) (%)	47.7	53.2	–5.5	57.8	54.1	3.7
Facility with drugs availability below the median (%)	53.6	46.7	6.9*	57.2	54.7	2.5
Facility with poorest catchment population (%)	49.5	45.7	3.8	61.6	52.5	9.1**
Facility with middle wealth catchment population (%)	55.5	45.7	9.8**	53.8	52.5	1.3
Facility in rural district (%)	50.8	47.9	2.9	56.1	55.3	0.8
Lower performer (below the median) (%)	37.3	63.5	–26.2***	44.0	68.9	–24.9***

Notes: We used a t-test to test the null hypothesis of a gap (column 3 and 6) equals to zero; Terciles classified in each arm separately were used for wealth status of the facility's catchment population; Availability of drugs include 37 drugs and analysis used a dummy variable classified in each arm separately based on baseline availability distribution (=1 for availability below the median/bottom half and 0, otherwise); Reference category for "NO" column in brackets: public (vs. non-public), dispensary (vs. health centre & hospital), with electricity and water supply at baseline (vs. none), baseline availability of drugs below the median/in bottom half (vs. top half), baseline lower performer/below the median (vs. higher performer); Similar pattern of results when hospitals are excluded for institutional delivery outcome; Overall baseline coverage in institutional deliveries was (84.7%, 86.8%) and IPT2 was (49.5%, 56.7%) for intervention and control arm respectively (Binyaruka et al. 2015); *** denotes significance at 1%, ** at 5%, and * at 10% level.

Table 6: Differential effects of P4P on service coverage outcomes

Characteristics	Description	Institutional deliveries		Provision of IPT2	
		Beta† **[95% CI]** (1)	**Beta†** **[95% CI]** (2)	**Beta†** **[95% CI]** (3)	**Beta†** **[95% CI]** (4)
Panel A: Facility-based characteristics					
Facility ownership	=1 for public owned (%)	2.8 [–6.8 to 12.3]	4.4 [–5.5 to 14.3]	4.7 [–13.3 to 22.7]	4.5 [–13.3 to 22.4]
Facility level of care	=1 for dispensary (%)	3.8 [–6.3 to 13.8]	2.7 [–6.7 to 12.0]	–7.9 [–20.3 to 4.6]	–9.6 [–22.5 to 3.4]
Availability of facility utilities	=1 for electricity & water supply (%)	–3.8 [–13.2 to 5.5]	–2.9 [–12.6 to 6.8]	0.5 [–11.6 to 12.6]	–0.2 [–12.9 to 12.6]
Availability of drugs	=1 for availability below the median (%)	6.5 [–2.7 to 15.8]	6.3 [–2.7 to 15.2]	–2.1 [–13.9 to 9.8]	–1.8 [–13.6 to 10.1]
Baseline coverage level	=1 for lower performer at baseline (%)	11.3*** [2.9 to 19.7]	13.0*** [4.1 to 21.9]	7.3 [–2.5 to 17.0]	7.5 [–2.4 to 17.3]
Panel B: Area-based characteristics					
Wealth status –tercile 1	=1 for poorest population (%)	2.5 [–7.9 to 12.9]	4.0 [–6.6 to 14.6]	6.3 [–7.1 to 19.7]	6.4 [–7.4 to 20.1]
Wealth status –tercile 2	=1 for middle wealth population (%)	12.6** [2.4 to 22.8]	14.3*** [4.1 to 24.4]	–6.9 [–21.7 to 7.9]	–6.4 [–21.2 to 8.4]
Facility location	=1 for facility in rural district (%)	6.6 [–2.8 to 15.9]	10.0** [1.0 to 19.1]	4.9 [–10.9 to 20.6]	5.2 [–11.1 to 21.5]
Adjusted for facility-level covariates		–	YES	–	YES
Number of observations (N)		300	300	300	300

Notes: †The Beta is the estimated differential effect between subgroups in percentage point after controlling for a year dummy, facility-fixed effects, and facility-level covariates (availability of utilities and wealth status of the catchment population) whenever specified (the adjusted model is the most robust and preferred); Each cell reports the "Beta [95% CI]" from a separate regression; Terciles classified in each arm separately were used for wealth status of the facility's catchment population; Availability of drugs include 37 drugs and analysis used a dummy variable classified in each arm separately based on baseline availability distribution (=1 for availability below the median/bottom half and 0, otherwise); Reference category in brackets: public (vs. non-public), dispensary (vs. health centre &

hospital), with electricity and water supply at baseline (vs. none), baseline availability of drugs below the median/in bottom half (vs. top half), baseline lower performer/below the median (vs. higher performer), rural (vs. urban district), poorest/ middle wealth (vs. least poor); No differential effects if oxytocics and antimalarials are respectively used for deliveries and IPT2 outcomes (Appendix A5); *** denotes significance at 1%, ** at 5%, and * at 10% level.

APPENDIX

Appendix A1: List of 37 essential drugs

Category/ type of classification	Number of items	List of medical commodities considered
Medicines combined	37	Antimalarials, antibiotics, antihypertensives, antidiarrheals, oxytocics, ARVs, vaccines, vitamin A, and family planning medicines
a) Antimalarials	3	Artemether-Lumefantrine (ALU), quinine and Sulfadoxine Pyrimethamine [SP (IPTp)]
b) Antibiotics	6	Cotrimoxazole, ampicillin, X-Pen injection, gentamycin, flagyl, and chloramphenicol
c) Antihypertensives	5	Magnesium sulfate, diazepam, aldomet, nifedipine, and hydralazine
d) Antidiarrheals	2	Oral rehydration salts (ORS) and zinc
e) Oxytocics	3	Oxytocin, misoprostol, and ergometrine
f) Antiretroviral therapy (ARTs)	7	Zidovudine, stavudine, lamivudine, lenofavir, nevirapine, efavirenz, and emtricitabine
g) Vaccines	5	BCG, OPV, DPT, measles and tetanus
h) Vitamin A	1	Vitamin A
i) Family planning medicines	5	Contraceptive pills, depo-provera, injectable, IUCD, and implants

Notes: For measurement, if a commodity was available on the day of the survey, the outcome was coded 1 and 0 otherwise.

Appendix A2: List of household ownership assets and characteristics for assessing wealth status

No.	Variable description
1.	Asset: electricity
2.	Asset: working radio
3.	Asset: working television (TV)
4.	Asset: working DVD

5.	Asset: working mobile phone
6.	Asset: working landline phone
7.	Asset: working iron
8.	Asset: working refrigerator
9.	Asset: working wall watch
10.	Asset: sewing machine
11.	Asset: table
12.	Asset: sofa coach
13.	Asset: cupboard
14.	Asset: motorcycle
15.	Asset: car
16.	Household member with a bank account
17.	Number of sleeping rooms
18.	Source of drinking water: piped water
19.	Source of drinking water: borehole/ covered well
20.	Source of drinking water: open well
21.	Source of drinking water: spring water
22.	Source of drinking water: river/ dam/pond/lake
23.	Toilet type: flush toilet
24.	Toilet type: pit latrine
25.	Toilet type: no/ other toilet
26.	Source of cooking energy: electricity
27.	Source of cooking energy: kerosene/paraffin
28.	Source of cooking energy: charcoal
29.	Source of cooking energy: firewood
30.	Source of light: electricity
31.	Source of light: solar
32.	Source of light: kerosene/ paraffin
33.	Source of light: candle/ firewood
34.	Source of light: torch or other source
35.	Floor material: sand/earth/dung
36.	Floor material: cement

37.	Floor material: other
38.	Wall material: grass/poles/mud wall
39.	Wall material: bamboo with mud wall
40.	Wall material: sundried/ burnt bricks
41.	Wall material: cement blocks
42.	Wall material: stones with mud

Table A3: Differential effects of P4P on institutional deliveries – (robustness check by excluding hospitals)

	All facilities (N=300)			**Excluding all hospitals (N=276)**	
	No covariates	Facility covariates		No covariates	Facility covariates
Characteristics	Beta†	Beta†		Beta†	Beta†
	(1)	(2)		(3)	(4)
Panel A: Facility-based characteristics					
Ownership (=1 for public)	2.8	4.4		0.5	3.3
level of care (=1 for dispensary)	3.8	2.7		2.1	1.8
Availability of utilities (=1 available electricity & water)	–3.8	–2.9		–1.6	–1.3
Availability of drugs (=1 below the median)	6.5	6.3		4.5	4.8
Drugs specific (Oxytocics availability)	1.9	1.6		0.8	1.1
Baseline coverage level (deliveries)	11.3***	13.0***		12.7***	14.3***
Panel B: Area-based characteristics					
Wealth status (=1 for lower wealth status; 0 for higher)	1.2	2.6		1.0	2.0
Wealth status (=1 for poorest population) –Tercile 1	2.5	3.9		0.7	1.9
Wealth status (=1 for middle population) –Tercile 2	12.6**	14.3***		11.9**	13.2**
Facility location (=1 for rural district)	6.6	10.0**		3.3	7.1
Adjusted for facility-level covariates (wealth & utilities)	–	YES		–	YES

Notes: 2-quantiles of wealth status (lower vs. higher) were used; Drugs specific (oxytocics) for deliveries were used; The differential effect by population wealth was insignificant when excluding the by-passers in the model which excludes all hospitals; *** denotes significance at 1%, ** at 5%, and * at 10% level.

Table A4: Differential effects of P4P on coverage outcomes – (robustness check by clustering at the district level)

	Institutional deliveries			Provision of IPT2	
	No covariates	Facility covariates		No covariates	Facility covariates
Characteristics	Beta†	Beta†		Beta†	Beta†
	(1)	(2)		(3)	(4)
Panel A: Facility-based characteristics					
Ownership (=1 for public)	2.8	4.4		4.7	4.5
level of care (=1 for dispensary)	3.8	2.7		–7.9	–9.6
Availability of utilities (=1 available electricity & water)	–3.8	–2.9		0.5	–0.1
Availability of drugs (=1 below the median)	6.5	6.3		–2.1	–1.8
Drugs specific availability (Oxytocics/Antimalarials)	1.9	1.6		4.2	3.4
Baseline coverage level (deliveries/IPT2)	11.3*	13.0*		7.3	7.5
Panel B: Area-based characteristics					
Wealth status (=1 for lower wealth status; 0 for higher)	1.2	2.6		6.1	6.5
Wealth status (=1 for poorest population) –Tercile 1	2.5	3.9		6.3	6.4
Wealth status (=1 for middle population) –Tercile 2	12.6	14.3		–6.9	–6.4
Facility location (=1 for rural district)	6.6	10.0*		4.9	5.2
Adjusted for facility-level covariates (wealth & utilities)	–	YES		–	YES
Number of observations (N)	300	300		300	300

Notes: Clustering at the district level with BOOTSTRAPPING method and used 400 reps.; 2-quantiles of wealth status (lower vs. higher) were used; Drugs specific (oxytocics/antimalarials) for deliveries and IPT2 respectively were used; *** denotes significance at 1%, ** at 5%, and * at 10% level.

Table A5: Differential effects of P4P on coverage outcomes – (robustness check with two quantiles of wealth status instead of terciles)

	Institutional deliveries			Provision of IPT2	
	No covariates	Facility covariates		No covariates	Facility covariates
Characteristics	Beta†	Beta†		Beta†	Beta†
Area-based characteristics					
Wealth status (=1 for lower wealth status; 0 for higher)	1.2	2.6		6.1	6.5
Wealth status (=1 for poorest population) –Tercile 1	2.5	4.0		6.3	6.4
Wealth status (=1 for middle population) –Tercile 2	12.6**	14.3***		–6.9	–6.4
Adjusted for facility-level covariates (wealth & utilities)	–	YES		–	YES
Number of observations (N)	300	300		300	300

Notes: 2-quantiles of wealth status (lower vs. higher) for below and above median were used; *** denotes significance at 1%, ** at 5%, and * at 10% level.

Table A6: Distribution of facility payout scores by wealth status of the catchment populations (n=75)

– Non-parametric test (Wilcoxon rank sum test) in column 4.

Payment Cycle	All	Area-based wealth status		Parametric	Non-parametric
	Mean [SD]	Higher status	Lower status	Gap (p-value)	Gap (p-value)
	(1)	(2)	(3)	(4)	(5)
CYCLE 1 (%)	50.1 [19.4]	53.9	46.3	**7.6 (0.089)**	7.6 (0.127)
CYCLE 2 (%)	50.3 [19.1]	56.6	43.9	**12.7 (0.003)**	**12.7 (0.003)**
CYCLE 3 (%)	64.6 [18.8]	69.7	59.6	**10.1 (0.019)**	**10.1 (0.005)**
CYCLE 4 (%)	67.5 [19.5]	68.5	66.5	2.0 (0.664)	2.0 (0.500)
CYCLE 5 (%)	74.5 [18.5]	75.8	73.3	2.5 (0.554)	2.5 (0.728)
CYCLE 6 (%)	69.6 [20.1]	73.9	65.4	**8.5 (0.063)**	8.5 (0.103)
CYCLE 7 (%)	77.7 [16.3]	79.1	76.3	2.8 (0.468)	2.8 (0.410)
Pooled–all cycles (1–7) (%)	64.7 [11.7]	68.2	61.8	**6.4 (0.015)**	**6.4 (0.028)**

Notes: P-values in column (5) are for testing the null hypothesis of zero gap [column (2) – (3)] using Wilcoxon rank sum (Wilcoxon rank sum test) test between two subgroups of wealth status; P-values in column (4) are from t-test; SD=Standard Deviation; Two subgroups of wealth status were generated with equal-size from intervention arm separately; Gap=Higher status–Lower status; Ratio=Higher status /Lower status.

III

Binyaruka *et al. International Journal for Equity in Health* (2018) 17:14
DOI 10.1186/s12939-018-0728-x

International Journal for Equity in Health

RESEARCH Open Access

Who benefits from increased service utilisation? Examining the distributional effects of payment for performance in Tanzania

Peter Binyaruka[1,2,3*], Bjarne Robberstad[1], Gaute Torsvik[3,4] and Josephine Borghi[5]

Abstract

Background: Payment for performance (P4P) strategies, which provide financial incentives to health workers and/or facilities for reaching pre-defined performance targets, can improve healthcare utilisation and quality. P4P may also reduce inequalities in healthcare use and access by enhancing universal access to care, for example, through reducing the financial barriers to accessing care. However, P4P may also enhance inequalities in healthcare if providers cherry-pick the easier-to-reach patients to meet their performance targets. In this study, we examine the heterogeneity of P4P effects on service utilisation across population subgroups and its implications for inequalities in Tanzania.

Methods: We used household data from an evaluation of a P4P programme in Tanzania. We surveyed about 3000 households with women who delivered in the last 12 months prior to the interview from seven intervention and four comparison districts in January 2012 and a similar number of households in 13 months later. The household data were used to generate the population subgroups and to measure the incentivised service utilisation outcomes. We focused on two outcomes that improved significantly under the P4P, i.e. institutional delivery rate and the uptake of antimalarials for pregnant women. We used a difference-in-differences linear regression model to estimate the effect of P4P on utilisation outcomes across the different population subgroups.

Results: P4P led to a significant increase in the rate of institutional deliveries among women in poorest and in middle wealth status households, but not among women in least poor households. However, the differential effect was marginally greater among women in the middle wealth households compared to women in the least poor households ($p = 0.094$). The effect of P4P on institutional deliveries was also significantly higher among women in rural districts compared to women in urban districts ($p = 0.028$ for differential effect), and among uninsured women than insured women ($p = 0.001$ for differential effect). The effect of P4P on the uptake of antimalarials was equally distributed across population subgroups.

Conclusion: P4P can enhance equitable healthcare access and use especially when the demand-side barriers to access care such as user fees associated with drug purchase due to stock-outs have been reduced.

Keywords: Inequality, Equity, Social determinants of health, Universal coverage, Distributional effects, Healthcare financing, Pay for performance, Tanzania

* Correspondence: pbinyaruka@ihi.or.tz
[1]Centre for International Health, University of Bergen, PO Box 7804, N-5020 Bergen, Norway
[2]Ifakara Health Institute, PO Box 78373, Dar es Salaam, Tanzania
Full list of author information is available at the end of the article

Introduction

Payment for performance (P4P) is a supply-side financing strategy which involves financial incentives being paid to health workers and/or facilities for reaching predefined performance targets. This approach started in high-income countries (HICs) with the aim of improving health care quality [24, 64, 65]. P4P is also increasingly being used in low- and middle-income countries (LMICs) to improve quality and use of health services, as well as to strengthen health systems [31, 57, 89]. The evidence base on the effectiveness of P4P is growing and suggests mixed effects with notable improvements for some incentivised indicators [9, 11, 17, 24, 26, 35, 61, 69, 73, 77].

However, most evaluations focus on average effects and pay little attention to distributional effects across provider or population subgroups [51]. There is, however, a growing awareness that average effects may mask important heterogeneous programme effects [12, 13, 19, 22, 38, 41, 51]. This study examines the heterogeneity of P4P effects on service utilisation across population subgroups. The overall goal is to display heterogeneous treatment effects, and specifically to check if the effects on population subgroups will reduce or enhance exiting inequalities in access to and utilisation of health care services.

Inequalities in access to and use of health services in favour of wealthier populations are still prevalent in many settings, with the greatest inequalities in the poorest settings [8, 15, 52, 56, 60, 68, 78, 79, 82, 84]. Factors referred to as "social determinants of health" such as economic status, education, location and age [21, 54, 60, 87], mostly drive these inequalities. From a theoretical point of view, it is hard to know how P4P will affect pre-existing inequalities. However, P4P can reduce inequalities in access to healthcare, for example, by encouraging providers to extend services to underserved groups (e.g. by reducing financial barriers to access care) in a bid to meet performance targets [31, 57]. On the other hand, P4P could also enhance inequalities in access to healthcare if providers cherry-pick the easier-to-reach patients in order to meet their performance targets [40].

Studies in HICs have found differential effects of P4P on healthcare quality between socioeconomic groups in favour of wealthier populations (pro-rich) but this effect declined over time. These studies have not found any differential effect with respect to age, sex and ethnicity [2, 14, 24, 80]. Evidence from LMICs is more limited and varied across service types [63]. For example, the effect of P4P on institutional delivery rates was greater among wealthier groups (pro-rich) in most settings [17, 46, 77] but there was an indication that it was greater among poorer groups (pro-poor) in Tanzania[11]. The effect of P4P on institutional deliveries was greater among women with health insurance in Rwanda [46] or a maternity care voucher in Cambodia [77] than their counterparts. The effect of P4P on family planning coverage was greater among wealthier groups (pro-rich), in Rwanda [46], and the effect on immunisation coverage was greater among poorer groups (pro-poor), in Burundi [17]. However, studies based on Rwanda Demographic Health Survey (DHS) data reported no differential effect by socioeconomic groups on the use of maternal care [62] and on child curative care seeking [72].

To date, most studies on differential effects of P4P have disaggregated the effect of P4P across population economic status particularly in LMICs, with little attention to other social determinants (e.g. education, occupation, and age), which are also known to affect the use of health services [4, 60], including maternal health services [30, 32, 71]. The assessment of programme differential effects across various social determinants in a broad perspective is crucial to inform universal access policies [28, 53, 60], and may help to understand how different service users are affected by a programme such as P4P [63]. In this paper, we examine the differential effect of P4P on service utilisation in Tanzania across a variety of population subgroups by stratified analyses according to various social determinants.

This paper proceeds as follows. The next section presents the conceptual framework, followed by the description of the P4P programme in Tanzania. The other sections include the methods and analysis, followed by the results, discussion and conclusion.

Conceptual framework

P4P programmes give providers incentives to change their behaviour to improve the quality of care in order to enhance utilisation and obtain financial rewards [66]. Based on this logic P4P can improve average service utilisation and the distribution of improved utilisation across population subgroups through the *supply-side response* (how providers respond to incentives) and the resulting *demand-side response* that triggers (how patients respond to supply side changes).

Supply-side response

To meet performance targets aimed at increasing the quantity of services provided, providers are likely to adopt strategies to attract more patients to facilities [31, 57]. One such strategy could be to make services more affordable [57], for example by reducing user fees, or by reducing drug stock-outs, avoiding patients having to procure drugs privately [10, 11]. Another strategy could be to improve responsiveness to service users, for example, by being kinder during service delivery [11]. However, providers might also attempt to cherry-pick patients or focus on easy-to-reach populations (i.e. underserved but easily reached) in order to meet the performance targets [25, 40], leaving the hard-to-reach (i.e. poorest with greatest need) underserved. In fact, providers may need to exert greater effort and time to serve the hard-to-reach [37]. The efficiency gains in that case can be reached but at the expenses of inequity [47].

Demand-side responses

According to Andersen's behavioural model of healthcare utilisation [3, 4], the use of health services is a function of patient's propensity to use services (predisposing factors), factors that facilitate or impede access and use (enabling factors), as well as perceived need for healthcare (need factors). These factors among others are also social determinants of health [21, 54, 74]. The interactions between a P4P programme (supply-side response) and social determinants (demand-side factors) may affect the use and distribution of health services. For example, reduced financial barriers to access care, resulting from provider response to incentives, may stimulate demand especially for poor and/or uninsured individuals, since they are more responsive to a change in healthcare costs consistent with demand theory [33, 49]. Demand for health services may also increase if the quality of care supplied is improved [1]; for example, through increased drug availability and better interpersonal care [10, 11]. Better-off populations (e.g. wealthier, educated, and urban residents) may also benefit more from quality improvements simply because they use services more than their counterpart populations [8, 15, 21, 32, 54, 68, 81].

Despite the potential interactions between the demand and supply-side response to P4P, the health care sector does not operate like a classic free market [6, 61]. For example, the demand-side response may be weak when some demand-side barriers to access care (e.g. cultural and information barriers) are unaffected by the supply-side response to incentives [27, 48, 61, 88].

P4P in Tanzania

In 2011, the Ministry of Health and Social Welfare (MoHSW) in Tanzania with support from the Government of Norway introduced a P4P scheme as a pilot in Pwani region. The scheme aimed to improve maternal and child health (MCH) and inform the national P4P roll out. Pwani is one of 30 regions in the country and has seven districts with more than 209 health facilities. It has a population of just over a million [59]. All health facilities providing MCH services in the region were eligible to implement the P4P scheme. The P4P scheme involved a series of performance targets for facilities that were set in relation to the coverage of specific services (e.g. institutional delivery) or for care provided during a service (e.g. uptake of antimalarials during antenatal care) (Table 1), as described in more detail elsewhere [11, 18]. Performance was rewarded based on two methods of target setting: single and multiple thresholds targets. The strategies to reach performance targets were left to the discretion of the health workers at the individual facilities. District and regional managers were also eligible to receive performance payouts based on the performance of the facilities in their district or region.

The extent to which facilities were successful in achieving performance targets determined the level of bonus payout they would receive as part of the programme. Full payment was made if 100% of a given target was achieved, and 50% of payment was made for 75% < 100% achievement, while no payment was made for lower levels of performance. The maximum payout if all targets were fully attained was USD 820 per cycle for dispensaries; USD 3220 for health centres and USD 6790 for hospitals. The payouts were additional to the funding facilities receive to cover operational costs and salaries of health workers. Incentive payouts at the facility-level included bonuses to staff (equivalent to 10% of their monthly salary if all targets were fully attained) and funds that could be used for facility improvement or demand creation initiatives (10% of the total in hospitals and 25% in lower level facilities). District and regional managers received bonus payments of up to USD 3000 per cycle.

To determine whether performance targets were met, performance data were compiled by facilities and verified by the P4P implementing agency every six months (one cycle) before distributing payouts.

The P4P programme was the subject of a process and impact evaluation. The impact evaluation showed a significant positive effect on two out of eight incentivised service indicators: institutional delivery rate and provision of antimalarial during antenatal care [11]. P4P was also associated with a number of process changes such as increased availability of drugs and supplies, increased supportive supervision, a reduced chance of paying user fees, and greater provider kindness during delivery care [5, 10, 11, 55].

Methods

Study design

Our study used data from a controlled before and after evaluation study of the P4P scheme in Pwani region, Tanzania, described elsewhere [11, 18]. All seven districts in Pwani region (intervention arm), and four districts from Morogoro and Lindi regions (comparison arm) were sampled. The comparison districts were selected to be comparable to intervention districts in terms of poverty and literacy rates, the rate of institutional deliveries, infant mortality, population per health facility, and the number of children under one year of

Table 1 Service indicators and performance targets for facilities implementing P4P in Tanzania

P4P service indicators	Method	Baseline coverage (previous cycle)				
		0–20%	21–40%	41–70%	71 – 85%	85%+
Coverage indicators						
% of institutional deliveries	Percentage point increase	15%	10%	5%	5%	Maintain 85%+
% of mothers attending a facility within 7 days of delivery.	Percentage point increase	15%	10%	5%	5%	Maintain 85%+
% of women using long term contraceptives	Percentage point increase	20%	15%	10%	Maintain above 71%	Maintain 85%+
% children under 1 year received measles vaccine	Overall result	50%	65%	75%	80%+	Maintain 85%+
% children under 1 year received Penta 3	Overall result	50%	65%	75%	80%+	Maintain 85%+
% of complete partographs	Overall result	80%	80%	80%	80%+	Maintain above 80%
HMIS reports submitted to district managers on time and complete	Overall result	100%	100%	100%	100%	100%
Content of care indicators						
% ANC clients receiving two doses of IPT	Overall result	80%	80%	80%	80%+	Maintain above 80%
% HIV+ ANC clients on ART	Overall result	40%	60%	75%	75%+	Maintain 85%+
% of children receiving polio vaccine (OPV0) at birth	Overall result	60%	75%	80%	80%+	Maintain 85%+

The United Republic of Tanzania, Ministry of Health and Social Welfare. 2011. The Coast Region Pay for Performance (P4P) Pilot: Design Document
85% + = 85% or more; 80% + = 80% or more; *HMIS* Health Management Information System, *ANC* Antenatal care

age per capita [18]. Baseline data collection was done in January 2012, with a follow-up survey 13 months later.

Sampling and data source

In the intervention arm, we included all 6 hospitals and 16 health centres that were eligible for the P4P scheme, and a random sample of 53 eligible dispensaries. A similar number of facilities were included in the comparison arm. Facilities were randomly sampled amongst those where P4P was implemented and matching comparison facilities were selected based on facility level of care, ownership, staffing levels, and case load [18]. To assess maternal and child health service utilisation in the population, we randomly sampled 20 households of women from the catchment area of each health facility who had delivered in the 12 months prior to the survey. In total, we surveyed 3000 households with eligible women in both arms at baseline, and a similar number in the follow-up survey. The household survey also collected information on maternal background characteristics (e.g. age, marital status, education occupation, religion, and number of births), and household characteristics (e.g. household size, health insurance status, and ownership of assets and housing particulars for assessing the household socioeconomic status).

Outcome variables

Our outcome variables include the two incentivised services which we know from prior analysis improved significantly as a result of P4P: institutional deliveries and uptake of two doses of *intermittent preventive treatment* (IPT2) for malaria during antenatal care [11]. These were measured as binary outcomes for whether a woman gave birth in a health facility and received IPT2 during antenatal care, respectively.

Generation of subgroups for distributional analyses

To examine the distribution of P4P effects on these two outcomes, we generated population subgroups based on individual and household-level characteristics, according to Andersen's behavioural model of healthcare utilisation [3, 4]. In this study we only considered predisposing and enabling factors since data on perceived illness was not available. "Perceived illness" could also be argued to be of less relevance for maternal service utilisation outcomes, since study participants were largely healthy.

Subgroups of predisposing factors include: marital status (married vs. none), maternal age (15–49) years (below vs. above the median age of 25), education (no education vs. primary level/above), occupation (farmer vs. non-farmer), religion (Muslim vs. non-Muslim), number of births/parity (parity 1 vs. parity 2/above), and household size (below vs. above the median size of 5 members). Subgroups of enabling factors include: health insurance status (any insurance vs. none), place of residence (rural vs. urban district), and household wealth status subgroups. The wealth subgroups were generated from wealth scores derived by the principal component analysis based on 42 items of household characteristics and asset ownership (Appendix 1: Table 5) [29, 83]. The household wealth scores were generated separately for baseline and follow-up samples, since participants differed over time. Households were ranked by wealth scores from poorest (low score) to least poor and classified into three-equal sized groups (terciles): poorest, middle and least poor. Subgrouping based on five-equal sized groups (quintiles) were also generated to examine the sensitivity of the findings to different wealth subgroupings.

Statistical analysis

We first compared the sample means of individual and household-level characteristics at baseline between intervention and comparison arms, and assessed whether the differences between arms were statistically significant by using t-tests. We then assessed the distribution of service utilisation outcomes at baseline across population subgroups by estimating the utilisation gap (i.e. a difference in average service use between two subgroups) [87]. We used t-tests to test whether the utilisation gaps were significantly different from zero.

To examine whether the effects of P4P on outcomes differed across population subgroups, we first performed subgroup analyses to identify the P4P effect on each subgroup, and then tested the significance of differential effects between subgroups through analysing the interaction effect. We identified the average effect of P4P on service utilisation by using a linear difference-in-differences regression model. This model compares the changes in outcomes over time between participants in the intervention and comparison arms as specified in Eq. (1):

$$Y_{ijt} = \beta_0 + \beta_1 (P4P_j \times \delta_t) + \beta_2 \delta_t + \beta_3 X_{ijt} + \gamma_i + \varepsilon_{ijt} \tag{1}$$

where Y_{ijt} is the utilisation outcome (institutional deliveries or uptake of IPT2) of individual i in facility *j's* catchment area and at time t. The intervention dummy variable $P4P_j$ takes the value 1 if a facility is in the intervention arm and 0 if it is in the comparison arm. The unobserved time invariant facility characteristics γ_j were controlled for through facility fixed-effects estimation; and included δ_t for year fixed effects. We also controlled for individual and household-level covariates X_{ijt} (age, education, occupation, religion, marital status, parity, insurance status, household size, and household wealth status) as potential confounders. The error term is ε_{ijt}. We clustered the standard errors at the facility level, or facility catchment area, to account for serial correlation of ε_{ijt} at the facility level. The effect of P4P on utilisation for each subgroup is given by β_1.

To test the significance of an eventual differential effect across subgroups, we included a three-way interaction term between the average treatment effect ($P4P_j \times \delta_t$) and a subgrouping variable G_i (based on predisposing and enabling factors). The associated two-order interaction terms were also included in the model. The coefficient of interest is β_4 which indicates the differential effect of P4P across subgroups as shown in Eq. (2):

$$\begin{aligned} Y_{ijt} = {} & \beta_0 + \beta_1 (P4P_j \times \delta_t) + \beta_2 \delta_t + \beta_3 X_{ijt} \\ & + \beta_4 (P4P_j \times \delta_t \times G_{ijt}) \\ & + \beta_5 (P4P_j \times G_{ijt}) + \beta_6 (G_{ijt} \times \delta_t) + \gamma_j \\ & + \varepsilon_{ijt} \end{aligned} \tag{2}$$

The use of the difference-in-difference approach to estimate the effect of P4P on outcomes relies on the key identifying assumption that the trends in outcomes would be parallel across study arms in the absence of the intervention [41]. While this can never be formally tested, we supported the assumption by verifying that the pre-intervention trends in utilisation outcomes at the household level were parallel across study arms as described elsewhere [11]. By surveying women who had delivered in the past 12 months at baseline, four longitudinal outcomes were generated and used to verify the assumption: share of institutional deliveries, caesarean section deliveries, women who breastfeed within one hour of birth, and women who paid for delivery care.

We further performed several robustness checks. First, we re-estimated the P4P differential effect by using wealth quintiles instead of wealth terciles to examine whether the results were sensitivity to wealth group classification. We also generated wealth status subgroups for each study arm and re-estimated the P4P differential effect by arm-based wealth subgroups to avoid the pre-existing baseline imbalance in wealth status between arms. Second, we re-estimated the regression model by including three-way interactions with categorical variable which gives multiple subgroups (e.g. education levels, occupation categories, parity groups and age groups) instead of interactions with binary variables (e.g. married vs. none). Third, we applied a non-linear logit model instead of linear model because of binary outcome variables. Fourth, we clustered the standard errors at the district level instead of facility level and used a bootstrapping method to adjust for the small number of clusters [20]. All the analyses were performed by using STATA version 13.

Results

The majority of individual and household characteristics were similar across intervention and comparison arms at baseline (Table 2). Exceptions were women in the

Table 2 Baseline individual woman and household characteristics by study arms

Characteristics	Description/subgroup	Intervention arm (*n* = 1376)	Comparison arm (*n* = 1468)	Difference
Panel A: Predisposing factors				
Marital status	=1 for married woman (%)	69.9	64.2	5.7[b]
Age	Mean maternal age (15–49) years [SD]	26.5 [6.7]	26.3 [6.5]	0.2
Age	=1 for younger below median age (25 years) (%)	50.9	50.5	0.4
Education	=1 for primary education/above (%)	80.3	80.2	0.1
Occupation	=1 for farming activities (%)	46.0	54.5	−8.5[b]
Religion	=1 for Muslim woman (%)	86.5	66.6	19.9[a]
Parity	Mean number of births [SD]	2.7 [1.8]	2.6 [1.7]	0.1
Parity	=1 for one birth (%)	32.4	31.6	0.8
Household size	Mean number of household members [SD]	4.7 [1.8]	4.8 [1.8]	−0.1
Household size	=1 for small/below the median size of 5 members (%)	51.1	50.5	0.6
Panel B: Enabling factors				
Health insurance status	=1 for insured woman (%)	8.6	8.5	0.1
Household wealth status	Mean household wealth index [SD]	−0.43 [2.7]	0.34 [3.3]	−0.77[b]
Wealth status –tercile 1	=1 for poorest household (%)	38.3	29.4	8.9[b]
Wealth status –tercile 2	=1 for middle wealth household (%)	33.6	33.3	0.3
Wealth status –tercile 3	=1 for least poor household (%)	28.1	37.3	−9.2[b]
Place of residence	=1 for rural district (%)	79.3	84.1	−4.8

SD=Standard Deviation; Subgroups of predisposing factors include: marital status (married vs. none), maternal age (15–49) years (below vs. above the median age of 25), education (no education vs. primary level/above), occupation (farmer vs. non-farmer), religion (Muslim vs. non-Muslim), number of births/parity (parity 1 vs. parity 2/above), and household size (below vs. above the median size of 5 members); Subgroups of enabling factors include: health insurance status (any insurance vs. none), place of residence (rural vs. urban district), and household wealth status subgroups (wealth terciles); [a]denotes significance at 1%, [b]at 5%, and [c]at 10% level

intervention arm, who were more likely to be married, non-farmers, and Muslim; and their households were more likely to be poor than their counterparts in the comparison arm.

The baseline rates of institutional deliveries in both arms were significantly lower for women in the poorest and middle wealth households, and for women who were illiterate, farmers, with parity greater than one than for their counterpart women (Table 3). The rate of institutional deliveries was also higher among intervention women with health insurance and from smaller households, as well as among urban women in the comparison arm than among their counterparts. The baseline uptake of IPT2 was generally similar across arms and population subgroups, except married women in the comparison arm, who were more likely to receive IPT2 than unmarried women (Table 3).

P4P significantly increased the rate of institutional deliveries among women in the poorest and in the middle wealth status households, but not among women in the least poor households (Table 4). However, when compared with the least poor subgroup, the effect of P4P was only marginally greater among women in the middle wealth status households only ($p = 0.094$ for differential effect) (Table 4). The effect of P4P on institutional deliveries was also significantly higher among women in rural districts compared to women in urban districts ($p = 0.028$ for differential effect), and among uninsured than insured women ($p = 0.001$ for differential effect). There were no differential effects of P4P on institutional deliveries among other subgroups, and no differential effects of P4P on the IPT2 outcome across any population subgroups (Table 4).

Our results were generally consistent following robustness checks. When we used wealth quintiles instead of terciles, the effect of P4P on deliveries was significantly higher in lower quintiles (indication of pro-poor) compared to the effect in the top quintile (least poor), but the results on IPT2 remained the same (Appendix 2: Table 6). When we used the arm-based wealth subgroups, the differential effect by quintiles on both outcomes remained broadly unchanged, but the differential effect by terciles on deliveries disappeared and appeared marginally for IPT2 (Appendix 2: Table 6). The effect of P4P on both outcomes remained equally distributed across categorical

Table 3 Baseline levels of service utilisation by subgroups across study arms

Outcome variable/ subgrouping variable	Intervention arm			Comparison arm		
	Yes	No	Gap	Yes	No	Gap
	(1)	(2)	(3)	(4)	(5)	(6)
OUTCOME 1: Institutional deliveries	(n = 1376)			(n = 1468)		
Predisposing factors						
Married woman (%)	84.8	84.7	0.1	86.7	87.0	−0.3
Woman below median age of 25 years/younger (%)	85.4	84.2	1.2	87.3	86.4	0.9
Woman with primary education/above (%)	85.9	80.4	5.5[b]	89.8	74.8	15.0[a]
Woman doing farming for occupation (%)	79.1	89.6	−10.5[a]	82.6	91.9	−9.3[a]
Muslim woman (%)	84.7	85.4	−0.7	87.5	85.5	2.0
Woman with one birth/parity 1 (%)	90.1	82.3	7.8[a]	92.5	84.3	8.2[a]
Household size below the median size of 5 members (%)	87.2	82.3	4.9[b]	87.3	86.4	0.9
Enabling factors						
Woman with any health insurance (%)	89.9	84.3	5.6[c]	83.3	87.1	−3.8
Household with poorest wealth status (Tercile 1) (%)	83.3	91.7	−8.4[a]	80.5	94.2	−13.7[a]
Household with middle wealth status (Tercile 2) (%)	80.8	91.7	−10.9[a]	84.2	94.2	−10.0[a]
Household in rural district (%)	83.9	88.0	−4.1	85.8	92.3	−6.5[c]
OUTCOME 2: Uptake of IPT2	(*n* = 1029)			(*n* = 1.199)		
Predisposing factors						
Married woman (%)	51.0	47.0	4.0	59.3	51.7	7.6[b]
Woman below median age of 25 years/younger (%)	48.7	51.1	−2.4	55.5	57.6	−2.1
Woman with primary education/above (%)	50.9	45.1	5.8	57.5	52.9	4.6
Woman doing farming for occupation (%)	48.5	51.1	−2.6	56.3	56.9	−0.6
Muslim woman (%)	49.9	50.4	−0.5	58.2	53.5	4.7
Woman with one birth/parity 1 (%)	48.0	50.8	−2.8	57.9	56.1	1.8
Household size below the median size of 5 members (%)	50.7	49.1	1.6	55.3	57.9	−2.6
Enabling factors						
Woman with any health insurance (%)	45.6	50.4	−4.8	61.6	56.1	5.5
Household with poorest wealth status (Tercile 1) (%)	47.8	49.6	−1.8	59.7	54.2	5.5
Household with middle wealth status (Tercile 2) (%)	52.6	49.6	3.0	56.9	54.2	2.7
Household in rural district (%)	50.4	48.1	2.3	56.7	56.4	0.3

We used a t-test to test the null hypothesis of a gap (column 3 and 6) equals to zero; Tercile 3 (least poor) was the reference category for Tercile 1 and 2; [a]denotes significance at 1%, [b]at 5%, and [c]at 10% level

subgroups of education, occupation, parity and age (Appendix 3: Table 7). Some changes in the results were noted with the use of a logit model, the pro-middle wealth and pro-rural effect on deliveries disappeared but all other results including the pro-uninsured effect remained the same (Appendix 4: Table 8). When standard errors were clustered at the district-level instead of at facility-level, the differential effect on deliveries by health insurance and wealth status disappeared, and women from larger households increased institutional deliveries more than their counterparts, but all other results including the pro-rural effect remained unchanged (Appendix 5: Table 9).

Discussion

This study examined the distribution of P4P effects on service utilisation outcomes across population subgroups in Tanzania. This is the first study in LMICs to examine who is really benefiting from the effects of P4P across a broad range of population characteristics which aligns with the social determinants of health framework. We found that P4P increased institutional deliveries more among women in middle wealth status households, among the uninsured, and among women living in rural areas than among wealthier, insured, and urban residing women. However, these differential effects were sensitive to the analytical specifications used during the robustness checks. The effect of P4P on IPT2 was equally

Table 4 Effect of P4P on service utilisation outcomes by population subgroups

Subgrouping variables	Institutional deliveries			Uptake of IPT2		
	Average subgroup effect		Differential effect test (*p*-value)	Average subgroup effect		Differential effect test (*p*-value)
	N	Beta		N	Beta	
Marital status						
Married	3869	7.7[a]	(*p* = 0.564)	3253	10.2[a]	(*p* = 0.927)
Unmarried	1878	9.1[b]		1504	9.1	
Maternal age						
Younger below the median age	2914	8.5[a]	(*p* = 0.553)	2336	9.6[b]	(*p* = 0.841)
Older above the median age	2833	7.2[b]		2421	9.8[b]	
Education						
Some education	4611	8.9[a]	(*p* = 0.378)	3877	9.3[a]	(*p* = 0.780)
No education/illiterate	1136	5.9		880	16.5[c]	
Occupation						
Farmer	2950	11.5[a]	(*p* = 0.133)	2434	16.0[a]	(*p* = 0.167)
Non-farmer	2797	5.6[b]		2323	5.6	
Religion						
Muslim	4376	9.7[a]	(*p* = 0.435)	3623	10.5[a]	(*p* = 0.562)
Non-Muslim	1371	3.9		1134	6.0	
Parity/births						
One birth	1886	9.7[a]	(*p* = 0.517)	1510	9.3[c]	(*p* = 0.882)
Two or more births	3861	7.6[a]		3247	10.3[a]	
Household size by members						
Small size (< 5)	2996	5.1[c]	(*p* = 0.173)	2476	7.7[c]	(*p* = 0.964)
Large size (≥5)	2751	10.4[a]		2281	9.9[b]	
Health insurance						
Insured	475	−7.6	(p = 0.001)	429	20.1[c]	(*p* = 0.932)
Uninsured	5272	9.7[a]		4328	10.4[a]	
Household wealth subgroups						
Tercile 1 (poorest)	1940	11.4[b]	(*p* = 0.232)	1559	14.5[b]	(*p* = 0.158)
Tercile 2 (middle)	1916	10.2[a]	(p = 0.094)	1576	16.2[a]	(*p* = 0.149)
Tercile 3 (least poor)	1891	3.7	Reference	1622	2.6	Reference
Place of residence						
Rural district	4694	9.9[a]	(p = 0.028)	3851	11.4[a]	(*p* = 0.349)
Urban district	1053	0.9		906	3.3	

Beta is the estimated P4P effect on a specific subgroup in percentage point after controlling for a year dummy, facility-fixed effects, and individual and household-level covariates (age, education, occupation, religion, marital status, parity, health insurance status, household size, and household wealth status); Each cell for Beta and differential effect reports the result from a separate regression; Differential effect test is a t-test of the null that the coefficient on the three-way interaction between the P4P effect and subgrouping indicator is zero; [a]denotes significance at 1%, [b]at 5%, and [c]at 10% level

distributed across population subgroups, and was robust across various analytical specifications. Our results show a declining trend in inequality to access institutional deliveries since service use improved most for subgroups which initially showed low utilisation rates; while the absence of inequality in uptake of IPT2 at baseline maintained after the introduction of P4P.

The greater impact of P4P on the use of institutional deliveries among women in the middle wealth households and uninsured than wealthier and insured respectively, is likely in part due to the increased adherence to user fee exemption policy among public facilities as well as the improved availability of drugs, minimising the need to pay for drugs in private pharmacies [5, 10, 11, 27, 39, 43, 45, 85, 86, 90]. The worse-off groups which experienced a greater P4P effect were also more responsive to a change in healthcare costs [33, 49]. This is consistent with our conceptual framework and demand theory, whereby the supply-side responses of reducing the financial barriers to access delivery care in turn stimulated the demand-side responses on service utilisation mostly among the disadvantaged population.

The finding that the increased uptake of IPT2 was similar across population subgroups may be explained by the already almost universal access to one antenatal care visit in Tanzania (above 97%) [11, 75, 76]. In an effort to achieve the IPT2 target, providers likely encouraged women to return for subsequent antenatal care visits to receive at least two doses of IPT. This represents a relatively easy task for most providers because continuation of care needs less effort than its initiation [34]. Although the provision of IPT is within the control of providers, it also depends on the available stock of antimalarial drugs for IPT. Another reason for the lack of differential effect on IPT2 may have been the pre-existing balance in the uptake of IPT2 across population subgroups at baseline. This is the first study to examine whether P4P had a differential effect on the uptake of IPT for malaria during antenatal care in LMICs. In Burundi, Bonfrer et al. [17] examined the differential effect of P4P on other contents of antenatal care and found a pro-rich effect on blood pressure measurement and a lack of differential effect on the uptake of anti-tetanus vaccination across socioeconomic groups.

The pro-middle wealth effect of P4P on institutional deliveries, as an indication of being pro-poor, is contrary to the pro-rich effect on deliveries reported in Burundi [17], Rwanda [46] and Cambodia [77]. The pro-rich effect in Cambodia was attributed to the lack of effective demand among the poorest women due to user fees [77]; whereas in Burundi it was attributed to other costs like transport because the user fees for deliveries were removed prior to P4P [17]. However, a pilot study in Burundi [16] and a study using demographic and health survey (DHS) data in Rwanda [62] found no differential effect on deliveries by household wealth status; and the results in the later study were attributed to low and uniform coverage of services at baseline. In the Democratic Republic of Congo providers implementing P4P negotiated user fees with communities and raised revenues without hurting the poorest [73], but the equity effects of this approach were not assessed empirically. Further evidence of a pro-poor effect of P4P has been shown on immunisation services in Burundi [17], and on quality of care improvement in high-income countries especially in the United Kingdom [2, 14, 23, 24, 80].

Moreover, our study found that institutional deliveries improved more in rural than in urban areas, while there was no differential effect on institutional deliveries by place of residence in Rwanda [62]. In Rwanda, the minimal number of urban clusters compared to rural clusters were thought to limit the power to detect the differential effect by place of residence [62], while our study had a slightly higher number of urban clusters compared to Rwanda (i.e. 28 versus 22 urban clusters). In the United Kingdom, the effect of P4P on quality of care was greater in urban areas than in rural areas [36, 42], while there was no differential effect of P4P on quality of care by rural–urban area in the United States [67].

We found a greater P4P effect on institutional deliveries among uninsured women, whereas a greater effect on deliveries was found among women with health insurance in Rwanda [46] and a maternity care voucher in Cambodia [77]. The findings from Rwanda and Cambodia were attributed to reduced financial barriers to access care [46, 77], and this could be the case with a stronger enforcement of fee exemptions in Tanzania [11].

However, another study in Rwanda based on DHS, as nationally representative data, found no differential effect on deliveries by health insurance status [62]. A greater P4P effect on deliveries among uninsured women in Tanzania, is partly because the baseline institutional delivery rate was already higher among insured than uninsured women in the intervention arm. A further reason could be that uninsured women were more responsive to reduced healthcare costs compared to insured women who were already covered. It is also likely that the statistical power to detect the effect among women with insurance was limited because few women are insured in Tanzania [58], compared to other countries like Rwanda [50, 70].

Furthermore, we found a similar distribution of institutional delivery rates and IPT2 uptakes across age groups prior to P4P, and the effect of P4P was equally distributed across age groups, which is contrary to P4P studies in high-income countries as they found inequalities in quality of care across age groups existed and persisted after the introduction of P4P [2, 14, 24, 80].

Overall our findings imply that when P4P results in supply side responses that reduce demand-side barriers to accessing care, it can enhance equity in service utilisation. P4P also appears less likely to show a differential effect when there is a similar level of service utilisation in a given indicator across population subgroups prior to an intervention. This study supports the argument that P4P can enhance equity in access for services where there is a pre-existing inequity in coverage, and where efforts to remove the demand-side financial barriers to access care have been made [28, 31, 44, 57, 86]. Thus, to ensure P4P reduces inequities in access to care, policy makers should consider introducing complementary measures to reduce demand-side access barriers. P4P is likely to be most effective at reducing inequities in settings where they offer free health services or there is high coverage of pre-payment schemes.

To make progress towards universal health coverage and achieve sustainable development goal three especially in LMICs, more efforts are needed to stimulate demand for and supply of healthcare services [57, 86, 90]. Further insights on how supply and demand side interventions interact and complement each other to affect outcomes are needed. Moreover, because the social determinants of health as sources of inequalities emerge from different sectors, strategies within the health sector alone cannot reduce inequalities in access and use of health services [21, 54].

This study has a number of limitations. First, our study may have been underpowered to detect the effect of P4P in some groups, for example among insured women and urban residents, possibly due to the more limited sample size within sub groups. Second, our results of differential effects on deliveries by wealth status, health insurance and place of residence, were not consistent across all analytical specifications used in robustness checks (i.e. non-linear model, and district level clustering of standard errors). However, the differential effects on deliveries for other subgroups of social determinants, and differential effects on IPT2, were robust to all analytical specifications used. Third, our finding that P4P reduces inequalities in service utilisation might be reflective of a regression to the mean principle (a random fluctuation rather than a true causal effect) because of having a short term evaluation [7]. Lastly, we restricted our distributional analysis to the outcomes which improved significantly under P4P. Although the inequalities in service use may happen with an outcome which showed insignificant P4P effect on average, our focus was limited to how the increased average utilisation effects were distributed across population subgroups.

Conclusion

In Tanzania, the effect of P4P on institutional deliveries was greater among women in middle wealth households, in rural areas and among the uninsured women than their counterparts. P4P effect on the uptake of IPT2 was equally distributed across population subgroups. Our finding suggests that P4P can enhance equitable healthcare access and use especially when the financial barriers to access care are reduced or removed.

Appendix

Appendix 1

Table 5 Items used to construct household wealth status score

No.	Variable description
1.	Asset: electricity
2.	Asset: working radio
3.	Asset: working television (TV)
4.	Asset: working DVD
5.	Asset: working mobile phone
6.	Asset: working landline phone
7.	Asset: working iron
8.	Asset: working refrigerator
9.	Asset: working wall watch
10.	Asset: sewing machine
11.	Asset: table
12.	Asset: sofa coach
13.	Asset: cupboard
14.	Asset: motorcycle
15.	Asset: car
16.	Household member with a bank account
17.	Number of sleeping rooms
18.	Source of drinking water: piped water
19.	Source of drinking water: borehole/ covered well
20.	Source of drinking water: open well
21.	Source of drinking water: spring water
22.	Source of drinking water: river/ dam/pond/lake
23.	Toilet type: flush toilet
24.	Toilet type: pit latrine
25.	Toilet type: no/ other toilet
26.	Source of cooking energy: electricity
27.	Source of cooking energy: kerosene/paraffin
28.	Source of cooking energy: charcoal
29.	Source of cooking energy: firewood
30.	Source of light: electricity
31.	Source of light: solar
32.	Source of light: kerosene/ paraffin
33.	Source of light: candle/ firewood
34.	Source of light: torch or other source
35.	Floor material: sand/earth/dung
36.	Floor material: cement
37.	Floor material: other
38.	Wall material: grass/poles/mud wall
39.	Wall material: bamboo with mud wall
40.	Wall material: sundried/ burnt bricks
41.	Wall material: cement blocks
42.	Wall material: stones with mud

Appendix 2

Table 6 Effect of P4P on service utilisation by different categories of wealth status and by arm-based wealth subgroups

Wealth subgrouping variables	Institutional deliveries			Uptake of IPT2		
	Average subgroup effect		Differential effect test (p-value)	Average subgroup effect		Differential effect test (p-value)
	N	Beta		N	Beta	
Panel A: Wealth subgroups						
Three wealth subgroups (Terciles)						
T1	1940	11.4[b]	(p = 0.232)	1559	14.5[b]	(p = 0.158)
T2	1916	10.2[a]	(p = 0.094)[c]	1576	16.2[a]	(p = 0.149)
T3	1891	3.7	Reference	1622	2.6	Reference
Five wealth subgroups (Quintiles)						
Q1	1170	13.8[b]	(*p* = 0.079)[c]	929	13.6[c]	(*p* = 0.166)
Q2	1158	8.8[c]	(*p* = 0.069)[c]	939	16.3[b]	(*p* = 0.102)
Q3	1143	8.2[c]	(*p* = 0.034)[b]	938	21.8[a]	(*p* = 0.120)
Q4	1146	11.4[a]	(*p* = 0.015)[b]	979	14.4[b]	(*p* = 0.175)
Q5	1130	−0.5	Reference	972	1.9	Reference
Panel B: Arm-based wealth subgroups						
Three wealth subgroups (Terciles)						
AT1	1917	10.2[b]	(*p* = 0.293)	1540	13.8[b]	(*p* = 0.117)
AT2	1913	9.2[b]	(*p* = 0.156)	1568	18.3[a]	(*p* = 0.084)[c]
AT3	1917	3.9[c]	Reference	1649	2.5	Reference
Five wealth subgroups (Quintiles)						
AQ1	1149	15.3[a]	(*p* = 0.089)[c]	914	16.8[b]	(*p* = 0.108)
AQ2	1151	6.6	(*p* = 0.230)	935	15.2[b]	(*p* = 0.139)
AQ3	1147	12.3[b]	(p = 0.001)[a]	949	14.6[b]	(p = 0.156)
AQ4	1152	9.9[b]	(*p* = 0.022)[b]	972	7.7	(*p* = 0.310)
AQ5	1148	0.3	Reference	987	0.5	Reference

[a] denotes significance at 1%, [b] at 5%, and [c] at 10% level; Beta is the estimated P4P effect on a specific subgroup in percentage point after controlling for a year dummy, facility-fixed effects, and individual and household-level covariates (age, education, occupation, religion, marital status, parity, health insurance status, household size, and household wealth status); Each cell for Beta and differential effect reports the result from a separate regression; Differential effect test is a t-test of the null that the coefficient on the three-way interaction between the P4P effect and subgrouping indicator is zero

Appendix 3

Table 7 Effect of P4P on service utilisation by subgroups for categorical variables

Subgrouping variables	Institutional deliveries			Uptake of IPT2		
	Average subgroup effect		Differential effect test (p-value)	Average subgroup effect		Differential effect test (p-value)
	N	Beta		N	Beta	
Education subgroups						
No education	1136	5.9	Reference	880	17.0[b]	Reference
Some primary	459	4.1	($p = 0.550$)	355	9.1	($p = 0.479$)
Primary/some secondary	3729	11.3[a]	($p = 0.157$)	3148	12.1[a]	($p = 0.965$)
Secondary/above	423	3.8	($p = 0.276$)	374	−9.8	($p = 0.144$)
Occupation subgroups						
Formal sector	113	−17.4	($p = 0.715$)	99	−5.1	($p = 0.329$)
Farmers	2950	11.6[a]	($p = 0.162$)	2434	15.9[a]	($p = 0.777$)
Self-employed	1167	7.7[b]	($p = 0.650$)	996	1.1	($p = 0.132$)
Unemployed	1517	3.9	Reference	1228	16.8[a]	Reference
Birth parity subgroups						
Parity 1	1886	9.8[a]	Reference	1510	9.3[c]	Reference
Parity 2	1353	3.4	($p = 0.215$)	1123	7.0	($p = 0.583$)
Parity 3	1029	10.9[b]	($p = 0.766$)	868	0.4	($p = 0.317$)
Parity 4	664	3.3	($p = 0.342$)	570	3.2	($p = 0.567$)
Parity 5+	815	13.3[c]	($p = 0.700$)	686	30.0[a]	($p = 0.038$)
Age subgroups						
Age (15–19) years	965	11.5[a]	Reference	726	19.2[b]	Reference
Age (20–24) years	1613	9.7[a]	($p = 0.366$)	1322	4.2	($p = 0.708$)
Age (25–29) years	1459	4.2	($p = 0.568$)	1232	7.3	($p = 0.820$)
Age (30–34) years	978	4.9	($p = 0.510$)	846	10.3	($p = 0.666$)
Age (35+) years	732	15.5[a]	($p = 0.446$)	631	20.4[b]	($p = 0.218$)

[a]denotes significance at 1%, [b]at 5%, and [c]at 10% level; Beta is the estimated P4P effect on a specific subgroup in percentage point after controlling for a year dummy, facility-fixed effects, and individual and household-level covariates (age, education, occupation, religion, marital status, parity, health insurance status, household size, and household wealth status); Each cell for Beta and differential effect reports the result from a separate regression; Differential effect test is a t-test of the null that the coefficient on the three-way interaction between the P4P effect and subgrouping indicator is zero

Appendix4

Table 8 Effect of P4P on service utilisation by subgroups –using the non–linear logit model

Subgrouping variables	Institutional deliveries			Uptake of IPT2		
	Average subgroup effect		Differential effect test (p-value)	Average subgroup effect		Differential effect test (p-value)
	N	(dy/dx)		N	(dy/dx)	
Marital status						
Married	3385	9.2[a]	(*p* = 0.503)	3253	9.2[a]	(*p* = 0.935)
Unmarried	1338	13.3[a]		1481	9.8[c]	
Maternal age						
Younger below the median age	2361	11.2[a]	(*p* = 0.492)	2336	9.2[b]	(*p* = 0.830)
Older above the median age	2325	9.1[a]		2421	9.5[b]	
Education						
Some education	4021	10.9[a]	(*p* = 0.070)	3877	8.6[a]	(*p* = 0.793)
No education/illiterate	900	9.1		816	16.5[c]	
Occupation						
Farmer	2638	13.4[a]	(*p* = 0.590)	2396	16.0[a]	(p = 0.149)
Non-farmer	2126	7.5[b]		2295	5.3	
Religion						
Muslim	3991	10.8[a]	(*p* = 0.497)	3614	9.7[a]	(*p* = 0.554)
Non-Muslim	980	5.6		1061	7.8	
Parity/births						
One birth	1180	15.2[a]	(*p* = 0.122)	1476	9.9[c]	(*p* = 0.939)
Two or more births	3436	9.3[a]		3247	10.0[a]	
Household size by members						
Small size (< 5)	2381	7.3[b]	(*p* = 0.320)	2464	7.6[c]	(*p* = 0.903)
Large size (≥5)	2299	12.8[a]		2281	9.1[b]	
Health insurance						
Insured	171	−20.7	(*p* = 0.012)	315	18.3	(*p* = 0.900)
Uninsured	4820	11.1[a]		4328	10.1[a]	
Household wealth status						
Tercile 1 (poorest)	1656	13.4[b]	(*p* = 0.894)	1508	13.2[b]	(*p* = 0.145)
Tercile 2 (middle)	1528	12.7[a]	(*p* = 0.737)	1539	17.1[a]	(*p* = 0.106)
Tercile 3 (least poor)	1066	8.2[b]	Reference	1599	2.4	Reference
Place of residence						
Rural district	4387	11.3[a]	(*p* = 0.152)	3851	11.2[a]	(*p* = 0.268)
Urban district	787	1.6		906	1.7	

Non-linear logit model with FE, covariates, clustering at HF level; Logit with FE cuts down the sample size; dy/dx is the estimated partial P4P effect on a specific subgroup in terms of marginal effect after controlling for a year dummy, facility-fixed effects, and individual and household-level covariates (age, education, occupation, religion, marital status, parity, health insurance status, household size, and household wealth status); Each cell for dy/dx and differential effect reports the result from a separate regression; Differential effect test is a t-test of the null that the coefficient on the three-way interaction between the P4P effect and subgrouping indicator is zero; [a] denotes significance at 1%, [b]at 5%, and [c] at 10% level

Appendix 5

Table 9 Effect of P4P on service utilisation by subgroups –using district-level clustering of Standard Errors

Subgrouping variables	Institutional deliveries			Uptake of IPT2		
	Average subgroup effect		Differential effect test (p-value)	Average subgroup effect		Differential effect test (p-value)
	N	Beta		N	Beta	
Marital status						
Married	3869	7.7	(*p* = 0.580)	3253	10.2^{a}	(*p* = 0.960)
Unmarried	1878	9.1		1504	9.1	
Maternal age						
Younger below the median age	2914	8.5	(*p* = 0.565)	2336	9.6^{b}	(*p* = 0.790)
Older above the median age	2833	7.2^{b}		2421	9.8	
Education						
Some education	4611	8.9^{c}	(*p* = 0.400)	3877	9.3^{a}	(*p* = 0.800)
No education/illiterate	1136	5.9		880	16.5	
Occupation						
Farmer	2950	11.5^{c}	(*p* = 0.140)	2434	16.0^{a}	(*p* = 0.060)
Non-farmer	2797	5.6		2323	5.6	
Religion						
Muslim	4376	9.7^{a}	(*p* = 0.600)	3623	10.5^{b}	(*p* = 0.440)
Non-Muslim	1371	3.9		1134	6.0	
Parity/births						
One birth	1886	9.7^{b}	(*p* = 0.455)	1510	9.3^{a}	(*p* = 0.895)
Two or more births	3861	7.6^{c}		3247	10.3	
Household size by members						
Small size (< 5)	2996	5.1	(*p* = 0.045)	2476	7.7^{c}	(*p* = 0.925)
Large size (≥5)	2751	10.4^{c}		2281	9.9^{a}	
Health insurance						
Insured	475	−7.6	(*p* = 0.225)	429	20.1^{a}	(p = 0.965)
Uninsured	5272	9.7^{b}		4328	10.4^{a}	
Household wealth status						
Tercile 1 (poorest)	1940	11.4	(p = 0.400)	1559	14.5^{a}	(p = 0.120)
Tercile 2 (middle)	1916	10.2^{c}	(*p* = 0.125)	1576	16.2^{a}	(*p* = 0.050)
Tercile 3 (least poor)	1891	3.7	Reference	1622	2.6	Reference
Place of residence						
Rural district	4694	9.9^{c}	(*p* = 0.080)	3851	11.4^{b}	(*p* = 0.430)
Urban district	1053	0.9		906	3.3	

Abbreviations

DHS: Demographic Health Survey; IPT: Intermittent preventive treatment; LMICs: Low- and middle-income countries; MCH: Maternal and child health; MoHSW: Ministry of Health and Social Welfare; P4P: Payment for performance; USD: United States dollar

Acknowledgements

We would like to thank the whole P4P evaluation research team, including data collectors and field coordinators. We thank Elinor Bartle for the language editing and Ottar Mæstad for reviewing the paper. We also acknowledge the support from all field respondents, and various officials such as Regional Medical Officers, District Medical officers, Health facility in-charges, and P4P implementing agency.

Funding

The Government of Norway funded the data collection for the program evaluation that was used in this paper (grant numbers: TAN-3108 and TAN 13/0005. http://www.norad.no/en/) and the UK Department for International Development (DFID) as part of the Consortium for Research on Resilient and Responsive Health Systems (RESYST) supported the data analysis and writing of this paper. This study is part of a PhD thesis at the University of Bergen for Peter Binyaruka, who is financially supported by the Norwegian State Education Loan Fund. The funding bodies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The data have been uploaded into a data repository. The DOI URL for the dataset is: https://doi.org/10.5281/zenodo.21709.

Authors' contributions
PB was involved in designing this sub-study with JB and oversaw data collection, and analysed the data. JB designed this sub-study and the impact evaluation within which it is embedded. PB wrote the first draft of the manuscript. BR, GT, JB involved in data interpretation, presentation, and revision of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Ethical approval for the evaluation study was obtained from the Institutional Review Board of the Ifakara Health Institute (approval number: 1BI1IRB/38) and the Ethics Review Board of the London School of Hygiene & Tropical Medicine. Study participants provided written consent to participate in this study, requiring them to sign a written consent form that was read out to them by the interviewers. This consent form was reviewed and approved by the ethics committees prior to the start of the research.

Consent for publication
Not applicable.

Competing interests
The authors of this manuscript have the following competing interests: two authors (PB and JB) were funded by the Government of Norway to undertake the data collection associated with this research. The Government of Norway also funded the P4P programme in Pwani region of Tanzania. The funder of the study had no role in data analysis, data interpretation, or writing of the manuscript.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Centre for International Health, University of Bergen, PO Box 7804, N-5020 Bergen, Norway. [2]Ifakara Health Institute, PO Box 78373, Dar es Salaam, Tanzania. [3]Chr. Michelsen Institute, PO Box 6033, Bergen, Norway. [4]Department of Economics, University of Oslo, PO Box 1095, Oslo, Norway. [5]Department of Global Health and Development, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London WC1H 9SH, UK.

Received: 14 September 2017 Accepted: 16 January 2018
Published online: 29 January 2018

References

1. Alderman H, Lavy V. Household responses to public health services: cost and quality tradeoffs. The World Bank Research Observer. 1996;11:3–22.
2. Alshamsan R, Majeed A, Ashworth M, Car J, Millett C. Impact of pay for performance on inequalities in health care: systematic review. J Health Serv Res Policy. 2010;15:178–84.
3. Andersen R. A behavioral model of families' use of health services. In: A behavioral model of families' use of health services; 1968.
4. Andersen R, Newman JF. Societal and individual determinants of medical care utilization in the United States. Milbank Mem Fund Q Health Soc. 1973;51:95–124.
5. Anselmi L, Binyaruka P, Borghi J. Understanding causal pathways within health systems policy evaluation through mediation analysis: an application to payment for performance (P4P) in Tanzania. Implement Sci. 2017;12:10.
6. Arrow KJ. Uncertainty and the welfare economics of medical care. Am Econ Rev. 1963;53:941–73.
7. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. Int J Epidemiol. 2005;34:215–20.
8. Barros AJ, Ronsmans C, Axelson H, et al. Equity in maternal, newborn, and child health interventions in countdown to 2015: a retrospective review of survey data from 54 countries. Lancet. 2012;379:1225–33.
9. Basinga P, Gertler PJ, Binagwaho A, et al. Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. Lancet. 2011;377:1421–8.
10. Binyaruka P, Borghi J. Improving quality of care through payment for performance: examining effects on the availability and stock-out of essential medical commodities in Tanzania. Tropical Med Int Health. 2017;22:92–102.
11. Binyaruka P, Patouillard E, Powell-Jackson T, et al. Effect of paying for performance on utilisation, quality, and user costs of health Services in Tanzania: a controlled before and after study. PLoS One. 2015;10:e0135013.
12. Bitler MP, Gelbach JB, Hoynes HW. What mean impacts miss: distributional effects of welfare reform experiments. Am Econ Rev. 2006;96:988–1012.
13. Bitler MP, Gelbach JB, Hoynes HW. Distributional impacts of the self-sufficiency project. J Public Econ. 2008;92:748–65.
14. Boeckxstaens P, Smedt DD, Maeseneer JD, Annemans L, Willems S. The equity dimension in evaluations of the quality and outcomes framework: a systematic review. BMC Health Serv Res. 2011;11:209.
15. Boerma JT, Bryce J, Kinfu Y, Axelson H, Victora CG. Mind the gap: equity and trends in coverage of maternal, newborn, and child health services in 54 countdown countries. Lancet. 2008;371:1259–67.
16. Bonfrer I, Soeters R, Van de Poel E, et al. Introduction of performance-based financing in burundi was associated with improvements in care and quality. Health Aff (Millwood). 2014a;33:2179–87.
17. Bonfrer I, Van de Poel E, Van Doorslaer E. The effects of performance incentives on the utilization and quality of maternal and child care in Burundi. Soc Sci Med. 2014b;123:96–104.
18. Borghi J, Mayumana I, Mashasi I, et al. Protocol for the evaluation of a pay for performance programme in Pwani region in Tanzania: a controlled before and after study. Implement Sci. 2013;8:80.
19. Brand JE, Xie Y. Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. Am Sociol Rev. 2010;75:273–302.
20. Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. J Hum Resour. 2015;50:317–72.
21. CSDH. 2008. Closing the gap in a generation: health equity through action on the social determinants of health.
22. Djebbari H, Smith J. Heterogeneous impacts in PROGRESA. J Econ. 2008;145:64–80.
23. Doran T, Fullwood C, Kontopantelis E, Reeves D. Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework. Lancet. 2008;372:728–36.
24. Eijkenaar F, Emmert M, Scheppach M, Schoffski O. Effects of pay for performance in health care: a systematic review of systematic reviews. Health Policy. 2013;110:115–30.
25. Ellis RP, McGuire TG. Hospital response to prospective payment: moral hazard, selection, and practice-style effects. J Health Econ. 1996;15:257–77.
26. Engineer CY, Dale E, Agarwal A, et al. Effectiveness of a pay-for-performance intervention to improve maternal and child health services in Afghanistan: a cluster-randomized trial. Int J Epidemiol. 2016;45:451–9.
27. Ensor T, Cooper S. Overcoming barriers to health service access: influencing the demand side. Health Policy Plan. 2004;19:69–79.
28. Evans DB, Hsu J, Boerma T. Universal health coverage and universal access. *Bull World Health Organ*. 2013;**91**:546.
29. Filmer D, Pritchett LH. Estimating wealth effects without expenditure data– or tears: an application to educational enrollments in states of India. Demography. 2001;38:115–32.
30. Fort AL, Kothari MT, Abderrahim N. Postpartum care: levels and determinants in developing countries. In: USAID; Calverton. Maryland: Macro International Inc.; 2006.
31. Fritsche G, Soeters R, Meessen B. Performance-based financing toolkit. Washington DC: The World Bank; 2014.
32. Gabrysch S, Campbell OM. Still too far to walk: literature review of the determinants of delivery service use. BMC Pregnancy Childbirth. 2009;9:34.
33. Gertler P, Locay L, Sanderson W. Are user fees regressive?: the welfare implications of health care financing proposals in Peru. J Econ. 1987;36:67–88.
34. Gertler P, Vermeersch C. Using performance incentives to improve health outcomes. In: NBER working paper no. 19046; 2013.
35. Gillam SJ, Siriwardena AN, Steel N. Pay-for-performance in the United Kingdom: impact of the quality and outcomes framework: a systematic review. Ann Fam Med. 2012;10:461–8.
36. Gravelle H, Sutton M, Ma A. 2008. Doctor behaviour under a pay for performance contract: further evidence from the quality and outcomes framework. CHE Research Paper 34.
37. Gwatkin DR. How much would poor people gain from faster progress towards the millennium development goals for health? Lancet. 2005;365:813–7.
38. Heckman JJ, Smith J, Clements N. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. Rev Econ Stud. 1997;64:487–535.

39. Idd A, Yohana O, Maluka SO. Implementation of pro-poor exemption policy in Tanzania: policy versus reality. Int J Health Plann Manag. 2013;28:e298–309.
40. Ireland M, Paul E, Dujardin B. Can performance-based financing be used to reform health systems in developing countries? Bull World Health Organ. 2011;89:695–8.
41. Khandker SR, Koolwal GB, Samad HA. Handbook on impact evaluation: quantitative methods and practices. Washington DC: The World Bank; 2010.
42. Kontopantelis E, Buchan I, Reeves D, Checkland K, Doran T. Relationship between quality of care and choice of clinical computing system: retrospective analysis of family practice performance under the UK's quality and outcomes framework. BMJ Open. 2013;3(8):1–11.
43. Kruk ME, Mbaruku G, Rockers PC, Galea S. User fee exemptions are not enough: out-of-pocket payments for 'free' delivery services in rural Tanzania. Tropical Med Int Health. 2008;13:1442–51.
44. Kutzin J. Health financing for universal coverage and health system performance: concepts and implications for policy. Bull World Health Organ. 2013;91:602–11.
45. Lagarde M, Palmer N. The impact of user fees on health service utilization in low- and middle-income countries: how strong is the evidence? Bull World Health Organ. 2008;86:839–48.
46. Lannes L, Meessen B, Soucat A, Basinga P. Can performance-based financing help reaching the poor with maternal and child health services? The experience of rural Rwanda. Int J Health Plann Manag. 2016;31:309–48.
47. Le Grand J. Equity versus efficiency: the elusive trade-off. Ethics. 1990;100:554–68.
48. Li J, Hurley J, DeCicca P, Buckley G. Physician response to pay-for-performance: evidence from a natural experiment. Health Econ. 2014;23:962–78.
49. Litvack JI, Bodart C. User fees plus quality equals improved access to health care: results of a field experiment in Cameroon. Soc Sci Med. 1993;37:369–83.
50. Lu C, Chin B, Lewandowski JL, et al. Towards universal health coverage: an evaluation of Rwanda Mutuelles in its first eight years. PLoS One. 2012;7:e39282.
51. Markovitz AA, Ryan AM. Pay-for-performance: disappointing results or masked heterogeneity? Med Care Res: Rev; 2016.
52. Marmot M. Social determinants of health inequalities. Lancet. 2005;365:1099–104.
53. Marmot M. Universal health coverage and social determinants of health. Lancet. 2013;382:1227–8.
54. Marmot M, Friel S, Bell R, Houweling TA, Taylor S. Closing the gap in a generation: health equity through action on the social determinants of health. Lancet. 2008;372:1661–9.
55. Mayumana I, Borghi J, Anselmi L, Mamdani M, Lange S. Effects of payment for performance on accountability mechanisms: evidence from Pwani, Tanzania. Soc Sci Med. 2017;179:61–73.
56. McIntyre D, Thiede M, Birch S. Access as a policy-relevant concept in low- and middle-income countries. Health Econ Policy Law. 2009;4:179–93.
57. Meessen B, Soucat A, Sekabaraga C. Performance-based financing: just a donor fad or a catalyst towards comprehensive health-care reform? Bull World Health Organ. 2011;89:153–6.
58. Mtei G, Makawia S, Masanja H. Monitoring and evaluating progress towards universal health coverage in Tanzania. PLoS Med. 2014;11:e1001698.
59. NBS. Tanzania population and housing census: population distribution by administrative areas 2012. Dar es Salaam: National Bureau of Statistics (NBS); 2013.
60. O'Neill J, Tabish H, Welch V, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. J Clin Epidemiol. 2014;67:56–64.
61. Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S. Does pay-for-performance improve the quality of health care? Ann Intern Med. 2006;145:265–72.
62. Priedeman Skiles M, Curtis SL, Basinga P, Angeles G. An equity analysis of performance-based financing in Rwanda: are services reaching the poorest women? Health Policy Plan. 2013;28:825–37.
63. Renmans D, Holvoet N, Orach CG, Criel B. Opening the 'black box' of performance-based financing in low- and lower middle-income countries: a review of the literature. Health Policy Plan. 2016;31:1297–309.
64. Roland M. Linking physicians' pay to the quality of care–a major experiment in the United kingdom. N Engl J Med. 2004;351:1448–54.
65. Rosenthal MB, Fernandopulle R, Song HR, Landon B. Paying for quality: providers' incentives for quality improvement. Health Aff (Millwood). 2004;23:127–41.
66. Rosenthal MB, Frank RG. What is the empirical basis for paying for quality in health care? Med Care Res Rev. 2006;63:135–57.
67. Ryan AM, Blustein J. The effect of the MassHealth hospital pay-for-performance program on quality. Health Serv Res. 2011;46:712–28.
68. Say L, Raine R. A systematic review of inequalities in the use of maternal health care in developing countries: examining the scale of the problem and the importance of context. Bull World Health Organ. 2007;85:812–9.
69. Scott A, Sivey P, Ait Ouakrim D, et al. The effect of financial incentives on the quality of health care provided by primary care physicians. Cochrane Database Syst Rev. 2011(9):Cd008451.
70. Sekabaraga C, Diop F, Soucat A. Can innovative health financing policies increase access to MDG-related services? Evidence from Rwanda. Health Policy Plan. 2011;Suppl 2(26):ii52–62.
71. Simkhada B, Teijlingen ER, Porter M, Simkhada P. Factors affecting the utilization of antenatal care in developing countries: systematic review of the literature. J Adv Nurs. 2008;61:244–60.
72. Skiles MP, Curtis SL, Basinga P, Angeles G, Thirumurthy H. The effect of performance-based financing on illness, care-seeking and treatment among children: an impact evaluation in Rwanda. BMC Health Serv Res. 2015;15:375.
73. Soeters R, Peerenboom PB, Mushagalusa P, Kimanuka C. Performance-based financing experiment improved health care in the Democratic Republic of Congo. Health Aff (Millwood). 2011;30:1518–27.
74. Solar O, Irwin A. A conceptual framework for action on the social determinants of health. Social determinants of health. In: Discussion paper 2 (policy and practice); 2010.
75. TDHS. Tanzania demographic and health survey 2010. Dar es Salaam: National Bureau of Statistics (NBS); 2011.
76. TDHS. Tanzania demographic and health survey and malaria indicator survey 2015–16. Dar es Salaam: National Bureau of Statistics (NBS); 2016.
77. Van de Poel E, Flores G, Ir P, O'Donnell O. Impact of performance-based financing in a low-resource setting: a decade of experience in Cambodia. Health Econ. 2016;25:688–705.
78. van Doorslaer E, Masseria C, Koolman X. Inequalities in access to medical care by income in developed countries. CMAJ. 2006;174:177–83.
79. Van Doorslaer E, Wagstaff A, Van der Burg H, et al. Equity in the delivery of health care in Europe and the US. J Health Econ. 2000;19:553–83.
80. Van Herck P, De Smedt D, Annemans L, et al. Systematic review: effects, design choices, and context of pay-for-performance in health care. BMC Health Serv Res. 2010;10:247.
81. Victora CG, Barros AJ, Axelson H, et al. How changes in coverage affect equity in maternal and child health interventions in 35 countdown to 2015 countries: an analysis of national surveys. Lancet. 2012;380:1149–56.
82. Victora CG, Requejo JH, Barros AJ, et al. Countdown to 2015: a decade of tracking progress for maternal, newborn, and child survival. Lancet. 2016;387:2049–59.
83. Vyas S, Kumaranayake L. Constructing socio-economic status indices: how to use principal components analysis. Health Policy Plan. 2006;21:459–68.
84. Wagstaff A. Poverty and health sector inequalities. Bull World Health Organ. 2002;80:97–105.
85. WHO. Equitable access to essential medicines: a framework for collective action. Geneva: World Health Organization; 2004.
86. WHO. The world health report: health systems financing: the path to universal coverage. Geneva: World Health Organization; 2010.
87. WHO. Handbook on health inequality monitoring: with a special focus on low- and middle-income countries. Geneva: World Health Organization; 2013.
88. Witter S, Fretheim A, Kessy FL, Lindahl AK. Paying for performance to improve the delivery of health interventions in low- and middle-income countries. Cochrane Database Syst Rev. 2012(2):Cd007899.
89. Witter S, Toonen J, Meessen B, et al. Performance-based financing as a health system reform: mapping the key dimensions for monitoring and evaluation. BMC Health Serv Res. 2013;13:367.
90. Yates R. Universal health care and the removal of user fees. Lancet. 2009;373:2078–81.

NORDIC ECOLABEL
241 PRINTED MATTER 731

Graphic design: Communication Division, UIB / Print: Skipnes Kommunikasjon AS

UNIVERSITAS BERGENSIS

uib.no

ISBN: 978-82-308-3637-8