Evaluating the Effects of Non-Financial Health System Policies on Health Care Access and Quality in Low- and Middle-Income Countries

Finbarr McGuire

PhD
University of York
Economics
September 2022

Abstract

This thesis consists of four independent chapters contributing to the empirical evidence on health care access and quality in low- and middle-income countries. The first two chapters examine issues of health care access and accessibility in Malawi, while the last two chapters evaluate topics on the quality of health care in South Africa, specifically examining policies which may act to increase the quality of health care supplied. The unifying theme of these chapters is that these system-wide policies do not lend themselves to evaluation through random assignment to treatment. In circumstances where experimental designs are not practicable or pragmatic, alternative methods of evaluation must be pursued. Therefore, all chapters exploit observational data to evaluate the respective health-system level policies examined. Chapter 1 examines distance as a determinant of obstetric health care utilisation in Malawi. Specifically, the study explores how the relationship between distance and health care utilisation may change across levels of distance. When combined with information on geographic population distributions this provides valuable evidence for health infrastructure planning. Additionally, attempts to address possible endogeneity of distance are made and whether the effect of distance differs across sub-groups. The results illustrate that distance continues to be a barrier to obstetric health care utilisation. Chapter 2 evaluates the effect of the availability of maternity waiting homes, a policy devised to overcome the distance barrier women face in accessing obstetric health care services in Malawi. Time and space variation in the construction of maternity waiting homes at health facilities are exploited, estimating various difference-in-difference specifications. The findings suggest no strong effect of maternity waiting homes on maternal health care utilisation or child health outcomes. Chapter 3 analyses the effect of a quality improvement programme implemented in primary health care facilities in South Africa. Specifically, the chapter explores whether the programme exacerbated pre-existing differences in facility quality. The findings suggest the quality improvement programme improved quality across all facilities, but may have increased variation in quality in the short run. Continuing the exploration of means to improve quality, Chapter 4 investigates whether facilities' quality responds to neighbouring peer facilities. Strategic interactions between health facilities are modelled using both a spatial econometric framework and instrumental variable approach. Despite the absence of material incentives, the results indicate facilities do respond to quality changes among their peer facilities.

Contents

List of Figures

List of Tables

Acknowledgements

First, I would like to express my immeasurable gratitude and thanks to my supervisors Noemi Kreif and Peter Smith. I could not have been luckier in having two supervisors who both combine an intimidating amount of knowledge and immeasurable quantity of empathy. Conversations had over the course of the PhD have provided a constant source of inspiration and will continue to influence my thinking for years to come. Their patient support throughout has been invaluable in getting to this point.

I would also like to thank the members of my advisory panel, Nigel Rice, Rita Santos, Marc Schurke and Andrew Mirelman, the thesis undoubtedly benefited for their insightful comments.

I am thankful to the Centre for Health Economics who funded my PhD and provided an amazing environment in which to undertake my studies. I would like to thank all the staff who contribute to creating an unrivalled level of collegiality which makes being part of the centre such a pleasure. Particularly, I would like to thank my fellow PhD students, including Laurie Rachet-Jaquet, Luis Fernandes, Gowokani Chirwa, Francesco Ramponi, David Glynn and Jacopo Gabani who, at various stages, I shared this journey with.

I dedicate this thesis to my family for their continuous love and support. In particular, I would like express my deep gratitude to Mireia, Christine, Marie and Elsapeth for their constant support. I am incredibly fortunate to be surrounded by such kind and caring people. A special thanks to my parents, Ali and Anne. I am incredibly grateful for all you have done for me; being a constant source of encouragement and always being there when I've needed. Finally, I am especially grateful to Lucie for being there throughout the PhD, providing endless encouragement and companionship during this cross-continental adventure. Without you all, this thesis would not be possible.

Declaration

I declare that this thesis is a presentation of original work and I am the main author.

Chapter 1 is co-authored with Noemi Kreif and Peter C. Smith. Previous versions of this chapter were presented and discussed at the European Union Health Economics Association Conference (Porto, 2019). A version of this chapter is published in *Health Economics*, 2021, Volume 30. Issue 9. pp. 2144-2167.

Chapter 2 is co-authored with Noemi Kreif, Peter C. Smith, Gowokani Chirwa and Gerald Manthalu. Previous versions of this chapter were presented and discussed at the 6th African Health Economics and Policy Association Conference (Kigali, 2022).

Chapter 3 is co-authored with Noemi Kreif, Peter C. Smith, Nicholas Stacey and Ijeoma Edoka. Previous versions of this chapter were presented and discussed at the 10th American Society of Health Economists Annual Conference, (Washington, 2021) and 14th International Health Economics Association (iHEA) Annual Conference (Cape Town, 2021).

Chapter 4 is co-authored with Noemi Kreif, Peter C. Smith, Rita Santos, Nicholas Stacey and Ijeoma Edoka. Previous versions of this chapter were presented and discussed at the 6th African Health Economics and Policy Association Conference (Kigali, 2022) and 3rd Spatial Health Econometrics Workshop (Venice, 2022).

I am the principal author for all chapters having defined the research questions, assembled the datasets, constructed the variables, defined the empirical model, undertaken the analysis, interpreted the results and written up the chapters. My co-authors advised on refinement of the research question, the empirical strategy, the interpretation of the results and were involved in editing the chapters.

I affirm that this thesis has not previously been presented for an award at this or any other university or educational institution. Any views expressed in this document are the exclusive responsibility of the author. All sources are acknowledged in the Bibliography.

Introduction

Strengthening health systems in low- and middle-income countries (LMICs) is a firm priority within global health and development. Specifically, there is a critical need to implement policies that increase access, raise utilisation and improve the quality of health care. It is estimated that over 400 million people globally lack access to at least one essential health care service (WHO, 2015). Furthermore, there are persistent socioeconomic disparities in service coverage. For example, in 2003 women in the richest quintile were 5.2 times more likely to have a supervised delivery than women in the poorest quintile (Gwatkin et al. 2003). Progress on tackling inequalities in health care access has been slow as utilisation of reproductive, maternal and new-born child health interventions continues to considerably increase with socioeconomic status (WHO, 2021). In addition to access issues and underutilisation, there continues to be major shortfalls in the quality of health care (Das et al. 2018). Without a basic standard of quality, improving access to and utilisation of health care will have little effect on health outcomes (Kruk et al. 2018). Estimates suggest 5.7-8.4 million deaths in LMICs, representing 10-15% of all deaths in these countries, are attributable to poor quality health care (National Academy of Sciences, Engineering, Medicine, 2018). Moreover, large variations in the quality of health care has also been cited as a factor exacerbating health care utilisation and outcome inequalities (Kruk et al. 2017). It is clear therefore, that there is an urgent need to remove remaining barriers in order to improve access to high-quality health care.

Since the onset of the Universal Health Coverage (UHC) agenda – broadly defined as ensuring everyone has access to the high-quality health care they need, without suffering undue financial hardship – a substantial effort has focused on health financing reforms aimed at increasing access to and the quality of health care in LMICs. In relation to access, there has been a growing push to ensure basic health care services are provided free at the point of utilisation. This has manifested in the widespread removal of user fees associated with publicly provided essential health care services (Powell-Jackson et al. 2014; Lagarde & Palmer 2008; Manthalu et al. 2016) or the provision of cash or vouchers (Lagarde et al. 2009; Van de Poel et al. 2014) among other schemes. The objective of these policies is to reduce the financial barriers to accessing health care, such that health care utilisation increases without compromising financial protection. Similarly, reforms relating to provider payment such as performance-based financing (PBF) are being implemented to improve access and quality (Basinga et al. 2011; Gertler & Vermeersch, 2012; Zeng et al. 2018; Chalkley et al. 2016; Witter et al. 2013). Therefore, a large focus has been on the responsiveness, of both consumers and providers, to financial incentives. While health financing policy is

undoubtedly central to past and future improvements in access to and the quality of health care in LMICs, non-financial health system policies also play a large role in determining these key components in moving towards UHC.

Three factors highlight why understanding non-financial determinants of health care access and quality, and the related potential policy prescriptions, are important. First, the policies of lowering the direct financial cost of accessing care and using monetary incentives to improve the quality of care have contributed to questions around the financial solvency and sustainability of health care systems in many LMICs (Gruber et al. 2014). Due to large informal sectors, government's ability to raise revenue has not kept pace with the desire to increase health service coverage and quality and improve social protection against health shocks (Banerjee et al. 2021). Second, there is mixed evidence on the effect of financial incentives on access to and the quality of health care. Theory and empirical evidence on the effect of reducing the monetary price of public health care on utilisation is modest, at least for curative care (Filmer et al. 2002). This does not suggest health financing reforms are unimportant and should not continue to be trialled and evaluated, nor undermine the potential for reducing the monetary price of health care to decrease out of pocket expenditure and improve financial risk protection. However, it does imply that other determinants of the demand for health care are also important. Similarly, evidence on the introduction of financial incentives as a mechanism to improve quality of health care provided in LMICs is far from conclusive (Binyaruka et al. 2020; Borghi et al. 2015; Paul et al. 2018), suggesting other factors warrant examination. Third, economic theory provides a well-developed understanding of the various determinants of the demand for health care (Grossman 1972; Gertler & van der Gaag, 1990) and the factors influencing public service provider behaviour (Benabou & Tirole, 2003; Besley & Ghatak, 2003; 2005; 2007; Delfgauuw & Dur, 2008; Besley & Burgess, 2002; Ferraz & Finan, 2008; Bloom et al. 2015) and specifically performance of health providers and quality of health care (De Geyndt, 1995; Leonard & Mæstad, 2016; Das & Hammer, 2014). From this literature a clear picture emerges on the important role of non-financial determinants of the demand for health care. However, there is a knowledge gap on potential policy-levers through which policy-makers can influence health-seeking behaviour and the performance of providers.

Specifically, robust econometric evidence on non-financial health system policies aimed at increasing health care access, utilisation and quality in LMICs are comparatively limited. The relative paucity of evidence may be methodologically driven, as many of these health system- and population-wide policies may be less amenable to randomised evaluations (Deaton & Cartwright,

2018). This makes the evaluation of such policies particularly susceptible to the relative lack of routine and administrative data and research capacity within LMICs. Whatever the cause, the above reaffirms the prospective importance of improving the evidence base on non-financial determinants of access and quality. This thesis examines the effect on health care access, utilisation and quality of a sample of non-financial policies of public health care systems in LMICs.

Chapters 1 and 2 examine non-financial health system policies which influence access to and the utilisation of obstetric health care in Malawi. Health-seeking behaviour can be influenced by constraints which affect individual's ability to utilise health care, and preferences which affect the willingness to utilise services (O'Donnell, 2007). A considerable body of work has examined the effect of demand-side financing interventions, providing monetary incentives to households and women to induce them to utilise obstetric health care (Witter, 2012; Ahmed & Khan, 2011; Lim et al. 2010; Powell-Jackson & Hanson, 2012; Grepin et al. 2019; Barber & Gertler, 2010; Gaarder et al. 2010; Kusuma et al. 2016). However, these strategies primarily address financial constraints – namely income and price – to accessing care. Furthermore, there is evidence of the utilisation of private health facilities for delivery services even in settings where public services are provided for free, suggesting that price may not be the only issue affecting utilisation (Silan et al. 2014). Specifically, Chapter 1 addresses the issue of geographical accessibility, examining how distance to a health facility effects the rate of institutional deliveries for pregnant women. Distance can act as a demand-side barrier to utilisation via physical accessibility issues and potential non-price costs associated with overcoming this, such as travel costs and foregone earnings. In rural areas in particular, distances to health facilities continues to impact health care utilisation. This can be exacerbated when transport options are limited and health conditions restrict mobility. While geographical accessibility is a well-known barrier to health care utilisation, there has been limited work on understanding how changes to geographical accessibility impacts health care utilisation. The majority of studies do not go beyond reporting associations between geographical accessibility and utilisation. Chapter 1 uses data from the Demographic Health Survey (DHS) and Service Provision Assessment (SPA) to identify household distances to the nearest health facility with delivery capacity. The chapter seeks to contribute to the existing evidence by addressing three methodological challenges: non-linear effects between distance and utilisation; unobserved heterogeneity through non-random distance 'assignment'; and heterogeneous effects of distance. Distance is considered as a continuous treatment variable with a Dose-Response Function of the distance-utilisation relationship estimated based on Generalised Propensity Scores. This allows for the exploration of non-linearities in the effect of an increment in distance at different distance

exposures. An instrumental variables approach is utilised to examine the potential for unobserved differences between women who reside at different distances to health facilities. The results suggest distance markedly reduces the probability of having a facility delivery, with some evidence of non-linearities in the effect. The negative relationship is shown to be particularly strong for women with poor health knowledge and lower socio-economic status. Additionally, there is evidence of potential unobserved confounding, suggesting that methods that ignore such confounding underestimate the true effect of distance on the utilisation of delivery services. The results can be combined to inform health infrastructure planning and other policy interventions which mitigate the effect of distance on health care utilisation.

Chapter 2 builds on this work, examining the effect of a specific policy aimed at reducing the constraint geographical accessibility poses, as well as potentially affecting women's preferences relating to the utilisation of services. Specifically, Chapter 2 examines whether opening Maternity Waiting Homes (MWHs) at health facilities increases the utilisation of obstetric health care services in Malawi. These structures enable women in late-stage pregnancy to wait at health facilities until delivery, thereby seeking to reduce some of the barriers faced by pregnant women in accessing care. MWHs are not a new policy in LMICs but, like many non-financial interventions aimed at improving access, have not been rigorously evaluated. The study exploits variation in the timing of MWH construction and opening at various health facilities to assess the impact on the utilisation of various maternal health care services and neonatal mortality. Again using the DHS and SPA, augmented with data collected on MWHs, we implement a difference-in-difference approach finding limited evidence that MWHs increase the utilisation of obstetric health care services. This result is unsurprising given the high utilisation rate for institutional delivery together with the relatively small sample of treated births. A conclusion from the results is that the MWH policy is unlikely to be a cost-effective use of resources.

Of course, the ultimate objective of increasing health care utilisation is to improve individual and population health. Therefore, focusing solely on issues related to the demand for health care without consideration of the quality of the health care supplied brings up two potential issues. First, there is little purpose in implementing policies aimed at increasing the demand for high-quality health care – sometimes referred to as effective health care – if it is not provided. The quality of health care provided is a strong determinant of demand and utilisation (Borah, 2006). Removing demand-side access constraints may not result in higher utilisation if the quality of services supplied do not align with population preferences. Quality deficiencies can result in a

number of sub-optimal health-seeking behaviours. In extreme cases it can cause individuals to forgo any form of health care utilisation. Alternatively, it may result in individuals utilising private care, or finally, it may result in bypassing (Leonard, 2014). The second issue of increasing the demand for care without consideration of the quality of health care supplied is, even if individuals utilise care, poor quality will moderate the health benefit, or worse there may be negative health effects. Das & Hammer (2014) note the prevailing evidence that although policies increasing the demand for health care may increase utilisation there are few studies showing any improvements in health outcomes. They go further by asking whether 'institutionalising births [is] institutionalising deaths?'. Andrew & Vera-Hernandez (2020) examine the impact of incentivising demand for institutional delivery in a supply-constrained setting, finding that the ultimate health effect strongly depends on the quality and capacity of health facilities. This clarifies the need, from a policy perspective, to simultaneously consider the demand for health care and the quality of health care, as it is the interaction of the two that determine health outcomes.

As noted, the primary policy-lever and much of the econometric work on improving the quality of health care supplied in LMICs has focused on financial incentives for providers, such as PBF schemes. This approach stems from the observation of a 'know-do' gap, whereby health workers and facilities are not providing health services as efficiently or effectively as they have demonstrated they are capable of (Mohanan et al. 2015; Leonard & Masatu, 2010; Das & Hammer, 2007). Therefore, the issue is frequently viewed as one of accountability or motivation rather than capacity constraints. However, financial incentives are not the only strategy to increase effort and support the delivery of high-quality care, and there is growing recognition that health workers are motivated by a range of factors (Lagarde et al. 2019; Ashraf et al. 2020). It is the combination of knowledge, equipment and effort which produce quality in health care (Leonard & Mæstad 2016). Therefore, depending on which input is primarily responsible for undermining quality, introducing or strengthening financial incentives may not be the appropriate policy response. Acknowledging this, chapters 3 and 4 examine non-financial health system policies in South Africa which might be leveraged to improve quality.

Chapter 3 examines the effect of the Ideal Clinic Realisation and Maintenance Programme (ICRMP). This system-wide supply-side quality improvement (QI) programme focuses on health care providers, primarily attempting to address capacity constraints – as opposed to a policy aiming at improving incentives for quality. The programme introduced a checklist, supportive supervision and funding, ensuring primary health care facilities have the foundational capacity in terms of

infrastructure, clear processes and equipment to provide quality care. The 'know-do' gap can be broken into the 'know-can' and 'can-do' gaps where 'can' reflects potential performance without capacity constraints and is primarily a function of health facility infrastructure and equipment (Ibnat et al. 2019)[1]. As opposed to focusing on the 'can-do' gap, which relates to motivation and effort, the ICRMP addresses constraints in capacity which might result in a 'know-can' gap. As noted, in addition to average low quality, high levels of quality variation is the second stylised fact about health care in LMICs. There is a concern that QI programmes and policies, while potentially improving quality on average, may not reduce inequities in access to high-quality care, or may further increase health inequalities. Despite this concern, few studies have examined the distributional impact of QI programmes. Therefore, chapter 3 specifically assesses whether the effects of the ICRMP are sensitive to previous quality performance. Variation in the timing of policy changes across facilities is exploited in order to implement a difference-in-difference-in-difference (DDD) approach to estimate treatment effects across subgroups defined by pre-treatment quality measures. Additionally, a changes-in-changes (CC) framework is employed to estimate the effect of the programme on quality across the distribution of past facility quality performance. The results suggest that while the programme improves quality measures for all facilities, the largest gains are realised by facilities with higher baseline quality. This finding is robust both across DDD and CC approaches, and a series of robustness checks that aim to account for possible endogenous selection of facilities in the programme. Therefore, this particular policy may have led to a worsening of pre-existing inequity in health care quality.

While the above has made the importance of identifying the primary source undermining quality clear, there is substantive evidence in many contexts that low motivation and effort are often responsible for low quality (Leonard & Masatu, 2010; Das & Hammer, 2007). A number of factors have a role to play in shrinking the 'can-do' gap. For instance, the 'can-do' gap is larger in public compared to private and non-profit health facilities and in facilities with more centralised management structures (Das & Hammer, 2007; Leonard et al. 2007). There is a well-developed literature regarding monitoring and public reporting increasing accountability and performance (Besley & Burgess, 2002; Ferraz & Finan, 2008), while various non-financial motivations and preferences of public sector workers have been acknowledged, specifically highlighting intrinsic motivation as a determinant of effort (Delfgauuw & Dur, 2008). Lagarde et al. (2019) highlight

[1] This terminology comes from the increasingly influential 'Three Gaps Model'. The principles can be pictured at an individual health worker level. A health worker with a significant 'know gap' has insufficient training, a health worker with a large 'know-can gap' has insufficient access to equipment and materials, and a health worker with a large 'can-do' gap has insufficient motivation.

policies targeting reputational concerns and intrinsic motivation as an alternative to financial incentives in LMIC health systems, and there is increasing evidence of its effect on the quality of health care in these settings (Leonard & Masatu, 2006; Leonard & Masatu, 2010). In fact, recognising these alternative determinants introduces the possibility that financial incentives can lead to reductions in quality (Ashraf et al. 2020).

Chapter 4 draws on these concepts examining whether strategic interactions between health facilities in South Africa have the potential to drive quality improvements, without the presence of financial incentives. A number of studies have examined the impact of provider competition and strategic interactions between health care providers on health care quality in high-income countries (Gravelle et al. 2014; Longo et al. 2017; Brekke et al. 2021; Moscelli et al. 2021). However, patient expectations of public health care services are very low in LMICs, making relying on demand-side pressures to stimulate quality improvements impracticable. Additionally, the responsiveness of health care demand with respect to quality reduces with socio-economic status, suggesting even if such a policy works it may worsen health inequalities (Lavy & Germain, 1994). Chapter 4 examines whether health facilities adapt their quality in response to changes in the quality of peer facilities, not based on the demand-side response and financial incentives, but on peer-to-peer comparisons and reputational concerns based on a type of intrinsic motivation. Using a national census of public primary health facilities, the study exploits data from the ICRMP on structural and process components of quality, examining how these measures change from 2015-2017. The study examines facilities' strategic interactions using both a spatial econometrics approach and a more traditional quasi-experimental approach exploiting the ICRMP QI programme as a source of exogeneous variation to estimate the response of facilities to changes in the quality of their peers. The results provide evidence of quality peer effects between primary health care facilities, with a 10-unit increase in average District facility quality causing facilities to increase their quality by 3.6 units. Given the lack of financial incentives, prosocial motivation and reputational concerns may be acting as the mechanism inducing facilities to respond to changes in peer quality. Importantly, these findings have significant policy implications suggesting the provision of relative performance information, allowing for peer comparisons, may induce a form of quality yardstick competition and be a credible quality improvement policy which may be considered alongside health financing reforms.

The thesis concludes with a discussion of the contribution of the studies outlined, including the significance of the findings for policy, and by presenting directions for future research.

Chapter 1

The Effect of Distance on Maternal Institutional Delivery Choice: Evidence from Malawi

1. Introduction

Low- and middle-income countries (LMICs) continue to face problems of underutilisation for basic health care (O'Donnell, 2007). The household cost of accessing health care can be broadly split into financial and time costs (Acton, 1973; Gertler & Van der Gaag, 1990). Consequently, countries pursue two broad policies to improve utilisation; reducing user fees, and setting geographical access policies. With an increasing number of LMICs removing pecuniary barriers to access, attention is switching to other determinants of health care utilisation, such as travel distance. Despite many LMICs improving the physical access of health care**,** travel time/distance is still frequently cited as a significant barrier (Tegegne et al. 2018; Karra et al. 2017; McLaren et al. 2014; Lohela et al. 2012; Hjortsberg, 2003).

Previous literature has identified a 'distance-decay rate', an inverse relationship between distance to health care and utilisation (Shannon et al. 1969; Lavy & Germain, 1994; Stock, 1983; Muller et al. 1998; Wong et al. 1987; Tanser et al. 2006; Sarma, 2009; Malqvist et al. 2010; Borah, 2006). The association between distance and health care utilisation has been found to be large relative to the effect of income, user fees and education (Buor, 2003; Thornton, 2008). Studies that examined the effect of travel time on health care utilization also found a negative relationship (Alegana et al. 2012; Blandford et al. 2012; Masters et al. 2013). It is well established that distance can influence health care seeking behaviours in expectant mothers (Thaddeus & Maine, 1994; Gabrysch & Campbell, 2009). Specifically, distance is found to be a significant determinant of having a facility delivery, even at small distances in circumstances with poor transport (Chowdhury et al., 2006; Yanagisawa et al., 2006; Nesbitt et al. 2016). In the few studies which found distance to have no impact on utilisation of delivery services (Duong et al., 2005; Paul & Rumsey, 2002), this might be explained with the relatively short average distances and high quality transport infrastructure in the settings evaluated. Some qualitative evidence has identified a contradictory effect of distance on utilisation of delivery services, with large distances stimulating women to seek facility deliveries due to the recognition of the impact of distance should complications arise during a home birth (Griffiths & Stephenson, 2001).

There is increasing recognition of the potential for unobserved confounding to bias estimates of the impact of distance on utilisation and health outcomes. Manang & Yamauchi (2018) exploit the opening of new facilities in a differences-in-differences design, and find that increased access, measured by number of local public health facilities, increases the probability of having a facility delivery. In a study that focuses on facility deliveries in India, Kumar et al. (2014) attempt to address the possibility that distance may be endogenously determined by instrumenting distance to health facility with an index capturing distance to non-health 'institutions of development'. Their instrumental variable (IV) estimates are five times larger than their ordinary least squares (OLS) estimates, with a one kilometre increase in distance resulting in between a 1.6-1.7 percentage point decrease in the probability of having a facility delivery. To the best of our knowledge, this represents the only current study examining the effect of distance on health care utilisation that attempts to account for potential endogeneity.

In this chapter, we investigate the impact of distance from the nearest health facility offering delivery services on the probability of having a facility delivery in rural Malawi. With a GDP per capita of US$338.50[2], Malawi is one of the poorest countries in the world (World Development Indicators, 2017). Unlike many LMICs, Malawi has not seen a significant change in urbanisation, with 84% of the population residing in rural households in 2018 (National Statistics Office, 2018). Further, there was in 2016 an average of 4.6 births per woman, one of the highest fertility rates in the world (World Development Indicators, 2019). Despite recent improvements, the country still suffers some of the highest rates of under-5 and neonatal mortality globally, 63/1,000 and 27/1,000 live births respectively in 2015/16 (Ministry of Health, 2017). The maternal mortality rate was 439/100,000 live births in 2016, a reduction from 675/100,000 in 2010 (Ministry of Health, 2017).

Like many LMICs, Malawi continues to aim to expand its health infrastructure and increase the physical access to health care services. However, despite the acknowledgement of a distance-decay effect, nuanced evidence to informatively guide infrastructure planning remains sparse. Distance is a continuous variable and populations are geographically distributed unevenly. Despite this, most studies treat the functional relationship between distance and utilisation as linear, limiting the potential to explore variation in the relationship across distance levels. As efficient investment in health infrastructure relies on understanding the full relationship between access and utilisation, it

[2] 2017 US dollars.

is important to identify differential impacts distance may have at different levels. Additionally, while many countries set minimum travel distance targets, it is unclear – equity arguments aside – whether such targets are appropriately set. Fundamentally, health systems have to provide a fair opportunity for populations to seek care. Therefore, from a health infrastructure planning perspective the effect of the minimum travel distance required to seek care on utilisation remains an important issue. Information on this relationship can guide minimum access thresholds and infrastructure planning. In order to provide more nuanced evidence to inform policy-making, we adopt a continuous treatment approach, estimating a dose-response function (DRF) that relates each level of distance to the probability of having a facility delivery using generalised propensity scores (Hirano & Imbens, 2004). We also examine the potential modifying effects household socio-economic status and mother's health knowledge may have on the distance-utilisation relationship. Additionally, we acknowledge that household and facility location may be strategically selected, and non-random sorting may result in distance to health facility being correlated with unobserved determinants of location of delivery such as health status or health-seeking preferences. To address this possibility, we employ an IV approach as an alternative estimation strategy. We view the methods applied as complementary, in that they address different challenges in estimating the distance-utilisation relationship: potential nonlinearity in the relationship, potential heterogeneity according to pre-specified subgroups, and potential unobserved confounding. In our conclusions, we synthesise the results from these methods to provide comprehensive information on how distance may effect health care utilisation and help inform policy decisions.

2. Methods

We briefly outline the contextual background and data used in the study before outlining a simple theoretical model of how distance impacts health care utilisation and the identification strategies applied.

2.1 Context

Primary health services in Malawi have been provided free at the point of access in public facilities since 1964. The services to which the population is entitled without user fees, including delivery services and maternal health care, were formalised with the introduction of the Essential Health Package (EHP) in 2004 (Ministry of Health, 2004). To further improve the utilization of maternal and child health services, service level agreements (SLAs) have been agreed with the Christian Health Association of Malawi (CHAM) since 2006. Under SLAs, CHAM facilities in catchment

areas where no government facilities exist provide the EHP without charging user fees (Manthalu, 2019).

The Government of Malawi owns 48% of the health facilities within the country while the CHAM owns 17% of the countries facilities, with most located in rural areas. The remaining facilities are either private-for-profit (22%), NGO owned (6%) or company facilities (7%) (Ministry of Health & ICF International, 2014).

The national access policy seeks to ensure that all households live within 5km of a health facility, reduced from a previous target of 8km (Ministry of Health, 2017). From 2011-16 twelve new health facilities were constructed. Despite this, the proportion of the population living within 8km of a health facility declined from 81% to 76% over the same period (Ministry of Health, 2017).

2.2 Data

Our analysis combines data primarily from two key sources. The Demographic and Health Survey (DHS) 2015/16, a population-based household survey, provides data on births and health care utilisation. It employs a two-stage cluster sampling design with 850 clusters identified in the first stage and approximately 30 households from each cluster selected in the second stage, resulting in a total sample population of 26,361 households. The survey provides self-reported birth histories for women up to 5 years before the survey.

We restrict analysis to rural households, as defined by the DHS. We focus on rural households because the determinants for health care utilisation likely differ between urban and rural women, based on systematic differences in their characteristics and environment and they should be treated as different sample populations. Furthermore, distance is unlikely to present a significant barrier for urban households. Births which took place prior to the woman residing in her current location were excluded as observed distances are not related to these births. Caesarean deliveries are excluded as they all take place in health facilities. Lastly, the analysis includes women only if the household they were surveyed in was their usual place of residence. The final analytical sample consists of 11,881 births to 9,250 women. See **Supplementary Appendix C1-1** for details on the construction of the analytical sample.

Facility data was obtained from the Service Provision Assessment (SPA) 2013/14, a census providing information on the availability and quality of health care services from all functioning

health facilities within Malawi. The survey captures the geographic coordinates of all facilities. Of Malawi's 977 facilities in 2013/14 only 540 had basic delivery capacity and 71 had capacity to perform caesarean sections (**Supplementary Appendix C1-2**).

The outcome of interest is mother-reported location of delivery, indicating whether a birth occurred at a health facility with delivery services or at a location without appropriate services, predominantly home births. We cannot identify the specific facility at which deliveries occur, and do not assume it to occur at the nearest facility. Consequently, we measure the impact of distance to the nearest health facility with delivery services on utilisation of any health facility with such capacity. In this sense, we examine the relationship between the minimum distance faced for the opportunity to access institutional delivery services and the fundamental decision of whether to utilise them or not. The DHS also provides information on reasons why women may not seek health care generally (**Table 1.1**). Distance is the second overall most cited problem in seeking health care. Unsurprisingly, a higher proportion of women residing at further distances cite distance as a significant problem.

Table 1.1: Self-reported reasons for not seeking health care (proportion citing reason)

	Full analysis sample	1st tertile (0.07-3.4km)	2nd tertile (3.4-5.8km)	3rd tertile (5.9-23.6km)
Permission	0.17	0.16	0.18	0.17
Financial	0.55	0.50	0.57	0.58
Distance to facility	0.61	0.47	0.65	0.72
Going alone	0.31	0.23	0.33	0.37
Concerned no female provider	0.26	0.21	0.28	0.28
Concerned no provider	0.53	0.50	0.53	0.56
Concerned no drugs	0.70	0.65	0.71	0.72

We generated the explanatory variable of interest, Euclidean distance to nearest health facility with delivery capacity, by spatially linking household clusters with health facilities using QGIS (QGIS Development Team, 2009). **Figure 1.1** shows the distribution of distance, with most women (62%) living within 6km of the nearest health facility offering delivery services and a mean distance of 4.98 kilometres. It should be noted that this represents the minimum distance women must travel in order to utilise an appropriate facility for delivery, not necessarily the facility in which the delivery ultimately took place.

The DHS geospatial data has geographic displacement procedures imposed to maintain respondent anonymity (Burgert et al. 2013). First, geographic coordinates are aggregated to a single point coordinate for each DHS cluster representing the cluster centroid. Second, a geo-masking

process displaces the aggregated cluster by a random-angle, random-distance process whereby 99% of rural clusters are uniformly displaced up to 5km and a further 1% are uniformly displaced up to 10km. Consequently, as geocoordinates are at DHS cluster centroids, distance is measured at the cluster-level. This results in the loss of within-cluster variation. Additionally, the displacement process introduces a random measurement error in the geographic coordinate data, biasing estimates towards zero (see **Supplementary Appendix C1-3**).

Figure 1.1: Distribution of distance

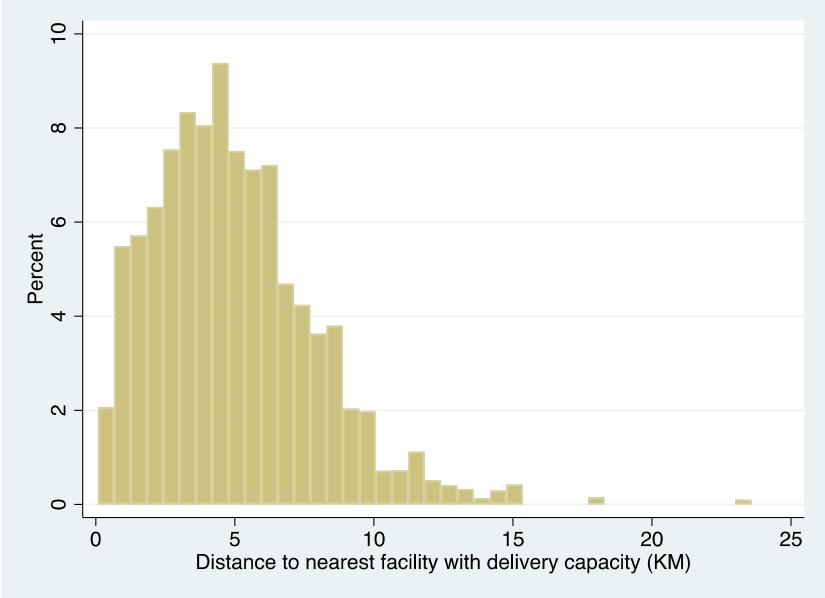


Table 1.2 provides summary statistics for the outcome and control variables for both the full analysis sample and across distance tertiles. On average, women in the sample had 1.5 births within the 5 year sample period. 92% of births take place at a health facility. Women who have a home birth on average reside further from a facility (6.2km) than those having facility deliveries (4.8km).

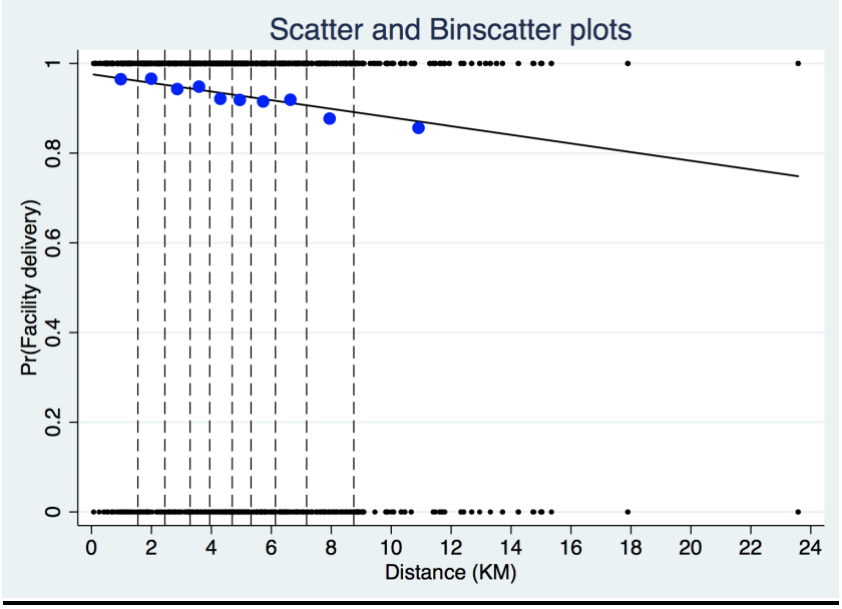
Table 1.2: Summary statistics

	Full analysis sample	1st tertile (0.07-3.4km)	2nd tertile (3.4-5.8km)	3rd tertile (5.9-23.6km)
Number of deliveries	11,881	3,907	4,062	3,912
Outcome variables				
Had facility delivery	0.92	0.96	0.92	0.89
Mother characteristics				
Age at delivery	26.5 (6.9)	26.3 (6.8)	26.6 (6.9)	26.7 (6.9)
Number of births in 5 years	1.5 (0.6)	1.5 (0.6)	1.5 (0.6)	1.5 (0.6)
Education level	1.0 (0.6)	1.1 (0.6)	1.0 (0.5)	1.0 (0.5)
Literacy	1.2 (0.9)	1.3 (0.9)	1.1 (0.9)	1.2 (0.9)
Health knowledge index	3.1 (0.8)	3.1 (0.8)	3.1 (0.8)	3.1 (0.8)
Gestation period	9.0 (0.5)	9.0 (0.5)	9.0 (0.5)	9.0 (0.5)
Frequency listen to radio*	0.7 (0.9)	0.8 (0.9)	0.7 (0.8)	0.7 (0.9)
Frequency read newspaper*	0.2 (0.5)	0.2 (0.5)	0.2 (0.5)	0.2 (0.5)
Frequency watch TV*	0.1 (0.5)	0.2 (0.6)	0.1 (0.4)	0.1 (0.4)
Years at current residents*φ	10 (7)	10 (7)	10 (7)	10 (7)
Have health insurance	0.00	0.00	0.00	0.00
Single child birth	0.96	0.97	0.95	0.97
Two years since previous birth	0.91	0.92	0.90	0.90
First child	0.21	0.23	0.21	0.20
Had Caesarean-section in last 5 years	0.01	0.01	0.01	0.01
HIV Positive*	0.07	0.08	0.08	0.07
Always lived at current residents	0.66	0.68	0.68	0.62
Child characteristics				
Birth year	2012.9 (1.4)	2012.9 (1.5)	2012.9 (1.4)	2013.0 (1.4)
Mother reported child birth size	2.8 (0.9)	2.8 (0.9)	2.8 (0.9)	2.8 (0.9)
Household characteristics				
Wealth index	2.8 (1.4)	2.9 (1.4)	2.7 (1.4)	2.7 (1.3)
Health care decision maker	2.2 (0.7)	2.1 (0.7)	2.2 (0.7)	2.2 (0.7)
Have bicycle	0.45	0.42	0.45	0.47
Have motorcycle	0.02	0.02	0.02	0.03
Have car	0.01	0.01	0.02	0.01
Female head of household	0.25	0.27	0.24	0.24
Environment characteristics				
Region	2.3 (0.7)	2.3 (0.8)	2.4 (0.7)	2.3 (0.8)
Rainy season birth	0.27	0.26	0.26	0.27
Number of HSAs in 1km	0.34 (0.57)	0.33 (0.56)	0.33 (0.56)	0.34 (0.57)
Facility characteristics				
Facility type	4.7 (1.0)	4.7 (1.0)	4.7 (1.1)	4.6 (1.0)
Managing authority	1.4 (0.8)	1.5 (1.0)	1.5 (0.9)	1.3 (0.6)
Other				
Bypassed nearest facility*	0.41	0.44	0.43	0.34

Notes: Variables marked with a * are not included in main specifications for inference for a variety of reasons but provide descriptive insight. Years at current residents is skewed upwards by the number representing 'always'. HSAs stands for Health Surveillance Assistant, Malawi's cadre name for community health workers. φ indicates variable is measured for sub-sample of individuals who have not always resided in the same location.

Figure 1.2 displays the observed locations of delivery, the mean outcome for 10 equally sized bins according to observed distances, and an unadjusted OLS line of distance on individual's probability of having a facility delivery[3], showing a clear negative association.

Figure 1.2: Non-parametric and parametric association of distance and facility delivery



2.3 Empirical strategy

2.3.1 Motivating model

Conventional models of demand view health care as an input in the health production function, influencing individuals' health (Fuchs, 1968; Grossman, 1972). The demand for health care is based on comparison of the relative costs and benefits of utilisation. Underutilisation is, therefore, a manifestation of the expected costs of utilisation exceeding perceived expected benefits. We outline a conceptual model of location of delivery choice based on mother and facility characteristics. We assume maternal utility is a function of what we call 'maternal capital', C, which can be thought of as wealth, and the outcome of the pregnancy, Y. That is, for woman $i = 1, \dots, N$:

$$U_i = U(C_i, Y_i)\ ;\ \frac{\partial U}{\partial C} \geq 0\ ;\frac{\partial U}{\partial Y} \geq 0 \qquad (1)$$

Capital is depleted by an amount dependent on the distance travelled to the location of delivery j, which includes all facilities with delivery capacity and home. Thus:

[3] 0 indicates a home birth and 1 a facility delivery. Binscatter plot produced from Cattaneo et al. (2019).

$$C_i^j = f\left(C_i^0, d_{ij}\right) ; \frac{\partial f}{\partial C_i^0} \geq 0 ; \frac{\partial f}{\partial d_{ij}} \leq 0 \quad (2)$$

where C_i^0 is the initial capital of mother i, C_i^j is the capital after delivery at location j and d_{ij} is the distance to location of delivery j. The outcome Y_i^j for mother i associated with delivery location j is assumed to be:

$$Y_i^j = g\left(Q^j, H_i^0, x_i\right) ; \frac{\partial g}{\partial Q} \geq 0 ; \frac{\partial g}{\partial H_i^0} \geq 0 \quad (3)$$

$g(.)$ can be thought of as the health production function of location of delivery j perceived by the mother, which depends on the quality of services, Q, the mother's underlying health, H^0, and a vector of individual characteristics x_i, which may include factors such as birth order, education level and health knowledge. By substituting (2) and (3) into (1), the utility V_i^j derived from delivery at location j can then be written as:

$$V_i^j = V\left(d_{ij}, Q^j, H_i^0, C_i^0, x_i\right) ; \frac{\partial V}{\partial d} \leq 0 ; \frac{\partial V}{\partial Q} \geq 0 \quad (4)$$

(4) is woman i's conditional indirect utility function for choice j which can be rewritten:

$$V_i^j = V\left(d_{ij}, Q^j, z_i\right) \quad (4.1)$$

where z_i is a vector of mother and household characteristics, including 'maternal capital' and health status. Thus we implicitly compare the utility associated with delivery at each feasible location, assuming women choose the location maximising V_i^j. Our dataset is at birth level, accordingly this optimisation problem occurs at each birth and woman i's strategy maximising utility may vary by birth. Empirical strategies and model specification choices derive from this simple theoretical model. Despite a set of j realisable utilities from alternative facility choices, we observe only the singular preference choice. The observed V_i^j is the revealed preference which can be formulated as a probabilistic choice model. Our first empirical strategies assume all elements of $V_i^j =$

$V(d_{ij}, Q^j, H_i^0, C_i^0, x_i)$ are captured, while our second acknowledges aspects of H_i^0 and x_i may be unobservable.

2.3.2 Identification strategy 1: Selection on Observables

Linear probability model (LPM)

We assume the linear predictor of the model to be additive separable and linear in its inputs, which relates to the conditional probability of the outcome with the link function $G(.)$:

$$\Pr(FD_B = 1) = G(\beta_0 + \beta_1 DIST_M + \beta_2 SES_M + \beta_3 HH_M + \beta_4 CB_B + \beta_5 FC) \quad (5)$$

where FD_B is a binary outcome indicating whether the birth occurred in a health facility or at home. $DIST_M$, the treatment, is distance to the nearest health facility from the woman's household. SES_M is a vector of socio-economic variables that may be related to women's choice of place of delivery such as age at delivery, education, literacy, health insurance, health knowledge and wealth index. HH_M includes household variables such as bicycle, motorcycle and car ownership. CB_B includes characteristics of the birth such as the gestation period, whether it was a single child birth and if it was the woman's first birth. Finally, FC includes facility characteristics such as the type of facility and the managing authority of the facility. A potential determinant of location of delivery is past exposure to the health care system, such as ante-natal care visits (ANC). However, ANC visits may themselves be affected by distance, and adjusting for it would cause bias in estimator of the treatment effect (Gelman & Hill, 2007), hence we do not include it as a covariate.

The causal interpretation of β_1 and the corresponding marginal effects hinge on two key conditions: exogeneity of distance and overlap. The former relies on including all confounders - variables that predict both distance and the location of delivery - in the regression model. Should a factor that influences a woman's decision to have a facility delivery (distance) not be observed, it must be assumed that this factor is independent of distance (choice of location of delivery). In the context of a continuous treatment such as distance, the latter condition, implies that for each level of treatment and combination of covariates, there is some non-zero probability that the treatment will be received (Cattaneo, 2010). The probability of poor overlap increases with the number of covariates adjusted for, and when the variable of interest is continuous, such as distance.

A problem with regression methods is that lack of overlap leads to a strong reliance of the specification of the regression model which extrapolates to regions with poor overlap (Ho et al.

2007). In our context, this involves correctly modelling the complex process that determines a woman's choice of birth location (Seljeskog et al. 2007). Specifically, correctly modelling the relationship between the distance and the outcome, as well as the relationship between the covariates and the outcome, through a linear predictor and the $G(.)$ link function. We first implement this regression approach, with main terms only in the linear predictor, and alternative (linear, probit and logit) link functions. Next, as a more flexible method, we implement a generalised propensity score approach, which moves the model specification task from the outcome regression function to the treatment assignment mechanism.

DRF Estimation

To relax the parametric assumptions of the regression function and flexibly explore the relationship between distance and the probability of having a facility delivery we treat distance as a continuous 'treatment', and estimate a DRF. We follow the Generalised Propensity Score (GPS) approach by Hirano and Imbens (2004)[4]. The attractive features of the GPS method relative to regression methods are that it requires only adjusting for a scalar variable to control for imbalance in observed covariates, and – in the more recent proposals implemented here – it is possible to use the GPS as a weight, without having to specify a parametric relationship between the treatment variable and the outcome.

We briefly outline the GPS method within the potential outcomes framework (Rubin, 1974). Given a random sample of individuals $i = 1,2, \dots, N$, for each unit i there exists a set of individual potential outcomes $Y_i(j)$ capturing i's response to treatment level j, known as the individual DRF. $j \in \Im$ denotes the treatment level – distance to nearest facility – where $\Im$ is the interval $[j_{min}, j_{max}]$, in this case $\Im$ is the interval $[0.07km, 23.58km]$. For every individual only one treatment J_i and one potential outcome is observed, $Y_i = Y_i(J)$. The causal effect of individual i moving from j to Δj is defined as, $Y_i(j) - Y_i(j + \Delta j)$ is unobservable. However, an estimate of the population average effect $E[Y_i(\Delta j)] - E[Y_i(j)]$ can be obtained. Calculated over the range of values $\Im$ this is known as the DRF, given as $\mu(j) = E[Y_i(j)]$ for all $j \in \Im$, measuring the relationship between the treatment, distance from the nearest health facility as the cause, and potential utilisation outcomes as the effect. The DRF, therefore, signifies the average response in the population if all women were at distance $J = j$. The marginal treatment effect estimation with respect to treatment level j is given as: $ATE = E[Y_i(j) - Y_i(j - \Delta j)] = \frac{E[Y_i(j)] - E[Y_i(j - \Delta j)]}{\Delta j}$.

[4] See also Kluve et al. (2012), Flores & Mitnik (2013), Egger & Ehrlich (2013) and Krief et al. (2015).

The approach relies on the weak unconfoundedness assumption, stating that, conditioned on the observed covariates, there is pairwise independence of the treatment level received with each of the potential outcomes : $Y_i(j) \perp J_i \mid X_i$ for all $j \in \Im$. (Hirano & Imbens, 2004). In our setting, this implies that distance to nearest facility is unrelated to unobserved covariates that themselves affect the probability of facility utilisation. For adjustment, we use the same covariates we used in the regression adjustment, specified in the previous section. Similar to the regression approach, identification with GPS relies on good overlap. This requires that the conditional density of the treatment is positive for any covariate values, $\Pr(r(j,x) > 0) = 1$, where $r(j,x) = f_{J|X}(j|x)$, is the conditional density function of the treatment given the covariates. However, an advantage of the GPS method is it enables a relatively straight-forward process, outlined below, of identifying women for whom it is difficult to construct counterfactual outcomes, allowing estimation to be restricted to comparable individuals.

The GPS is defined as $r(J,X)$, the probability that individual i belongs to the distance at which they are observed. Assuming the conditional distribution of treatment has been correctly specified, the GPS has a balancing property: within strata with the same value of the GPS evaluated at a given treatment level, $r(j,X)$, the probability that the treatment received equals this treatment level, $J = j$, does not depend on the values of the covariates.

Informed by statistical tests (Akaike information criteria (AIC), Bayesian information criteria (BIC) and Modified Park tests), we estimate the GPS using GLM with Gamma distribution and log link function for the conditional distribution of distance. Following Hirano and Imbens (2004), we perform balance checks on the estimated GPS, dividing the sample into three mutually exclusive intervals according to the 33rd and 66th percentile of the distribution. Within each interval the GPS is computed at the median distance. Each interval is divided into 5 blocks by the quintiles of the GPS evaluated at the median. Within each block, covariates difference in means are calculated for individuals who have a GPS such that they belong to that block but belong to a different treatment interval. T-statistics are used to assess the differences in the GPS-weighted means between each treatment interval and the pooled means of the remaining two intervals.

To estimate the DRF without the need to specify an outcome regression as a function of the GPS (Bia et al. 2011; Bia & Mattei, 2012) we implement a non-parametric inverse-weighting (IW) estimator (Flores et al. 2012). The approach corresponds to implementing a local linear regression

of the outcome on the treatment levels, using a global bandwidth that is chosen data-adaptively (Fan et al. 1996). More detail on the IW estimation approach is outlined in the **Supplementary Appendix C1-4**.

We restrict estimation to areas of common support with respect to the estimated GPS, using a method proposed by Flores (2007): again we split the treatment at the 33rd and 66th percentile, and evaluate the GPS at the median treatment of each group for the whole sample, we then compare the distribution of the GPS for observations that belong to one group versus the other two groups pooled, doing this for all three groups[5].

Heterogeneity analysis

Studies indicate that health knowledge and socio-economic status (SES) affect engagement with health systems, particularly in LMICs (Budhathoki et al. 2017; van Doorslaer & Masseria 2004; Wagstaff & van Doorslaer, 2000). The effect of distance may also vary along these dimensions. Households with higher SES may mitigate the impact of distance with their ability to pay for public transport. Greater health knowledge may reduce the disincentive effect of distance through better awareness of the benefits of health care. We generate a measure of mother's health knowledge through the cumulative score of a set of questions including whether the mother has heard of oral rehydration solution, Tuberculosis and natural birth complications and whether they know that HIV is spread by sexual activity. DHS rural-specific wealth index is used creating wealth quintiles. We adapt equation (5) to estimate the average marginal effects of distance across these subgroups. We undertake sub-group analysis using a regression approach due to the larger sample requirements of the GPS framework.

2.3.3 Identification strategy 2: Selection on Unobservables

If mother- or community-specific unobservable characteristics are correlated with distance to health facility and health care utilisation, this would bias the estimated effect of distance on utilisation (Schultz, 2004). Several mechanisms may result in such a scenario arising. Selective migration may lead individuals with stronger need or preferences for health care to relocate to communities with better access to health facilities. HIV positive individuals in rural Malawi have over two times greater odds of migration than HIV negative individuals (Anglewicz et al. 2016). Furthermore, placement of facilities may be influenced by lobbying from local communities or

[5] DRF analyses are implemented using the STATA packages *gpsscore*, *doseresponse* and *drf* by Bia et al. (2008) and Bia et al. (2014).

other political pressures (Todd, 2007). Rosenzweig & Wolpin (1986) showed endogeneity in development programme placement can be a source of significant bias. Such targeting has been identified for reproductive health services (Schultz, 2005; Strupat, 2017). The number of health facilities has expanded in Malawi, with 575 facilities in 2003, rising to 606 in 2010 (Ministry of Health 2010). Should facility placement be linked to health care demand, this would violate the assumption of exogeneity of distance. Finally, facilities may be located in areas of higher population density. Such areas may suffer higher rates of communicable disease or other risk factors – the HIV prevalence rate was 17.4% in urban settings compared to 8.9% in rural areas in 2010 (National Statistics Office. Malawi Demographic and Health Survey 2010/2011) – resulting in lower health status and increased need for health care utilisation. It has been noted generally that the argument for exogeneity is weak in cases where distance is not fixed, but responsive to incentives (Basker, 2007).

To attenuate concerns about potential endogeneity, we also employ an IV approach. The candidate IVs for distance must meet the standard conditions; (1) they have a strong correlation with distance to health facility, and (2) they do not have any effect on location of delivery, other than through their relationship with distance to health facility. Several studies have used instruments based on distance to other infrastructure when concerned about the endogeneity of distance to a specific service (Lavy, 1996; Mukhopadhyay & Sahoo, 2016; Kumar et al. 2014). Kumar et al. (2014) use distance to 'non-health institutions of development'[6] as an instrument for distance to health facility. Following this approach we first use (a) distance to nearest school and (b) distance of nearest school to closest trading centre, as instruments.

However, distance to other types of institutions may be correlated with community-level variables, which in turn, may influence the demand for health care. Lavy et al. (1996) suggests a community's local infrastructure may measure the degree to which the village leadership supports public service provision and, generally, the degree of community 'progressivity'. Therefore, institutions, health-related or otherwise, may be subject to the same non-random sorting, bringing into question the credibility of distance to other infrastructure as a valid instrument. Hence, we identify two instruments that we believe more appropriate than distance from other infrastructure: (c) number of qualified teachers at nearest school (d) number of students at nearest school.

[6] Non-health institutions of development include towns, district headquarters, railway stations and bus stops.

Local context underpins our rationale for why these instruments (c) and (d) are potentially less likely to suffer from violations of the exclusion restriction. We contend that local institution quality – proxied by number of qualified teachers – is less under the direct control of communities than access, and should be less related to unobserved community heterogeneity which may also influence distance to health facility. In Malawi, remote schools often face difficulty recruiting and retaining qualified teachers. Teachers themselves are important decision makers, using formal and informal channels to influence placement to avoid remote schools, where there are also high teacher attrition rates (Asim et al. 2017). Therefore, while 'progressive' communities may increase public service delivery in their locality, they may have less influence on the quality of those services, in this case number of qualified teachers. Hence, while number of qualified teachers captures an aspect of remoteness, this aspect is less directly related to the factors that jointly determine distance to health facility/public infrastructure and health care utilisation. Likewise, the number of students at the nearest school will generally capture 'remoteness' of households with schools in more remote communities having less students. However, as households may be located close to or far from the nearest school, we expect it to be a measure of 'remoteness' less related to preferences than distance measures. Because the household may in fact be located close to or far from the nearest school (and health facility) we expect that this instrument captures a measure of distance isolated from preferences which may also drive utilisation. Additionally, we contend these instruments are unrelated to selective migration due to the high informational requirements and the localised nature of migration in Malawi (Anglewicz et al. 2017).

We extract data on the instruments from a comprehensive World Bank dataset on teachers and schools. This dataset included information on the placement of all teachers in Malawi's 5,700 schools, linked with data on school facilities and locations and geo-spatial coordinates of commercial centres.

We implement several alternative IV approaches: 2SLS, IVprobit and two-stage residual inclusion (2SRI). The latter, 2SRI approach is our preferred specification, due to its relatively good performance in non-linear models (Terza et al. 2007; Terza, 2018; Wooldridge, 2014). In all specifications, we use the following linear model to estimate the first stage:

$$DIST_M = \beta_0 + \beta_1 Z + \beta_2 SES_M + \beta_3 HH_M + \beta_4 CB_B + \beta_5 FC + \varepsilon_B \qquad (6)$$

where Z is a vector of each pair of the above specified instruments. For the 2SRI method, we use the predicted residuals ($\hat{\varepsilon}_B$), and include them in the regression of distance on place of delivery in addition to the original endogenous variable using the following regression model:

$$\Pr(FD = 1) = \Phi(\beta_0 + \beta_1 DIST_M + \beta_2 SES_M + \beta_3 HH_M + \beta_4 CB_B + \beta_5 FC + \beta_6 \widehat{\varepsilon_B^{2SRI}}) \quad (7)$$

where Φ is a probit link function. The β_1 coefficient and corresponding marginal effect in the second-stage equation reflects the causal effect of distance. Following Hausman (1978), we test the coefficient of the first stage residuals to test for the presence of endogeneity. We obtain corrected standard errors in the second stage via bootstrapping.

2.4 Robustness checks

In addition to varying specifications of our primary models, we undertake robustness checks examining a different measure of distance and a more detailed metric capturing facility quality. Euclidean distance may not always best represent the realistic travel distance women travel in order to reach the nearest health facility (Guagliardo, 2004; Nesbitt et al. 2016). Therefore, we also calculate the road-network distance to test the robustness of the results to alternative distance measures.

Poor service quality or perceptions of quality also play a role in utilisation decisions (Mwabu et al. 1993; Macarayan et al. 2018). Further, quality may be associated with distance, with more remote facilities being of worse quality on average, which could inflate the estimated impact of distance if not adequately controlled for. Our baseline specifications included facility type and ownership which are frequently used proxies for facility quality. However, the SPA captures more detailed facility information allowing us to better control for potential variation in relevant aspects of facility quality.

3. Results

3.1 LPM results

Estimation of equation (5) by OLS shows a significant inverse relationship between distance and the probability of having a facility delivery (**Table 1.3**). Comparing the unadjusted correlations with those including the full set of controls the relationship remains almost unchanged. This can be explained by the small and insignificant coefficients of the covariates in the outcome model

with few variables strong predictors of location of delivery. A kilometre increase in distance to nearest facility reduces the probability of having a facility delivery on average by between 1.1-1.3 percentage points significant at the 1% level.

Table 1.3: Regression of location of delivery (home vs. facility)

	LPM (1)	LPM (2)	LPM (3)
	Coeff	Coeff	Coeff
Distance to nearest relevant facility	-0.011***	-0.013***	-0.013***
	-0.002	-0.002	-0.002
Number of Observations	11,881	11,375	11,375
Number of Clusters	677	676	676
Mother, household, environment controls		x	x
Facility characteristics		x	x
Birth year trend			x
District fixed effects		x	x

Notes: Controls refers to the inclusion of the full set of mother, child, household, environment controls as well as regional fixed effects. Average marginal effects are calculated following probit and logit specifications.

Results remain largely unchanged when using non-linear model specifications and when run on several relevant sub-samples (**Supplementary Appendix C1-5**).

3.2 DRF results

Table 1.4 shows the improvement in balance of the covariates once they have been adjusted by the calculated GPS using the procedure outlined above. There is an initial lack of balance with 31 of 66 t-statistics greater than |1.96|, indicating significant differences in means between treatment intervals. After adjusting for the GPS this is reduced to 22 of 66. Further, balancing for the GPS causes the average absolute t-statistic to decrease from 2.63 to 1.89[7]. Although not achieving perfect balance, adjusting for the GPS does improve comparability across distance.

Assessment of overlap in the distribution of the estimated GPS among individuals in the three treatment tertiles shows a high level of common support (**Supplementary Appendix C1-6**). We restrict DRF estimation to observations within the common support for all three treatment groups simultaneously[8].

[7] We also perform a likelihood-ratio test to check balance reaching similar conclusions. Table available upon request.
[8] This drops only 16 observations.

Table 1.4: Balance given GPS: t-statistics for equality of means						
	Unadjusted			Adjusted for GPS		
Variable	**1st tertile (0.07-3.4km)**	**2nd tertile (3.4-5.8km)**	**3rd tertile (5.9-23.6km)**	**1st tertile (0.07-3.4km)**	**2nd tertile (3.4-5.8km)**	**3rd tertile (5.9-23.6km)**
Mother characteristics						
Mother age at delivery	**2.1**	-0.3	-1.8	1.5	-0.6	-0.7
No. of births in last 5 years	**6.2**	**-2.9**	**-3.2**	**2.7**	**-4.9**	**2.3**
Education level	**-7.3**	**2.2**	**5.1**	**-2.1**	**4.0**	-1.2
Literacy	**-5.9**	**3.2**	**2.8**	**-2.6**	**4.5**	-1.3
Mother health knowledge	**-4.2**	**2.3**	1.9	-1.6	**3.3**	-1.4
Gestation period	-1.3	0.8	0.5	-1.6	0.7	1.0
Health insurance	0.1	-0.3	0.2	0.4	-0.3	0.4
Single child birth	-1.2	**2.8**	-1.6	-1.1	**2.8**	-1.5
Two years since previous birth	**-3.9**	0.7	**3.2**	-1.8	**2.3**	-0.7
First child	**-3.1**	0.7	**2.4**	-0.9	1.3	-0.4
C-section in last 5 years	-0.8	1.5	-0.6	-1.1	1.4	-0.4
Child characteristics						
Birth year	0.3	0.9	-1.2	0.0	0.6	-0.7
Mother reported child birth size	-0.8	1.6	-0.8	-1.2	1.5	0.0
Household characteristics						
Wealth index	**-8.4**	**4.0**	**4.4**	**-3.7**	**5.3**	-0.8
Bicycle	**3.9**	-0.2	**-3.7**	1.1	-1.4	-0.4
Motorcycle	1.5	-0.2	-1.3	0.9	-0.5	-0.5
Car	0.8	**-2.4**	1.5	1.2	**-2.3**	1.1
Female headed household	**-3.5**	1.3	**2.2**	-1.2	**2.2**	-0.9
Environment characteristics						
Region	**2.5**	**-10.2**	**7.6**	**6.1**	**-9.1**	**2.5**
Rainy season birth	-0.1	1.8	-1.7	-0.6	1.6	-1.1
Facility characteristics						
Facility type	0.5	-1.8	1.3	-0.8	**-2.7**	**3.3**
Managing authority	**-8.1**	**-4.1**	**12.2**	**-2.6**	**-4.4**	**8.2**

We estimate the effect of distance at values between 1-20km, as small samples at the extreme distances prevent meaningful estimates being obtained at the largest distances observed. We present estimates with 95% confidence bands obtained by 1,000 bootstrap replications. While the confidence bands grow after 10km, from the DRF it is clear the probability of having a facility delivery is a negative function of distance, with a 96.9% probability of facility delivery at 1km, falling to 74.1% at 20km (**Figure 1.3**). We find clear non-linearities in the marginal effect of a 1km increment in distance on the probability of having a facility delivery at different distance exposures (**Figure 1.4**): the estimated treatment effect of an additional kilometre appears, largely, to increase with the level of distance. A movement from 1km to 2km leads to a fall of 0.9 percentage points in the probability of having a facility delivery, while moving from 14km to 15km, the furthest distance for which we have a statistically significant effect, causes the probability to fall by 2.9

percentage points. **Appendix Table 1.6** in **Supplementary Appendix C1-7** shows estimates and confidence bands relating to figures.

To check the robustness of our estimates we also follow Hirano and Imbens (2004) in estimating the DRF. This approach estimates the conditional mean of the outcome given the observed treatment level and the probability of receiving that value by parametrically fitting a linear regression function on the treatment and the estimated GPS: $E[Y_i\,|J_i, \widehat{GPS}_i] = \beta_0 + \beta_1 J_i + \beta_2 \widehat{GPS}_i$ (**Supplementary Appendix C1-8**). All specifications gave results similar to our IW estimates.

Figure 1.3: Dose-response function

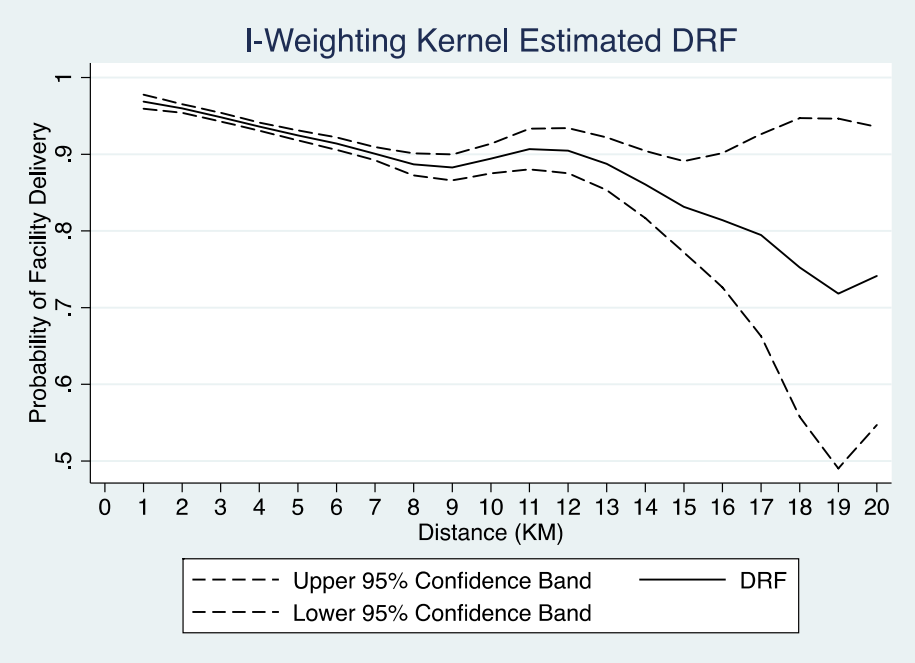
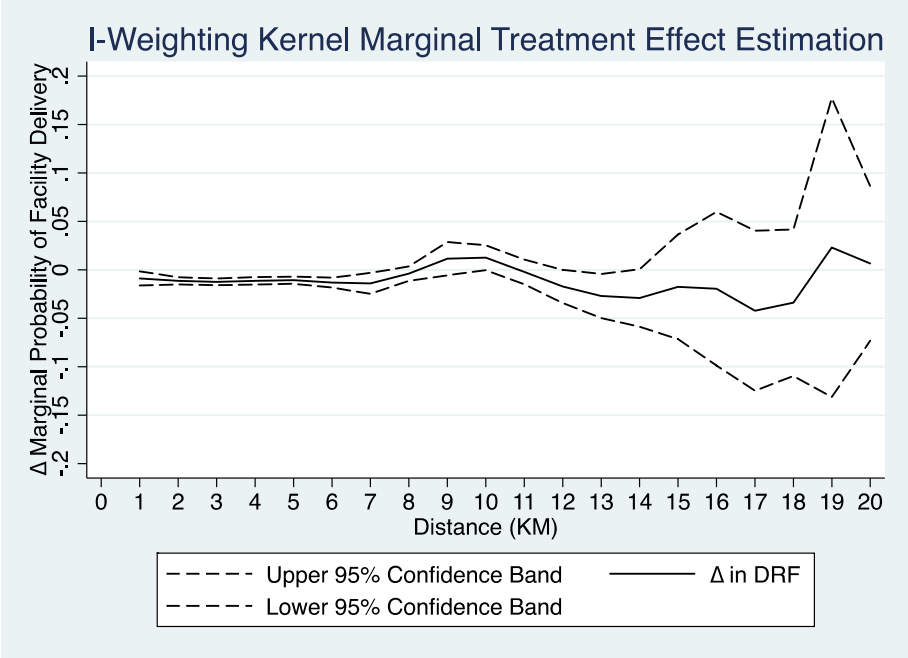


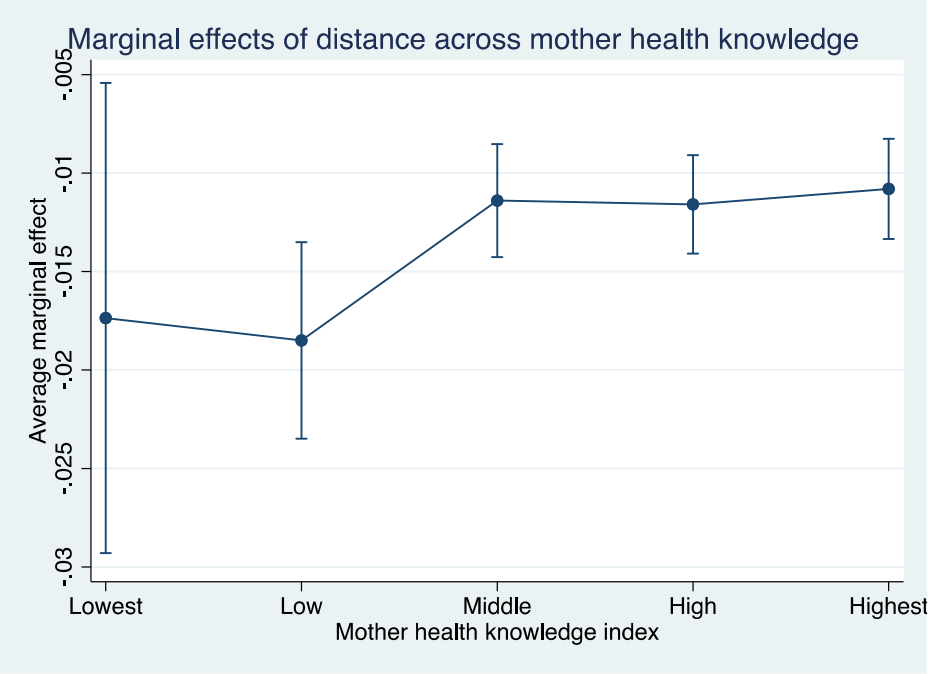
Figure 1.4: Derivative of dose-response function



3.3 Heterogeneity analysis results

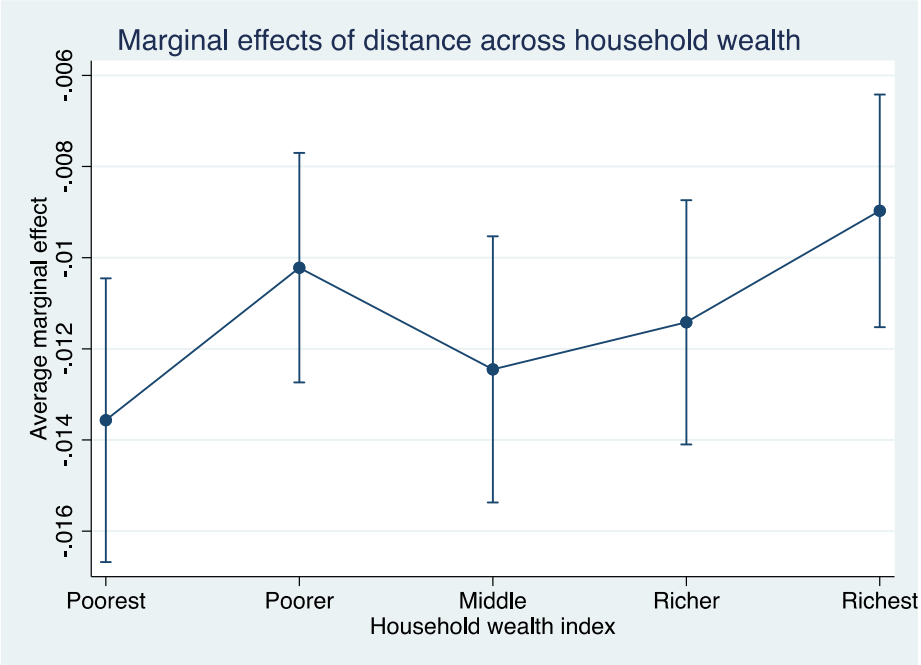
While distance has a significant negative effect on the probability of having a facility delivery across all levels of health knowledge, as expected, there is evidence that distance has a smaller reductive effect on the probability of having a facility delivery for women with higher health knowledge (**Figure 1.5**). There appears to be a threshold level of health knowledge at which the modifying effect disappears.

Figure 1.5: Heterogeneous effect of distance across levels of mother health knowledge



Similarly, the effect of distance is largest for the poorest households and smallest for the richest (**Figure 1.6**).

Figure 1.6: Heterogeneous effect of distance across levels of household wealth



3.4 IV results

All proposed instruments have statistically significant predictive effects on distance to health facility (**Table 1.5**). These first-stage partial correlations are in line with anticipated effects of the excluded instruments on distance to health facility. The Kleinbergen-Paap rk Wald F-statistics exceed the instrument relevance rule of thumb >10 for all instruments individually. Tests checking the endogeneity of distance validating the IV approach and tests for the exogeneity of the proposed instruments further validate the IV approach and instruments (**Supplementary Appendix C1-9**).

Table 1.5: First-stage LPM-IV results

	IV: Distance to nearest school	IV: Number of qualified teachers at nearest school	IV: Number of students in nearest school	IV: Distance of nearest school to trading centre	IV: All
	Linear IV model	Linear IV model	Linear IV model	Linear IV model	Linear IV model
First-stage					
Distance to nearest school	0.057***	-	-	-	0.042***
	0.017				0.016
Number of qualified teachers at nearest school	-	-0.082***	-	-	-0.005
		0.012			0.022
Number of students in nearest school	-	-	-0.001***	-	-0.001*
			0.00		0.00
Distance of nearest school to trading centre	-	-	-	1.666***	0.257***
				0.276	0.042
Number of Observations	11,375	11,375	11,375	11,375	11,375
Number of Clusters	676	676	676	676	676
Mother, household, environment controls	x	x	x	x	x
Facility controls	x	x	x	x	x
Birth year trend	x	x	x	x	x
District fixed effects	x	x	x	x	x
Post-estimation results					
Kleibergen-Paap rk Wald F statistic	11.3	34.2	32.2	36.6	23.74
Prob > K-P F	0.00	0.00	0.00	0.00	0.00
Effective F statistic (Montiel Olea & Pflueger, 2013)	11.3 < 37.4	34.2 < 37.4	32.2 < 37.4	36.6 < 37.4	25.3 > 23.6
Partial R-squared	0.025	0.049	0.044	0.099	0.177

Notes: Robust standard errors clustered as the DHS cluster level in parentheses. Statistical significance levels are denoted by *** p<0.01, ** p<0.05 and * p<0.10. All model specifications include the full set of covariates as in specification 3 of Table 5. All effective F-stats are compared to the critical value given for 5% of the 'worst-case' bias and for 5% significance levels. Both first- and second-stage equations include the full set of control covariates. In specifications with one instrument the effective F-stat collapses to the K-P F-stat.

Table 1.6 shows the results of our IV estimates. Using the distance-based instruments, the estimates suggest the impact of distance to be similar to those estimated using the LPM/GLMs. However, using our preferred instrument specification, it appears previous estimates underestimate the effect. The LPM-IV and 2SRI estimates suggest a kilometre increase in distance reduces the probability of having a facility delivery by between 2.3-2.5 percentage points.

Table 1.6: Instrumental variable estimation results

	IV: Distance to nearest school & Distance of nearest school to nearest trading centre			IV: Number of students at nearest school & number of qualified teachers at nearest school		
	LPM-IV	IVprobit	2SRI (Raw residuals)	LPM-IV	IVprobit	2SRI (Raw residuals)
First-stage						
IV: Distance to nearest school	0.041*** -0.016	0.041*** -0.016	0.041*** -0.016	-	-	-
IV: Number of qualified teachers at nearest school	-	-	-	-0.057** -0.023	-0.055** -0.023	-0.057** -0.023
IV: Number of students in nearest school	-	-	-	-0.000 0.000	-0.000 0.000	-0.000 0.000
IV: Distance of nearest school to trading centre	0.283*** -0.042	0.282*** -0.041	0.283*** -0.041	-	-	-
Second-stage						
Distance to nearest relevant facility	-0.012*** 0.004	-0.012*** 0.003	-0.012*** 0.002	-0.025*** -0.006	-0.011*** -0.006	-0.023*** -0.005
Number of Observations	11,375	11,456	11,375	11,375	11,456	11,375
Number of Clusters	676	675	676	676	675	676
Mother, household, environment controls	x	x	x	x	x	x
Facility characteristics	x	x	x	x	x	x
Birth year trend	x	x	x	x	x	x
Region fixed effects	x	x	x	x	x	x

Notes: 2SRI first-stages estimated by GLM and second-stage estimated via probit. AMEs for Ivprobit and 2SRI are reported. Standard errors are bootstrapped for 2SRI estimates to account for estimated residuals included in second-stages. Second-stage reported estimates are calculated average marginal effects and represent changes in the probability of having a facility delivery from a unit increase in distance.

When including residuals from the first-stage equation in the second-stage, the positive sign suggests individuals who reside at distances further from facilities have unobservable characteristics which increase the probability of having a facility delivery.

3.5 Robustness check results

The effect of a change in road-network distance on the probability of having a facility delivery is smaller than that of Euclidean distance (**Supplementary Appendix C1-10**). That the size of the relationship between road-network distance and the utilisation of facility delivery services is smaller than that of Euclidean distance suggests the former may be less representative of true travel distances. This is highly possible due to the number of informal paths and roads and travel habits

in Malawi. Therefore, we view this as a validation of the use of Euclidian distance as the preferable measure of travel distance.

We use principal component analysis to create a composite quality index and categorise facilities by level of quality. We then divide the sample according to the quality category of the nearest health facility offering delivery services (low, medium, high). Our results are unchanged when controlling for the quality of the nearest facility with this more detailed measure of quality (see **Supplementary Appendix C1-11**).

4. Discussion

This chapter examines the effect of geographical access to health care, measured by distance to nearest health facility, on the utilisation of delivery services. The results suggest that distance to health care still represents a significant constraint to utilising maternal health care services in Malawi. Our findings go beyond previous studies which have also identified similar negative distance-utilisation relationships. Unlike previous models which have parametrically constrained the impact of distance, our DRF estimates allow full exploration of heterogeneity in the effect of distance on facility delivery along the values of distance observed. The expected probability of having a facility delivery falls with distance: from 96.9% at 1km, to 74.1% for women residing at 20km from their nearest facility. We find non-linearities, with the marginal effect of distance increasing at greater distance levels. Ceteris paribus, this suggests targeting health infrastructure development towards households at marginally further distances will result in larger utilisation gains. Such information can be combined with population distribution data to significantly improve evidence-based infrastructure planning. Additionally, we find that distance has a more adverse effect on the utilisation rates of women with lower levels of health knowledge and household wealth. Hence, even in circumstances where the direct money price of health care is zero, such as in our setting, heterogeneous responses to distance may maintain inequities in access. This suggests that population characteristics should also be considered in both infrastructure development and attempts to increase utilisation rates and improve equity. Finally, we find the estimated impact of distance on the probability of facility delivery substantially increases when accounting for the endogeneity of distance, with estimates over twice as large as when not accounting for unobserved differences in women across distance. The finding that distance is endogenous with respect to utilisation suggests the results from the LPM and DRF models should be interpreted as lower bounds of the impact of distance.

Although the models utilised should be considered as complementary, the DRF is our preferred model specification due to the valuable information provided beyond an average effect, which can be readily used to guide health infrastructure development. The findings suggest that health infrastructure policy may benefit from considering factors beyond counts of populations within distance thresholds. Non-linearities in the effect of distance should be considered in facility openings/closings, and can be used with population distribution data to model potential utilisation impacts. Information on population background characteristics can also be used to target health infrastructure towards localities with populations for whom distance has a relatively larger reductive effect on utilisation. Considering such information in health infrastructure development could not only result in greater utilisation rate improvements than targeting solely based on population levels, but could also reduce within country utilisation inequalities.

Understanding of the mechanisms through which distance reduces utilisation is also of significant importance for designing effective policy to mitigate the impact. Policy responses to increase utilisation rates may either be supply-side, aimed at increasing the availability and quality of maternal health care services, or demand-side, aimed at increasing individuals' demand for services. A range of policies including travel vouchers (Ommeh et al. 2019), improving referral transport (Samai & Sengeh, 1997) and cash on delivery (Grepin et al. 2019) have been trialled in various LMICs to increase maternal health care utilisation. In Malawi a presidential initiative stated an intention to build 130 maternity waiting homes in Malawi (Presidential Initiative on Safe Motherhood, 2012). When the physical obstacle of distance is the primary disincentive to seek care, policy should focus on expanding access or improving transport. However, in circumstances where other factors, such as a baseline preference for a home delivery, are important, such policies might have less impact[9]. The hypothesised mechanisms behind the unobserved confounding and the higher utilisation rates in women with greater health knowledge suggests the disincentive effect of the physical obstacle of distance may not be the primary issue. This suggests alternatives to increasing the physical accessibility of health facilities may be effective in reducing the impact of distance. General improvements in health information or facility quality should increase utilisation among women not currently seeking facility deliveries. Therefore, although increasing health infrastructure will increase utilisation rates, Malawi has a range of policy alternatives which should be considered. Further research on the factors mediating the effect of distance on utilisation can provide useful insights to assist policy-makers in ensuring access policies target the specific factors dissuading women from seeking care.

[9] In addition to likely not being cost-effective in high baseline utilisation contexts such as Malawi.

Relatedly, additional research is required into the source of distance's endogeneity with respect to utilisation. Kumar et al. (2014) speculate that residing closer to health facilities results in improved health behaviours. Individuals residing further from facilities may, therefore, have worse health status inducing individuals to seek delivery in a health facility to offset higher risks of complications. An alternative explanation is that women at further distance differ in the information and treatment by health workers. As almost all women in our sample attend ANC visits, some exposure to health care during pregnancy is consistent across distances. Pressure may be put on women who live in remote communities by health workers who understand the consequences of complications in remote settings. This is similar to the qualitative findings of Griffiths & Stephenson (2001) who found remote women understood the extra importance of preventive action. This could be through strategic relocating during the final stages of pregnancy or simply extra effort to ensure a facility is reached. The different mechanisms suggest differential appropriate policy responses.

Finding ideal instruments for a variable such as distance – which is related to remoteness – is inherently difficult, due to the relation of remoteness with a large number of other factors. One alternative strategy, and an approach for consideration in future research, could be to exploit facility openings and closings. However, to date data limitations have precluded this approach. A complicating factor of utilising panel data on this topic is that in the presence of measurement error it can significantly magnify attenuation bias (Griliches & Hausman, 1986). Therefore, IV approaches may represent the optimal strategy without access to more accurate spatial data. However, exploration with alternative methods could also shed light on whether local average treatment effects (LATE) of IV estimates are closer to the policy relevant parameter than OLS (Heckman et al. 2006). Future research would also benefit from data providing measures of distance to health facility at household level. This would allow for controlling of unobserved village differences that are common to women within a village that may be correlated with health facility accessibility and utilisation rates (Kondylis & Manacorda, 2012). Methods which could account for such unobserved heterogeneity should be combined with models allowing the exploration of the heterogeneity in the effect of distance to provide the best information to inform health infrastructure planning.

Like any study, we faced several limitations. We are unable to identify the specific facility women utilised. This constraint partially informed our research question as, had we observed delivery

location, it would be possible to construct a full patient choice model, mapping women's preferences among the full set of alternative health care providers (McFadden, 1973). In the absence of this information, we examine the impact of distance to the nearest health facility on the utilisation of *any* health facility for delivery services. A distinction has been made between 'passive' and 'active' patients in LMICs, where unlike the former, the latter do not necessarily seek health care at the lowest distance/cost provider (Leonard, 2014). While this chapter does not attempt to address issues surrounding facility choice, our data suggests that a non-trivial proportion of women in Malawi may bypass the nearest health facility (**Table 1.2** and **Supplementary Appendix C1-5**). Future research would benefit from data that definitively matched individuals with where care was sought, allowing for the development of accurate facility choice models. To date this has only been examined in an urban environment where distance is a much weaker determinant of facility choice (Cronin et al. 2017). Such research would allow for the examination of the impact of distance compared to other facility-level characteristics and how these are traded-off. Use of administrative data linking individuals to facilities where care was sought would also circumvent the potential for recall bias faced when using mother-reported historical birth location.

We rely on having adequately controlled for all observable confounders. It is likely that rurality, to which distance is related, is highly correlated with factors that may also impact utilisation, for example SES. The DHS does not contain information on income or consumption expenditure, therefore, we use proxy variables to capture variation in SES. Our primary specifications include measures of women's education and literacy and a household wealth index. While this is standard practise, the veracity of the use of wealth indices to proxy for income is debatable (Filmer & Prichett, 2001). We also checked specifications including further information on the characteristics of women's husband which may have further captured variation in household income (husband's occupation, husband age, husband educational level, woman's earning relative to husband etc.) with the effect of distance unchanged by their inclusion[10]. Further, contextual factors reduce concerns about missing variation in SES and its potential importance in modelling health care utilisation in Malawi. Work has examined the income/expenditure distribution in Malawi to identify households in need of targeted cash transfers (International Labour Organisation, 2016). This work has exposed the difficulty in identifying households due to a lack of variation in many common measures of SES. Given this high degree of homogeneity in income and expenditure in the rural population, this reduces the risk of variation in income/expenditure explaining different

[10] These results are available upon request.

in health care utilisation in our sample. The absence of user fees also reduces the role of income in directly determining utilisation. Finally, in addition to the use of variables proxying for SES, the IV approaches should alleviate concerns of confounding from this source.

For most household surveys collecting sensitive information and geographic data, some form of 'geo-scrambling' is undertaken to maintain individual 'anonymity'. In our context, two distinct geographic displacement procedures introduce measurement error to the constructed distance variable. First, due to the aggregation of households located within the same cluster to a single point coordinate representing the DHS cluster centroid results, we are unable to measure within-cluster distance to health facility variation. However, this within cluster variation is unlikely to be substantive. DHS clusters are related to Enumeration Areas (EA) defined for the Population and Housing Census in Malawi. EAs are the lowest administrative area within Malawi and therefore represent small geographic areas: "Since the EAs are delineated for the purpose of census enumeration, they generally have a relatively small number of households (for example, between 80 and 120 households, which is a practical size for the listing operation. It is important that the EAs have well-defined boundaries, which are generally defined on census maps." (Department of Economic and Social Affairs: United Nations, 2016). Second, the displacement of cluster centroids results in random noise being added to the measure of distance. We contend that both procedures bias our estimates of the causal effect of distance towards 0 (**Supplementary Appendix C1-3**). Simulation work attempting to quantify the impact of such displacement on empirical estimates has suggested that the coefficient on distance may be 36% smaller for the circumstances most similar to ours faced (Elkies et al. 2015). Therefore, the implications of this measurement error imparted from the 'geo-masking' are predictable and can be accounted for in the interpretation of our results. Acknowledging these effects of the measurement error (in addition to the dissection of the endogeneity) we interpret our estimate as a lower bound of the effect of distance on utilisation. On the other hand, a strength of our study is that we avoid another sources of measurement error, such as expert elicitation or household estimates of distance, or only using a sample of health facilities. These approaches have been shown to introduce relatively more severe effects than cluster displacement, as they have an ambiguous effect on the sign and size of the bias compared to measurement error originating from known geographic displacement formulae (Schoeps et al. 2011, Skiles et al. 2013).

There is a growing debate about the perceived trade-off between increasing health care accessibility and improving the quality of health care (Kruk et al. 2018). Improvements in access to and

utilisation of health care services have not always translated into improved health outcomes. This has led some to suggest LMICs should consider relocating certain health care services – including delivery services – to higher levels of care, such as specialist hospitals (Gage et al. 2019). Given that such policies would result in longer travel distances to utilise services, the potential benefits of improved quality must be considered in the context of potentially significant reductions in utilisation, as suggested by our findings. Our results can provide important inputs to this quality-access trade-off debate in order to model welfare changes from different policy choices.

Chapter 2

A Solution Looking for a Problem: The Effect of Maternity Waiting Homes on Health Care Utilisation and Child Health Outcomes

1. Introduction

Despite vast improvements, the global burden of mortality and morbidity due to complications related to childbirth remains exceptionally high. In 2019, an estimated 2.4 million neonatal deaths occurred (United Nations Inter-agency Group for Child Mortality Estimation, 2020). This is in addition to stillbirths of which an estimated 2.6 million took place in 2015 (Blencowe et al. 2016). Finally, approximately 295,000 women died during and following pregnancy and childbirth in 2017 (World Health Organisation, 2019). However, it has been noted that the true cost of maternal mortality may be even greater, stemming from the additional long-term consequences that maternal deaths can cause including future infant and child mortality, poverty and general wider effects on families and communities (Miller & Belizan, 2015).

A majority of these deaths occur in low- and middle-income countries (LMICs) and are considered preventable. Since the 1987 Safe Motherhood Initiative great significance has been placed on reducing maternal mortality rates (MMR) and neonatal mortality rates (NMR) in LMICs. Targeting reductions have been key dimensions of both the Millennium Development Goals (MDGs) 2000-2015 and Sustainable Development Goals (SDGs) 2015-2030. However, 60 countries are currently on course to miss their SDG targets for neonatal mortality by 2030 (Hug et al. 2019).

As most obstetric complications occur around the time of delivery and cannot be predicted, it is generally accepted that a key strategy to reducing pregnancy related mortality is to ensure women deliver in health facilities under the supervision of trained health care professionals (Campbell & Graham, 2006). However, access to and the utilisation of quality health care remains a problem in a large number of LMICs (O'Donnell, 2007). A number of studies have examined the factors determining whether women have a skilled health facility based delivery (Thaddeus & Maine, 1994; Gabrysch & Campbell, 2009). A diverse array of policies have been targeted towards the perceived barriers of accessing obstetric health care and increasing the quantity of deliveries taking place in

facilities. For example, a large number of LMIC health systems provide maternal and child health care services free at the point of use (Powell-Jackson et al. 2014; Lepine et al. 2018). However, accessing health facilities can have significant travel costs associated as well as foregone earnings resulting from time spent seeking and receiving care. Indeed, a number of studies have shown how non-price costs associated with accessing health care act to reduce utilisation (Acton, 1973; Mwabu et a. 1993; Lavy et al. 1996; Thomas et al. 1996). Distance has been shown to be a significant determinant of health care utilisation (Wong et al. 1987; Borah, 2006; Hjortsberg, 2003; Sarma, 2009; Kumar et al. 2013; Karra et al. 2017; Manang & Yamauchi, 2018; McGuire et al. 2021). Additionally, distance can be further complicated by poor road conditions and transport options. It has been shown that lower socio-economic status individuals are more sensitive to travel distances and time, suggesting that it may also contribute to inequities in health care utilisation and health status. (Gertler & van der Gaag, 1990).

A policy proposal which has seen renewed emphasis to increase the utilisation of obstetric health care services in the hope of improving maternal and neonatal health outcomes is the construction of Maternity Waiting Homes (MWHs). These are structures built adjacent to health facilities which assist in ensuring all births take place in suitable health facilities by reducing some of the barriers faced by pregnant women in accessing care. Specifically, MWHs are shelters to which women can relocate and stay in late-stage pregnancy. Subsequently, at the onset of labour, or in the circumstance of requiring health care, this eliminates some of the barriers that women would face in accessing that care. This should increase the proportion of facility deliveries through enabling women who would not or could not travel to a facility in late stage pregnancy or labour. Additionally, MWHs should increase the timeliness of the utilisation of obstetric health care, even among those women who would deliver in a health facility regardless. Furthermore, women may be able to stay at MWHs after delivery to receive postpartum care.

The concept of MWH is not new, at the beginning of the 20th century similar models existed in Europe, Canada and the United States to serve women in remote geographic areas (Aday et al. 1974). In Africa one of the first countries to introduce the concept was Nigeria in the 1950s (Poovan et al. 1990). The inclusion of MWHs in national strategies to increase facility-based deliveries is increasingly common. The concept has been adopted by a number of countries in Africa (Lesotho, Malawi, Ethiopia, Uganda, Ghana, Kenya, Liberia, South Africa, Zimbabwe, Eritrea, Namibia, Zambia), Latin America (Cuba, Guatemala, Honduras, Nicaragua, Peru) and Asia (Lao PDR, Nepal, Timor-Leste) (Penn-Kekana et al. 2017; Ngoma et al. 2019; WHO, 2016;

Partners in Health, 2013). Despite the newfound enthusiasm for the concept of MWHs, there is scarce evidence on their effect on utilisation rates of obstetric health care services and on health outcomes.

In this chapter we explore the effect of MWHs in Malawi. Specifically, we evaluate the impact of MWHs on women's utilisation of pre-natal, post-natal and delivery health care services. Additionally, we examine whether there is any discernible impact on neonatal mortality. Our empirical strategy relies on exploiting differential timing in the opening of MWHs across Malawi.

The chapter is structured as follows. Section 2 outlines the previous literature examining the impact of MWHs. Section 3 provides the institutional background. Section 4 presents the empirical strategy. Section 5 describes the data. Section 6 presents the main econometric results and includes a discussion of robustness checks. Section 7 offers concluding comments.

2. Previous Literature

Previous quantitative research evaluating the effect of MWHs have utilised case-control or before-and-after study designs. Zuanna et al. (2019) compare perinatal mortality for women delivering in a rural hospital in Ethiopia. They find the risk of perinatal mortality was half for women entering the hospital from the MWH compared to women admitted to the hospital directly. Wild et al. (2012) found that the proportion of births taking place in health facilities to women from two districts in Timor-Leste did not increase after the construction of MWHs. Scott et al. (2018) designed the first large scale assessment of MWHs using more rigorous quasi-experimental methods in Zambia. 20 health facility clusters were assigned to a MWH model with 20 control clusters. 2,400 women are sampled at baseline (2016) and endline (2018). Lori et al. (2019) present cross-sectional analysis from the baseline sample, suggesting women who used MWHs were more likely to attend 4 or more ANC visits, as well as more likely to attend all PNC visits and to take measures to avoid pregnancy.

Most research to date on MWHs has been qualitative. Uny (2017) found that women faced issues when waiting in MWHs in Malawi such as hardships in obtaining food, which is often not provided at MWHs, and how husbands responses to their prolonged departure from the household acted as a disincentive to utilisation. As such, while nominally free, MWHs may have associated costs of attendance. Singh et al. (2018) conducted interviews of MWH users and non-MWH users at two health facilities in Malawi. They found women with less pregnancy experience, women who were

aware they were at higher risk status, women of lower SES and women who lived further from facilities were more likely to use MWHs.

3. Institutional Context

A majority of Malawi's population continue to reside in rural locations, 84% as of 2018 (National Statistics Office, 2018). Additionally, Malawi has one of the highest fertility rates in the world with an average of 4.6 births per woman in 2016 (World Development Indicators, 2019). Despite recent improvements, the country still suffers some of the highest rates of under-5 and neonatal mortality globally, 63/1000 and 27/1000 live births respectively in 2015/16 (Ministry of Health, 2017). The maternal mortality rate was 439/100,000 live births in 2016, a reduction from 675/100,000 in 2010 (Ministry of Health, 2017).

Maternal and child health care interventions including obstetric related services are delivered free as part of the Essential Health care Package (EHP) (Ministry of Health, 2017). The national access policy seeks to ensure that all households live within 5 km of a health facility, reduced from a previous target of 8 km (Ministry of Health, 2017). However, the proportion of the population living within 8 km of a health facility declined from 81% to 76% over the same period (Ministry of Health, 2017).

Malawi has a markedly uneven utilisation of obstetric health care services. 91% of births occurred in health facilities between 2010-2015 (National Statistics Office, 2017). However, it has been demonstrated that utilisation rates are significantly lower for women residing further from facilities (McGuire et al. 2021). The Focused Antenatal care (FANC) model of four visits recommended by the World Health Organization (WHO) for mothers with low-risk pregnancies was adopted in 2003. Despite this, between 2000-2013 90% of pregnant women did not access ANC in the first trimester, while only 51% had four or more ANC visits during pregnancy (Kuuire et al. 2017). However, there were improvements over time as in 2013 women were 32% more likely to utilise ANC in the first trimester (Kuuire et al. 2017). A similar underutilisation of post-natal care (PNC) services is seen. Between 2010-2015 only 3% of women received maternal PNC within 24 hours of delivery and 16% within the first week. Only 3% of new-borns received PNC within 24 hours and 26% within a week of delivery (Kim et al. 2019).

Malawi's MWH construction has not been part of one single cohesive policy over time. The earliest known MWH was constructed in 2009 in Malawi. However, over the years plans for the

construction of large numbers of MWHs have been developed. In 2012, Malawi officially adopted MWH construction and use as a policy as part of the Presidential Initiative on Maternal Health and Safe Motherhood (PIMHSM). Accordingly, the development of 130 MWHs was planned across the country.

The PIMHSM had a number of components in addition to MWH construction. These included community mobilisation and training of chiefs and training of community midwife assistants.

4. Empirical Strategy

4.1. Difference-in-Difference and Panel Event Study

We estimate the effect of MWHs by comparing the various outcomes in periods pre- and post-construction, capitalising on the staggered timing of the construction of MWHs at health facilities. In the base case analysis we employ the standard Two-Way Fixed-Effects (TWFE) estimator of the difference-in-difference (DiD) model, which has been used extensively to evaluate policies progressively introduced over time[11].

$$Y_{ivt} = \alpha_v + \gamma_t + \beta^{DD} D_{vt} + X_{bvt} + X_{mvt} + \varepsilon_{ivt} \qquad (\mathbf{1})$$

With $i = \{b,\ m\}$ indicating a birth or mother variable. For instance, Y_{bvt} could measure the probability of having a facility delivery in village v in month-year t, while Y_{mvt} would measure number of ANC visits by a mother. α_v and γ_t are vectors of village and period dummy variables accounting for village and calendar year fixed effects, X_{bvt} and X_{mvt} are a set of time-varying variables reflecting mother and birth characteristics, and ε_{kt} is the error term. D_{vt} is a dummy variable equal to one in the periods after village v has a MWH constructed at the nearest health facility and zero otherwise. The parameter of interest, β^{DD}, captures how exposure to a MWH changes the expected outcomes: the mothers' utilisation of obstetric health care services or birth outcomes.

[11] It should be noted that even for binary outcomes we estimate Eq. (1) using linear two-way fixed effects resulting in a linear probability model. Although methods for estimating non-linear DiD exist (Blundell & Dias, 2009; Athey & Imbens, 2006) this remains the conventional approach. Similarly there is disagreement of the interpretation when modelling DiD in a non-linear framework (Ai & Norton, 2003; Puhani, 2012; Karaca-Mandic et al. 2011). Another specific reason we don't estimate conditional logit models is that the outcome is constant within a number of villages which cannot be modelled via maximum likelihood, causing a number of observations to be dropped.

We also estimate event studies, allowing us to flexibly check the assumption of parallel pre-treatment trends and to examine treatment effect dynamics. We might expect the effect of MWHs to grow over time as knowledge of the service increases.

$$Y_{ivt} = \alpha_v + \gamma_t + \sum_{\tau=-20}^{-2} \rho_\tau D_{v\tau} + \sum_{\tau=0}^{10} \theta_\tau D_{v\tau} + X_{bvt} + X_{mvt} + \varepsilon_{ivt} \qquad (\mathbf{2})$$

Where τ defines event time – measured in year-quarters – rather than calendar time with MWH construction occurring at $\tau = 0$. Each coefficient on $D_{v\tau}$ is a canonical 2x2 DiD estimator with the year-quarter period just before the MWH is constructed as the 'before' period and the period of the coefficient ρ_τ or θ_τ as the 'after' period. The omitted category is $\tau = -1$, the year-quarter prior to MWH construction, capturing baseline outcome differences between facilities where MWH were and were not constructed. As such, ρ_τ and θ_τ denote the change in outcomes in villages where MWHs were constructed at the nearest health facility relative to villages where no MWHs were constructed at the nearest health facility, measured with respect to the year-quarter just prior to construction[12]. In the absence of treatment, it is assumed that villages with MWHs constructed and control villages would have maintained similar differences as in the baseline period. However, like many public programmes, choices of where to construct MWHs may not be random (Rosenzweig & Wolpin, 1986). Todd (2007) illustrated how placement of facilities may be influenced by lobbying. Such policy endogeneity may violate the parallel trends assumption. The event studies provide evidence on the likelihood of parallel trends holding.

4.2. New Methods for Estimating Effects with Staggered Treatment Timing

Recent literature has examined settings with staggered treatment timing and suggested the standard TWFE approach may be biased (Bacon-Goodman, 2021; Chaisemartin & D'Haultfœuille, 2020; Sun & Abraham, 2021). The estimated treatment effect is the weighted average of all possible two-group and two-period DiD estimators. In our case, these 2x2 DiDs are of three types of comparisons: (a) villages where MWHs were constructed with villages where no MWH was constructed as controls; (b) early-construction villages with late-construction villages as controls

[12] It is convention to use the first period prior to treatment as the omitted category. Because we have to coarsen the time period data from birth dates (month/year) to year-quarter periods to increase the units per lead and lag, some births occurring in the first treated year-quarter occur just prior to MWH opening. Specifically we have 28 births occurring in the same year-quarter as MWH opening in treated villages. Therefore, we may expect the first treated period to be a slight underestimate of any treatment effect.

(treated with not-yet-treated villages); and (c) late-construction villages with early-construction villages as controls (treated with previously treated villages)[13]. Bias arises if heterogeneous treatment effects over time are present. Specifically, this creates issues with using previously treated units as controls i.e. type (c). This has led to this comparison of using already treated units as controls being referred to as the 'forbidden comparisons' (Borusyak et al. 2022). In our setting, the effects of MWHs over time may be heterogeneous for a number of reasons. It may take time for information on the opening and availability of MWHs to disseminate into communities, or confidence in the services may increase over time as a growing number of women utilise the service. In addition to heterogeneous treatment effects over time potentially being caused by time-varying treatment effects, if treatment effects are heterogeneous between villages and adoption/construction is related to this essential heterogeneity (Heckman et al. 2006) this also creates a situation where treatment effects are heterogeneous over time. We might, therefore, expect the effect of a MWH to grow over time as women start to utilise the service and local information sharing occurs. If present, heterogeneous treatment effects over time could lead to an underestimate of the average treatment effect on the treated (ATT). Conversely, if experiences are negative the opposite might occur with less women utilising a MWH over time and a potential overestimate of the ATT.

To examine the possibility of bias we compute both the Bacon-Goodman (2021) and Chaisemartin & D'Haultfœuille (2020) decompositions to check the weights of the unit-specific treatment effects used to construct the difference-in-difference parameter, β^{DD}. Essentially, the decompositions allow us to check the relative influence of each of the type of 2x2 DiD comparisons – (a), (b) and (c) from above – on the computation of the β^{DD}. The primary potential cause for concern is the role of the 'forbidden comparisons'. This check enables us to know if the TWFE estimates need to rely on the assumption of time-invariant treatment effects or construction of MWHs not related to essential heterogeneity.

Finally, we implement the Callaway & Sant'Anna (2021) estimator which is robust to the potential issues effect heterogeneity and staggered treatment timing can cause. This approach estimates a generalisation of the standard two-group two-period ATT referred to as the 'group-time average treatment effects on the treated', $ATT(g,t)$, for each group g (indicating the timing of treatment) at each time t. Each $ATT(g,t)$ is nonparametrically defined as:

[13] The TWFE estimator can be decomposed into K^2 $2x2$ DiDs where K is the number of timing groups i.e. the number of different periods at which villages are treated.

$$ATT(g,t) = E\left[\left\{\frac{G_g}{E[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{E\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right\}(Y_t - Y_{g-1})\right] \quad (3)$$

Where G_g equals 1 if a village belongs to group g i.e. has a MWH constructed in period t, C equals 1 for villages where MWHs are never constructed. $p_g(X)$ is the generalised propensity score giving the probability of being first treated as part of group g i.e. having an MWH constructed in the respective period t, conditional on covariates and either being part of group g or being a unit for which C equals 1 i.e. never having an MWH constructed[14]. Intuitively this takes differences between control (never-treated) observations and units in group g, where controls are weighted based on the similarity in their characteristics to units in group g. Similar to many of the new methods, this ensures the ATT is computed only using comparisons which cannot be affected by heterogeneous treatment effects. Using this approach calculates multiple $ATT(g,t)$, which can then be aggregated in various ways following the general formula:

$$ATT_{CS} = \frac{\sum(w_{gt} * ATT(g,t))}{\sum(w_{gt})} \quad (4)$$

Where w_{gt} is a general weight capturing how much information was used in estimating $ATT(g,t)$, with w_{gt} increasing in the number of observations.

5. Data

We combine information on the construction of MWHs at health facilities with retrospective data on births, birth outcomes and obstetric health care utilisation utilising several data sources. The Malawi Demographic and Health Survey (DHS) 2015/16 provides nationally and regionally representative cross-sections of women aged 15–49. The survey captures retrospective self-reported information on births and associated health care utilised for surveyed women taking place over the last 5 years[15]. Geographic information is available on the small area level at which surveyed mothers reside, DHS clusters, which we refer to as villages due to them approximately proxying

[14] Assumptions required include parallel trends conditional on covariates, irreversibility of treatment and propensity score overlap.
[15] Specifically, we use the Birth Recode DHS Survey Data.

for similar areas. We use the Malawi Service Provision Assessment (SPA) 2013/14 to provide information on health facilities including their geographic location. This facility census allows us to link women surveyed to their nearest health facilities.

We compiled data on the development of MWHs including information on the health facility at which they are constructed and the date of competition and opening. Despite being part of official government health policy, information on the location and opening of MWHs is not well documented. As such, various sources were used to ensure we capture information on all operational MWHs within the country. Malawi's Ministry of Health (MoH) website hosts an incomplete list of MWHs which was used as a starting point[16]. Information on MWHs not captured and data on the opening dates of MWHs at health facilities was web-scraped from a number of sources, including newspaper archives. Due to the significance of the MWHs policy, their development was often accompanied by ceremonies for the laying of the foundation stones and official openings. Plaques attached to the structures provide information on the exact date of openings. Verification of the data compiled was undertaken by the Reproductive Health Unit of the MoH and the programme coordinator of the MWH project.

We merge the MWH information with the SPA dataset and then link the DHS villages data to the health facility data, allowing identification of the closest health facilities for each village and associated births from resident mothers[17]. Therefore, the birth-specific treatment is defined at the village-year level as any birth taking place in a village where the nearest health facility has an operational MWH at the time of the birth. As we only have data on the month-year of births, while we have exact opening dates of MWHs, we take the conservative choice of only designating women who gave birth in the month following the opening of a MWH at their nearest health facility or later as having a MWH available as an option.

5.1. Inclusion Criteria

In total the DHS captures data on births from 850 clusters – 677 rural clusters (villages) and 173 urban clusters – with information on 17,286 births over the 5 year period (2010-16). We restrict the analytical sample to births from a subset of women for a number of reasons. We exclude births

[16] https://www.health.gov.mw/index.php/2016-01-06-19-58-23/maternity-waiting-home [accessed 20 August 2020].
[17] Because the SPA doesn't have facility names – which is how the MWH facilities are identified – an intermediary step was required for the linking health facilities in the SPA with MWHs. We spatially join the Central Monitoring & Evaluation Department (CMED)/UNICEF facility census list which does have facility names with the SPA facilities to map SPA facility IDs to facility names. Spatial joining is done on an overlap basis in QGIS.

which took place prior to women residing in their current location, as we do not possess information on previous residence location. For a similar reason we only include women who were surveyed in their usual place of residence. Caesarean deliveries are excluded as these pregnancies and births likely significantly differ in their characteristics which impacts choices around delivery location. Finally, we exclude births occurring in villages where Mchinji District Hospital is the closest facility as a MWH was operating since 2009, prior to birth sample. This leaves a final analytical sample consisting of 13,744 births to 10,801 women occurring between 2010-2016 (**Table 2.1**).

Table 2.1: Analytical sample construction

	Number			Percent		
Criteria	**Births (N)**	**Women**	**Cluster**	**Births (N)**	**Women**	**Cluster**
Births in last 5 years (full sample)	17,286	13,448	850	100%	100%	100%
of those Lived in same location during time of birth	14,764	11,600	850	85%	86%	100%
of those Non-caesarean	13,839	10,883	849	80%	81%	99.9%
of those Have information on place of delivery	13,839	10,883	849	80%	81%	99.9%
of those Usual residents of household (de jure resident)	13,778	10,883	849	80%	81%	99.9%
of those Not resident in village nearest to Mchinji District Hospital	13,744	10,801	843	80%	80%	99.2%

5.2. Outcome Variables

Due to the primary objective of MWHs being to increase the rate of facility delivery, our primary outcome is a binary indicator of whether a mother reported a birth as having occurred at a health facility with delivery capacity. It should be noted the DHS does not provide information on the exact facility in which women deliver (or utilise any other kind of health care). Therefore, while women delivering in circumstances where their nearest health facility with delivery capacity has a MWH in operation are considered as treated, we do not make inferences about the specific location of delivery. We also consider a number of secondary outcomes including the number of antenatal care visits (ANC), whether the baby received postnatal care (PNC) within 2 weeks of delivery and neonatal mortality. For ANC visits we restrict treated women to those who reached completion of a pregnancy at least 9 months after the opening of a MWH. Finally, we consider whether MWHs affected rates of facility bypassing, whereby women seek institutional delivery care but not at their nearest facility. A woman is reported as bypassing if the mother-reported facility type at which the delivery occurred differs from the type of the nearest facility. The raw trends of these outcomes are presented in **Figure 2.1**[18]. We observe some contrasting trends in health care

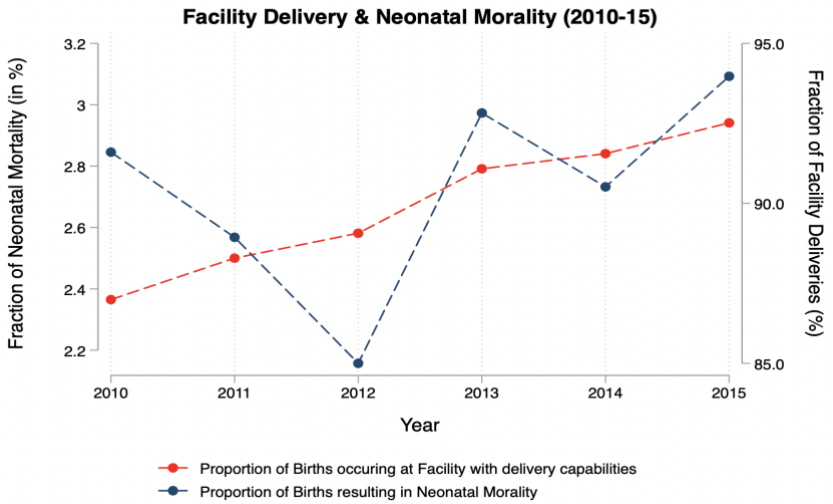
[18] We omit some observations from 2016, as the survey ended at the start of the year and therefore only a small number of observations from this year are available.

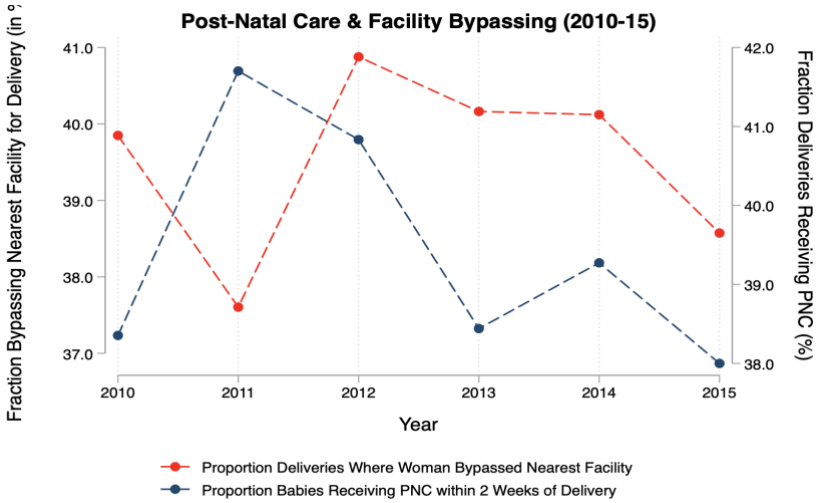
utilisation over the period. While there is an increase in the proportion of deliveries taking place at appropriate health facilities over the period, there is a decrease in the average number of ANC visits per pregnancy.

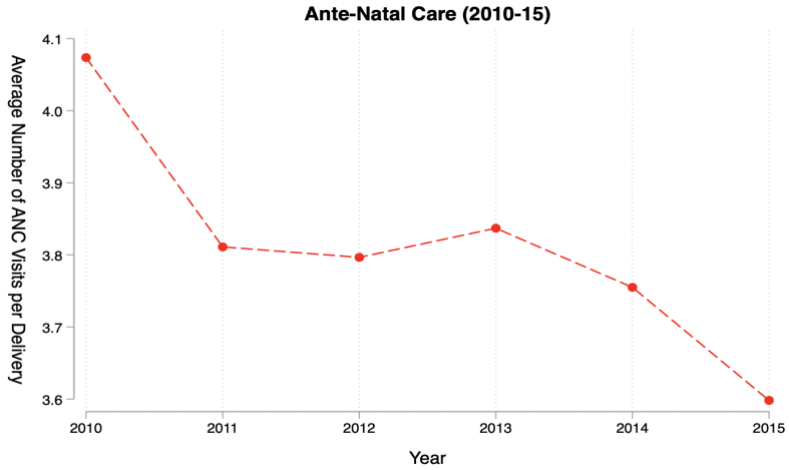
5.3. Control Variables

Although treatment occurs at the village-level, and as such the random unit is the village rather than the individual, we have a number of birth-level controls. Therefore, to maximise precision, our primary strategy is to estimate the model at the birth-level including birth-specific controls and cluster standard errors at the village level. Specifically, we include controls capturing information on the mothers ethnicity, age at delivery, education and literacy levels, level of health knowledge, health insurance status, number of births in the last 5 years, if mother had a caesarean section in the last 5 years. Then birth information such as if this was a first child, whether it had been two years since previous birth, whether it was a single child birth, gestation period, birth weight and whether it occurred during rainy season. Finally, we control for household transport options and the region of residents.

Figure 2.1: Evolution of Outcomes 2010-2015







5.4. Descriptive Statistics

Of Malawi's 977 facilities in 2013/14, only 540 had basic delivery capacity and 71 had capacity to perform caesarean sections (**Table 2.2**).

Table 2.2: Health Facilities in Malawi

	No delivery capacity	With delivery capacity	Total	% type with delivery capacity	In analysis sample	% type with delivery capacity in sample
Central hospital	0	4	4	0.7%	2	0.5%
District hospital	0	24	24	4.4%	22	5.5%
Rural/community hospital	0	41	41	7.6%	36	8.7%
Other hospital	18	29	47	5.4%	20	4.8%
Health centre	53	420	473	77.8%	322	77.4%
Maternity unit	0	4	4	0.7%	3	0.7%
Dispensary	47	0	47	0.0%	0	0.0%
Clinic	299	18	317	3.3%	12	2.9%
Health post	20	0	20	0.0%	0	0.0%
Total	**437**	**540**	**977**	**100%**	**417**	**100%**

Supplementary Appendix C2-1 provides information collected on MWHs operating in Malawi. We capture opening information on 15 MWHs. However, only 10 MWHs opened at health facilities relevant for our sample[19]. **Table 2.3** shows the types of facilities MWHs were constructed at[20].

Figures 2.2 and 2.3 show the number of births over time and how the share and number of births occurring in villages with MWHs constructed at the nearest health facility increases over the period 2010-2016. Despite the gradual opening of more MWHs, the fraction of 'treated' births only peaks at 2.34% of annual births[21].

[19] This is because one MWH was constructed prior to 2010 and so villages linked to this health facility were dropped. Some MWHs were constructed at health facilities which were not linked to villages i.e. they were not the closest facility for any of the DHS Clusters. Finally, some MWHs were constructed and opened just after the survey was completed.
[20] **Appendix A** presents a smaller table with more concise information of the names of the facilities at which MWHs were constructed, including information on the opening date for the subset of 15 facilities most relevant to the analysis sample.
[21] In 2016 the fraction is 6.25% but this is also related to the small number of observations available in this year.

Table 2.3: MWHs Constructed by Type of Health Facility		
	Health Facilities captured in analysis sample without MWH constructed at Nearest Health Facility	**Health Facilities captured in analysis sample with MWH constructed at Nearest Health Facility**
Central hospital	2	0
District hospital	16	6
Rural/community hospital	33	3
Other hospital	20	0
Health centre	317	5
Maternity unit	3	0
Dispensary	0	0
Clinic	12	0
Health post	0	0
Total	**403**	**14**

Notes: As outlined we do not include villages connected to Mchinji District hospital in the sample hence a total of 14 instead of 15 facilities summarised here. As also noted, 4 of the MWHs attached to these 14 facilities were constructed after the sampling period and therefore not included in inference.

Summary statistics comparing villages where MWHs are constructed to villages where they are not constructed, for years prior to the first MWH construction, are presented in **Table 2.4**[22]. Villages where MWHs are constructed at the nearest facility have relatively better obstetric health care utilisation rates and outcomes in years prior to construction. Relatedly, treated villages are disproportionately urban and therefore based at average lower distances to the nearest facility. **Table 2.4** reaffirms the finding that MWHs in the same were disproportionately constructed at urban based District Hospitals.

[22] For constructing **Table 3** we include information on 14 MWHs rather than just the 10 that are relevant in inference. In other words, we calculate summary statistics on differences in villages also using 4 facilities which have MWHs constructed after the first quarter of 2016.

Figure 2.2: Analysis Sample Births across Year-Quarters

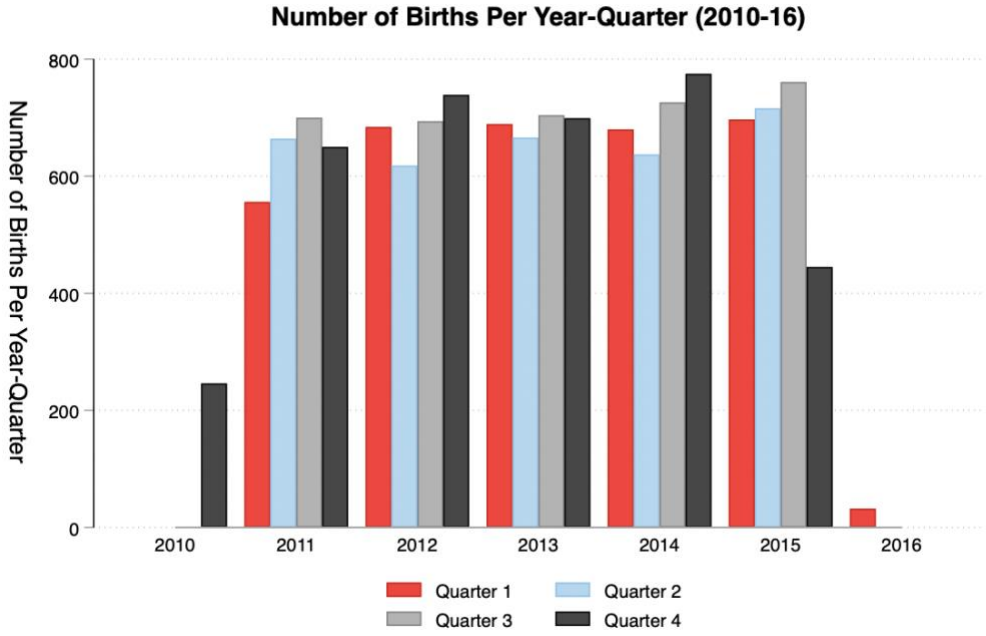


Figure 2.3: Fraction and Number of Births covered by MWHs

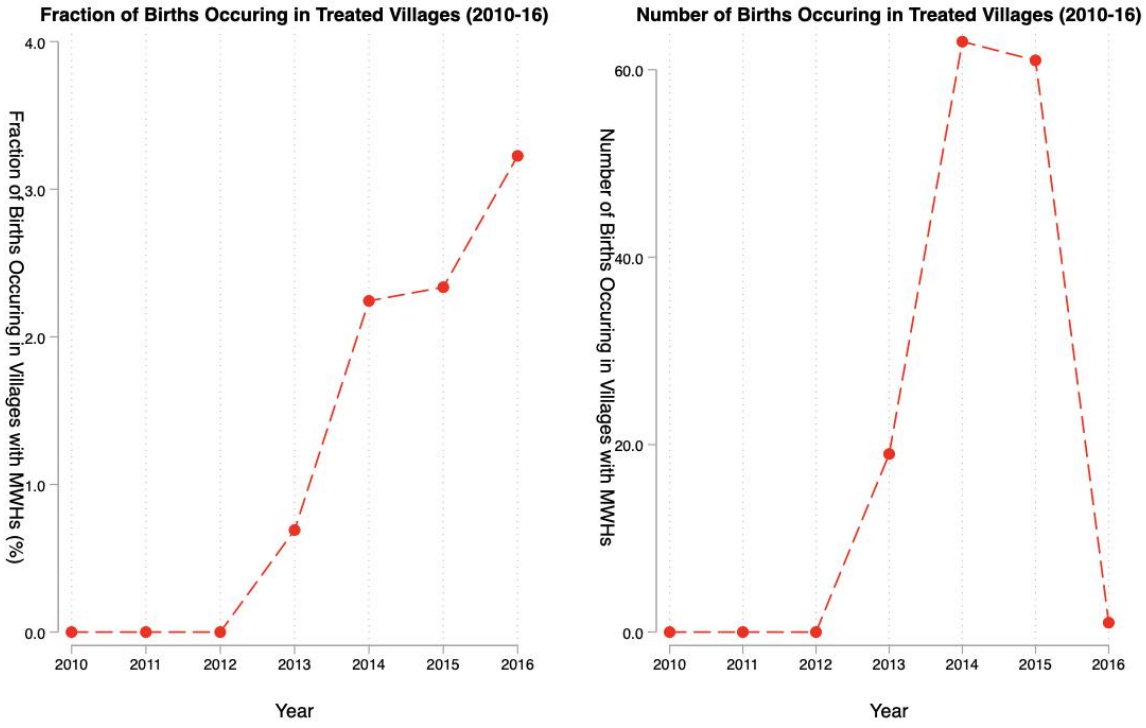


Table 2.4: Summary Statistics 2010-2012		
Variable	**Villages without MWH constructed at Nearest Health Facility**	**Villages with MWH constructed at Nearest Health Facility**
Observations	798	45
Outcome Variables		
Proportion of facility delivery	0.90 (0.16)	0.93 (0.12)
Average number ANC visits	3.80 (0.98)	4.02 (1.04)
Proportion of deliveries where baby had PNC within 2 months	0.42 (0.35)	0.51 (0.37)
Proportion of deliveries where nearest facility bypassed	0.41 (0.36)	0.34 (0.39)
Proportion deliveries resulting in neonatal mortality	0.02 (0.06)	0.003 (0.02)
Village Variables		
Average distance to nearest Health Facility (KM)	4.31 (2.94)	3.59 (3.03)
Region		
Northern	19%	22%
Central	32%	53%
Southern	49%	24%
Type of Place of Residence		
Urban	18%	53%
Rural	82%	47%
Average number of HSAs within 5km	7.5 (4.3)	7.9 (3.4)
Household, Mother and Birth Variables (Average village level)		
Average mother education level*	1.08 (0.40)	1.27 (0.36)
Ethnicity	6.86 (14.00)	4.84 (5.46)
Proportion bicycle ownership*	0.45 (0.49)	0.38 (0.49)
Proportion motorbike ownership*	0.04 (0.10)	0.03 (0.07)
Proportion car ownership*	0.02 (0.09)	0.03 (0.12)
Proportion female headed household	0.25 (0.23)	0.26 (0.25)
Frequency Read Newspaper*	0.24 (0.34)	0.29 (0.28)
Frequency Listen to Radio*	0.78 (0.50)	1 (0.55)
Frequency Watch TV*	0.28 (0.46)	0.52 (0.57)
Average baby birth weight (KG)	3.28 (0.41)	3.27 (0.37)
Proportion births w/HIV+ mother*	0.09 (0.22)	0.13 (0.26)

	2011.46	2011.57
Year of Birth	(0.26)	(0.23)
	0.54	0.34
Distance cited as significant access barrier	(0.32)	(0.31)
	1.49	1.34
Number of Births in Last 5 years	(0.31)	(0.25)
	2.84	2.85
Wealth Index*	(0.87)	(1.16)
	3.18	3.24
Mother Health Knowledge*	(0.42)	(0.34)
	26.44	26.93
Mother Age at Delivery	(3.20)	(3.03)
	0.32	0.20
Illiterate	(0.27)	(0.20)
	0.21	0.20
First Child	(0.19)	(0.18)
	0.97	0.98
Single Child Birth	(0.08)	(0.08)
	0.28	0.23
Rainy Season	(0.21)	(0.21)
	0.90	0.89
Two Years Since Previous Birth	(0.13)	(0.15)
	2.76	2.58
Mother Reported Child Size at Birth	(0.43)	(0.46)
Facility Variables		
Facility Type		
Central Hospital	1.25%	0%
District Hospital	12.67%	71.11%
Rural/Community Hospital	8.28%	8.89%
Other Hospital	5.90%	0%
Health Centre	68.26%	20%
Maternity Unit	0.50%	0%
Clinic	3.14%	0%
Managing Authority		
Government / Public	69.76%	100%
Christian Health Association of Malawi (CHAM)	24.47%	0%
Other	5.77%	0%

Notes: Mean and standard deviation calculated as average of 2010-2012. Note that Observations illustrates the maximum number of observations going into the construction of summary statistics. Real number of observations may be smaller due to distribution of birth timing across villages or missing data. As the survey occurred in 2015/16 some variables are used as proxies for levels in the years 2010-2012 but may have changed over time. These are labelled with *.

6. Results

6.1. Estimation Results

Table 2.5 shows the results of Eq. **(1)**. To account for serial correlation over time we cluster standard errors at the village (DHS cluster) level (Bertrand et al. 2004; Abadie et al. 2020)[23][24]. These results suggest MWH construction at the nearest health facility increases the probability of having

[23] As we have far more than 50 clusters we can rely on the asymptotic validity of using a cluster-robust variance-covariance estimator (Cameron & Miller, 2015).

[24] Because we include fixed effects at the village (DHS cluster) level, Abadie et al. (2020) note that heterogeneity in the treatment effect (across villages) is a requirement for using clustered standard errors to be necessary.

a facility delivery by 2.8 percentage points (p<0.10). However, no significant effect is observed for any other utilisation measures or for neonatal outcomes.

Table 2.5: Difference-in-Difference – TWFE estimates

Outcome	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Facility Delivery	0.036*	0.014	0.017	0.013	0.017	0.031**	0.028*
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)
N	13,778	13,778	13,778	13,350	13,778	11,591	11,291
Facility Delivery – Rural Households Only	0.041	0.02	0.023	0.016	0.023	0.032*	0.027
	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)
N	12,095	12,095	12,095	11,697	12,095	10,045	9,770
ANC visits	-0.390*	-0.261	-0.245	-0.256	-0.261	-0.270	-0.280
	(0.21)	(0.22)	(0.22)	(0.23)	(0.22)	(0.19)	(0.20)
N	10,492	10,492	10,492	10,215	10,492	9,129	8,920
ANC visits – Rural Households Only	-0.515*	-0.394	-0.386	-0.391	-0.404	-0.384	-0.384
	(0.28)	(0.29)	(0.29)	(0.30)	(0.28)	(0.24)	(0.25)
N	9,080	9,080	9,080	8,823	9,080	7,810	7,620
Baby had PNC within 2 months	0.013	0.038	0.038	0.048	0.038	0.050	0.060
	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
N	10,465	10,465	10,465	10,159	10,465	9,037	8,815
Baby had PNC within 2 months - Rural Households Only	-0.059	-0.035	-0.037	-0.034	-0.036	-0.035	-0.032
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)
N	9,054	9,054	9,054	8,771	9,054	7,727	7,525
Bypassed nearest facility	-0.012	-0.018	-0.013	-0.029	-0.013	0.002	-0.021
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
N	10,648	10,648	10,648	10,308	10,648	9,464	9,205
Bypassed nearest facility – Rural Households Only	-0.012	-0.018	-0.013	-0.029	-0.013	-0.002	-0.021
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
N	10,648	10,648	10,648	10,308	10,648	9,464	9,205
Neonatal mortality	0.006	0.001	-0.003	-0.006	-0.003	0.004	0.004
	(0.01)	(0.010	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
N	13,778	13,778	13,778	13,350	13,778	11,591	11,291
Neonatal mortality – rural	0.014	0.01	0.007	0.002	0.009	0.015	0.016
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)
N	12,095	12,095	12,095	11,697	12,095	10,045	9,770
Controls							
Year Fixed Effects		x	x	x	x	x	x
Mother characteristics			x	x	x	x	x
Mother characteristics (incl. vars w/ missing data)				x			x
Birth characteristics					x	x	x
Birth characteristics (incl. vars w/ missing data)						x	x

Notes: All coefficients relate to Post MWH construction relative to outcome. Standard errors in parentheses *p<0.10, **p<0.05, ***p<0.01. Standard errors clustered at the village (DHS cluster) level.

Given the expectation that MWH construction should only realistically result in an improvement in the majority of the outcomes[25], **Table 2.6** shows the same results for a one-tailed t-test, where the null is that the coefficients are less than or equal to zero for facility delivery, ANC and PNC visits, and that the coefficients are greater than or equal to zero for bypassing and neonatal mortality. This assists in overcoming the statistical power issues stemming from the relatively small number of treated births.

Table 2.6: Difference-in-Difference – TWFE estimates (one-side hypothesis p-values)

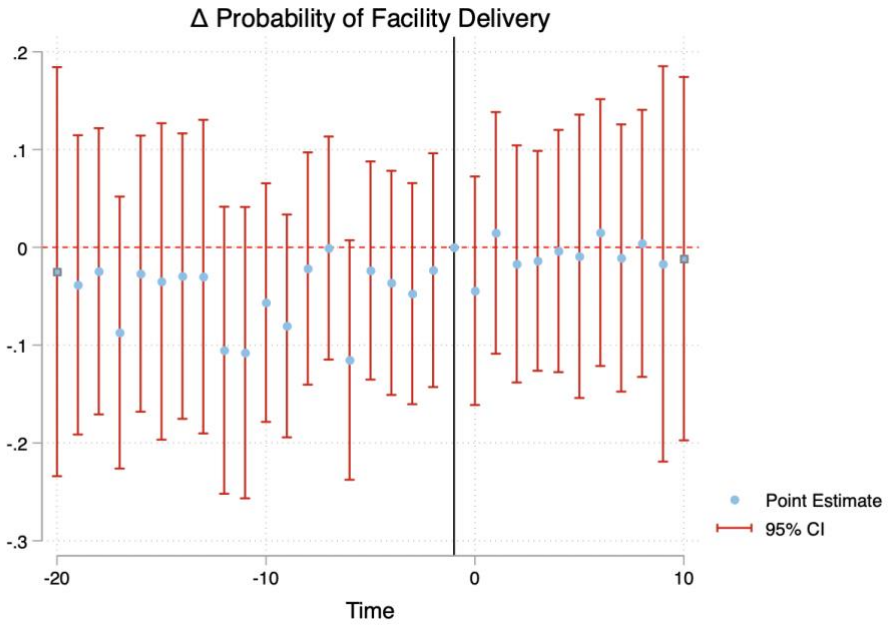
Outcome	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
Facility Delivery	0.036**	0.014	0.017	0.013	0.017	0.031**	0.028**
	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.01
N	13,778	13,778	13,778	13,350	13,778	11,591	11,291
Facility Delivery – Rural Households Only	0.041*	0.02	0.023	0.016	0.023	0.032**	0.027*
	-0.03	-0.03	-0.03	-0.02	-0.03	-0.02	-0.02
N	12,095	12,095	12,095	11,697	12,095	10,045	9,770
ANC visits	-0.39	-0.261	-0.245	-0.256	-0.261	-0.27	-0.28
	-0.21	-0.22	-0.22	-0.23	-0.22	-0.19	-0.2
N	10,492	10,492	10,492	10,215	10,492	9,129	8,920
ANC visits – Rural Households Only	-0.515	-0.394	-0.386	-0.391	-0.404	-0.384	-0.384
	-0.28	-0.29	-0.29	-0.3	-0.28	-0.24	-0.25
N	9,080	9,080	9,080	8,823	9,080	7,810	7,620
Baby had PNC within 2 months	0.013	0.038	0.038	0.048	0.038	0.05	0.06
	-0.04	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05
N	10,465	10,465	10,465	10,159	10,465	9,037	8,815
Baby had PNC within 2 months - Rural Households Only	-0.059	-0.035	-0.037	-0.034	-0.036	-0.035	-0.032
	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.06
N	9,054	9,054	9,054	8,771	9,054	7,727	7,525
Bypassed nearest facility	-0.012	-0.018	-0.013	-0.029	-0.013	0.002	-0.021
	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05
N	10,648	10,648	10,648	10,308	10,648	9,464	9,205
Bypassed nearest facility – Rural Households Only	-0.012	-0.018	-0.013	-0.029	-0.013	-0.002	-0.021
	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05
N	10,648	10,648	10,648	10,308	10,648	9,464	9,205
Neonatal mortality	0.006	0.001	-0.003	-0.006	-0.003	0.004	0.004
	-0.01	(0.010	-0.01	-0.01	-0.01	-0.01	-0.01
N	13,778	13,778	13,778	13,350	13,778	11,591	11,291
Neonatal mortality – rural	0.014	0.01	0.007	0.002	0.009	0.015	0.016
	-0.02	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01
N	12,095	12,095	12,095	11,697	12,095	10,045	9,770
Controls							
Year Fixed Effects		x	x	x	x	x	x
Mother characteristics			x	x	x	x	x
Mother characteristics (incl. vars w/ missing data)				x			x
Birth characteristics					x	x	x
Birth characteristics (incl. vars w/ missing data)						x	x

Notes: All coefficients relate to Post MWH construction relative to outcome. Standard errors in parentheses *p<0.10, **p<0.05, ***p<0.01. Standard errors clustered at the village (DHS cluster) level.

[25] This is potentially slightly more contentious for the neonatal mortality outcome where MWH accessibility might in some cases have the effect of reducing women from seeking care at higher levels. However, given almost all MWHs were constructed at District Hospitals, this issue does not affect our setting.

The event studies examine changes in the outcomes – for example, the probability of having a facility delivery for instance – over year-quarters in villages where MWHs were constructed at the closest facility relative to villages without MWH construction, before and after construction. The event study plots show the results of Eq. **(2)**. **Figure 2.4** presents the event study plot for changes in the probability of a birth being a facility delivery (see **Supplementary Appendix C2-2** for plots for other outcomes).

Figure 2.4: Event Study Plot Facility Deliveries



Although there are some year-month outliers for various outcomes, there are no systematic patterns indicating differences in pre-treatment trends in outcomes where MWHs were constructed at the nearest health facility and where no MWH was constructed. That no lead or lag coefficient is significant is a strong suggestion of the comparability of villages where MWHs were constructed at the closest facility and control villages, particularly considering there are some leads and lags with relatively few units which commonly leads to over-rejection of the null (Mackinnon & Webb, 2017). We also replicate the event study aggregating birth data to year-half (6 month) periods. Doing so reduces the variation in changes in the outcome overtime confirming that deviations in lead and lag coefficients between treated and control villages largely result from small

per relative time period samples rather than reflecting any trends[26]. Further, the graphs reaffirm the previous finding, that effects of MWH were limited.

6.2. Decomposition Results

The Bacon decomposition breaks the components of a TWFE into specific groups according to treatment times and examines all the possible 2x2 permutations. This requires that the panel is strongly balanced[27]. Given our data is at the birth level, with each observation signifying an event, this requires reformatting the data. Therefore, we aggregate birth data to the village-level across years[28]. As such, instead of estimating model **(1)** at the individual (birth) level we estimate it on village-year means. In this case, births are considered treated if they occur in the calendar year after the construction of a MWH at the nearest health facility. See **Supplementary Appendix C2-3** for information on implications for observation weighting. Despite the specification differences, results from the decompositions should provide indicative evidence on whether results from the single-coefficient DiD model provides a valid estimate of the ATT. We only examine the decomposed estimates and weights using the outcome of whether a woman had a facility delivery, as the implications of the tests for a single outcome should hold across all outcomes.

Ex-ante we speculate that use of TWFE will not pose a problem in our setting due to the proportion of untreated villages/births relative to treated villages/births. Therefore we would expect the weights of late-construction villages using early-construction villages as controls will not be large enough to significantly impact our estimates, even in the presence of heterogeneous treatment effects over time.

Figure 2.5 presents the set of 2x2 DiD estimates of the effect of MWH construction on the probability of having a facility delivery. Three key points are worth noting. First, the treated vs. never treated comparisons constitute almost all the weight in the estimation of the aggregate DiD coefficient. Second, the estimates for the treated vs. never treated comparisons are relatively closer to zero than other group comparisons. We present the global (Bacon) decompositions without and with covariates for all outcomes in **Supplementary Appendix C2-4**. Here we see that the treated vs. never treated comparisons constitute no less than 98.7% of the weight. As would be

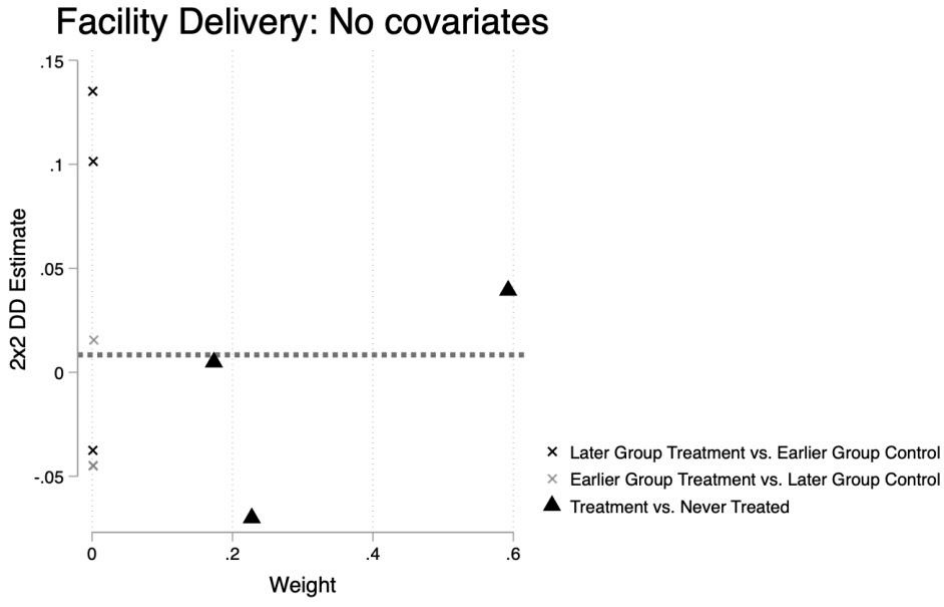
[26] Results available upon request.
[27] Similarly, the Chaisemartin & D'Haultfœuille decomposition requires a balanced panel.
[28] Even then the panel is not balanced due to there not being at least one birth in every village across each year. Therefore, we restrict our sample to the years 2011-2015 such that we have $N = 757$ villages observed over $T = 5$ years. In reformatting the data in this way we lose only 1 treated village.

expected, this finding holds for all outcomes and when controls are included. Third, the estimates between the timing-group comparisons are quite distinct from the treated vs. never treated comparisons. This may result from the relatively small number of treated villages, but illustrates the value of the number of never-treated villages in our sample.

Similarly, using the procedure outlined by Chaisemartin & D'Haultfœuille (2020) we observe that none of the weights associated to the component 2x2 DiDs aggregating to the ATT are negative[29]. As such, our fixed effect estimate must have the same sign as the village-specific average treatment effects. Finally, we also test whether $\hat{\beta}^{DD}$ from $(\mathbf{1})$ provides an unbiased estimate of the ATT. As suggested by Chaisemartin & D'Haultfœuille (2020) checking $H_0: \beta_{fe} = \beta_{fd}$ can be used as a diagnostic to check that the weights associated to each unit- time-specific average treatment effect (ATE) and the respective ATEs are uncorrelated. This is due to β_{fe} and β_{fd} using different weights. It should also be noted the weights discussed here are different to the Goodman-Bacon weights. If $H_1: \beta_{fe} \neq \beta_{fd}$ holds at least one is a biased estimate of the ATT. We fail to reject the null of $H_0: \beta_{fe} = \beta_{fd}$ ($t - stat = 0.1$).

Combined, the results of the Bacon decomposition and Chaisemartin & D'Haultfœuille tests reassure that the standard TWFEs are robust to biases driven by heterogeneous effects.

Figure 2.5: Goodman-Bacon Decomposition



[29] Results available upon request.

Table 2.7 reports the Callaway & Sant'Anna Estimates, ATT_{CS} for each outcome for models estimated at the village-level both without and with inverse-variance weighting (constant).

Table 2.7: Callaway & Sant'Anna Estimates

	Village-Year Level	
Outcome	**Unweighted**	**Weighted: constant (average)**
Facility Delivery	-0.027	0.009
	(0.046)	(0.02)
N	3,489	3,489
ANC visits	-0.401	-0.469
	(0.35)	(0.38)
N	3,336	3,336
Baby had PNC within 2 months	0.122*	0.152*
	(0.07)	(0.091)
N	3,325	3,325
Bypassed nearest facility	-0.171	-0.130
	(0.12)	(0.13)
N	2,801	2,801
Neonatal mortality	0.008	-0.001
	(0.03)	(0.03)
N	3,489	3,489
Controls		
Mother characteristics	x	x
Mother characteristics (incl. vars w/ missing data)	x	x
Birth characteristics	x	x
Birth characteristics (incl. vars w/ missing data)	x	x

Notes: All coefficients relate to Post MWH construction relative to outcome. Standard errors in parentheses *p<0.10, **p<0.05, ***p<0.01. Standard errors clustered at the village (DHS cluster) level.

Focusing on the weighted specification, the sign of the results is mostly similar. The effect on the probability of having a facility delivery is reduced and no longer significant, while MWHs increase the probability of having a PNC focused on the baby within 2 weeks of delivery by 15.2 percentage points (p<0.10).

6.3. Sensitivity

While the decompositions and Callaway & Sant-Anna estimator reassure that the TWFE and event study estimates are not biased due to possible heterogeneous treatment effects over time, we undertake some further sensitivity checks of our results. First, we include region-by-year fixed effects – $\alpha_r * \gamma_t$ – in Eq. $(\mathbf{1})$. This ensures that the control villages come from the same Region as treated villages, with effects identified from deviations from region-specific trends.

Supplementary Appendix C2-5 shows the results are almost identical to the standard TWFEs estimates[30].

Second, villages where an MWH is constructed at the nearest health facility may differ from villages where no MWH is constructed. Although we observed no obvious differences in outcome pre-trends in the event studies, there are some clear differences in characteristics between villages where MWHs were built. Therefore, we estimate Eqs. **(1)** and **(2)** using the sample of treated villages only, excluding the never adopters from contributing any identifying variation. Therefore, identification relies solely on the timing of treatment rather than whether a village has a MWH constructed at all.

Supplementary Appendix C2-5 shows that using the sub-sample of ever treated villages, having an MWH constructed at the nearest facility increases the probability of having a facility delivery by 4.5 percentage points ($p<0.10$) an effect size 60% larger than previously estimated. Similarly, there is a larger negative effect on the number of ANC visits which is now significant ($p<0.10$) than previously reported. Finally, the effect on child outcomes is now negative, reducing the probability of neonatal mortality by 4.2 percentage points ($p<0.10$). These changes are not totally unsurprising due to the magnitude of never-treated units dropped. The respective event study plots in **Supplementary Appendix C2-5**, which are arguably slightly more robust to the previously outlined issues of using early adopters as subsequent controls, show less obvious differences to the static DiD specification[31]. This sensitivity check, therefore, does not provide further evidence that the primary results aren't driven by differences between ever-treated villages and never-treated villages, as it would if the results had been similar to **Table 2.5**. However, the previously presented specifications utilising all the data remain our preferred specifications, due to the issues of relying exclusively on ever-treated villages and the potential impact of 'forbidden comparisons' on the result.

7. Discussion

We examine the impact of opening MWHs at health facilities on the utilisation of obstetric health care services and child health outcomes. This is done from liking historical birth information from a household survey which overlaps with the construction and opening of a number of MWHs in Malawi. This represents, to our knowledge, the first national scale evaluation of MWHs using a

[30] We also allowed for different urban-rural time fixed effects without any impact on the results.
[31] Available upon request.

robust quasi-experimental evaluation design. Using a difference-in-differences approach we find some limited evidence that MWHs increase the probability of having a facility delivery. Our estimates of effects on other outcomes are not significantly different from zero at conventional levels. This is unsurprising given that we were only able to detect an effect in the outcome for which a priori we would expect the largest impact at a 10% significance level. Given the relatively small number of treated units the study may be underpowered to detect effects on the additional outcomes. Finally, due to the relative proportion of births/villages where an MWH never opens at the nearest health facility, we conclude our results are likely robust to the potential issues of using TWFEs in the presence of staggered treatment timing and possible effect heterogeneity. Specifically, this is because the so called 'forbidden comparisons' play only a minor role in the estimation of the ATT.

Further, descriptive analysis suggests it is unsurprising that large and significant effects are not easily identifiable for the range of outcomes examined. The proportion of women having facility deliveries, the primary targeted outcome, was high prior to the implementation of the MWH policy. This suggests women were already overcoming the health care access issues faced, limiting the MWHs' potential effect. Additionally, although the MWH policy is intended to address issues in accessing health care, the implementation of the policy may have been sub-optimal in this regard. A substantive proportion of MWHs to date have been constructed at urban District Hospitals (7/15 for which we have full data) rather than where they may have the potential to have greatest impact at Health Centres with significant rural populations. The lower average distances to the nearest health facility of villages where MWHs were constructed compared to where they were not illustrates that MWHs may have, so far, been constructed in environments where access problems and distance are not as strong factors in determining utilisation (McGuire et al. 2021). Given the perceived poor locational choices of the MWHs constructed to date, it might straightforwardly be expected that the observed effect represents a lower bound should the policy continue to be implemented. However, it has also been observed that facility bypassing is a significant phenomenon in Malawi. Specifically, there appears to be a revealed preference for delivery at higher level facilities. Therefore, despite the greater access issues, it is unclear how the effect of continued MWH construction at Health Centres would relate to the effect currently observed.

Malawi's continued pursuit of the construction of MWHs[32], the small effect on facility deliveries and the null and near zero point estimate on the effect on neonatal mortality is potentially important. It suggests the policy, to date, has not had a large effect on the number of births taking place at health facilities and an even smaller subsequent downstream effect on health outcomes. Given the not insignificant cost of constructing health infrastructure, it is therefore highly unlikely that in its current form this represents a cost-effective use of health care resources. Given the re-emergence and re-popularisation of MWHs this possibly represents the adoption of a policy with growing global popularity without due consideration of the suitability to the local prevailing context. We suggest a concerted effort to gain further evidence is prudent before continued pursuit of and construction of further MWHs.

As always, our study has a number of possible limitations. While care has been taken to try and collate and verify the list of MWHs operating in the country, it is possible that the construction and opening of MWHs was missed. Despite being highlighted as a flagship policy of the PIMHSM, documentation and record keeping of its implementation has been poor.

The prevalence and nature of bypassing has significant potential implications for our results. The predominant nature of bypassing observed, whereby women forgo seeking care at their nearest facility to utilise higher level facilities (as shown in **Appendix Table 1.5)**, combined with most MWHs being constructed at District Hospitals, suggests the potential for contamination effects. Specifically, while the construction of MWHs has the potential to induce 'local' women (i.e. residents of treated villages) to utilise health facilities where they may otherwise not have, it also has the potential to induce 'non-local' women to change their care seeking behaviour. Specifically, if women who reside in non-treated villages (i.e. villages where the nearest facility does not have an MWH constructed), our control group, are induced to seek care by the construction of an MWH at a further facility, this will reduce our ability to observe a treatment effect among women residing in treated villages. This contamination issue has the potential to attenuate the estimated treatment effect towards zero[33].

As outlined in Chapter 1, the DHS suffers from a geographic displacement procedure. Unlike chapter 1, we utilise both urban and rural residents in the analysis. Therefore, the geo-masking

[32] It is unclear whether this is occurring in practise but the policy has never been officially revised since its adoption as part of the PIMHSM.

[33] One potential solution would be to restrict the sample to 'non-bypassers' and estimate the effect of MWH construction on this group. This would eliminate from the sample those women who reside closest to a Health Centre but deliver in a District Hospital. And provide a sort of LATE estimate for this sub-group.

procedure displaces aggregated urban clusters by a random-angle, random-distance between 0-2km and 99% of rural clusters by 0-5km and 1% rural clusters by 0-10km. This displacement can potentially lead to incorrectly assigning villages (DHS clusters) to health facilities which in reality are not geographically the nearest facility. Most importantly for our study, this may result in some villages being indicated as having a MWH constructed at their nearest facility, while in actuality their nearest facility did not have an MWH built (and vice versa). This potential measurement error and misallocation of treatment status adds noise to any potential relationship between MWH construction and health care utilisation, making treatment effects – should they exist – more difficult to identify. However, one advantage of the limited number of MWHs constructed in our data is to limit the risk of DHS clusters being assigned the wrong treatment status as a result of the displacement.

Chapter 1 showed that distance had a relatively larger effect on sub-groups of women. Unfortunately, the relatively small sample of treated births prevents the meaningful undertaking of heterogeneity analysis, such as examining if women residing at further distances from their closest facility are more impacted by MWHs. Finally, as noted we do not have information on whether women actually utilised MWHs. Therefore, the parameter observed may be considered similar to an intention to treat effect.

Chapter 3

Do Health Care Quality Improvement Policies Work for All? Estimating Distributional Effects According to Baseline Quality Levels

1. Introduction

Poor quality has acted to undermine the impact of expansions in the supply and utilisation of health care seen in most low- and middle-income countries (LMICs) (Banerjee et al. 2004; Das et al. 2016). Increasing the supply and demand for care without attention to quality may do little to improve health outcomes or protect against the financial risk of health expenditures (Kruk et al. 2018). Further, wide variations in the quality of care provided have been identified in a number of LMICs, often creating and exacerbating health inequalities (Kruk et al. 2017). Just as there are access deserts, quality deserts exist whereby although health care is technically provided, it is not of an adequate standard. This is acknowledged in the WHO concept of 'effective coverage' (World Health Organisation, 2015). Consequently, many LMICs have attempted to improve health care quality through a range of quality improvement (QI) programmes. QI programmes encompass a wide range of distinct interventions and policies, including public-private contracting, payment reforms and introducing accreditation standards targeting different components of the quality framework (Rowe et al. 2018; Rowe et al. 2019). A specific type of policy we focus on in this chapter relates to national accreditation schemes and supportive supervision (see Section 2). The former are becoming increasingly common in attempts to improve quality standards (Mate et al. 2014), with more than 70 accreditation programmes identified globally in 2013 (Saleh et al. 2013).

Evaluations across a range of QI programmes have found mixed results. Bukonda et al. (2002) noted the positive effect of an accreditation programme in Zambia on facilities compliance with outlined standards, but the high associated costs led to the discontinuation of the programme. Liberia introduced an accreditation system for all facilities linked to funding eligibility, however large deficiencies in facility standards were identified and follow-up surveys never completed (Cleveland et al. 2011). A systematic review of the impact of accreditation schemes identified only a modest effect with a median of 7.1 percentage point increase in quality outcomes (Rowe et al.

2018). Bosch-Capblanch et al. (2011) undertook a review of the impact of managerial supervision to PHC facilities, defined as routine supervision visits of health care providers by higher-tier or district health workers[34]. They find nine studies examining various types of managerial supervision schemes across LMICs. The studies looked at a diverse range of outcomes from drug stock management to adherence to standard treatment guidelines with limited evidence of effects found, however the quality of the evidence was deemed to be poor.

These evaluations, typically focussing on a single average treatment effect however, may mask variation in the impact of QI programmes, which is important for two reasons. First, it has been noted that heterogeneity[35] in the impact of QI programmes – both within and across programmes – may contribute to explaining the mixed results found in evaluations of QI programmes (Binyaruka et al. 2020). A small number of studies have examined heterogenous effects in Performance Based Financing (PBF) schemes in LMICs[36]. Primarily, differences in effects have been explored across patient sub-populations (Lannes et al. 2016; Binyaruka et al. 2018; Van de Poel et al. 2015) and facility sub-groups (Sherry et al. 2017; Binyaruka et al. 2018). Sherry et al. (2017) find heterogeneous responses to a P4P programme in Rwanda, specifically effects varying by baseline levels of facility quality, with the largest improvements seen in the medium-quality tier for both rewarded and unrewarded services. They found high-quality facilities saw the greatest increase in provision of services with the largest associated financial reward, as they had the highest marginal incentive for doing so. However, this variation is at least partly induced by differential incentives, as programme payments were scaled by a general quality multiplier, introducing variation in the incentives for facilities of different baseline quality. In many cases, observed heterogenous policy impacts derive from incentive design effects. Binyaruka et al. (2018) are able to distinguish between incentive design effects and structural effects of a P4P programme in Tanzania due to different performance target features used across quality measures. They find the effect of P4P on institutional deliveries, for which facilities face different threshold targets and

[34] They distinguish this from clinical supervision which have more of a medical education agenda and are focused specifically on clinical practise. Managerial supervision is more wide-ranging and an important part of the link between peripheral health care providers and district-level policy makers. It is often part of regular district management procedures and examines administrative and managerial activities as well as clinical procedures.

[35] It should be noted that heterogeneous treatment effects is not a singular concept. Here we restrict exploration to when treatments interact with pre-treatment variables, sometimes referred to as treatment-covariate interactions. Subsequently, we narrow our discussion to assessment of heterogeneity across pre-defined, discrete sub-groups. However, another source of heterogeneity may be apriori partial knowledge of potential gains, resulting in possible correlation between the treatment effect size and the probability of receiving treatment (Manski, 1990; Heckman et al. 2006). The presence of the latter leads to questions on the validity of standard treatment effect estimates. See **Supplementary Material A** for a more formal distinction between sources and types of treatment effect heterogeneity.

[36] Also known as pay for performance (P4P), Results based financing (RBF).

therefore differential incentives, is largest among baseline low performers, reducing performance inequalities among facilities. For the provision of Intermittent Preventive Treatment for Malaria, for which facilities face identical threshold targets, the effect of P4P was constant across facilities. They note that this goes against much of the literature which predicts failing to account for variation in baseline performance should lead to increases in performance inequality (Rosenthal et al. 2005). Finally, they find that larger facilities and facilities with more supplies received greater P4P pay-outs. This shows that, while context and programme specific factors clearly play a role, examining how the effects of QI programmes vary across units with different baseline characteristics may help understanding the circumstances where QI programmes may be effective. These studies all focused on PBF-style programmes. To our knowledge, there is no current evidence of heterogeneity in the effects of other forms of QI programmes.

Second, QI programmes have the potential to address distributional concerns by reducing variations in health care quality which contribute to health inequalities. When distributional concerns are important, effect heterogeneity, may have equally important policy implications as average effects. Consequently, it is important to characterise the distributional impacts of QI programmes: did those at the lower end of the quality distribution gain relatively more from the programmes? As such, standard evaluation methods examining average impacts, implicitly based on fundamental utilitarian principles, may be less appropriate as a normative basis for assessing policy success than evaluations enabling assessment of equity objectives. In addition to equity objectives, there may be efficiency reasons to examine heterogeneity in the effect of QI programmes. Mortality due to the provision of low quality health care remains a significant burden (de Savigny et al. 2004; Kruk et al. 2018). QI programmes could be integral in reducing the health burden attributable to poor quality health care. Despite these important potential benefits, few studies have examined whether QI programmes have contributed to reductions in variations in the quality of health care provided.

In this chapter, we examine treatment effect heterogeneity of a QI programme – the Ideal Clinic Realisation and Maintenance Programme (ICRMP) – being implemented in primary health care (PHC) facilities in South Africa (SA). Our identification strategy exploits the staggered roll-out of the ICRMP which led to facilities across a wide range of pre-treatment quality levels being present in both treated and control group. We employ two primary econometric strategies to identify possible heterogeneous treatment effects. First, we implement a Difference-in-Difference-in-Difference approach enabling identification of different effects of the ICRMP on sub-groups

defined by baseline facility quality levels. We then estimate a Changes-in-Changes model which overcomes the issue of any heterogeneity identified being a function of model specification choices, by relaxing some of the DD assumptions, and by explicitly allowing for the identification of distributional impacts (Athey & Imbens, 2006). A previous study of the ICRMP found a significant positive effect of the programme on quality checklist scores, while no effect was identified on a range of further non-ICRMP quality process measures (Stacey et al. 2021). However, this average impact may conceal a range of impacts across facilities with important policy implications.

The remainder of the chapter is organised as follows. Section 2 provides contextual background and information on the ICRMP. Section 3 summarises the data. Section 4 presents the methods employed. In Section 5 we present our central results. Section 6 examines the validity of these methods, outlining possible violations in assumptions and assesses the robustness of our results. Sections 7 and 8 provide discussion and conclusion.

2. Ideal Clinic Realisation and Maintenance Programme

The ICRMP was established as part of SA's strategy to roll-out National Health Insurance by 2025. The programme was conceived following a 2012 national facility audit, which identified significant short-comings in the quality of care provided by health facilities in the country. The objective of the programme is to set a quality standard for PHC facilities in an attempt to improve health care quality across multiple quality domains. PHC facilities range in size from one-room clinics housing only two staff in rural areas to facilities with over ten rooms and ten plus staff in urban densely populated areas. Each PHC facilities should have a PHC facility manager in place who is responsible for the running of the facility (NDoH, 2019).

The ICRMP has two major components; a checklist and a QI programme. The ICRMP checklist is a national standardised list of quality indicators against which facilities are assessed and scored, comprised of approximately 200 indicators separated into 10 components: administration, integrated clinical services management, medicines supplies and laboratory services, human resources, support services, infrastructure, health information management, communication, and stakeholder engagement (**Supplementary Appendix C3-1**). Facilities are designated 'Ideal' by achieving a weighted average score across indicators, tiered as 'vital', 'essential' and 'important', above a single universal threshold value, with scoring undertaken annually. The QI programme consists primarily of supportive supervision designed to assist facilities in achieving 'Ideal Clinic'

status. Supportive supervision entails district-level Perfect Permanent Teams for Ideal Clinic Realisation and Maintenance (PPTICRM) providing assistance to facilities to improve checklist scores with the objective of achieving 'Ideal Clinic' status.

Facilities are largely expected to improve checklist scores within existing resources budgeted for routinely as part of provincial Health Department budget allocations. However, if deficiencies are identified in facilities' infrastructure or equipment, additional financial resources were provided to QI programmes address this (Hunter et al. 2017). Until the establishment of a National Health Insurance Fund (NHIF), financial resources required for the implementation of the ICRMP come from the National Health Insurance (NHI) Indirect Grant. The NHI Indirect Grant is aimed at preparing the South African health system for the eventual implementation NHI. Funding for assisting facilities to improve ICRMP quality scores falls under the 'Non-Personal Services Component'[37]. However, a specific Ideal Clinic sub-component budget item was only introduced in the 2016/17 FY (10m Rand) (Kabane, 2016)[38]. Prior to that, a second component of the NHI Indirect Grant, the 'Health Facility Revitalisation Component', also aiming to improve, rehabilitate and upgrade facilities could be used to improve ICRMP quality scores, which had an allocation of 612m Rand in 2015/16 FY[39]. The Indirect Grant is allocated by the National Department of Health (DoH) to Provincial DoHs[40]. Despite this, interviews with health facility staff highlight persistent financial resource challenges as an issue in improving ICRMP quality (Muthathi & Rispel, 2020). Additionally, procurement delays were cited by PHC managers as a constraint to quality score improvements, as in most cases facility managers place orders with District Health Offices (Muthathi et al. 2020). These supply chain issues contributed to consistent annual underspends on the NHI grants (NDoH, 2021). However, despite noting that some indicators were beyond their direct control, most facility managers acknowledged a relatively high degree of capacity and autonomy in influencing ICRMP indicators (Muthathi et al. 2020).

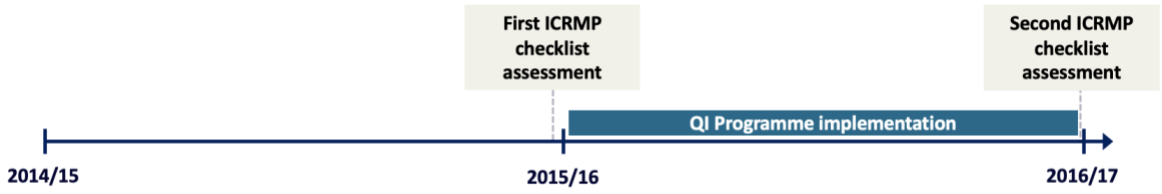
At the start of the 2015/16 Fiscal Year, all PHC facilities in SA undertook an ICRMP checklist self-assessment. A structured roll-out of the QI programme was planned with PHC facilities allocated years in which they would receive support, with those prioritised starting immediately

[37] https://static.pmg.org.za/220323RHAP_-_Submission_to_Appropriations_Committee_11.pdf & https://data.vulekamali.gov.za/dataset/2f19beb5-73fb-41d4-ab92-2fc74b80352f/resource/d4f8bc66-db47-4eb9-ad41-fcec280e6c34/download/2021-21-national-health-insurance-indirect-grant.pdf - accessed 19th Jan 2023.
[38] This Ideal Clinic Sub-Component was scaled up to 26m Rand in 2017/18 (National Department of Health, Annual Report 2017/18).
[39] https://www.da.org.za/2019/10/r5-billion-wasted-on-failed-nhi-pilots-could-have-improved-health-system
[40] Additionally, Provinces had the NHI Direct Grant.

(2015/2016 FY). All PHC facilities were scheduled to receive support over a 3-year period. Although the ICRMP is a National-level initiative, the scheduling for facilities' enrolment in the QI programme took place at Province-level by Provincial DoHs. However, the initial prioritisation of facilities to receive support was largely arbitrary and a systematic process was not followed[41].

Although all facilities were initially assigned a year for QI programme enrolment at the start of 2015/16 FY, some facilities were unenrolled and the scheduled enrolment year of others changed from the start of 2016/17 FY. Therefore, from 2016/17 FY, we cannot rule out facilities' past outcomes impacting their subsequent receipt of support. Consequently, we restrict analysis to the checklist scores corresponding to the start and end of 2015/16 FY. For ease of exposition we refer to the ICRMP assessment taking place at the start of 2015/16 as the 2015 assessment and the one occurring at the end as the 2016 assessment. Additionally, we refer to facilities receiving the ICRMP QI programme supportive supervision from April 2015 through to March 2016 (2015/16 FY) as enrolled. Likewise, all facilities not receiving supportive supervision in this period are referred to as non-enrolled.

Figure 3.1: Chronology of ICRMP implementation



There are numerous ways through which the ICRMP QI programme may have interacted with pre-existing health system features, influencing both overall programme success and leading to a heterogeneity in benefits[42]. Specifically, a number of programme characteristics provide strong priors for why effect heterogeneity may be found according to baseline quality. All facilities enrolled into the programme are provided with additional financial resources, where required, for quality improvement while the nature of the ICRMP quality indicators is such that population characteristics and demand-side factors should not play a strong determining role in quality scores.

[41] This conclusion was reached through numerous indirect discussions with programme managers, which made clear that no explicit criteria was used in the prioritisation of facilities for enrolment.
[42] In any given study there are numerous sub-populations across which samples can be split to test for treatment effects when there is a belief these may be heterogeneous. This is one of several ways issues associated with multiple hypothesis testing can arise increasing the possibility of type I error (List et al. 2019).

As such, whereas heterogeneous treatment effects observed in previous studies may stem from differences in marginal costs of quality improvement or differential incentive sizes, in the SA context these should be largely constant across QI programme recipients. Assuming that quality is a function of capacity and effort, pre-existing quality variation may be caused by variation in facility capacity or facility staff efforts. If capacity is the constraint, we would expect that we may see a greater treatment effect among low baseline performers due to the QI programme provision of guidance materials, support and additional financial resources to meet deficiencies in infrastructure and equipment. However, if pre-existing quality variation is caused by differential effort, then we might expect a larger treatment effect among high baseline performers. Therefore, the ICRMP's idiosyncratic features suggest heterogeneous treatment effects according to baseline performance may be expected. In addition heterogeneity identified may potentially reveal information on the factors determining variation in facility quality, although data constraints prevent a comprehensive investigation of mechanisms.

3. Data

Our primary dataset is the ICRMP data collected during routine self-assessments. This facility-level data provides information on ICRMP QI programme enrolment and ICRMP checklist score information. **Table 3.1** shows that 3,433 PHC facilities were in operation across the nine Provinces of SA in this period. All Provinces – with the exception of Western Cape – had PHC facilities that were both enrolled and not enrolled in the ICRMP during the 2015/16 FY.

Table 3.1: PHC facilities and QI enrolment by Province

Province	Number of PHC Facilities	Proportion of Total PHC Facilities	Not enrolled in QI 2015/16	Enrolled in QI 2015/16	Proportion of PHCs enrolled
Eastern Cape	763	22%	529	234	31%
Free State	221	6%	121	100	45%
Gauteng	370	11%	195	175	47%
Kwazulu-Natal	597	17%	394	203	34%
Limpopo	473	14%	300	173	37%
Mpumalanga	284	8%	198	86	30%
North West	305	9%	203	102	33%
Northern Cape	160	5%	101	59	37%
Western Cape	260	8%	260	0	0%
Total / Average	**3,433**	**100%**	**2,301**	**1,132**	**33%**

The diverse range of ICRMP components provide an overview of structural and process indicators of PHC quality (Donabedian, 1988). These indicators are separated into 10 components (see **Supplementary Appendix C3-1** for full list of indicators). A score of 0 or 1 is assigned to each

indicator based on whether the facility has achieved or passed each measure. The assessments share many characteristics with globally undertaken Service Availability and Readiness Assessments (WHO) and Service Provision Assessments (USAID). Our primary outcome variable is the aggregated ICRMP checklist scores. We aggregate facilities scores and divide by the total potential score i.e. score if facility achieved all indicators. This provides a composite quality index between 0-100 symbolising the percentage of ICRMP quality indicators each facility has satisfied, providing a measure of facilities' ability to deliver quality health care. ICRMP checklist score data is available on 2,381 facilities[43].

Supplementary Appendix C3-2 shows the full distributions of the ICRMP quality scores by baseline quality quartiles for the 2015 and 2016 assessments respectively. They show a high degree of variation in changes in ICRMP quality scores both across and within baseline stratum.

In addition to the ICRMP-specific data we utilise a number of other data sources. The District Health Information System (DHIS) routinely compiles monthly facility-level activity data. For our purposes we primarily use a measure of monthly patient headcount and a measure of facility labour supply; number of clinic nurse work days per month. Both are aggregated to provide annual counts. The South Africa Index of Multiple Deprivation provides socio-demographic characteristics of areas surrounding health facilities (Noble et al. 2013). Finally, we use spatial population distribution data from Afripop (Linard et al. 2012) to create population densities surrounding health facilities[44].

Table 3.2 presents descriptive statistics between enrolled and non-enrolled facilities for the full sample as well as within baseline quality strata. Facility and local-area level characteristics are mostly similar across enrolled and non-enrolled. Further, comparability is even stronger when examining characteristics of enrolled and non-enrolled facilities within stratum of baseline score.

[43] There is general missing data in addition to Western Cape not participating during 2015/16 FY and therefore not having checklist information.
[44] This was done using QGIS.

4. Methods

4.1. Difference-in-Difference-in-Difference

We first estimate the effect of the QI programme on facilities of differing pre-treatment quality within a Difference-in-Difference-in-Difference (DDD) framework. DDD has been used to improve the validity of standard DD models (Rosenbaum, 1987; Yelowitz, 1995; Long et al. 2010), but can also allow the estimation of sub-group effects of a treatment (Stokes et al. 2017). We estimate both stratified regression models across the sub-samples defined by baseline quality and, to preserves the full statistical power of the available sample (Wang & Ware, 2013), a single regression allowing interactions between treatment and baseline quality strata. Identification in both approaches assumes that after controlling for facility-level covariates and fixed unobserved effects, QI programme assignment can be considered random. There is no evidence of geographical clustering of ICRMP enrolment or baseline quality levels (see **Supplementary Appendix C3-3**). This restricts concern to facility-level time-varying heterogeneity. So there may be unobservable time-varying differences between facilities who are enrolled in the QI programme and those that aren't[45].

[45] **Supplementary Appendix C3-3** shows the geographical distribution of facilities by ICRMP enrolment and baseline quality strata across Districts in SA. The table illustrates that facilities of different baseline quality are well distributed geographically with limited evidence of spatial clustering. Further, ICRMP enrolment is well distributed across baseline quality strata within districts. These factors are important for two reasons. If facility ICRMP enrolment were spatially clustered by Districts we may be concerned that these Districts systematically vary in not only their levels of facility quality but also their ability to impact changes in facility quality in ways that may be unrelated to ICRMP enrolment. Second, if Districts were enrolling facilities in the ICRMP by baseline stratum, concerns would arise that authorities may systematically focus on these facilities in other ways. Given there appears to be limited evidence of either issue taking place, suggests geographical controls may not be fundamental to any identification strategy.

Table 3.2: Descriptives Statistics

	All facilities		Lowest Base Q		Low Base Q		High Base Q		Highest Base Q	
	QI enrolled facilities	Non-enrolled facilities	QI enrolled facilities	Non-enrolled facilities	QI enrolled facilities	Non-enrolled facilities	QI enrolled facilities	Non-enrolled facilities	QI enrolled facilities	Non-enrolled facilities
Municipal socio-demographics										
Population <15 years (,000)	202 (282)	194 (275)	211 (299)	178 (271)	177 (256)	182 (270)	199 (282)	208 (279)	221 (291)	212 (281)
Population >60 (,000)	57 (82)	54 (81)	60 (87)	49 (80)	50 (75)	51 (80)	56 (82)	58 (82)	61 (83)	59 (83)
Household size	3.4 (0.5)	3.5 (0.5)	3.4 (0.4)	3.5 (0.5)	3.4 (0.4)	3.5 (0.5)	3.4 (0.5)	3.5 (0.4)	3.5 (0.5)	3.5 (0.5)
Proportion with no schooling	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Proportion population with primary education	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
Proportion population with secondary education	0.18	0.18	0.17	0.17	0.17	0.18	0.18	0.19	0.18	0.18
Proportion with no income	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
Proportion of population black	0.9	0.9	0.9	0.9	0.8	0.9	0.9	0.9	0.9	0.9
Proportion population urban dwelling	0.5	0.4	0.5	0.4	0.5	0.5	0.5	0.4	0.5	0.4
Proportion households with flush toilet	0.5	0.4	0.4	0.4	0.5	0.4	0.5	0.4	0.4	0.4
Proportion households with piped water	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
PHC local geography										
Distance to closest PHC (km)	6.3 (8.2)	6.2 (8.6)	5.1 (5.0)	6.8 (10.0)	6.4 (8.4)	5.7 (8.3)	6.6 (9.6)	6.3 (9.4)	6.7 (8.2)	5.7 (5.1)
Number of PHCs in 10km	6.2 (7.8)	5.4 (7.0)	7.6 (9.7)	5.2 (7.4)	5.9 (7.6)	5.4 (6.8)	5.8 (7.2)	5.6 (6.9)	5.8 (7.2)	5.5 (6.9)
Population within 10km	189,314 (339,965)	170,135 (309,225)	234,052 (421,229)	170,655 (333,963)	169,057 (334,803)	162,653 (299,313)	170,583 (303,540)	169,475 (290,557)	195,905 (317,867)	179,006 (304,656)
Number of PHCs in Local Municipality	23 (18)	26 (22)	26 (20)	26 (24)	25 (18)	23 (20)	24 (20)	27 (23)	20 (14)	25 (22)
DHIS										
Monthly professional nurse clinic working days	103 (102)	84 (95)	102 (90)	74 (70)	98 (97)	80 (72)	107 (118)	93 (144)	105 (97)	90 (80)
Monthly patient headcount	3,373 (3,287)	2,594 (2,366)	3,221 (2,968)	2,298 (1,918)	3,038 (2,683)	2,530 (2,230)	3,544 (3,833)	2,634 (2,347)	3,588 (3,373)	3,045 (2,980)
Monthly new fully immunised <1 year	22.8 (22.8)	18.8 (17.7)	21.5 (22.5)	18.0 (18.3)	20.8 (19.8)	17.9 (14.8)	23.4 (23.6)	19.2 (18.4)	24.8 (24.6)	20.3 (19.0)
Monthly antenatal 1st visit before 20 weeks (%)	64.9 (11.7)	65.2 (11.4)	63.9 (12.0)	64.4 (11.5)	64.7 (12.2)	65.1 (11.3)	65.5 (11.5)	65.3 (11.9)	65.3 (11.3)	66.6 (10.9)
Monthly cervical cancer screening	19.4 (20.1)	16.1 (16.6)	16.9 (14.7)	14.3 (14.3)	17.7 (17.5)	16.8 (19.3)	21.0 (23.3)	16.2 (15.9)	21.0 (21.4)	17.9 (16.9)
Monthly HIV positive new eligible client initiated on IPT	11.3 (12.6)	10.6 (15.5)	9.9 (9.4)	9.4 (12.9)	9.7 (9.8)	10.1 (12.8)	12.2 (14.9)	10.6 (15.5)	12.7 (13.8)	13.0 (20.6)
Monthly Tracer item stock out rate	25.0 (30.7)	22.9 (29.9)	24.7 (30.8)	23.1 (31.3)	27.4 (32.1)	23.6 (29.8)	25.3 (30.9)	25.5 (31.1)	22.8 (29.1)	18.7 (26.2)
South African Index of Multiple Deprivation										
SES 1st quantile	0.06	0.05	0.03	0.08	0.07	0.05	0.09	0.03	0.05	0.02
SES 2nd quantile	0.09	0.08	0.10	0.08	0.09	0.08	0.10	0.09	0.08	0.07
SES 3rd quantile	0.13	0.13	0.14	0.12	0.10	0.14	0.13	0.12	0.15	0.15
SES 4th quantile	0.23	0.28	0.21	0.24	0.23	0.29	0.19	0.28	0.27	0.34
SES 5th quantile	0.49	0.46	0.52	0.48	0.51	0.44	0.48	0.49	0.45	0.42

We estimate regression models for each of the strata defined by baseline quality g as:

$$Y_{fgt} = \alpha_{fg} + \gamma_g QI_{fgt} + \eta_{gt} + \beta_g \boldsymbol{X}_{fgt} + \varepsilon_{fgt} \quad (1)$$

Where $f = 1, \ldots, F$ indexes facilities, and $t = 1, \ldots, T$ indexes time periods. We split the sample of facilities into $g = 1, \ldots, 4$ groups, corresponding to the quartiles of baseline quality. As enrolment occurs at facility-level we have enrolled and non-enrolled facilities within each stratum[46]. Y_{fgt} represents the aggregate ICRMP quality score for each facility f part of strata g for every period t; QI_{fgt} denotes enrolment in the ICRMP QI programme; γ_g is the treatment effect of interest for a given group g, $\boldsymbol{X}_{fgt}$ represents a vector of time-varying facility- and small area-level covariates that may affect ICRMP quality score attainment and enrolment in the QI programme. α_{fg} are time-invariant facility-specific omitted factors impacting quality scores. We also include η_{gt} to capture the general secular trend in the ICRMP quality scores, controlling for unobserved variables that evolve over time across all facilities. This may capture differences in information about the calculation of quality scores facilities are given across years. These models allow the effect of all the control variables on the outcome to differ by baseline quality.

For the interaction model, we estimate:

$$Y_{ft} = \alpha_f + \gamma QI_{ft} + \sum_{g=2}^{4} \gamma'_g (BaseQul_f * QI_{ft}) + \eta_t + \sum_{g=2}^{4} \phi_g (BaseQul_f * \eta_t) + \beta \boldsymbol{X}_{ft} + \varepsilon_{ft} \quad (\mathbf{2})$$

All variables remain the same as (1) but instead of estimating over g samples, we include $BaseQul_f$ representing facilities' pre-enrolment ICRMP quality score. $BaseQul_f$ is time-invariant and therefore its main effect cannot be estimated by FE, but we can estimate it's interaction the effect with the QI programme enrolment, on facility quality scores[47]. γ is the treatment effect for the lowest baseline quality strata. γ'_g is then a second-order interaction outlining the difference in the effect of the QI programme for quality strata relative to the effect

[46] **Supplementary Appendix C3-4** shows there is sufficient treatment variation across within each strata of baseline quality to allow for estimation of the effect of ICRMP enrolment for all strata.

[47] In most models with interaction terms it is important to include main effects for interpretation purposes but when estimating FE models this is not possible.

on the lowest baseline strata. We avoid the common strong assumption of linear interaction effects in regression-based multiplicative interaction models by estimating separate parameters for the effect of the QI programme for each baseline quality quartile (Hainmueller et al. 2019)[48][49]. Finally, we allow for differential time trends in the outcome by baseline quality strata as implied by **Table 2**[50]. Identification via DDD takes the change in the non-enrolled facilities as the counterfactual change for enrolled facilities. Therefore, allowing for differential time trends between strata allows unobserved variables that evolve over time to differ by strata of baseline quality. For instance, facility-level funding changes over time may be related to relative quality levels rather than a constant change across all facilities. However, this specification still requires that the counterfactual trends of facilities enrolled in the QI programme and non-enrolled facilities within the same quality strata are parallel

An important distinction should be made between our models and lagged dependent variable (LDV) specifications[51]. Although we include the outcome as an explanatory variable, a key distinction is that baseline quality is time-invariant. Consequently, the models do not suffer from Nickell bias associated with dynamic models with fixed effects (Nickell, 1981) (See **Supplementary Appendix C3-5** for detail). Because only facilities whose treatment status changes contribute to the estimation of the treatment effects, γ and γ_g', our estimates provide the average treatment effect on the treated (ATT) for each sub-group. The model assumes no heterogeneity in the effects of the covariates included in the model but not interacted. Our DDD model, therefore, does not estimate full conditional average treatment effects on the treated (CATTs) as they do not allow the effects of all covariates to vary across the four strata (Gibbons et al. 2019).

Two sets of confounding factors must be considered to make causal claims with interaction analysis; treatment-outcome and moderator-outcome confounders (VanderWeele, 2015). However, if only potential confounding factors of the relationship between enrolment in the QI programme and the ICRMP checklist score are considered, this is sufficient for identifying effect

[48] Specifically, the linear interaction effect assumes the effect of treatments varies at a constant rate across levels of the moderator examined. In this case this would imply $\partial Y/\partial QI = \alpha + \gamma BaseQul_f$.

[49] Hainmueller et al. (2019) also highlight the requirement of common support over moderators with such interactive models. However, common support issues are less pressing with discrete moderator variables and reassuringly **Table 4** showed strong common support of the treatment across the moderator strata.

[50] Time-invariant covariates whose effects vary over time must be included in the model as FE estimation only controls for time-invariant variables with time-invariant effects (Allison, 2009).

[51] Unlike an LDV specification, our dependent variable remains the within facility change in quality score.

heterogeneity across the strata examined. The coefficients on the interaction terms can be considered the causal effects of the QI programme within each stratum defined by baseline quality score and the differences a measure of heterogeneity in this causal effect. We control for facility's patient headcount and a measure of facility staffing (nurse working days) as **Table 3.2** revealed differences between enrolled and non-enrolled facilities in facility size by these metrics. It is possible that facility size, patient volume, staffing levels and related activity intensity may influence the ability of facilities to attend to quality deficiencies highlighted in the ICRMP assessments.

4.2. Changes-in-Changes

A limitation with the above approach is the counterfactuals, and therefore treatment effects identified, are functions of how subgroups are composed i.e. treatment effect heterogeneity is a parametric function of the number of groups, allowing identification of $E[Y_i^T - Y_i^C|g] = \gamma_g$ with constant treatment effect within these groups imposed. Therefore, while the DDD estimates give an indication of the pattern of treatment effect heterogeneity, it cannot identify the treatment effect of the QI programme across the full distribution of quality scores in our data. The DDD model also relies on strong linear additive separability assumptions. Although our DDD models allow the impact of time and treatment effect to vary, this is done in a restricted way, across strata of pre-treatment quality. Additionally, additivity assumptions imply that facility returns to the QI programme are not affected by unobserved facility characteristics such as staff effort or managerial quality. Further, the model still implies additive separability between the treatment and unobservables. Therefore, we still require conditional mean independence of the unobservables and enrolment status within pre-treatment strata.

To avoid the impact of modelling specifications on heterogeneity identified and relax the additive separability assumptions, we implement the Changes-in-Changes (CC) model proposed by Athey and Imbens (2006). With CC, we are able to estimate the full counterfactual distribution of quality scores that QI enrolled facilities would have achieved if they had not been enrolled. The CC model is based on a single non-separable equation allowing for arbitrary interactions between treatment and unobservable characteristics through a structural function $h(.)$. This allows the distribution of unobservables to be arbitrarily different across enrolled and non-enrolled facilities i.e. $F_U|QI = 1$ does not need to be the same as $F_U|QI = 0$.

The intuition for identification and estimation for the CC model is similar to DD, with some important distinctions. Estimation of the CC model requires comparing the quality score

cumulative distribution functions (CDFs) of the four treatment-by-period groups, rather than just the first moments. The change in the distribution of the quality scores for unenrolled facilities over 2015/16 is used to estimate the counterfactual CDF of the quality scores enrolled facilities would have achieved over the same period had they not been enrolled, $F_{Y_1^T(0)}$. Specifically, $F_{Y_1^T(0)}$ is identified by (Athey & Imbens, 2006):

$$F_{Y_1^T(0)}(Y) = F_{Y_0^T}\left(F_{Y_0^C}^{-1}\left(F_{Y_1^C}(Y)\right)\right)$$

Where $F_{Y_0^T}$ is distribution of the pre-treated outcomes for the treated group, $F_{Y_0^C}^{-1}$ is the inverse of the distribution of the pre-treated outcomes for the control group and $F_{Y_1^C}$ is the distribution of the post-treatment outcomes for the control group. As the three distributions on the RHS are observable, the LHS is identified.

Having constructed the full counterfactual distribution of quality scores, we can estimate the full set of quantile treatment effect on the treated (QTT). For example, the QTT for the 20th percentile is the difference in the potential outcome distributions for enrolled facilities at the 20th percentile of the quality score distribution without QI enrolment. Consequently, the QTT at each quantile q is calculated as the difference between the inverse of the constructed counterfactual CDF, $F_{Y_1^T(0)}(Y)$, and the inverse of the observed CDF for the enrolled facilities post-treatment:

$$QTT(q) = F_{Y_1^T(1)}^{-1}(q) - F_{Y_1^T(0)}^{-1}(q) \qquad (\mathbf{3})$$

While the ATTs given by the DDD provide treatment effect estimates for clearly defined sub-groups, CC identifies treatment effect heterogeneity across quantiles of the counterfactual outcome distributions[52]. Borrowing a phrase from Djebbari & Smith (2008) the QTT reflects 'impacts *at* quantiles rather than *on* quantiles'.

[52] Because of how the counterfactuals are identified with the CC, it is not possible, without strong assumptions, to clearly identify units or groups to which the treatment effect relates and reference can only be made to points in the distribution. In order for the QTT to reflect the treatment effect for a particular unit an implausibly strong assumption of rank preservation needs to hold. This would require that each facility maintains its rank across the (potential) outcome distributions.

The key assumptions underlying construction of the counterfactual distribution are largely generalisations of DD assumptions. Quality scores (QS), are assumed to be generated by an unknown non-separable function; $QS = h(U,T)$. Where U is a vector capturing unobservable facility characteristics and T is time. As QS does not depend on enrolment status, enrolled and non-enrolled facilities in the pre-treatment period with the same quality score, QS', must have identical $U = u$. There are two primary assumption difference between DD and CC (see **Supplementary Appendix C3-6** for full technical assumptions). CC requires the function $QS = h(U,T)$ be strictly monotone increasing, i.e. $\Delta h(U) > 0$, so higher unobservables result in strictly higher outcomes. This is non-restrictive in our case as it is natural to assume that greater effort or capacity result in higher quality scores. Additionally, the distribution of unobservable facility characteristics are time invariant within both enrolled and non-enrolled groups, $U \perp T|QI$. As such, quality scores may change over time through the previously outlined production function, $QS = h(U,T)$, but because the within group distribution of U is time-invariant, the change can only reflect a time effect. Therefore, non-enrolled facilities at the same quantile q of their respective outcome distribution pre-, $F_{Y_0^C}$, and post-treatment, $F_{Y_1^C}$, may have different outcomes due to an effect of T but must have identical $U = u$. Therefore the evolution in outcomes for non-enrolled facilities at quantile q provides a counterfactual for the evolution in outcomes for enrolled facilities with pre-treatment quality scores, QS', had these facilities not been enrolled. The QTT can then be calculated for the full support of quality scores of enrolled facilities post-enrolment.

Although pre-treatment we may have $F_{Y_0^T} \neq F_{Y_0^C}$ due to $F_{U_T} \neq F_{U_C}$, the assumed time-variance of the distribution of unobservables – $F_{U_{T0}} = F_{U_{T1}}$ and $F_{U_{C0}} = F_{U_{C1}}$ – means, in the absence of treatment, both groups would have seen the same growth in quality score; $QS = h(.,T)$. Consequently, unlike DD frameworks, the CC identifying assumption is invariant to transformations of the outcome variable as common growth in the outcome is assumed rather than parallel trends (Lechner, 2011). Specifically, the assumption is that the change in quality scores over 2015/16 is the same for facilities with pre-treatment quality QS' in both enrolled and non-enrolled groups, in the absence of the QI programme.

Not including relevant time-varying covariates can cause differences in the production functions between the enrolled and non-enrolled groups that map the unobservables to outcomes in a given period, and would lead to inconsistent estimates of QTTs. Following Melly and Santangelo (2015),

we use an extension to the CC model allowing the incorporation of covariates. This estimation approach is based on a semi-parametric quantile regression with identification of the counterfactual distribution following from the Athey & Imbens (2006) approach described above (**Supplementary Appendix C3-7**)[53]. We estimate the near full set of QTTs, resulting in 99 QTTs.

5. Results

5.1. Difference-in-Difference-in-Difference

Table 3.3 presents average facility quality scores by baseline quartiles, illustrating the large pre-existing variation in scores. In 2015, the average score for facilities in the 75th percentile of the baseline quality score distribution was 79% higher than the average score of facilities in the 25th percentile. However, average quality over the period converges across facilities with different baseline scores, with the average score of facilities in the bottom 25th percentile at baseline increasing over 15 points by 2016 while scores of facilities in the 75th percentile decreased.

Table 3.3: ICRMP Checklist Scores Trends

	Lowest baseline quality facilities	Low baseline quality facilities	High baseline quality facilities	Highest baseline quality facilities	All facilities
Average (SD) aggregate quality 2015	40.1 (6.0)	51.7 (2.4)	59.8 (2.4)	71.8 (6.2)	55.9 (12.5)
Average (SD) aggregate quality 2016	55.7 (14.6)	59.9 (13.9)	63.0 (16.1)	67.5 (17.3)	61.5 (16.1)
Average aggregate score change	15.6	8.2	3.2	-4.3	5.6
Observations	596	595	595	595	2,381

Table 3.3 also reassures that we do not need to be concerned about potential ceiling effects on further quality improvements for the highest baseline performers.

In lieu of data to examine pre-treatment trends in ICRMP quality scores for enrolled and non-enrolled facilities, we note the similarities in pre-treatment levels of the scores and facility characteristics (**Supplementary Appendix C3-8**). The largest within quartile pre-treatment difference in average quality scores between enrolled and non-enrolled facilities is 1.6, for the lowest baseline performing quartile. Additionally, the distributions of within quartile quality scores are almost identical.

[53] We use the cic STATA command (Melly & Santangelo, 2015) to obtain conditional CC estimates.

Tables 3.4 and **3.5** present the results of equations (1) and (2) respectively. For all specifications, we clustered our standard errors at the facility-level to address concerns of serial correlation (Bertrand et al. 2004). The estimated coefficients capturing the effect of the QI programme for the stratified and interaction regressions are, as expected, almost identical[54]. From **Table 3.4** we see that the effect of the ICRMP is positive and significantly different from 0 for all baseline quality strata. The impact of the ICRMP increases with the facilities' baseline quality. The effects represent 29% of the average score for the highest performing and 15% of lowest performing strata at baseline (**Table 3.3**), showing the estimated treatment effects are regressive even from a proportional perspective. The POLS and FE estimates are almost identical (**Table 3.5**), indicating that controlling for facility fixed effects appears to have little impact on effect estimates.

Notably, there are vastly differential time trends, as for facilities with lowest/low baseline quality, there is a general increase in the quality score over time, while the opposite is true for facilities with high/highest baseline quality. The estimated time trend is small when examining all facilities together (**Table 3.4**). Subsequently, the ratio of the effect of the QI programme and the time effect would lead to the conclusion that the QI programme has an exceptional effect size. However, once the time trend is allowed to vary by baseline quality it becomes clear that the small average common trend is composed of large and opposing trends across facility strata types. It is clear the time effect parameter is large in absolute size within these strata. This suggests sizeable over time changes in quality scores may not be uncommon and that more evidence on the variance of intra-facility quality may be beneficial.

Table 3.4: FE Stratified Regressions

	All facilities	**Lowest Baseline Quality**	**Low Baseline Quality**	**High Baseline Quality**	**Highest Baseline Quality**
QI programme	10.06***	6.103***	6.596***	15.77***	20.75***
	(0.668)	(1.318)	(1.115)	(1.153)	(1.190)
Year	1.092*	13.70***	5.066***	-4.327***	-15.62***
	(0.504)	(0.730)	(0.753)	(0.834)	(0.951)
R^2	0.175	0.539	0.318	0.27	0.394
F	170.3	173.9	74.61	60.18	84.09
Observations	4,727	1,178	1,181	1,181	1,187

[54] A benefit of the stratified approach is the ease of interpretation of coefficients which are given in absolute terms rather than relative to the baseline strata (lowest baseline quality).

Table 3.5: Multiplicative interaction models		
	POLS	FE
QI prog Lowest Baseline Quality (γ)	7.702***	6.064***
	(1.197)	(1.312)
QI prog Low Baseline Quality relative to γ	-0.251	0.651
	(1.592)	(1.722)
QI prog High Baseline Quality relative to γ	7.665***	9.671***
	(1.612)	(1.747)
QI prog Highest Baseline Quality relative to γ	12.21***	14.74***
	(1.578)	(1.769)
Year lowest baseline quality	13.21***	13.57***
	(0.712)	(0.734)
Year low baseline quality	-8.204***	-8.378***
	(1.012)	(1.044)
Year high baseline quality	-17.27***	-17.99***
	(1.073)	(1.109)
Year highest baseline quality	-28.10***	-29.10***
	(1.122)	(1.202)
R^2	0.547	0.403
F	358.3	151.5
Observations	4,727	4,727

5.2. Changes-in-Changes

Figures 3.2(a) and 3.2(b) presents results of the CC model with covariates[55], evaluating the QTT at every percentile point of the distribution (without covariates presented in **Supplementary Appendix C3-9**)[56]. The estimated effect of the QI programme is positive across the whole of the quality score distribution, with the estimated confidence bands including zero only for the lowest fraction of the distribution. The pattern of treatment effects reflects that of the DDD results, with the effect of the QI programme increasing along the distribution of pre-treatment quality scores. Including covariates reduces the range of QTEs observed from 2.4 – 12.6 to 4.4 – 11.8. Despite this marginal impact on the point estimates, in the model with covariates we cannot reject the hypothesis that all QTEs are equal to the median QTE, whereas without covariates this was rejected at the 1% level. Accompanying tables for the CC figures and a direct comparison of the QTEs with and without covariates is found in **Supplementary Appendix C3-10**. All figures include 90% confidence bands based on a non-parametric bootstrap with 1,000 replications.

Although not directly comparable, the 90% confidence bands from the CC estimates include the point estimates of the treatment effect for all but the final stratum from the DDD results.

[55] Namely the average monthly nurse working days and average monthly patient headcounts at facility.
[56] Note the distribution is that of the pre-treatment quality for the enrolled facilities, therefore, the quantiles along which the treatment effects (QTTs) are calculated and presented are with respect to this CDF.

However, due to the different estimands examined this restricts the value in making direct effect comparisons across the methods.

Figure 3.2(a): Quantile treatment effects - CC model with covariates

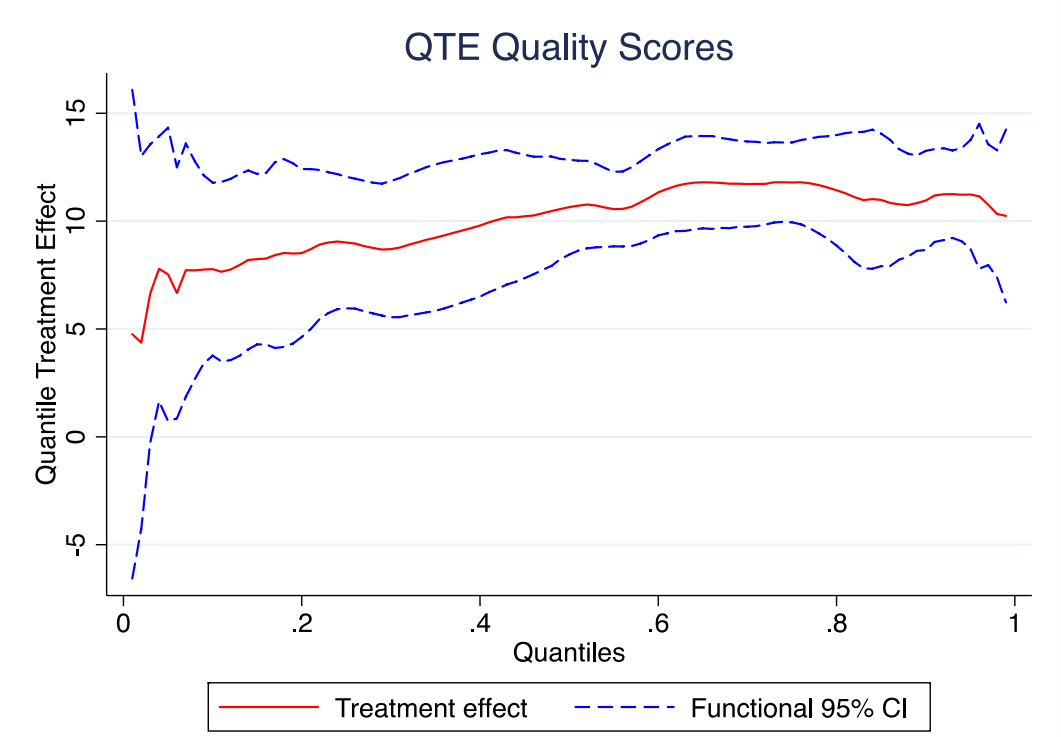
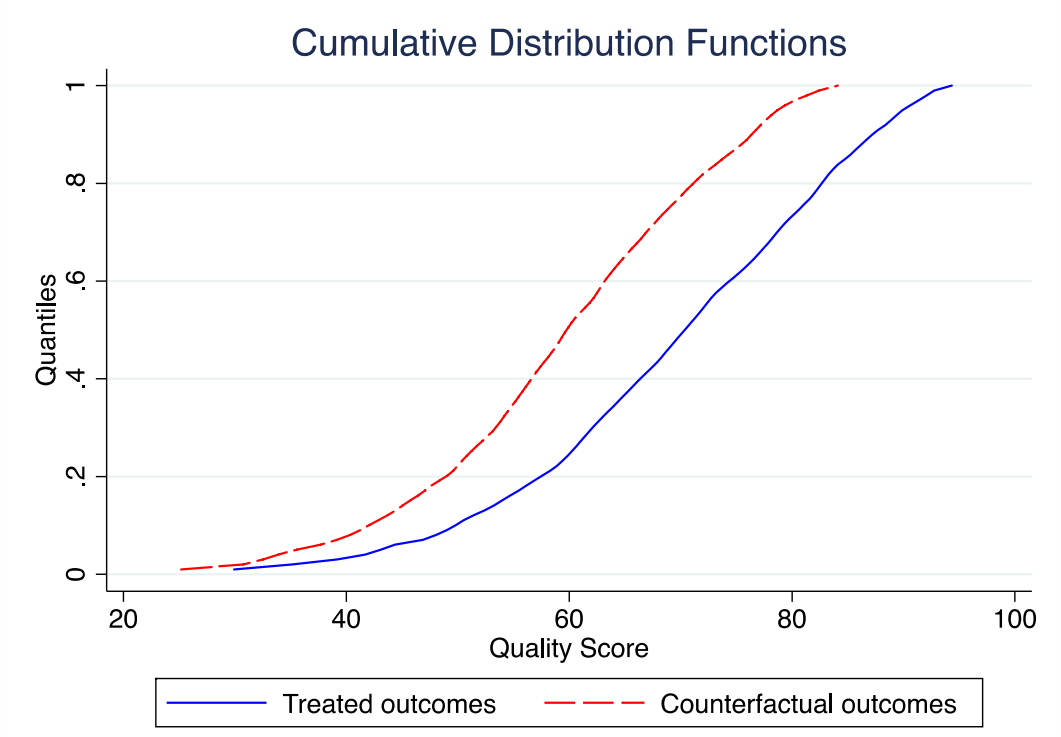


Figure 3.2(b): Cumulative distribution functions - CC model with covariates



6. Sensitivity Checks

Policy endogeneity may undermine the ability to exploit policy variation for identification (Besley & Case, 2002). Informal interviews with ICRMP managers suggest timing of QI programme enrolment may have been influenced by views on the performance of facilities. Facilities viewed as performance improving may have been prioritised for enrolment in the programme, partly to encourage continued quality gains. If already improving facilities are enrolled in the QI programme, this will upward bias the estimated effects of the programme. Zeldow & Hatfield (2021) illustrate how the parallel trends assumption can be violated if 1) the unobserved fixed effects are not balanced and the effects of the unobserved fixed effects are not constant or 2) there are unobserved time-varying confounders with differentially evolving trends or effects. In our setting, if there is a prioritisation of facilities based on recent performance improvements, resulting from higher effort levels or capacities, this will result in unobserved differences between enrolled and unenrolled facilities. Below we examine three approaches that allow for, control or reveal the potential for self-selection and ensuing unobserved differences between enrolled and unenrolled facilities.

6.3. Lagged Dependent Variable Model

Due to these questions around the parallel trend assumption, we estimate a lagged dependent variable (LDV) model which allow the effects of unobserved confounders to change over time (O'Neil et al. 2016; Ding & Li, 2019). LDV models assume selection based on past outcomes and, therefore, imply unconfoundedness conditional on the lagged values of the outcome:

$$Y_1^T(0), Y_1^C(0) \perp QI_f | Y_{f0}, X_{ft}$$

If it is the case that enrolment in the QI programme is determined by lagged dependent variables then fixed effects estimates are not consistent (Angrist & Krueger, 2000). We run an OLS which controls for the lagged outcome level and as before, allows for differential policy effects in the four strata based on pre-policy quality score quartiles:

$$Y_{ft} = \alpha + \sum_{g=1}^{4} \gamma_g (BaseQul_f * QI_{ft}) + \theta Y_{ft-1} + \beta X_{ft} \quad (4)$$

The estimated effects from the LDV models are very similar to those of the DDD models (**Table 3.6**). It has been noted that DDD and LDV estimates have a bracketing relationship in linear models (Angrist & Pischke, 2009)[57]. Applying these bounding properties we have $6.06 \leq \gamma_1 \leq 7.43$, $6.72 \leq \gamma_2 \leq 6.95$, $15.19 \leq \gamma_3 \leq 15.74$, $19.77 \leq \gamma_4 \leq 20.80$[58].

[57] Because $0 < \theta < 1$ and $\bar{Y}_{ft}^{T} > \bar{Y}_{ft}^{C}$, the estimated effect of the QI programme is larger from the LDV model than DDD. If treatment assignment is positively selected on (unobserved) fixed effects then $plim\ \hat{\gamma}_{LDV} \geq \gamma$, where $\hat{\gamma}_{LDV}$ is the estimated effect from the LDV model and γ is the true treatment effect. Similarly, if treatment assignment is positively selected on lagged outcomes $plim\ \hat{\gamma}_{DD} \leq \gamma$. Therefore, $plim\ \hat{\gamma}_{DD} \leq \gamma \leq plim\ \hat{\gamma}_{LDV}$. If $\bar{Y}_{ft}^{T} < \bar{Y}_{ft}^{C}$ the reverse relationship holds. In our setting, these properties will hold for each strata based on pre-treatment quality.

[58] For γ_1 and γ_2 we have $plim\ \hat{\gamma}_{LDV} > plim\ \hat{\gamma}_{DD}$ while for γ_3 and γ_4 the inverse holds.

Table 3.6: Lagged Dependent Variable estimation

	(1)	(2)	(3)	(4)
QI prog Lowest Baseline Quality (γ)	7.432***	6.818***	6.618***	7.559***
	(1.226)	(1.158)	(1.159)	(1.150)
QI prog Low Baseline Quality relative to γ	-0.481	0.0279	0.157	-0.239
	(1.676)	(1.582)	(1.580)	(1.563)
QI prog High Baseline Quality relative to γ	8.217***	8.009***	7.996***	7.627***
	(1.672)	(1.580)	(1.579)	(1.561)
QI prog Highest Baseline Quality relative to γ	13.25***	12.99***	13.06***	12.21***
	(1.672)	(1.578)	(1.575)	(1.559)
Lagged ICRMP Quality Score	0.289***	0.249***	0.248***	0.256***
	(0.0614)	(0.0581)	(0.0580)	(0.0573)
Proportion Population no education		13.08	13.62	7.420
		(10.36)	(10.38)	(10.39)
Proportion Population no income		-45.41***	-48.95***	-44.58***
		(10.34)	(10.37)	(10.34)
Proportion Population Urban		4.540	4.538	4.107
		(3.115)	(3.141)	(3.179)
Household Size		7.239***	7.296***	7.371***
		(0.886)	(0.889)	(0.920)
Proportion Households with Flush/Chemical Toilet Access		22.50***	21.82***	19.26***
		(3.936)	(3.974)	(3.998)
Proportion Households with Piped/Borehole Water Access		-16.75***	-16.69***	-14.61***
		(2.170)	(2.176)	(2.182)
SES quintile		1.491***	1.302***	1.508***
		(0.267)	(0.278)	(0.297)
Mean Monthly Nurse Working Days			0.00574*	0.00610*
			(0.00321)	(0.00317)
Mean Monthly Patient Headcount			0.000106	-0.0000688
			(0.000134)	(0.000138)
Distance to closest other PHC facility (km)				0.0363
				(0.0354)
Number of PHC facilities within 10 km				-0.766***
				(0.108)
Population within 10 km				0.0000189***
				(0.00000252)
District				0.0326
				(0.0199)
Constant	41.76***	29.60***	31.19***	28.52***
	(2.524)	(5.796)	(5.807)	(5.785)
R-sq	0.264	0.349	0.352	0.370
F	106.5	84.26	75.00	65.52
N	2381	2373	2365	2365

Standard errors in parentheses, *p<0.1, **p<0.05, ***p<0.01

6.4. Matching on Pre-treatment Quality Performance

However, as we only have outcome data for one pre-treatment period available, this restricts our ability to examine and condition on past outcomes and doesn't provide evidence to examine if

enrolled facilities were experiencing a pre-treatment increase in quality[59][60][61]. In order to gain more insight into pre-treatment facility performance we utilise DHIS data on facility activity. This represents the most comprehensive formal data on facility performance prior to the implementation of the ICRMP checklist. **Table 3.7** outlines the DHIS facility activity variables.

Table 3.7: DHIS variables (data from January 2013 – June 2015)
Monthly children <1 years fully immunised
Monthly patient head count
Monthly patients seen by professional nurse
Monthly professional nurse days at facility
Monthly rate of ANC 1st visit before 20 weeks
Monthly number of cervical cancer screenings >30 years
Monthly number of Measles 1st dose
Monthly number of RV 2nd doses for <1 years
Monthly number of HIV+ new client initiated on IPT
Monthly tracer item stockout rate

As a means of addressing possible non-parallel trends between enrolled and non-enrolled facilities we match facilities on pre-baseline DHIS variables under the assumption that facilities with similar trends in these observables have time-varying unobservables which evolve similarly[62]. If we believe these pre-treatment measures of activity are correlated with factors that may impact facility quality measures then matching improves comparability of facilities. Specifically, returning to potential concerns around differences in facility effort or capacity, matching on these activity factors should increase comparability between enrolled and non-enrolled facilities, reducing reliance on the assumption that the effect of these unobservables is constant over time.

Because matching cannot distinguish systematic trend differences from short-term fluctuations due to random shocks, we aggregate monthly DHIS variables to quarterly averages (Linder and McConnell, 2018). **Supplementary Appendix C3-11** presents graphs of the trends in the pre-treatment DHIS variables for both enrolled and non-enrolled facilities. Although enrolled facilities perform a slightly larger number of services for each of the activities listed, the trends in activity are largely identical.

[59] Further, if there is differential trends in pre-treatment quality and this is caused by facility managerial quality or some other facility unobserved effects, these are now no longer controlled for.
[60] This also prevents us from pursuing a modelling strategy which simultaneously controls for unobserved fixed effects and past outcomes. Although this saves us having to address the known challenges with such models (Nickell, 1981).
[61] It is worth clarifying that although the DDD specifications captured a measure of the baseline quality measure in the RHS (namely a categorical variable indicating pre-treatment quartile) this only allowed the parameters of the treatment and time trend to vary across these groups and remains distinct from the LDV specification.
[62] This is a slight adaptation of the popular approach which matches on pre-treatment outcomes then applies DD which is used to address non-parallel trends (Blundell & Dias, 2009).

Given the impracticality of matching on numerous covariates, we match on propensity scores (PS) calculated from these activity measures (Rosenbaum and Rubin, 1983). We estimate PS's using both levels and trends in the DHIS variables (see **Supplementary Appendix C3-12** for detail). Balance of the propensity scores and overlap is tested using the standard block method (Imbens, 2004; Becker & Ichino, 2002) and shows good balance between enrolled and non-enrolled facilities. Further, covariates are shown to be largely balanced between enrolled and non-enrolled facilities within blocks of the propensity scores. We follow Heckman, Ichimura and Todd (1997), implementing a kernel propensity-score matching difference-in-difference estimator. Estimation augments the standard DD estimator, whereby instead of controlling for covariates, X, in a regression framework, each enrolled facility is matched to the whole sample of non-enrolled facilities based on weights defined by the propensity score.

The results of the matched DDD analysis, presented in **Table 3.8** suggest that controlling for facilities' previous performance, as measured by DHIS indicators, does not substantially alter the estimated effect of the effect of the QI programme.

Table 3.8: Kernel Propensity Score Matching Difference-in-Difference

	Matched on Levels				Matched on Trends			
	Lowest quality	Low quality	High quality	Highest quality	Lowest quality	Low quality	High quality	Highest quality
2015								
Non-enrolled	40.20	51.57	59.726	72.02	40.06	51.46	59.963	71.499
Enrolled facilities	41.20	51.85	59.85	71.754	41.20	51.85	59.85	71.754
Difference	1.006	0.281	0.125	-0.266	1.148*	0.393	-0.113	0.256
	(0.677)	(0.282)	(0.267)	(0.710)	(0.674)	(0.289)	(0.282)	(0.603)
2016								
Non-enrolled	53.95	57.45	58.57	57.894	53.56	56.49	58.42	58.731
Enrolled facilities	61.12	63.87	71.459	77.267	61.12	63.87	71.459	77.267
Difference	7.163***	6.418***	12.889***	19.373***	7.553***	7.380***	13.04***	18.536***
	(1.881)	(1.585)	(2.083)	(1.389)	(1.660)	(1.538)	(1.695)	(1.326)
DD	6.157***	6.137***	13.638***	19.639***	6.406***	6.987***	13.152***	18.28***
	(1.929)	(1.656)	(2.105)	(1.606)	(1.721)	(1.595)	(1.771)	(1.388)
Observations	818	956	994	1,052	798	942	964	1,042

Standard errors in parentheses, *p<0.1, **p<0.05, ***p<0.01

6.5. Alternative control group

At the start of 2015/16 FY, every facility was assigned a year for enrolment in the QI programme[63]. If enrolment year decisions were based consistently on beliefs about which facilities were most likely to benefit, this would imply that facilities assigned for enrolment in the subsequent year (2016/17) would most resemble facilities enrolled during our period of analysis in relevant unobservables. Therefore, these facilities should constitute a more valid control group than all non-enrolled facilities. If this were the case, and our previous analyses had not adequately captured the aspects of the unobservables which impact enrolment and quality scores, we would expect estimates of the effect of the QI programme with this more targeted control group to be smaller than our previous estimates. Unlike the other specification tests, this test utilises policy-makers own prioritisation criteria in order to maximise the comparability of enrolled and non-enrolled facilities. We rerun equation **(2)** with this targeted control group.

As can be seen from **Table 3.9**, although the estimated effects of the QI programme using the restricted control group are marginally smaller compared to when estimated using the full sample, this difference is minor. This suggests facilities selected for prioritisation are not significantly different than those earmarked for later enrolment. This either allays concerns that prioritised facilities had some unobserved characteristics related to both enrolment and changes in quality scores, or that if these unobservables are present, such as differences in facility capacity, they do not have a significant impact on facility ICRMP quality scores.

[63] Originally the plan was for all facilities to be enrolled within 3 years. Therefore, approximately 1,000 facilities were assigned for enrolment per year between 2015/16-2017/18.

Table 3.9: Multiplicative interaction model with restricted sample		
	FE (full sample)	**FE (restricted sample)**
QI prog Lowest Baseline Quality (γ)	6.064***	5.431***
	(1.312)	(1.501)
QI prog Low Baseline Quality relative to γ	0.651	0.328
	(1.722)	(2.034)
QI prog High Baseline Quality relative to γ	9.671***	8.615***
	(1.747)	(2.060)
QI prog Highest Baseline Quality relative to γ	14.74***	13.68***
	(1.769)	(2.077)
Year lowest baseline quality	13.57***	14.18***
	(0.734)	(1.036)
Year low baseline quality	-8.378***	-8.075***
	(1.044)	(1.507)
Year high baseline quality	-17.99***	-16.974***
	(1.109)	(1.556)
Year highest baseline quality	-29.10***	-28.010***
	(1.202)	(1.625)
R^2	0.403	0.413
F	151.5	117.5
Observations	4,727	3,458

*R-squared reported is with-in value.

7. Discussion

7.1. Main Findings

The examination of heterogeneous treatment effects are particularly important in circumstances where policy objectives value reducing inequality, and therefore weight is given to the distribution of an outcome. There is increasing recognition of the importance of addressing the large variations in the quality of health care provision and improving equity in access to high-quality health care. Understanding distributional treatment effects of QI programmes is vital to guide the design, implementation and adjustment of programmes to ensure they contribute towards improving equitable access to high-quality health care. While a common objective of QI programmes is to reduce variation in the quality of health care, the literature on evaluations exploring effect heterogeneity and the consequences for variation in the quality of health care is currently limited. This is more surprising given the growing recognition that a potential unintended consequence of programmes can be to exacerbate pre-existing disparities in health care quality, with the design of a number of QI programmes including features intended to reduce this risk (Eijkenaar, 2013).

This chapter explores heterogeneous effects of a QI programme in SA, attempting to illuminate the distributional consequences for health care quality. First, we employ a DDD design providing insight into the existence and direction of heterogeneous effects. We then estimate the full

counterfactual distribution of the enrolled facilities across baseline quality levels using the CC framework, allowing for a full assessment of effect heterogeneity. The chapter fits into the growing literature estimating quantile treatment effects (Bitler et al. 2006; 2008; Dammert, 2009; Powell, 2020; Callaway et al. 2018). Key advantages of the CC model are the provision of more information on the distributional effects of the QI programme as well as removing the unrealistic assumption of additivity. If the distribution of effort or facility unobserved characteristics are different between enrolled and non-enrolled groups, and these unobservable facility characteristics are related to responsiveness to the QI programme, the assumptions of the CC model enables a more accurate reflection of how quality scores are determined. However, this benefit is achieved at the expense of being able to identify the relevant units treatment effects relate to. The scale-independence of CC is not a significant benefit in our case due to the similar distribution of pre-treatment quality scores between enrolled and non-enrolled facilities reduces its importance in our particular case (Meyer, 1995).

Despite the differences in the interpretation of the treatment effect heterogeneity identified by the DDD and CC methods, the results point towards similar conclusions. All facilities, regardless of pre-treatment quality score, benefit from the QI programme. However, the QI programme disproportionately benefits facilities with higher baseline quality. A key objective of the ICRMP is to increase the quality of health care provision in SA across all PHC facilities to a set quality threshold. Our analysis suggest that the programme will promote this objective, as well as improving average quality. However, this is occurring at the expense of increased variation in the quality of care. Consequently, the programme may – at least in the short term – exacerbate pre-existing variations in health care quality and inequalities in the provision of quality health care.

As highlighted, there is a general convergence in the quality scores among facilities across baseline score quartiles for unenrolled facilities (**Table 3.3** and **Table 3.8**). For facilities with higher baseline quality scores, the QI programme had a protective effect against observed quality score reductions observed among the non-enrolled facilities, while for facilities of lower baseline quality, the QI programme only marginally added to the over-time improvements of their unenrolled peers. This suggests the QI programme may be off-setting other factors that are causing variation in quality over time for facilities of different baseline quality scores. While this pattern of quality changes over time among low and high baseline performing facilities resembles a regression to the mean (RTM) effect, unlike traditional regression to the mean, this does not impact the internal validity of our results. The problems of RTM in DD settings are well known (Daw & Hatfield,

2018; Ryan, 2018; Daw & Hatfield, 2018; Chabe-Ferret 2015). These studies show problems are caused by pre-treatment outcome levels being correlated with treatment assignment. This is a potentially common issue as health policies are frequently targeted and the standard assumption is that DD can be applied even in circumstances where there are baseline outcome differences between treated and control groups. In our case, the QI programme enrolment is distributed across facilities of all baseline quality levels and the baseline quality scores are very similar between enrolled and unenrolled facilities. In the absence of the QI programme, any RTM would cause enrolled and unenrolled facilities to regress back to the same common level. Therefore the treatment effects estimated for each quartile/quantile approximates the true treatment effects of the QI programme, even in the presence of differential dynamic trends in quality across facilities. Therefore, in the short term at least, the QI programme actively worked against other phenomena which would have reduced variation in the quality of health care. The programme traded-off reduced variation in the quality of care offered by facilities across SA with an improvement in the aggregate score compared to the counterfactual situation where the programme was not implemented.

7.2. Mechanisms Behind Heterogeneous Treatment Effects

The focus of this chapter is to examine possible distributional effects of the ICRMP QI programme. As we do not explicitly examine mechanisms, partially due to data limitations, we are cautious in our interpretation around the mediating causes for the differences in the observed effect of the programme. However, understanding the mechanisms behind the observed heterogeneity has obvious policy importance. Previous studies of QI programmes, notably PBF schemes, have hypothesised factors that may determine both past and present health care quality and the effectiveness and responsiveness of health facilities to improvement programmes. Disentangling the effect of various factors on the impact of QI programmes is difficult due incentive structures often interacting with facility characteristics, and various determinants of past quality likely impacting facilities ability to respond to programmes[64]. This introduces challenges in separating the impacts of the various determinants of past performance have on responsiveness to QI programmes (Markovic & Ryan, 2017). In our setting, idiosyncratic programme features – specifically the quality indicators measured and the additional resources – may provide insight into

[64] Due to concerns around factors influencing past performance also influencing PBF responsiveness, many PBF-style programmes implement prior facility readiness assessments and structural and process quality equalising investments.

factors impacting both past quality performance and QI programme responsiveness[65]. Unlike PBF programmes, where theory predicts baseline low performers face higher marginal costs in responding to programmes, there is limited rationale for differential responses to the ICRMP QI programme (Mullen et al. 2010). If we observed 'progressive' treatment effects of the ICRMP QI programme, this may have hinted towards previous capacity constraints (e.g. financial resources) faced by low baseline quality facilities being responsible for past poor performance, as the programme relaxing this constraint would allow previous low performers to improve quality. However, because the observed treatment effects are 'regressive', this suggests that effort may have been a determinant of both past ICRMP quality scores and responsiveness and subsequent effectiveness of the programme. Therefore, while we do not directly examine between facility variation in pre-treatment quality, our results enable us to speculate on factors explaining observed pre-existing facility quality variation, and which may be important in modifying facility's QI programme responsiveness. While our analysis is largely restricted to attempts to identify the impact of the QI programme for facilities and how this impact may vary across facilities of different pre-treatment quality performance, this illustrates how examining effects beyond the mean may provide introductory insights regarding important effect mediators. In order to formally assert such claims about mechanisms, future research might do a full mediation analysis.

7.3. Limitations

Our identification strategy is susceptible if there is selection into QI enrolment based on underlying time trends in facility quality, which would bias estimates of the ATT. We have attempted to control for the various sources which may contribute to such differential trends with the DDD and CC identification strategies ruling out differences in time-invariant factors contributing to differential trends. Additionally our sensitivity checks ensure enrolled and unenrolled facilities are as comparable as possible on a range of pre-treatment observables and related unobservables. Although we try to control for key time-varying facility-level variables that may determine QI enrolment and have an independent influence on quality scores, it is not only contemporaneous differences in changes across time-varying inputs which may be problematic. If facilities viewed as

[65] Unlike previous studies, observed heterogeneity should not be unduly influenced by health facility's individual capacities, notably financial resources. This is due to a the indicators included in the ICRMP checklist which relate primarily to structural and process measures of health care quality which lie largely within the influence of facilities own actions. Additionally, for the select indicators which do require external assistance such as direct financing, this was provided as part of the QI programme. Differences in the ability of facilities to attain targets is what causes programmes to introduce differential incentives and targets and further complicates learning about the determinants of part performance and responsiveness to QI programmes.

improving had received increases in their inputs pre-treatment, the full effect of these inputs may occur over a period of time. Therefore, pre-treatment changes in facility inputs or circumstances with a lagged effect may also impact quality scores. Our matching sensitivity analysis should account for this by comparing facilities with similar pre-treatment facility activity accounting for changes in pre-treatment observables with a lagged effect and correlated unobservables.

We only have two-periods and two-groups in which to assess the impact of the QI programme. In such cases, inference is dependent on the structure of uncertainty (Lechner, 2011). This restricts inference to assessment of short-term effects of the QI programme. On the other hand, analysis on only two periods ensures that any bias imparted from a violation of the parallel trends assumption is bounded by the single year maximum difference in the trends, as opposed to a cumulative function of trend differentials, and therefore increases our confidence that the true pattern of heterogeneity is close to that observed.

Finally, there is ongoing debate regarding measures of quality of health care (Akachi & Kruk, 2017). The quality measures used in the ICRMP are largely restricted to structural and process measures of quality (McIntyre & Ataguba, 2018; Donabedian, 1966). While there is evidence that accreditation type QI programmes may improve such structural and process measures, there is less evidence that this translates into improve health outcomes. However the focus of our study, on emphasising the importance of measuring distributional effects and examining how this may be done, is equally applicable to all QI programme indicators types.

8. Conclusion

Inequality in access to high-quality health care is a prominent issue in many LMICs. Efforts to promote higher quality health care, including QI programmes, are being increasingly implemented. Key objectives of QI programmes often include promoting minimum quality standards or reducing variation in the quality of health care provided, with strong efforts dedicated in programme design to promote such objectives. Even when equity is not an explicit objective, the equity consequences of a policy should be reported. Larger effects in reducing negative health outcomes attributable to low-quality care are likely to be observed when programmes disproportionately impact suppliers at the bottom end of the quality distribution. We therefore contend that future evaluations of QI programmes should not limit themselves to the examination of mean impact. The case for focusing solely on mean impacts is that undesirable distributional aspects of policies are either unimportant or can be offset by transfers (Heckman et al. 1997).

Neither of these are true in the case of QI programmes in health care. Consequently, evaluations should be undertaken with equality objectives in mind and programmes constructed to target the sources of these inequalities.

Chapter 4

Health Facility Quality Peer Effects: Are Financial Incentives Necessary?

1. Introduction

Low quality health care continues to afflict many LMICs. A contributing factor is that many health systems offer limited incentives for quality improvement. Consequently, many countries are pursuing health financing reforms in an effort to increase the material rewards and incentives for health care providers to improve quality. However, evidence on the effectiveness and cost-effectiveness of these schemes is mixed (Binyaruka et al. 2020; Borghi et al. 2015). An important question therefore, and the focus of recent research is, other than traditional financial incentives what are the potential drivers of health care quality (Lagarde et al. 2019).

Several studies have examined the impact of provider competition and strategic interactions between health care providers on health care quality (Cooper et al. 2012; Gaynor et al. 2013; Gravelle et al. 2014; Longo et al. 2017; Brekke et al. 2021; Moscelli et al. 2021). The mechanism driving the relationship in these studies is the demand-side response to quality changes. Health care providers are seen as demand substitutes, with patients able to switch to a provider if its quality is increased and away from it if a rival's quality is increased. Therefore, the extent to which competition or changes in peers' quality impacts quality depends on the elasticity of demand for health care with respect to quality. Three conditions are necessary for this mechanism to operate. First, financial incentives linked to provider reimbursement must be present, such that providers are adequately rewarded for the quantity of services supplied i.e. the marginal patient is profitable. Second, patients must have sufficient information on providers, allowing informed choice and adaptation of choices to changes in provider characteristics. Third, a low cost of switching between providers for patients must also be present, without which providers have a large relative quality range to operate within before a utility maximising patient is induced to switch between providers.

However, it is unlikely that similar market mechanisms will work effectively to drive health care quality in LMICs. To date, most LMICs lack the prospective payment systems necessary to induce health care providers to compete over patients. Even in cases where prospective payment systems are implemented, patients lack sufficient information allowing informed utilisation choices.

Indeed, Pickett (2016) found that in circumstances where strong information systems are not present, and the perceived quality of providers relies on word-of-mouth, an increase in provider density/competition may not lead to quality improvements. Therefore, in settings of low information, the relationship between demand and quality may be weak, even in circumstances with adequate competition among providers facing financial incentives.

While financial incentive schemes remain a central policy intervention in circumstances where effort is a strong determinant of service output quality, a growing body of research has examined the role of wider motivations and non-financial incentives in determining public service provider behaviour (Benabou & Tirole, 2003; 2006; Besley & Ghatak, 2005). Alternative drivers of provider behaviour suggested include; professionalism and professional identity whereby individuals self-regulate in order to abide by professional norms and formal and informal agreed upon standards (McWilliams, 2020; Madara & Burkhart, 2015); altruism where an individual's own utility is partially dependent on the well-being of others (Papanicolas & Smith, 2015; Galizzi et al. 2015); and reputational concerns, social prestige and prosocial motivation where individuals care about external perceptions (Ashraf et al. 2014; Leonard & Masatu, 2006; Leonard & Masatu, 2010)[66].

Growing empirical evidence on these non-financial determinants of provider performance has led a number of countries to introduce non-financial incentive schemes for health care providers to improve quality (Kairies & Krieger, 2013). Performance monitoring and feedback reporting has been shown to result in quality improvements among health care providers (Bjorkman et al. 2009; Bevan et al. 2019; Godager et al. 2016), while selecting candidates with traits deemed desirable for health care workers can improve outcomes (Dal Bó et al. 2013; Deserranno, 2019; Ashraf et al. 2020).

A number of these non-financial drivers of provider performance suggest that strategic interactions between health care providers may occur without relying on demand-side responses or financial incentives. It's been noted in many settings that by observing peer's performance

[66] Intrinsic motivation is a commonly used term, with no single agreed upon definition, to refer to various non-financial motivations. Lagarde et al. (2019) define it as "the satisfaction derived from undertaking actions that benefit other people or society" explicitly distinguishing it from reputational concerns which are viewed as a non-pecuniary extrinsic motivation. Meanwhile Kolstad (2013) uses the term to "refer to incentives unrelated to profit and model it as a function not only of quality itself but of the ability to observably perform well relative to a reference group." Leonard & Masatu (2017) describe prosocial motivation as "caring about the welfare of others or caring about the opinion of others". Brock et al. (2015a) separate intrinsic and prosocial motivation, the latter viewed as a type of the former. Specifically, prosocial motivation suggests individuals care about external perceptions with reputational concerns entering their utility function.

individuals and firms often increase their own effort, not only due related to information gains on the marginal cost and returns to effort but also the aforementioned social preferences (Villeval, 2020). For instance, if providers have reputational concerns the availability of relative performance information has the potential to induce or stimulate quality peer effects between health care providers. Therefore, unlike attempts to improve health care quality through financing reforms, quality improvement policies targeting non-financial drivers of provider behaviour may require relatively modest changes and expenditure commitments. Such policy interventions have important implications for LMICs struggling with resource limitations and poor quality performance among health care providers, as they can potentially improve quality while minimally impacting costs.

In this chapter, we examine whether health facilities' quality (partially) depends on and responds to the quality of peer facilities, even without material incentives. The existence of such quality spill-over effects without an accompanying financial incentive suggests not only the presence of non-pecuniary determinants of quality, but they're ability to stimulate beneficial strategic interactions. We utilise data from the ICRMP quality checklist introduced in South Africa and previously outlined in chapter 3. Due to the idiosyncrasies of the context studied, our examination of peer effects is synonymous with testing for the presence of non-financial incentives for health care quality improvements. We use plausibly exogenous variation in peer facility quality shocks induced by a quality improvement programme as instruments for average peer quality. We compare our results from our preferred Policy-based IV specification with results from Spatial Econometrics approaches, which are commonly applied to such settings. If facilities do adapt their quality to changes in peers' quality, this has important policy implications as it suggests measurement and public reporting of quality indicators may be sufficient to induce a quality response, even in the absence of financial incentives. Additionally, peer effects, if present, are an important consideration in the assessment of the costs and benefits of certain policies. QI policies may be more cost-effective if there is a feedback effect across facilities.

The rest of the chapter is structured as follows. Section 2 outlines the related literature with Section 3 reviewing the contextual background. Section 4 outlines the empirical strategy and data. Finally, Sections 5 and 6 provide the results and discussion.

2. Literature

As noted, a significant volume of work has examined the impact of provider competition and strategic interactions between health care providers on health care quality (Cooper et al. 2012; Gaynor et al. 2013; Gravelle et al. 2014; Longo et al. 2017; Brekke et al. 2021; Moscelli et al. 2021). An increasing number of studies have examined the impact of intrinsic motivation in health care and the role this can play in the quality of care provided (Kolstad, 2013; Lagarde et al. 2014). While both literatures examine means of improving health care quality, they are distinct in the hypothesized mechanisms of action and the methodologies used. We briefly outline both literatures below.

Research has examined the effect changes in market structure impacting competition has on health care quality in high-income countries (Cooper et al. 2012; Gaynor et al. 2013). These studies exploit exogeneous policy changes increasing the degree of competition between health care providers to examine how this impacts providers' subsequent quality. The results of these studies generally show that introducing provider competition through reforms enabling patient choice increases health care quality and improves health outcomes[67]. A related body of work has directly examined strategic interactions between health care providers. Rather than estimating a reduced form assessing how quality responds to changes in market structure, these studies estimate health care provider reaction functions, directly examining how providers respond to changes in other providers' quality. Therefore, instead of examining how changes in market structures influence how providers interact, the market structures are assumed to be fixed and examination focuses on strategic interactions between providers.

Adopting the latter approach, Gravelle et al. (2014) examine whether the quality of hospitals in the UK is influenced by the quality of other hospitals operating in the same 'market'. They find that 7 of 16 hospital quality measures are strategic complements. A 10% increase in peer hospital's quality increases a hospital's quality by 1.7% to 2.9%. Longo et al. (2017) examine the same phenomenon. Employing panel data within a similar spatial econometric framework, they find except for overall hospital mortality rates, neither hospitals' quality or efficiency respond to changes in peer hospitals' quality or efficiency. However, unconditionally a spatial correlation is observed, suggesting previously observed positive associations may have resulted from correlated spatial effects, rather than a true peer effect.

[67] This is based on quality competition occurring in a fixed price context.

As noted, there is growing evidence of the importance of non-financial motivations and determinants of public service provider effort at both the individual- and organisational-level (Gneezy, 2011; Bowles, 2016; Delfgauuw & Dur, 2008). One well-known manifestation is the Hawthorne effect, whereby effort increases when individuals know they are being observed (Leonard & Masatu, 2010b).

A number of empirical studies have examined both the existence of prosocial motivation and how variation in prosocial incentives – often through monitoring and public reporting – can increase prosocial motivation and ultimately health care quality. Brock et al. (2015b) investigate the roles of esteem in determining the quality of care provided by health care practitioners in Tanzania. They outline how practitioners' perceptions of how others view them can motivate health care quality. Bjorkman et al. (2009) show a similar quality effect induced by community-based monitoring of public primary health care (PHC) facilities in Uganda. This illustrates how the opinion of patients or community can induce an increase in effort via prosocial motivation.

Leonard & Masatu (2017) show that performance measurement via peers results in quality improvements which are sustained even after observations have ended, suggesting the opinion of peers matters in effort exerted, but also that reaffirmation of expectations and norms may sustain the effect. These shared expectations and norms are known and understood as the professional standards. Leonard & Masatu (2006; 2010a) established a significant know-do gap also showing that peer observation significantly reduces the difference between best-practise and actual performance (Leonard & Masatu, 2010b). Olivella & Siciliani (2017) show how reputational concerns result in less altruistic health care workers mimicking their more altruistic peers.

This suggests public reporting can act as a prosocial incentive based on a desire to be seen as conforming to prespecified shared professional norms, as adhered to by peers. Kolstad (2013) specifically illustrates how having access to information enabling performance comparisons between peers can induce health care quality improvements directly from prosocial motivation. Specifically, he finds new performance information on relative quality results in only a small increase in quality originating from a profit-motive while the change in quality deriving from prosocial motivations is four times larger.

The presence of the know-do gap combined with growing evidence of the potential effect of prosocial motivation, and the poor standard of both the quality of health care provision and the measurement and public reporting of quality indicators, suggests that the introduction of routinely measured and standardised quality indicators may induce a yardstick quality competition effect (Shleifer, 1985). Health care facilities may act to ensure they closer conform to the quality standards of peer facilities, even in the absence of material incentives for improving quality. This is seen by Kolstad (2013), who shows how practitioners who learned they are performing worse than they previously believed prior to new information, significantly increased their quality. Yardstick competition has also been shown to occur through informal local information spill-overs (Bordignon et al. 2004). However, the introduction of national-level standardised performance measures has also been shown to weaken local performance comparisons and localised peer effects, as in the presence of national performance evaluation systems, the relevance of the performance of local peers may decrease (Revelli, 2006).

We explore strategic interactions between health care providers in South Africa (SA), examining how public PHC providers respond to changes in peer providers' quality, with reporting coming from a national standardised quality index. To our knowledge, this is the first examination of strategic interactions between health care facilities in a LMIC. Whilst our study fits into the literature investigating how health care providers respond to changes in the behaviour and signals of their peers, we distinguish our study from previous work in the area. While previous studies have tested whether financial incentives linked to provider reimbursement cause strategic interactions, we explore whether these interactions and quality spillovers may exist in a context without financial incentives.

3. Institutional Background

3.1. The South African health system

SA operates a decentralised health care system (White Paper for the Transformation of the Health Sector in South Africa, 1997). The National Department of Health (NdoH) guides health policy, with nine Provincial Departments of Health (PdoH) and 52 health Districts focused on implementation. District Health Management Offices (DHMOs) in SA are responsible for PHC service delivery (Government of SA, 2003). These services are provided largely through a network of approximately 3,500 PHC facilities, free at the point of use. An estimated 120 million PHC facility visits took place in 2015, 2.2 visits per capita (UNICEF, 2019).

The National Health Care Facilities Baseline Audit (2011/12) – led by the Office of Standards Compliance within the NdoH – highlighted that the quality of PHC facilities remained a significant issue. For example, only 23% of PHC facilities had all tracer medicines available and only 6% of relevant facilities had all essential maternity ward equipment available and functional (Health Systems Trust, 2013).

3.2. The Ideal Clinic Realisation and Maintenance programme (ICRMP)

As a result of the identified quality deficits, the NdoH introduced the Ideal Clinic Realisation and Maintenance programme (ICRMP) in 2015. The programme attempts to identify, measure, and improve a diverse range of quality indicators seen as foundational for facilities to be able to offer high quality health care. The first component of the programme is a checklist of standardised inputs and processes all public PHC facilities are expected to meet. The checklist contains ~150 indicators separated into 10 components: administration, integrated clinical services management, medicines supplies and laboratory services, human resources, support services, infrastructure, health information management, communication, and stakeholder engagement (See **Supplementary Appendix C3-X** for full indicator list). Facilities' ICRMP quality scores are self-assessed annually with external verification of a sample of facility scores. Scores are hosted on a NdoH website, allowing Provincial and District DoHs and facilities to observe performance as well as make inter-facility comparisons. Prior to introduction of the ICRMP checklist, the District Health Information System (DHIS) 2, measuring volume of health care services delivered, provided the only set of standardised indicators for PHC facilities. The quality reporting is available only to PHC facilities and higher-level health Departments, and are not observable by the wider population. Therefore, unlike the previous literature examining provider quality competition reliant on demand-side responses, the potential quality peer effect here is reliant to a purely supply-side response.

To operationalise the programme, in addition to quality monitoring and reporting, a Quality Improvement (QI) programme was implemented, aimed at assisting facilities to improve their ICRMP checklist quality scores. The ICRMP QI programme provides facility-level support via so-called district-based Perfect Permanent Team for Ideal Clinic Realization and Maintenance (PPTICRM). These teams provided support to enrolled facilities to improve their ICRMP checklist quality scores. Importantly, while most indicators were expected to be met within existing facility resources budgeted for routinely, facilities enrolled in the QI programme were provided with

supplementary financing to address equipment and infrastructure deficits (Hunter et al. 2017). The ICRMP began at the end of the 2014/15 Fiscal Year (FY) with facilities receiving the ICRMP QI programme in a staggered enrolment. Scheduling for facility's enrolment in the ICRMP QI programme was planned such that 1,000 facilities were enrolled each year.

Crucially, while long-term plans for reforms to link facilities' ICRMP quality scores to official facility accreditation and financing have been discussed, currently the ICRMP offers no material incentive stimulating facilities to improve quality scores.

Due to the nature of the indicators, primarily structural and process measures of health care quality (Donabedian, 1966), facility efforts are expected to be a relatively stronger determinant of ICRMP quality compared to if the programme captured outcome quality i.e. patient morbidity and mortality.

4. Methods

4.1. Empirical Strategies

The econometric problems associated with identifying and estimating peer effects are well known (Manski, 1993; Angrist, 2014). Manski (1993) outlines three phenomena that can lead to individual outcomes not being independent that, in our setting, could lead to correlations in quality scores among peer PHC facilities. The first, endogenous peer effects, represent the strategic interactions between PHC facilities, the focus of our investigation. Another source could be exogenous (contextual) peer effects, if PHC facility quality varies with the background characteristics of peers. Finally, a common issue is the presence of correlated effects as peers face the same shared environment. There are two identification issues: i) separately identifying endogenous and exogenous peer effects (referred to as the 'reflection problem' (Manski, 1993)) and ii) distinguishing social interactions – endogenous and exogenous effects – from correlated effects. Without addressing these issues, any correlation between group and individual outcomes may be spuriously identified as endogenous peer effects.

Previous empirical studies examining health facility strategic interactions with respect to quality have utilised spatial econometric models to estimate facility reaction functions. Such spatial econometric approaches overcome the inherent identification issues by estimating peer effects via maximum likelihood methods (Ord, 1975; Anselin, 1988; Lee, 2004; Elhorst, 2003 & 2010; Lee et

al. 2010) or generalised method of moments (Kelejian & Prucha, 1998; Lee, 2007). However, these methods operate conditional on the assumption that the spatial econometric model estimated represents the true data generating process, an important aspect of which is that peer groups are correctly specified. This has been demonstrated by Lee (2009), who illustrates the bias arising from peer group misspecification.

Gibbons and Overman (2012) discuss the limitations of spatial econometric approaches, suggesting more credible estimation of peer effects requires an exogenous source of variation in the explanatory variable of interest, as has become the norm in econometric policy evaluation. This chapter is the first to examine health facility peer effects using a quasi-experimental design as we exploit the ICRMP QI programme as a source of exogenous variation to estimate the quality response of facilities to changes in the quality of their peers in an instrumental variables (IV) framework. We compare our estimates from this policy-based IV approach with standard spatial econometrics methods that the previous literature examining health facility peer effects have adopted. Before describing the primary empirical strategy in detail, we briefly discuss the standard spatial econometrics methods.

4.2. Spatial Econometrics Approach

We follow Gravelle et al. (2014) and Longo et al. (2017) who estimate health facility reaction functions, presenting quality as a function of the spatially weighted quality of other PHC facilities, using standard spatial econometric methods. We adopt their baseline model, a linear Spatial Autoregressive model (SAR) of the form:

$$y_i = \rho \sum_{(i)j} w_{(i)j,i} y_{(i)j} + \beta' X_i + \varepsilon_i \qquad (1a)$$

, where $(i)j$ is the leave-out mean indicating facility i's own quality is determined by their peers' quality in addition to their own characteristics. This can be simplified by presenting the matrix form analogue commonly used in spatial econometric models:

$$Y = \rho WY + \beta X + \varepsilon \qquad (1b)$$

Where Y is an $N\ x\ 1$ vector of facility quality and $\boldsymbol{W}$ is a $N\ x\ N$ spatial weight matrix (SWM) describing facilities' peer groups. $\boldsymbol{W}$ is constructed to have zero elements on the leading diagonal, ensuring WY excludes facility i (itself). X is a facilities own characteristics, namely average monthly patient headcount and average monthly nurse working dates. ρ is a scalar spatial dependence parameter, where $\rho > 0$ suggests that PHC quality increases with peer quality. As outlined, there are a number of phenomena which could result in $\rho > 0$, other than a true peer effect. Without addressing these we might spuriously identify correlations between WY and Y as reflecting true endogenous peer effects.

Correlation in quality scores between PHC facilities may result from geographically concentrated characteristics i.e. correlated effects. If facility quality scores are correlated due to prevailing factors facing facilities within a geographic area, the above model could lead to a spurious conclusion that strategic interactions are responsible for observed similarities. For instance, there may be geographic differences in facility financing or staff competencies due to differences in area desirability. Likewise, perceptions of the importance of ICRMP quality scores may vary geographically. Such phenomenon would result in an unaccounted interdependence in peer facilities' error terms. Therefore, we estimate the following Fixed Effects Spatial Autoregressive model (SAR-FE) to control for time-invariant unobservable factors:

$$Y = \rho WY + \beta X + \alpha_i + \lambda_t + \varepsilon_{it} \qquad (2)$$

Now Y is an $N\ x\ T$ matrix of facility quality, with WY and X also having a time dimension, where t refers to the Fiscal Year. In addition to controlling for time-invariant facility-level factors, facility fixed effects, α_i, may also account for the possibility that the same external peer review teams assess quality for facilities across the years. Facility fixed effects will control for this assuming evaluators scoring behaviour is consistent across years. However, if common evaluators assess several facilities within a geographic radius this may overlap with peer group definitions. Therefore, there is a chance that the presence of common evaluators results in a correlation in scores between peers. For example, if there are differences in evaluators' application of criteria across years and a preference for reducing within year score variances, then this may result in the perception of peer effects driving similarities while it is actually driven by facilities facing common evaluators. We include year fixed effects, λ_t, as all PHC health care facilities were provided with the ICRMP quality checklist at the start of the 2015/16 FY, regardless of whether they were enrolled in the

ICRMP QI programme. As such, we may expect a general increase in facility quality over time simply resulting from an increased awareness of the indicators.

Finally, contextual effects may be important in determining PHC quality, in which case a Spatial Durbin model (SDM) is estimated:

$$Y = \rho WY + \beta X + \delta WX + \alpha_i + \lambda_t + \varepsilon_{it} \quad (3)$$

Where $WX = \sum_{(i)jt} w_{(i)j,i} x_{(i)jt}$. For instance, District-level health financing budgets and resources are finite, therefore, if facility i's peers receive high financing this results in less financing available for facility i. If there are large relative differences in the levels of resources across peer facilities this may impact quality scores beyond facilities own absolute resources.

4.2.1 Peer Group Structure (W construction)

Identification using spatial econometric methods relies on the choice of peer group structures. There are numerous ways in which a facility's peers could be defined. Both Gravelle et al. (2014) and Longo et al. (2017) defined hospitals within a 30km radius as peers. This is informed by studies examining patient choice sets, as patient demand is the proposed mechanism through which the strategic interactions operate. However, in our case the proposed mechanism driving interactions is PHC facilities' own references regarding who they view as peers against who they may evaluate their quality performance. We define $\boldsymbol{W}$ using two different strategies described below.

a) *District classification ($\boldsymbol{W^D}$):* This peer group definition utilises information on aspects of the SA health care system and implementation of the ICRMP. A number of recent studies have suggested that institutions may be important in peer group definitions (Atella et al. 2014; Arbia et al. 2009; Guccio & Lissi, 2016). As noted, in SA's decentralised health care system DHMOs are responsible for the delivery of primary health care services. Subsequently, facility-level support for the ICRMP was provided by district-level teams. Therefore, facilities within the same District share a common administrative and governance structures, suggesting facilities' knowledge and interactions may be more intense within Districts. This suggests facilities and higher levels health authorities plausibly view facilities within the same District as relevant comparators.

In this case, peer groups are discrete mutually exclusive groups, resulting in $\boldsymbol{W^D}$ being block-diagonal with facilities treating all other facilities within their respective Districts as equal peers.

$$w^D(i,j) = \begin{cases} 0 & \text{if } D_i \neq D_j \\ 1 & \text{if } D_i = D_j \end{cases}$$

The weight fixed to each peer, $w^D(i,j)$, indicates the assumed strength of the spatial interaction between the facilities. Because of the prevailing institutional structures, the District classification represents our preferred peer group definition.

b) *K Nearest Neighbour (KNN) ($\boldsymbol{W^{KNN}}$):* We construct peer groups with a distance-based definition of neighbours. While institutional factors may be important, it is possible that a proximity-based definition of peers may better reflect to whom facilities compare themselves. Unlike Gravelle et al. (2014) and Longo et al. (2017), there are no obvious geographic constraints in our hypothesized mechanism potentially driving quality competition. Therefore, facilities may view their peers as their K closest 'neighbouring' facilities. There is little guidance on the number of facilities which may be considered 'neighbours'. However, beyond a certain range, facilities may be viewed as facing different prevailing local conditions. Given this uncertainty, we analyse a range of $\boldsymbol{W^{KNN}}$ specifications. The average number of facilities per district in our sample is 51. We analyse a somewhat arbitrary range of $K = \{3, 7, 10, 13, 17\}$, under the assumption facilities are mindful of their more immediate peers.

$$w^{KNN}(i,j) = \begin{cases} 0 & \text{if } K_j > K \\ \dfrac{1}{dist(i,j)} & \text{if } K_j \leq K \end{cases}$$

Where K_j is the ordinal neighbour ranking of facility j with respect to facility i, and K is the selected $\boldsymbol{W^{KNN}}$ specification. Like Gravelle et al. (2014) and Longo et al. (2017), in our proximity-based specification we assume the strength of the relationship between facilities classified as peers is inversely proportional to the distance.

Unlike peer groups defined using $\boldsymbol{W^D}$, defining peers by $\boldsymbol{W^{KNN}}$ does not lead to discrete mutually exclusive groups, but network structures. Networks can be distinguished from

group structures by the topology of interactions. While in group structures every member of a group affects every other member of that group, in networks each individual has their own unique reference group (if i is the peer of j, and j is the peer of h this does not necessarily imply h is the peer of i).

We take peer groups to be exogenously determined, as is often the case with geographically defined peer groups. Following common practise, we row standardise all SWMs by transforming $w(i,j)^* = \frac{w(i,j)}{\sum_{j=1}^{n} w(i,j)}$ such that for each facility i $\sum_{j=1}^{n} w(i,j)^* = 1$. It is this transformation that leads to the interpretation of WY as the average quality of all peer PHC facilities. Additionally, peer groups defined are constant over time as no facilities were known to open or close during our analysis period. The peer information of the $\boldsymbol{W}$ specifications examined is found in **Table 1**.

$\boldsymbol{W^D}$ and $\boldsymbol{W^{KNN}}$ lead to different peer group and therefore network structures. The source and conditions for identification using the above spatial econometric methods varies depending on the characteristics of the peer groups chosen. Two primary approaches have been used to estimate the above models: maximum likelihood (ML) methods and generalised method or moments (referred to as spatial two-stage least squares (S2SLS)). Specifics of the estimation approaches and identification conditions are outlined in **Supplementary Appendix C4-1**.

As noted, identification in spatial econometric methods rely on the choice of $\boldsymbol{W}$, and strong functional form assumptions, opening the possibility of misspecification of the interaction structure and, often arbitrary, model assumptions affecting identification. Therefore, peer effect estimation utilising spatial econometric methods may be considered somewhat of a black box compared to quasi-experimental methods. Gibbons & Overmans (2012) suggest that to be credible, estimation of peer effects should come from quasi-experimental approaches where known sources of exogenous variation in peers' outcomes are exploited. In our case this relies on having an exogenously determined variable that explains the quality scores of a facilities' peers but has no direct effect on a facility's own quality score.

Although we try to justify our construction of the SWM in addition to reporting results from several specifications, we do not have data on which health facilities interact and therefore may be considered peers from a quality reference perspective. There is also little guidance in the literature on how institutions may define peers given the hypothesised mechanism of intrinsic motivation/yardstick competition which would drive any potential strategic interactions. This is

important as incorrect specification of the SWM potentially results in the exclusion restrictions which enable identification of endogenous peer effects being violated (Gibbons et al. 2014).

4.3. Policy-based IV

The ICRMP QI programme previously described arguably provides such a source of exogenous variation. Here – unlike S2SLS – our instrumental variable approach relies on knowledge of the policy context, removing the identification burden on correct specification of the peer network structure. As such, our Policy-based IV approach doesn't rely on assumptions on the spatial structure of interactions to generate the instrument.

We rewrite (3) as:

$$y_{it} = \rho\bar{y}_{(i)jt} + \beta_1{}'X_{it} + \beta_2\bar{X}_{(i)jt} + \beta_3 z_{it} + \alpha_i + \lambda_t + \varepsilon_{it} \qquad (4)$$

Which is a standard linear-in-means (LIM) type-model, where $\bar{y}_{(i)j} = \frac{1}{|d(i)|}\sum_{j\epsilon d(i)} E[y_j|\psi_i]$ is the expected quality of all other j facilities in facility i's district, $d(i)$. Here $\bar{y}_{(i)j}$ is the leave-out mean of District average ICRMP quality scores, representing the expectation of peer quality scores $E[y_j|\psi_i]$, conditional on the information known to facility i, ψ_i. ρ is the endogenous peer effect, the parameter of interest.

This clarifies the previously implicit assumption that facilities hold rational expectations of their peers' quality. Similar to the spatial econometric models, we also include facility characteristics, X_{it}, and z_{it} an indicator of whether the facility was enrolled in the ICRMP QI programme, as well as contextual effects $\bar{X}_{(i)jt}$, defined as the District average of the facility level characteristics. Facility fixed effects control for time-invariant unobserved heterogeneity, which can be arbitrarily correlated with facility quality. The above outlines the policy-based IV approach using the District classification of peer groups, where $\boldsymbol{W^D Y} = \frac{1}{|d(i)|}\sum_{j\epsilon d(i)} E[y_j|\psi_i]$. However, this approach is equally applicable utilising $\boldsymbol{W^{KNN}}$. We estimate both in our analysis.

As before, OLS estimation of (4) cannot identify the endogenous peer effects, ρ, due to the 'reflection problem'. To overcome this, we exploit the sporadic enrolment of facilities in the ICRMP QI programme, resulting in only a subset of facilities in each District being enrolled in the

QI programme at one time. This allows us to instrument average peer facility quality, $\bar{y}_{(i)jt}$, with the fraction of peers enrolled in the QI programme, $\bar{z}_{(i)j} = \frac{1}{|d(i)|}\sum_{j \epsilon d(i)} z_j$. However, the effect of QI programme enrolment on facility quality is not restricted to the year of enrolment, there may be lagged quality effects resulting from QI enrolment in the previous year. Therefore, our IV specification uses both the fraction of peer facilities enrolled in the QI programme in the current FY, $\bar{z}_{(i)jt}$, and the fraction enrolled in the previous FY, $\bar{z}_{(i)jt-1}$, as instruments.

To see identification using this approach, we present a stylised example where there are $d = 1, \dots, D$ Districts each with $n = 1,2$ facilities. Assume facility 1 is enrolled in the ICRMP QI programme and facility 2 is not:

$$y_{1t} = \alpha_1 + \lambda_t + \rho\bar{y}_{2Dt} + \beta_1{}'X_{1t} + \beta_2\bar{X}_{2Dt} + \beta_3 z_{1t} + \varepsilon_{1t} \qquad (5a)$$

$$y_{2t} = \alpha_2 + \lambda_t + \rho\bar{y}_{1Dt} + \beta_1{}'X_{2t} + \beta_2\bar{X}_{1Dt} + \varepsilon_{2t} \qquad (5b)$$

Where $\bar{y}_{1Dt} = y_{1Dt}$ and $\bar{y}_{2Dt} = y_{2Dt}$ i.e. the quality of the other facility in the district. This makes clear how the QI programme effects peer facility quality ($\bar{y}_{1Dt} = y_{1Dt}$), enabling us to instrument facility 1's outcome, $\bar{y}_{1Dt} = y_{1Dt}$, with its ICRMP QI enrolment z_{1t}. This amounts to instrumenting average peer quality scores with average peer QI programme enrolment. Therefore, identification of a facility's strategic response stems from the random shock in quality a facility's peer group gets from enrolment in the QI programme.

4.3.1 Validity of Policy-based IV Identification Strategy

The ability of the Policy-based IV approach to consistently estimate endogenous peer effects relies on the satisfaction of standard IV criteria. First, the fraction of peers enrolled in the QI programme in the current and previous FY must be strong determinants of the average peer quality score (relevance condition)[68]. Second, the enrolment of facility i's peers in the QI programme should only affect facility i's quality indirectly through its impact on peer facilities quality scores. Therefore, the fraction of peers enrolled in the QI programme, $\bar{z}_{(i)j}$, must be uncorrelated with the error term in (4) (exogeneity condition).

[68] The relevance condition also implies the importance of including a facilities own current and past QI programme enrolment status in the outcome equation (von Hinke et al. 2019).

We test the relevance condition using the LM Test (Kleibergen & Paap, 2006). However, previous studies have already identified the positive effect of the ICRMP QI programme on facility quality scores (Stacey et al. 2021; McGuire et al. Unpublished Manuscript). Additionally, we report the joint F-statistic of the excluded instruments to reassure we do not face a weak instrument problem (Staiger & Stock, 1997).

The exclusion restriction condition is the key difference between the typical spatial econometric and Policy-based IV approaches. Unlike the spatial econometric approaches to identification and estimation, which heavily rely on knowledge of the appropriate model functional forms and spatial weights (network structure), the policy-based IV strategy relies on programmatic information about the exogeneity of proposed instruments.

While we can be fairly certain of our maintained assumption that average peer facility's QI enrolment does not have a direct influence on own facility quality (i.e. $\bar{z}_{(i)j}$ does not directly impact y_{it}), we cannot definitively test if facility enrolment in the QI programme, and therefore $\bar{z}_{(i)j}$, is not correlated with the error term, ε_{it}. We perform both a balancing test and Hansen J tests (Hansen, 1982), which provide suggestive evidence of the conditional exogeneity of the instruments. Moreover, including facility fixed effects strengthens the validity of our proposed instrument. If QI programme enrolment is related to facility-level time invariant unobserved heterogeneity related to facility quality, then peer enrolment status will provide information on own facility quality. Including facility fixed effects prevents this possible violation of the exclusion restriction.

Timing of individual facilities' enrolment in the ICRMP QI programme was decided by PdoHs, based on a somewhat opaque administrative process, suggesting potentially up to 8 different allocation criteria being used. However, there remains some threats to the validity of the instruments. It has been suggested that facilities viewed as performance improving may have been prioritised for enrolment in the QI programme, partly to encourage continued quality gains. If facility-level enrolment decisions were based on unobservable facility characteristics (not controlled for by fixed effects) this would result in $Cov(z_{it}, \varepsilon_{it}) \neq 0$[69]. However, this type of

[69] While there is evidence that ICRMP QI programme enrolment decisions are credibly unrelated to a facility's own ICRMP score in the first year of implementation, the picture is less clear during the programme's second year. Although a schedule outlining enrolment years for all PHC facilities QI programme was specified on initial implementation of the programme, this was not strictly followed after the first year of implementation. Broadly, each facility was scheduled for enrolment over a three year period with approximately 1,000 facilities enrolled in the QI

facility-level policy endogeneity would not necessarily result in a violation of the exogeneity condition in our instrumental variables, which only occurs if the fraction of peers enrolled in the QI programme is correlated with the error, $Cov(\bar{z}_{(i)j}, \varepsilon_{it}) \neq 0$. Even in the presence of $Cov(z_{it}, \varepsilon_{it}) \neq 0$, it is difficult to imagine a situation resulting in $Cov(\bar{z}_{(i)j}, \varepsilon_{it}) \neq 0$. For the fraction of facilities in Districts enrolled in the QI programme to be correlated with the error term would require higher performing facilities – and such, facilities with higher propensity for enrolment – being strongly clustered by Districts. This could result in Districts with a high fraction of enrolled facilities systematically differing from Districts with a low fraction of facilities enrolled, and therefore the fraction of peer facilities enrolled providing information on a facility's own quality. For instance, if QI programme saturation is higher in Districts with distinct determinants of quality changes – promising Districts with more capacity for improvements for instance – then we may mistakenly infer the presence of peer effects due to both non-enrolled and enrolled facilities seeing quality improvements while quality remained static in Districts with low QI programme saturation. Even in such a case, this would suggest that higher level authorities paid little attention to the geographical distribution of QI programme implementation. Therefore, even in the presence of $Cov(z_{it}, \varepsilon_{it}) \neq 0$, $Cov(\bar{z}_{(i)j}, \varepsilon_{it}) \neq 0$ would require the unobservables which facility QI programme enrolment is potentially based on to be geographically concentrated, *and* decision makers not being overly concerned with the geographic distribution of facility-level enrolment. We consider these factors unlikely but also examine the clustering of facilities by performance within Districts and the geographic dispersion of the QI programme (**Appendix Figure 4.1**).

4.4. Data

We primarily draw on data from the ICRMP which collects information on ICRMP quality scores and ICRMP QI enrolment status from a census of public PHC facilities in SA. We use a 3-period facility-year panel from 2015/16-2017/18. As noted, the ICRMP checklist includes 10 components. We construct the dependent variable as the aggregate ICRMP quality score resulting in a quality score measured in integers between 0-100. We construct peer groups using information on the District to which facilities belong or point data with geographic coordinates for each facility, specifying the location within SA.

programme each year until all PHC facilities were enrolled. **Table 4.5** shows how facilities' enrolment status may change over time.

We include a number of control variables which largely remain consistent across the models estimated, although there are minor differences across some specifications. We include individual facilities' past and current ICRMP QI programme enrolment status, their average number of nurse working days per month across the year and the average number of patients seen per facility across the year which measure some degree of variation in facility financing.

5. Results

5.1. Descriptives

Table 4.2 provides an overview of the number and distribution of PHC facilities across SA. As a result of SA's policy of Provincial devolution, our analysis does not include Western Cape, which initially chose not to participate in the ICRMP programme[70]. Within the 8 Provinces used for analysis there are 46 Districts.

Table 4.1: Spatial Peer Information

	Mean Peers	Minimum Peers	Maximum Peers	Facilities
Nearest neighbour SWM (KNN=3)	3	3	3	2,368
Nearest neighbour SWM (KNN=7)	7	7	7	2,368
Nearest neighbour SWM (KNN=10)	10	10	10	2,368
Nearest neighbour SWM (KNN=13)	13	13	13	2,368
Nearest neighbour SWM (KNN=17)	17	17	17	2,368
District	51	9	116	2,368

Note: There are no openings and closing of facilities over the analysis period resulting in a time-invariant SWM for all specifications.

Table 4.2: PHC facilities in South Africa

Province	Number of PHC Facilities	Proportion of Total PHC Facilities
Eastern Cape	763	22%
Free State	221	6%
Gauteng	370	11%
Kwazulu-Natal	597	17%
Limpopo	473	14%
Mpumalanga	284	8%
North West	305	9%
Northern Cape	160	5%
Western Cape	260	8%
Total	**3,433**	**100%**

[70] Western Cape represents an outliner within SA in many aspects due to being run by the opposition political party since 2009.

As can be seen from **Table 4.1**, in addition to not including Western Cape we exclude some facilities in our analysis due to missing data issues. 2,368 of the full relevant sample of 3,173 facilities had 3 years of ICRMP quality score data i.e. full outcome data. Due to practical complications with varying SWMs across years, we restrict our primary analysis sample to the 2,368 facilities which report ICRMP quality scores for the full 3 years. Of these 2,368 facilities, there were 872 observations with missing covariate data. We use multiple imputation on the missing covariates to provide complete data for the 2,368 facilities which constitute our primary analysis sample. Further information on the missing data and our solutions can be found in **Supplementary Appendix C4-2**.

Table 4.3 presents descriptive statistics.

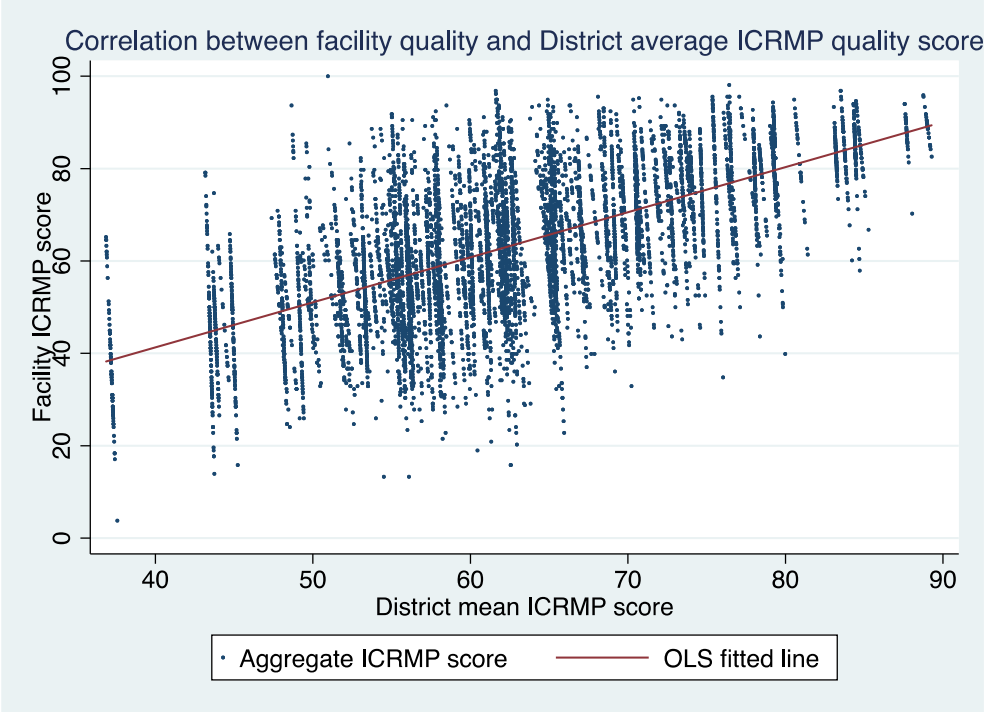
Table 4.3: Descriptive Statistics

		Std. dev.				
Variable	**Mean**	**Overall**	**Between**	**Within**	**Min**	**Max**
Outcome						
ICRMP Quality Score	59.6	14.8	9.9	11.1	3.8	100
Quality Improvement Programme						
ICRMP QI programme enrolment	0.32	0.47	0.27	0.38	0	1
Peers ICRMP QI programme enrolment	0.32	0.31	0.13	0.28	0	1
PHC facility characteristics						
Facility type						
Clinic	0.88	0.32	0.32	0	0	1
Community health centre	0.12	0.32	0.32	0	0	1
Facility location						
Rural	0.82	0.38	0.38	0	0	1
Peri-urban	0.01	0.08	0.08	0	0	1
Urban	0.17	0.38	0.38	0	0	1
Monthly patient headcount	2176.5	1647.8	1617.8	315.2	18.7	18,623.5
Monthly nurse working days per month	80.8	88.6	77.6	42.9	1	2,744.5
Municipality characteristics						
Population within 10km	66,137	88,491	88,511	0	7.8	770,000
Average Household size	3.4	0.3	0.3	0	2.5	4.5
Proportion with no schooling	0.10	0.04	0.04	0	0.02	0.20
South African Index of Multiple Deprivation						
Quintile 1	0.07	0.26	0.26	0	0	1
Quintile 2	0.11	0.31	0.31	0	0	1
Quintile 3	0.12	0.32	0.32	0	0	1
Quintile 4	0.28	0.45	0.45	0	0	1
Quintile 5	0.42	0.49	0.49	0	0	1

Notes: We use the District SWM classification ($\boldsymbol{W^D}$) to calculate peers ICRMP QI programme enrolment. These statistics will vary according to the SWM specification used.

Figure 4.1 shows the correlation between facility-level ICRMP scores and the leave-out mean District ICRMP score. Although there is a positive correlation, this tells us little about the presence of endogenous peer effects as facilities are affected by the same District-level random shocks. However, **Figure 4.1** does suggest smaller variation in ICRMP scores across facilities from better performing Districts.

Figure 4.1: Correlation of facility quality and average District quality



5.2. Spatial Econometrics Results

Table 4.4 presents the results of the panel spatial econometric models. In all models using the $\boldsymbol{W^{KNN}}$ peer group specification, once time fixed effects are included the results become indistinguishable from 0. However, with the $\boldsymbol{W^{D}}$ specification there appears to be strong positive peer effects, with a 1 unit increase in average peer quality leading to a 0.74-0.78 increase in facility quality. The difference in the results between the $\boldsymbol{W^{KNN}}$ and $\boldsymbol{W^{D}}$ may be due to the different estimation procedures required for the different peer group specifications. While $\boldsymbol{W^{KNN}}$ uses S2SLS, the network structure of $\boldsymbol{W^{D}}$ requires quasi-maximum likelihood estimation (see **Supplementary Appendix C4-1**).

Table 4.4: Spatial Panel Models						
Quality Indicator	**Model 1 (SAR-FE)**	**Model 2 (SAR-FE)**	**Model 3 (SAR-FE)**	**Model 4 (SAR-FE)**	**Model 5 (SDM-FE)**	**Model 6 (SDM-FE)**
ICRMP Aggregate Quality Score (District)	0.78*** (0.01)	0.74*** (0.01)	0.74*** (0.01)	0.74*** (0.02)	0.78*** (0.01)	0.76*** (0.01)
ICRMP Aggregate Quality Score (KNN=3)	0.27*** (0.01)	0.15*** (0 .01)	0.15*** (0.01)	-0.02 (0.01)	0.06*** (0.01)	-0.02 (0.01)
ICRMP Aggregate Quality Score (KNN=7)	0.38*** (0.01)	0.22*** (0.01)	0.22*** (0.01)	-0.02 (0.02)	0.07*** (0.02)	-0.02 (0.02)
ICRMP Aggregate Quality Score (KNN=10)	0.42*** (0.01)	0.26*** (0.02)	0.26*** (0.02)	-0.02 (0.02)	0.08*** (0.02)	-0.02 (0.02)
ICRMP Aggregate Quality Score (KNN=13)	0.45*** (0.01)	0.28*** (0.02)	0.28*** (0.02)	-0.02 (0.02)	0.07*** (0.02)	-0.02 (0.02)
ICRMP Aggregate Quality Score (KNN=17)	0.48*** (0.01)	0.31*** (0.02)	0.31*** (0.02)	-0.02 (0.02)	0.08*** (0.02)	-0.02 (0.02)
Controls						
ICRMP Quality Improvement Programme	x	x	x	x	x	x
Past ICRMP Quality Improvement Programme		x	x	x	x	x
Facility Patient Headcount			x	x	x	x
Facility staff levels			x	x	x	x
Year				x		x
Lagged covariates (SDM)					x	x
Observations	7,104	7,104	7,104	7,104	7,104	7,104

Notes: Row-standardised District and nearest neighbour (KNN) SWMs.

5.3. Main Results: Policy-based IV

We first present descriptive information on the ICRMP QI programme to check the validity of our IV strategy. **Table 4.5** shows District-level information on the saturation of the ICRMP QI programme enrolment. Due to no opening or closing of facilities taking place the composition of the peer groups stays constant across years[71]. The table also shows the average saturation of the ICRMP QI programme across Districts by year. There is significant variation in the intensity of ICRMP QI programme enrolment within Districts.

[71] Additionally, this implies that controlling for facility-level fixed effects also controls for District-level fixed effects.

Table 4.5: District-level facility and QI enrolment information				
District Information	**2015/16**	**2016/17**	**2017/18**	**All years**
Average number of facilities per District (standard deviation)	51 (28)	51 (28)	51 (28)	-
Minimum facilities per District	9	9	9	-
Maximum facilities per District	116	116	116	-
Average proportion of facilities enrolled in ICRMP QI programme per District (standard deviation)	0 (0.00)	0.49 (0.25)	0.48 (0.25)	0.32 (0.13)
Maximum proportion of facilities enrolled in ICRMP QI programme per District	0	1	1	0.61
Minimum proportion of facilities enrolled in ICRMP QI programme per District	0	0.03	0.18	0.13

Notes: Data on all years represents the average proportion of facility-years which are QI enrolled i.e. on average Districts have 32% of facility-years with facilities enrolled in the QI programme.

Supplementary Appendix C4-3 (Appendix Figure 4.1) shows the roll-out of the QI programme including the distribution of the programme across Districts for 2016/17 and 2017/18. This visually illustrates the variation in the fraction of facilities enrolled in the QI programme across Districts. Although some Districts have a high saturation of enrolled facilities, for many, the fraction enrolled is relatively similar. This provides suggestive evidence that even if QI enrolment decisions were made based on facility-level unobservables, these unobservables are geographically spread. We further test this by looking at the relationship between the District saturation of QI enrolled facilities and previous District average quality score. A 1 percentage point increase in the proportion of facilities enrolled in the QI programme in a District is associated with a 0.05 point increase in a District's average quality score in the previous year[72]. Although significant, the trivial magnitude suggests that District-level saturation of the QI programme was not strongly related to the previous quality level of facilities within Districts. Together these provide suggestive evidence that, even in the case of selective facility-level QI programme enrolment, this does not indicate that Districts with high saturation of the QI programme are systematically different from Districts with low saturation, providing confidence in the instruments.

Table 4.6 provides results of the balancing tests for the instruments, which assess the correlation of the instruments with facility characteristics (Lavy & Schlosser, 2011; Bifulco, 2011). Following the logic of Altonji et al. (2005) if the instruments are uncorrelated with observable characteristics

[72] We regressed Districts average quality score on the fraction of facilities enrolled in the District in the next fiscal year, controlling for a number of factors that may impact average District quality scores.

which may determine facility quality, then this increases confidence in them being uncorrelated with unobservable determinants of quality[73]. The results show the instruments are conditionally

Table 4.6: Balancing Tests for association between instrumental variables and facility- and small area-level characteristics

Outcome Specification	Facility type	Facility location	Average monthly patient headcount (2015)	Average monthly nurse working days (2015)	Population within 10km	Municipality average household size	Municipality proportion with no schooling	Municipality index of multiple deprivation
District fraction of facilities enrolled in QI programme	-0.00*	0.00	-4.58**	-0.00	10.8	0.00	0.00	0.00
	(0.00)	(0.00)	(1.84)	(0.08)	(117.1)	(0.00)	(0.00)	(0.00)
Observations	7,104	7,104	7,104	7,104	7,104	7,104	7,104	7,104
F-statistic	3.52	0.12	6.2	0	0.01	0.21	0.16	0.8
F-p-value	0.06	0.73	0.01	0.98	0.93	0.65	0.69	0.37
Past District fraction of facilities enrolled in QI programme	-0.00*	0.00	-8.59***	-0.19*	27.39	-0.00	-0.00	0.00
	(0.00)	(0.00)	(3.04)	(0.11)	(107.5)	(0.00)	(0.00)	(0.00)
Observations	7,104	7,104	7,104	7,104	7,104	7,104	7,104	7,104
F-statistic	3.87	0.20	7.96	2.97	0.06	0.04	0.7	2.04
F-p-value	0.05	0.66	0.00	0.08	0.80	0.84	0.40	0.15

Notes: Additional controls in each regression include own facilities current and past QI enrolment status and fixed effects. All regressions are estimated by OLS and include a constant. Heteroscedasticity robust standard errors in parentheses. ***p < 0.01; **p < 0.05; *p < 0.1. Table produced using District SWM specification.

Uncorrelated with most facility characteristics. As we run 16 tests we would expect approximately 2 statistically significant results by chance. Our tests result in 5 statistically significant results but the magnitude of the effects are incredibly small, to the degree they are not particularly economically meaningful. Therefore, the above tests and information reaffirm the assumption that the fraction of facilities in a District enrolled in the QI programme is a credible source of exogenous variation affecting average District quality scores but not facilities own quality scores.

Table 4.7 reports the final instrument test and the main results from the study based on IV fixed effects estimates of Eq. **(4)**[74]. Although we reject the null of the Hansen J test for our preferred specification (column J), we perform two subsequent checks on this[75]. The Hansen J test obtains multiple estimates of the endogenous peer effects based on subsets of the instruments (i.e. using current and past enrolment separately), testing whether estimated effects are similar. If all instruments are uncorrelated with the error term, all subsets should (asymptotically) return the

[73] Therefore this approach uses the degree of selection on observables as a guide to the degree of selection on the unobservables. The approach provides confidence in the instruments to the extent that the unobservables affecting facility quality scores are correlated with the observable determinants. Of course, this provides suggestive rather than conclusive evidence of the exogeneity of the instruments.

[74] We use the xtivreg2 command for the Fixed Effects Instrumental Variable analysis (Schaffer, 2010).

[75] It is also worth noting that the joint null hypothesis of the test is that the instruments are valid instruments, i.e., uncorrelated with the error term, and that the excluded instruments are correctly excluded from the structural equation.

same peer effect estimate[76]. However, even if instruments are valid, in the presence of heterogeneous treatment effects instruments may identify different local average treatment effects (LATEs) causing instrument subsets to estimate different effects and reject the null (Baiocchi et al. 2014). Therefore, this rejection might instead suggest the constant effects model assumption is not appropriate. We check this in **Supplementary Appendix C4-6**. As outlined, we do not believe there to be any substantive reason for a violation of the orthogonality of the instruments and the error, $Cov(\bar{z}_{(i)j}, \varepsilon_{it}) \neq 0$. However, the test may also suggest violations of the exclusion restriction. Therefore, we also check the robustness of our results to violations of the exclusion restriction in **Section VI.4.1**.

We report results from the LIM model before presenting estimates from our instrumental variable specification. First stage results are reported in **Supplementary Appendix C4-4**. As expected, a strong and significant effect is found from the LIM models (columns A-E). However, due to the identification issues discussed, we do not treat these as credible estimates of endogenous peer effects. The primary results are Columns (F-J) of **Table 4.7**, using current and past District-level saturation of the ICRMP QI programme as an instrument for District leave-out average of ICRMP quality. For our preferred specification (column J), a 1-unit increase in average District peer quality causes own facility quality to increase by 0.36 ($p<0.01$)[77]. Unlike the spatial econometric models, using the nearest neighbour peer group specification, $\boldsymbol{W^{KNN}}$, the endogenous peer effects are also positive and significant (from KNN=7). A similar pattern was seen in models 1-3 of the spatial econometric approach (**Table 4.4**). Due to the likely overlap in peer facilities between the peer group specifications (i.e. many of the nearest neighbours will be within the same district for many facilities) it is not possible to speculate at which level peer effects may be most prominent. Additionally, the limited effect observed for small nearest neighbour peer specifications might be due to less variation in QI enrolment and quality scores in smaller geographic areas.

[76] Therefore, technically the J-statistic does not allow for testing of instrument validity, which is an untestable identifying assumption, but only tests whether the different instruments identify different parameters.

[77] All results are largely the same as those obtained when running the models on the available data prior to multiple imputation. Imputation was necessary to allow utilisation of the full set of facilities for which we had 3 years of outcome data. Results available upon request.

Table 4.7: Panel policy-based models

SWM model	**Linear-in-means model (Manski model)**					**Fixed-Effects Instrumental Variable**				
ICRMP Aggregate Quality Score	**(A)**	**(B)**	**I**	**(I(E)**	**(F)**	**(G)**	**(H)**	**(I)**	**(J)**	
District Classification	0.98*** (0.01)	0.86*** (0.01)	0.86*** (0.01)	0.82*** (0.02)	0.88*** (0.03)	0.99*** (0.02)	0.79*** (0.02)	0.79*** (0.02)	0.60*** (0.03)	0.36*** (0.09)
Nearest Neighbour Peer Classification (KNN = 3)	0.55*** (0.01)	0.42*** (0.01)	0.42*** (0.01)	0.30*** (0.02)	0.16*** (0.02)	0.84*** (0.02)	0.62*** (0.02)	0.62*** (0.02)	0.42*** (0.03)	-0.04 (0.06)
Nearest Neighbour Peer Classification (KNN = 7)	0.70*** (0.01)	0.52*** (0.02)	0.51*** (0.02)	0.33*** (0.02)	-0.04 (0.03)	0.93*** (0.02)	0.72*** (0.02)	0.71*** (0.02)	0.54*** (0.03)	0.16* (0.09)
Nearest Neighbour Peer Classification (KNN = 10)	0.75*** (0.01)	0.56*** (0.02)	0.56*** (0.02)	0.37*** (0.02)	-0.03 (0.03)	0.95*** (0.02)	0.73*** (0.02)	0.73*** (0.02)	0.56*** (0.03)	0.30*** (0.09)
Nearest Neighbour Peer Classification (KNN = 13)	0.78*** (0.01)	0.59*** (0.02)	0.58*** (0.02)	0.40*** (0.02)	-0.04 (0.03)	0.96*** (0.02)	0.74*** (0.02)	0.74*** (0.02)	0.58*** (0.03)	0.41*** (0.10)
Nearest Neighbour Peer Classification (KNN = 17)	0.80*** (0.01)	0.62*** (0.02)	0.61*** (0.02)	0.42*** (0.02)	-0.04 (0.04)	0.97*** (0.02)	0.75*** (0.02)	0.75*** (0.02)	0.60*** (0.03)	0.54*** (0.10)
Controls										
ICRMP Quality Improvement Programme Enrolment		x	x	x	x		x	x	x	x
Past ICRMP Quality Improvement Programme Enrolment				x	x				x	x
Year					x					x
Facility Patient Headcount			x	x	x			x	x	x
Facility staff levels			x	x	x			x	x	x
R-squared	0.49	0.54	0.54	0.54	0.55	0.50	0.53	0.53	0.53	0.49
Kleibergen–Paap F	5391	3240	1569	1283	941	3739	2284	1093	911	728
Hansen J (p-value)	-	-	-	-	-	0.29 (0.59)	125 (0.00)	126 (0.00)	79 (0.00)	80 (0.00)
Observations	7,104	7,104	7,104	7,104	7,104	7,104	7,104	7,104	7,104	7,104

Note: Coefficients present effect of leave-one-out mean ICRMP quality scores of PHC facilities on facility quality. ***p < 0.01; **p < 0.05; *p < 0.1. Model summary statistics relate to District Classification estimates. R-squared calculated as the average of inference on the M=10 MI datasets.

5.4. Sensitivity

5.4.1 Potential violations of the exogeneity of fraction of peers enrolled in the QI programme

We assess the robustness of our Policy-based IV results to possible violations of the exclusion restriction using an approach proposed by Conley et al. (2012). The approach allows the instruments – $\bar{z}_{(i)jt}$ and $\bar{z}_{(i)jt-1}$ in our case – to enter the outcome equation:

$$y_{it} = \rho\bar{y}_{(i)jt} + \beta_1{}'X_{it} + \beta_2\bar{\bar{X}}_{(i)jt} + \beta_3 z_{it} + \gamma_1\bar{z}_{(i)jt} + \gamma_2\bar{z}_{(i)jt-1} + \alpha_i + \lambda_t + \varepsilon_{it} \qquad (7)$$

Where $\gamma \neq 0$, instead of the dogmatic IV exclusion restriction that imposes $\gamma = 0$. In many cases γ may be close to but not exactly 0, $\gamma \approx 0$. Using information on the possible extent of the violation of the exclusion restriction allows us to produce bounds on the size of ρ, the endogenous peer effects. Conley et al. (2012) present a number of ways of incorporating possible deviations from the key identifying assumption, through either imposing support restrictions or assuming reasonable prior distributions on γ. We use Conley's Union of Confidence Intervals (UCI) method to specify minimum and maximum values which γ may take. Because concern around the potential violation of the exclusion restriction is for the current and past fraction of peers enrolled in the QI programme (the instruments) to be conditionally negatively related to facility quality scores (outcome), we set $\gamma_{max} = 0$ and $\gamma_{min} = \{0, -0.01, -0.02, -0.04, -0.08\}$[78][79]. The methods are ideally suited to our context as the bounds produced are most informative when the instruments used are strong.

[78] Specifically these γ_{min} values refer to the values assigned to γ_1 in equation (7). We assign respective values to γ_{2min} corresponding to $\gamma_{2min} = 0.25 * \gamma_{1min}$ under the realistic assumption that the effect of peers past QI enrolment having a lesser effect on current average peer quality score.

[79] In the ideal implementation of the method we would have a consistent estimate of the direct effect of the fraction of peers enrolled in the QI programme on facility quality score. Van Kippersluis & Rietveld (2018) suggest obtaining this input from a subsample for which the instrument does not affect the treatment variable, which they call the 'zero-first stage subsample'. However, in our case we do not have such a subsample.

Figure 4.2: Conley-Hansen-Rossi Bounds

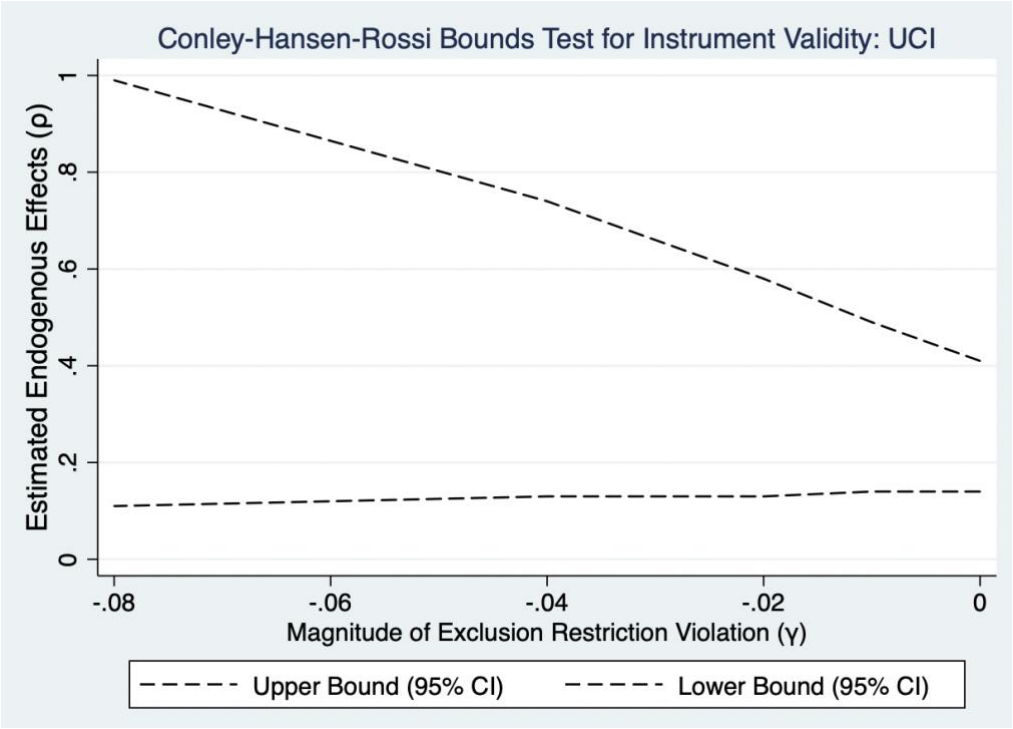


Figure 4.2 presents Conley-Hansen-Rossi bounds on the estimate of the endogenous peer effects[80]. This constitutes a strong test of our inference using the Policy-based IV, showing that our results are robust to weak violations of the exclusion restriction. Moreover, the results show our results are likely to represent a lower bound, as the stronger the violation of instrument exogeneity, the larger the potential magnitude of the estimated peer effects. We also calculate Conley-Hansen-Rossi bounds using their local-to-zero (LTZ) method which provides similar results (**Appendix C4-5**)[81].

5.4.2 Falsification Test

If peer effects operate through the peer groups as outlined above, we should not be able to obtain our results with random peer allocation. Therefore we undertake a falsification test to strengthens the credibility of our peer effects identified within the observed peer groups by ruling out the presence of peer effects within randomly generated peer groups. Following the procedure of Roychowdhury (2019), we randomly allocate facilities to Districts creating pseudo peer groups and

[80] We calculate these bounds using the *plausexog* command Clarke & Matta (2018). Again we 'concentrate out' the fixed effects in this method using Frisch-Waugh-Lovell theorem.
[81] This method also provides point IV estimates at different levels of violation of the exclusion restriction, but requires imposing more assumptions on the parameters γ_1 and γ_2 as it requires imposing a full prior distribution.

estimate equation (4). We obtain estimates of endogenous peer effects on 100 iterations of the pseudo peer groups.

Figure 4.3: Histogram of falsification tests

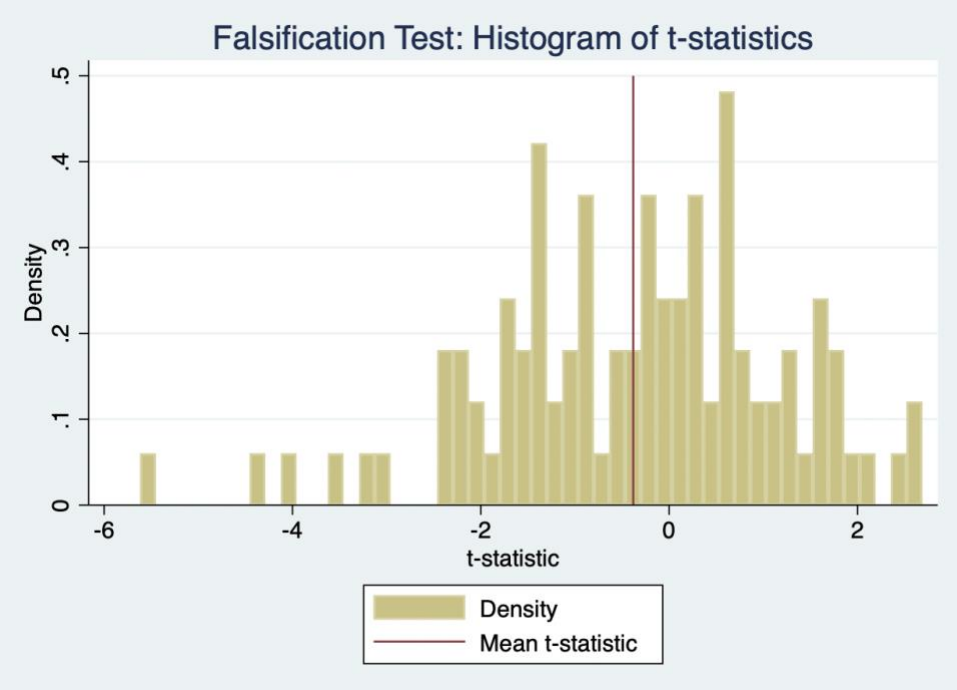


Figure 4.3 illustrates the empirical distribution of t-statistics corresponding to the peer effects estimated from 100 replications. As we cannot reject the null on the hypothesis of peer effects on the randomised peer groups in a majority (82/100) of simulations, this strengthens the credibility of the peer effects identified for true peers. The mean t-statistic is -0.38.

6. Discussion

In this chapter we exploit a QI intervention to examine the impact peer effects may have on the quality of health care provided by public PHC facilities. Our results suggest there are positive strategic interactions, with facilities improving quality in response to changes in peers' quality, even in lieu of an external financial incentive to do so. The magnitude of the peer effects in our preferred specification (0.36) are somewhat comparable to those identified by Gravelle et al. (2013) who find a 10% increase in peer hospital quality results in a 1.7-2.9% increase in own quality. The slightly larger magnitude of our findings might result from the quality measures examined in our context which are arguably more strongly determined by health facility efforts and less susceptible to external factors.

To our knowledge, this is the first attempt to examine strategic interactions between health facilities in a LMIC. It is also the first study to examine the topic of peer effects between health facilities using a quasi-experimental approach based on a policy change. While our primary empirical strategy is closely related to the S2SLS approach, the identification condition comes from a policy change impacting facility quality scores, rather than assumptions around facilities' interaction structure and model functional form. Finally, while previous studies have tested whether financial incentives linked to provider reimbursement may cause strategic interactions between health facilities, we explore whether these interactions may exist even without material incentives. Research on non-financial motivations and determinants of health care provider effort and quality offer alternative channels through which peer effects may operate, whereby significance is placed on being seen as following peer group norms and standards. Our empirical findings are, therefore, consistent with a growing recent literature suggesting that non-financial incentives can be utilised – such as public measurement and reporting – to improve the quality of health care services.

The potential policy-importance of supply-side reputational peer effects is clear when the scale of the payment reforms required to implement more traditional financial competition between health care facilities in LMICs is considered. However, even if prospective payment systems are implemented, to induce quality competition among providers relies on information being available to patients allowing informed choice and patients' ability to interpret quality metrics. This has been shown to be difficult in high-income contexts, suggesting potential greater difficulties in LMICs where the elasticity of demand for health care with respect to quality is likely lower due to access constraints. Comparatively, in many settings, introducing measurement and peer-to-peer reporting of metrics of health care facility quality may require relatively modest reforms. Therefore, if quality information reported on and to health care suppliers can induce quality competition, even without the accompanying material incentives, this may be a relatively straight-forward policy option when attempting to improve health care quality. It has also been noted that it may be desirable to vary the information included in public reporting depending on if the intended recipient are peers or patients (Kolstad, 2013). Additionally, research has also shown how providers' personal characteristics can impact responses to incentives, with a trade-off in quality determined by intrinsic motivation and quality determined by financial incentives (Bowles & Polania-Reyes, 2012; Donato et al. 2017; Ashraf et al. 2020). As such, policies based solely on public reporting may act as complements to policies based on financial incentives.

In addition to the direct policy relevance of whether measurement and public reporting can stimulate peer effects there are wider reasons why understanding quality peer effects is important. The existence of a social multiplier effect, if present, suggests the aggregate impact of quality improvement policies may be larger than the sum of the effects on individual facilities. If peer effects were present and unenrolled facilities were adjusting quality in response to changes in peers' quality then traditional evaluations of the impact of the ICRMP QI programme, not accounting for this, would understate the benefit of the ICRMP QI programme, leading to inaccurate cost-effectiveness assessments. Therefore, examination of the hypothesis of potential spill-over effects from a policy such as the ICRMP QI programme can also complement more standard policy evaluations by checking the validity of the assumption of "unpolluted" counterfactuals.

6.1 Limitations

Facility quality scores are calculated through self-assessments. This opens up the potential for systematic measurement error in facility scores. However, both the threat of and actual verification of a sample of facilities by PPTICRM reduces concerns around 'fraudulent' quality self-assessments. Additionally, even if quality scores do not accurately reflect true facility quality, the observed peer effect suggests that facilities respond in their reported quality, still suggesting a prosocial motivation effect.

Our study suffers from a common issue in the peer effects literature in that peers are assumed, and, to an extent, determined by data limitations. Ideally, we would have access to more detailed information on how facilities define their peers. Such information could be gathered through subjective assignment of peers by facilities themselves or through observing certain pathways through which facilities may interact. We addressed this by examining two distinct ways in which peer groups may be composed – geographical and institutional – as well as undertaking sensitivity analysis around peer group definitions. There is a literature examining endogenous peer group formation and the difficulties of separating peer selection from peer influence in order to obtain valid casual estimates of peer effects (Johnson & Roger Moon, 2021). However, due to the set placement of facilities we do not believe this problem to be present in our setting. It may be the case that none of the specifications examined correspond exactly to the true manner in which facilities classify their peer group. Based on **Table 4.3**, 68% of facilities have monthly patient headcounts between 559–3,793 and 3.2-158.4 nurse working days per month (i.e. one standard deviation from the mean). This indicates that despite the analysis only examining PHC facilities, even within this category facility sizes varies significantly. Given these large differences in these

characteristics of facilities, it is possible other criteria may determine who facilities view as their peers. Additionally, all our models assume peer effects are bidirectional within peer groups[82]. However, it may be the case that certain facilities are viewed as 'leaders' and others 'followers'. This would counter the assumptions of bidirectionality and that peer group mean quality is of central concern to facilities. Likewise, a single rule has been outlined for defining peer groups across the whole of SA, however, it is equally possible that different facilities may have different means of defining their respective peer groups, either through different parameters within a specification or via different means entirely. Any deviation from the true peer group specification would attenuate our estimates of peer effects.

All the specifications, including our preferred – equation (4) – made two important assumptions. First, facilities at different points of the quality distribution respond similarly to changes in the average quality of their peers. Second, facilities care about and respond to mean peer quality. We examine the potential for impact heterogeneity of changes in average peer quality at different points of the quality distribution using instrumental variable quantile regression (Lee, 2007; Chernozhukov et al. 2010; 2015). To test whether facilities care about changes in the average of peer quality we split the sample into 'high' and 'low' performers and re-run equation (4). We find facilities at a lower point in the quality distribution have a slightly larger reaction to changes in peer quality while we find in general that facilities respond to changes in quality of 'high' performing peers and react less to quality changes among 'low' performers. Both these results, to an extent, corroborate the idea that peer effects in our context are based on prosocial motivation involving norms and standards, whereby facilities do not want to be seen as lagging behind their peers (see **Supplementary Appendix C4-6**).

The ICRMP quality checklist focuses on structural and process measures of quality. Although outcome measures of health care quality may be most relevant from a policy perspective, they require risk-adjusting in order to capture the effect of health facilities' efforts and ensure they are a comparable measure of quality across facilities. Because structure and process measures are more significantly determined by health facilities efforts than outcome measures this suggests observed

[82] We focus attention on PHC facilities in SA, and characteristics of this facility type are relatively comparable. Due to the relative homogeneity of facility type this strengthens the assumption that relationships should be bidirectional and averages of peers might be more important that certain specific 'leader' facilities.

changes in quality may more closely match changes in facility efforts. Therefore, this potentially reduces noise in examining peer effects in health care quality[83].

Although a sizeable private health care sector exists in SA, we contend that – due to the proposed mechanism for facility responsiveness – the presence and quality of private facilities should not impact our relationship of interest.

Given that all facilities are aware they would eventually be enrolled in the ICRMP QI programme, there is a possibility this may reduce the incentive for unenrolled facilities to exert efforts to increase quality. This potential unintended consequence of the ICRMP QI programme may therefore alter unenrolled facilities responsiveness to quality changes of peers. Such an anticipatory effect among the unenrolled facilities cannot be ruled out.

Finally, while we have identified the occurrence of quality peer effects without the presence of financial incentives, more work is required to gain insights to the specific mechanism stimulating facilities to respond to peers. While we have proposed how various non-financial motivations and determinants of provider effort and quality may act to induce provider strategic interactions and quality spill-overs in lieu of financial incentives, we are unable to pinpoint which specific mechanism is in action. The peer effects may result from competitive pressures such as prosocial motivation and reputational concerns or a sense of professionalism and professional pride, or they may be collaborative and social learning related, with better performers able to identify and reach out to assist poor performing peers, or poor performers able to learn from their higher quality peers.

6.2 Conclusion

In the search for effective policies to improve the quality of health care supplied in LMICs, a number of potential alternatives to material incentives have been examined; including observation and measurement (Leonard & Masatu, 2017), community engagement (Bjorkman et al. 2009), encouragement (Lee, 2018), clinical guidelines and protocols (Papanicolas et al. 2015), and public reporting (Smith et al. 2010). Policies may have the potential to leverage prosocial motivation to induce provider quality competition, strengthening the case for improving data collection and M&E in health systems in LMICs. While in the long-term health systems should continue to

[83] One potential related consideration is that as quality scores may be viewed as reflective of efforts this may also encourage greater awareness and care for relative scores compared to circumstances where convoluted processes can be cited as contributing to differences in health outcome quality metrics.

pursue payment and purchasing reforms, policies that do not require extensive financial system reform may be considered as short-term substitutes and long-term complements to such reforms.

Discussion

This thesis has investigated the impact of various non-financial health system policies on access to and the utilisation of health care services and the quality of health care in LMICs.

The first two chapters deal with determinants of the demand for and utilisation of obstetric health care. In Chapter 1, we examine the effect of distance on the utilisation of institutional delivery services. We explore heterogeneity in the effect of distance on facility delivery along the values of distance observed. Additionally, we attempt to identify the causal effect of a change in distance on the probability of having a facility delivery. The key results are that distance continues to act as a barrier to obstetric health care utilisation – even in an environment with high rates of institutional delivery – and that this effect is more pronounced in women with poor health knowledge and lower socio-economic status. Further, estimates of the impact of distance on utilisation not addressing unobserved confounding risk underestimate the negative effect of distance on utilisation.

This chapter contributes to a large literature examining the association between geographical accessibility and health care utilisation. However, to date, health infrastructure planning in LMICs has lacked information to guiding geographic access policies and informing health infrastructure planning. The impracticality of randomising geographical access to health care services is likely partially responsible for this evidence gap. By going beyond simply identifying a distance-utilisation relationship, the evidence we generate can be combined with population distribution data to significantly improve evidence-based infrastructure planning.

Chapter 2 examines the effect of MWHs, a policy aimed at reducing the geographical access burden for women residing at further distances from facilities. The results suggest that MWHs do not have a significant effect on obstetric health care utilisation rates. This null effect should be interpreted with caution due to limited power of the analysis. However, this power issue in part stems from a relatively low ceiling in the potential effectiveness of the policy due to pre-existing high delivery rates. As such, there are potential questions regarding the contextual relevance of the policy. Additionally, the implementation of the policy – whereby most MWHs were constructed at urban District Hospitals – may also have stunted its effectiveness.

Future research should focus on understanding the mechanisms through which distance reduces utilisation. Distance may pose a physical obstacle to accessing care or women residing in more remote settings may have different health-seeking preferences or need. The hypothesised mechanisms behind the unobserved confounding and the higher utilisation rates in women with greater health knowledge in Chapter 1 hint that the latter might be more of an issue in this setting. This suggests that MWHs, a policy response aimed at increasing the accessibility of maternal health care services, may not be the optimal policy prescription. Research detailing the factors mediating the effect of distance on utilisation can provide useful insights to assist policy-makers in ensuring access policies target the specific factors dissuading women from seeking care.

An additional finding of Chapter 1 is that individuals frequently bypass the nearest facility, seeking care at higher levels. This is in-line with the growing empirical support for the active patient model, whereby patients seek out high-quality care rather than the closest or lowest cost provider (Leonard, 2014). Combined with the limited effects of MWHs, this points to a fruitful area for further research being the development of more comprehensive health facility choice models, providing a better insight into the relative preferences of patients which determine whether and where they seek care. However, gaining a fuller understanding of population health-seeking behaviours in LMICs will require overcoming a number of data limitations, many of which are faced in Chapters 1 and 2.

Both chapters suffer from a lack of accurate information on household location due to the practice of geo-masking true locations to maintain respondent anonymity. While the rationale for geo-masking is clear, there are a number of simple procedures which could maintain anonymity while improving the accuracy of analysis. For example, the producers of DHS data have accurate household location information prior to displacement. There is little preventing the data curators calculating a number of useful indicators using this information, such as distance to nearest facility, prior to geo-masking or undertaking more sophisticated geo-masking procedures which balance anonymity concerns and limiting consequences for ensuing analysis (Arbia et al. 2015).

A further potential source of measurement error is the reliance on household surveys for birth records and child outcome histories. The reliance on such data stems from the lack of administrative patient data and records in most LMICs. Future improvements in routine administrative data at the patient level made available to researchers would drastically improve the quality of work on similar topics.

Relatedly, the inability to accurately link households with the facilities at which health care is sought restricts the scope of research questions which can be addressed. Specifically, modelling individuals' facility choices requires matching individuals to facilities at which care is ultimately sought. Only with this data will it become possible to develop a better understanding of how populations evaluate trade-offs between price, quality and other characteristics of health care when making utilisation decisions (Cronin et al. 2016). Currently, very few surveys collect this type of data. South Africa's National Income Dynamics (NIDS) survey is one example but is not publicly released and therefore not practically available to most researchers.

More work is also required on understanding how the effect of geographic accessibility impacts the utilisation of health care for different health care services. Much of the literature examining the impact of geographical accessibility has focused on obstetric health care. Although there are obvious reasons why pregnant women may be a particularly relevant group for examining this relationship, it is also relevant to know the impact of these access issues on preventative or children's health care.

Finally, greater efforts should be made to move beyond estimating correlations in the relationship between geographic accessibility and the demand for health care. Future work should aim to identify additional plausible sources of exogeneous variation. For example, the few studies that have attempted to infer a causal relationship between geographic accessibility and health care utilisation, have all relied on instruments (Kumar et al. 2013; McGuire et al. 2021). Information on health facility openings and closings – similar to that in Chapter 2 – might be sought and survey instruments designed to capture the subsequent response of affected households and communities.

Chapters 3 and 4 investigate questions of how to stimulate the quality of care supplied in settings without strong financial incentives. Vast and growing variations in the quality of health care supplied within LMICs have been observed (Kruk et al. 2017). While QI programmes offer a potential solution to these variations, if not well-designed, they also have the potential to exacerbate the problem. Chapter 3 examines heterogeneity in the effect of the ICRMP, a QI programme aiming to improve PHC facilities' structural and process quality. The results suggest that while the programme was successful in improving quality, it may have increased variation in the quality of care offered by facilities in South Africa. Given the growing policy importance placed on reducing inequalities in health care utilisation and health outcomes, examination of

heterogeneous treatment effects of QI programmes should be increasingly undertaken as standard practise. Even in cases where equity concerns are not the primary consideration, variations in treatment effects may impact QI programme's effectiveness in improving health outcomes. As noted, a significant burden of mortality has been attributed to poor quality care (National Academies of Sciences, Engineering, Medicine, 2018). This suggests larger health effects of QI programmes are likely to be observed if programmes disproportionately impact health care providers at the bottom end of the quality distribution. Therefore, given we found the opposite, the ICRMP QI programme's distribution of effects on quality may also be resulting in reduced effects on health outcomes.

Future research should seek to address the limitations of this chapter by using a longer panel than the 2 periods we were restricted to. Additionally, while we speculate on the mechanisms which impact past facility performance and responsiveness to the QI programmes, more work is needed to understand the causes of variation in quality at health facility level. The results of such work are crucial to ensuring subsequent policies are able to target the correct constraints causing certain facilities to fall behind their peers.

Chapter 4 examines whether PHC facilities strategically interact with peer facilities to improve their (ICRMP) quality, even in the absence of material incentives for doing so. While there is a growing literature examining strategic interactions between health care providers in HICs, to our knowledge, this is the first investigation of strategic interactions between health facilities in a LMIC. The findings suggest that quality peer effects are present suggesting financial competition between facilities may not be the only way to induce a form of quality yardstick competition.

A growing number of LMICs are improving accountability through strengthening monitoring and reporting systems (Bjorkman et al. 2009; Berlan et al. 2012; Duke et al. 2015). Chapter 4 illustrates an additional aspect of such policies which can potentially be exploited to create competitive pressures between providers. A key distinction between this study and previous work exploring strategic competition between health care providers in HICs is the recognition that most LMICs currently lack the health system and financing architecture to enable financial pressures to drive quality (Cooper et al. 2012; Gaynor et al. 2013; Gravelle et al. 2014; Longo et al. 2017). Therefore, the work utilises the growing evidence on intrinsic and prosocial motivation (Ashraf et al. 2020) to suggest that this non-financial determinant of quality might be leveraged to induce quality competition among health care facilities. The findings that alternative non-financial mechanisms

might induce similar behaviours as financial incentives has important policy implications. The results strengthen the arguments for improving standards of quality monitoring, but also suggest the need to think about how indicators should be reported. While most LMICs will and should continue to pursue health financing reforms in order to provide incentives which can help shrink the difference between competence and performance, these findings suggest policies that act as prosocial incentives such as measurement and public reporting may represent low-hanging fruit to improve quality in the short-term.

Chapters 3 and 4 raise a number of points which future work examining similar research questions and quality of care generally should consider. A general limitation of research examining health care quality relates to the measures of quality used. Health care quality is grouped into three types: structural, process and outcome quality (Donabedian, 1988). As is common in much research on health care quality, the metrics used in these chapters are largely restricted to structural and process measures of quality (see **Supplementary Appendix C3-1**). While there is growing evidence that a wide range of QI programmes can and do improve structural and process quality, there is less evidence that this translates into improved outcome quality. As improving health outcomes is the ultimate objective of these programmes, greater efforts should be made to capture indicators measuring aspects of outcome quality. The lack of outcome quality data continues to be a common problem when assessing many global health policies. The lack of policy evaluations with endpoints such as mortality and morbidity, which fall under outcome quality, has been noted as problematic (Cohen & Easterly, 2009). The ultimate impact of QI programmes which do not capture measures of outcome quality on population health will remain uncertain without information on the relation between changes in structural and process quality on mortality and morbidity. An additional benefit of capturing such endpoints is that it would improve the comparability of the effectiveness of different health system policies.

Future research on health facility peer effects would benefit from careful study design to allow the examination of the effect of the introduction of measurement and public reporting of quality metrics in contexts with weak accountability and limited financial incentives. This would require separating out measurement and reporting such that indicators would be observable to the researcher prior to being reported to the target audience, or clever design of how new quality information affects behaviours such as the approach used by Kolstad (2013).

Additionally, as health facility quality is the aggregation of the competence, capacities and efforts of the health workers staffing them (Ibnat et al. 2019), similar competitive pressures which can be leveraged towards quality improvement may exist at the staff level. This warrants exploration as an alternative means of quality improvement.

Designing a health system that delivers high-quality health care is an exceptionally complex task, requiring a diverse range of health system inputs and ideally rigorous evidence-informed decisions. Evaluation of population- and system-wide health system policies, such as those examined in this thesis, is frequently infeasible using randomised controlled trials (RCTs). Even if it is possible to undertake smaller pilots of such programmes and policies more amenable to RCTs, there is often little information to suggest results would hold after they are scaled up. However, population- and system-wide health system policies play a large role in determining levels of access to and quality of health care. Therefore, it is necessary to generate evidence on these aspects of health systems with the understanding that future novel data or improvements in methods can add to the knowledge base. With this in mind, this thesis has evaluated a number of non-financial population- and system-wide health system policies that impact the utilisation and quality of health care in LMICs, providing evidence which can inform health system planning, while also clarifying questions requiring further research.

Bibliography

Abadie, A. Athey, S. Imbens, G. Wooldridge, J. (2020). Sampling-Based versus Design-Based Uncertainty in Regression Analysis. *Econometrica.* 88.1. 265-296.

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies.* 72.1. 1-19.

Acton, J. (1973). Demand for Health Care when Time Prices Vary More than Money Prices. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/reports/R1189.html. Also available in print form.

Ahmed, S. Khan, M. (2011). Is demand-side financing equity enhancing? Lessons from a maternal health voucher scheme in Bangladesh. *Social Science and Medicine.* 72. 1704-1710.

Akachi, Y. Kruk, M. (2017). Quality of Care: Measuring a Neglected Driver of Improved Health. *Bulletin of the World Health Organisation.* Vol. 95. No. 6. 465-472.

Alegana, V. Wright, J. Pentrina, U. Noor, A. Snow, R. Atkinson, P. (2012). Spatial modelling of healthcare utilisation for treatment of fever in Namibia. *International Journal of Health Geographics.* 11(6).

Altonji, J. Elder, T. Tabor, C. (2005). Selection on observed and unobserved variables: assessing the effectiveness of catholic schools. *Journal of Political Economy*. 113. 151–184.

Andrew, A. Vera-Hernandez, M. (2020). Incentivising demand for supply-constrained care: Institutional birth in India. *IFS Working Paper.*

Anglewicz, P. VanLandingham, M. Manda-Taylor, L. Kohler, H. (2016). Migration and HIV Infection in Malawi. *AIDS*. 30. 2099-2105.

Anglewicz, P. VanLandingham, M. Manda-Taylor, L. Kohler, H. (2017). Cohort profile: internal migration in sub-Saharam Africa – the Migration and Health in Malawi (MHM) study. *BMJ Open*. 7.

Angrist, J. (2014). The Perils of Peer Effects. *Labour Economics.* 30. 98-108.

Angrist, J. Krueger, A. (2000). Empirical Strategies in Labour Economics. In Handbook of Labour Economics. Ed. Ashenfelter, O. Card, D. Amsterdam, Elsevier. 1277-1366.

Angrist, J. Pischke, S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Anselin, L. (1988). Spatial Econometrics: Methods and Models. Kluwer Academic, Dordrecht.

Arbia, G. (2014). A Primer in Spatial Econometrics. Palgrave Texts in Econometrics.

Arbia, G. Battisti, M. Di Vaio, G. (2009). Institutions and geography: empirical test of spatial growth models for European regions. *Economic Modelling.* 27. 12–21.

Arbia, G. Espa, G. Giuliani, D. (2015). Measurement errors arising when using distances in microeconometric modelling and the individuals' position is geo-masked for confidentiality. *Econometrics.* 3(4). 709-718.

Ashraf, N. Bandiera, O. Lee, S. (2020). Losing prosociality in the quest for talent? Sorting, selection, and productivity in the delivery of public services. *American Economic Review.*

Asim, S. Chimombo, J. Chugunnov, D. Gera, R. (2017). Moving Teachers to Malawi's Remote Communities: A Data-Drive Approach to Teacher Deployment. *Policy Research Working Paper.* Education Global Practise Group. World Bank.

Atella, V. Belotti, F. Depalo, D. Piano Mortari, A. (2014). Measuring spatial effects in the presence of institutional constraints: The case of Italian Local Health Authority expenditure. *Regional Science and Urban Econoomics.* 49(C). 232-241.

Athey, S. Imbens, G. (2006). Identification and Inference in Non-Linear Difference-in-Difference Models. *Econometrica.* Vol. 74. No. 2.

Ai, C. Norton, E. (2003). Interaction terms in logit and probit models. *Economics Letters.* 80.1. 123-129.

Aday, L. A. Anderson, R. (1974). A Framework for the Study of Access to Medical Care. *Health Services Research.* 208-220.

Baiocchi, M. Cheng, J. Small, D. (2014). Tutorial in Biostatistics: Instrumental Variable Methods for Causal Inference. *Statistical Medicine.*

Banerjee, A. Deaton, A. Duflo, E. (2004). Health, Health Care, And Economic Development: Wealth, Health, and Health Services in Rural Rajasthan. *American Economic Review.* Vol. 94. No. 2.

Banerjee, A. Finkelstein, A. Hanna, R. Olken, B. Ornaghi, A. Sumarto, S. (2021). The Challenges of Universal Health Insurance in Developing Countries: Experimental Evidence from Indonesia's National Health Insurance. *American Economic Review.* Vol. 111. No. 9.

Barber, S. Gertler, P. (2010). Empowering women: how Mexico's conditional cash transfer programme raised prenatal care quality and birth weight. *Journal of Development Effectiveness.* 2.1. 51-73.

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

Basker, E. (2007). When Good Instruments Go Bad: A Reply to Neumark, Zhang, and Ciccarella. *SSRN.* DOI: 10.2139/ssrn.980988.

Becker, S. Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *Stata Journal.* Vol. 2. No. 4. 358-377.

Benabou, R. Tirole, J. (2003). Intrinsic & Extrinsic Motivation. *The Review of Economic Studies.* 70(3).

Benabou, R. and J. Tirole (2006), Incentives and Prosocial Behaviour'. *American Economic Review*. 96(5), 1652–78.

Berlan, D. Shiffman, J. (2012). Holding health providers in developing countries accountable to consumers: a synthesis of relevant scholarship. *Health Policy & Planning*. 27: 271–80.

Bertrand, M. Duflo, E. Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics*. 119 (1):249-275.

Besley, T. Burgess, R. (2002). The Political Economy of Government Responsiveness: Theory and Evidence from India. *Quarterly Journal of Economics*. 1415-1451.

Besley, T. Case, A. (2002). Unnatural Experiments? Estimating the Incidence of Endogenous Policies. *The Economic Journal*. Vol. 110. No. 467. 672-694.

Besley, T. Ghatak, (2003). Incentives, choice, and accountability in the provision of public services. *Oxford Review of Economic Policy*. 19.2. 235-249.

Besley, T. Ghatak, (2007). Reforming Public Service Delivery. *Journal of African Economies*. 16.1. 127-156.

Besley, T. Ghatak, M. (2005). Competition and incentives with motivated agents. *American Economic Review*. 95 (3). 616-636.

Bevan, G. Evans, A. Nuti, A. (2019). Reputations count: why benchmarking performance is improving health care across the world. *Health Economics, Policy and Law*. 14. 141-161.

Bia, M. Mattei, A. (2012). Assessing the effect of the amount of financial aids to Piedmont firms using the generalized propensity score. *Statistical Methods and Applications*. 21(4). 485-516. DOI: 10.1007/s10260-012-0193-4.

Bia, M. Flores, C. Mattei, A. (2011). Nonparametric Estimators of Dose-Response Functions. No 2011-40, *LISER Workinng Paper Series*. LISER.

Bia, M. Flores, C. Flores-Lagunes, A. Mattei, A. (2014). A Stata Package for the Application of Semiparametric Estimators of Dose–Response Functions. *The Stata Journal*. 14(3). 580-604. DOI: 10.1177/1536867X1401400307.

Bia, M. Flores, C. Mattei, A. (2008). A Stata package for the estimation of the dose–response function through adjustment for the generalized propensity score. *The Stata Journal*. 8(3). 354-373. DOI: 10.1177/1536867X0800800303.

Bifulco, R. Fletcher, J. Ross, S. (2011). The effect of classmate characteristics on individual outcomes: Evidence from the add health. *American Economic Journal: Economic Policy*. 3. 25–53.

Binyaruka, P. Lohmann, J. De Allegri, M. (2020). Evaluating performance-based financing in low-income and middle-income countries: the need to look beyond average effect. *BMJ Global Health*. 5.

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. *International Journal of Equity in Health.* 17. 14.

Binyaruka, P. Robberstad, B. Torsvik, G. Borghi, J. (2018). Does payment for performance increase performance inequalities across health providers? A case study of Tanzania. *Health Policy and Planning.* 33. 1036-1036.

Bitler, P. Gelbach, B. & Hoynes, W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*. 96, 988– 1012.

Bitler, P. Gelbach, B. & Hoynes, W. (2008). Distributional impacts of the Self-Sufficiency Project. *Journal of Public Economics.* 92, 748–765.

Bivand, R., Hauke, J. and Kossowski, T. (2013). Computing the Jacobian in Gaussian Spatial Autoregressive Models: An illustrated Comparison of Available Methods. *Geographical Analysis*. 45. 150–179.

Bjorkman, M. Svenson, J. (2009). Power to the people: evidence from a randomised field experiment on community-based monitoring in Uganda. *Quarterly Journal of Economics.* 124:735-69.

Blandford, J. Kumar, S. Luo, W. MacEachren, A. (2012). It's a long, long walk: accessibility to hospitals, maternity and integrated health centres in Niger. *International Journal of Health Geographics*. 11(24). DOI: 10.1186/1476-072X-11-24.

Blencowe H, Cousens S, Jassir FB, Say L, Chou D, Mathers C. (2016). National, regional, and worldwide estimates of stillbirth rates in 2015, with trends from 2000: a systematic analysis. *Lancet Global Health.* 4:e98-e108.

Bloom, N. Propper, C. Seiler, S. Van Reenen, J. (2015). The impact of competition on management quality: evidence from public hospitals. *The Review of Economic Studies.* 82.2. 457-489.

Blundell, R. Dias, M. (2009). Alternative Approaches to Evaluation in Empirical Microeconomics. *Journal of Human Resources*. 44.3. 565-640.

Borah, B. (2006). A mixed logit model of health care provider choice: analysis of NSS data for rural India. *Health Economics.* 15. 915-932. DOI:10.1002/hec.1166.

Bordignon, M. Cerniglia, F. Revelli, F. (2004). Yardstick competition in intergovernmental relationships: theory and empirical predictions. *Economics Letters.* 83, 325–333.

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

Borusyak, K. Jaravel, X. Spiess, J. (2022). Revisiting Event Study Designs: Robust and Efficient Estimation. *Working Paper.*

Bosch-Capblanch, X. Liaqat, S. Garner, P. (2011). Managerial supervision to improve primary health care in low- and middle-income countries. *Cocchrane Database of Systematic Reviews.* 7.

Bowles, S. (2016). The Moral Economy: Why Good Incentives Are No Substitute for Good Citizens. Yale University Press.

Bowles, S. Polania-Reyes, S. (2012). Economic incentives and social preferences: substitutes or complements? *Journal of Economic Literature.* 368–425.

Bramoullé, Y. Djebbari, H. Fortin, B. (2009). Identification of Peer Effects through Social Networks. *Journal of Econometrics.* 150(1). 41–55.

Brekke, K. Chiara, C. Siciliani, L. Odd Rune, S. (2021). Hospital competition in a national health service: Evidence from a patient choice reform. *Journal of Health Economics.*

Brock, J. Lange, A. Leonard, K. (2015b). Esteem and social information: On determinants of prosocial behaviour of clinicians in Tanzania. *Journal of Economics Behaviour & Organisation.* 118. 85-94.

Brock, J. Lange, A. Leonard, K. (2015a). Generosity and Prosocial Behaviour in Healthcare Provision: Evidence from the Laboratory and Field. *Journal of Human Resources.* 51. 1.

Budhathoki, S. Pokharel, P. Good, S. Limbu*, S*. Bhattachan, M. Osborne, R. (2017). The potential of health literacy to address the health related UN sustainable development goal 3 (SDG3) in Nepal: a rapid review. *BMC Health Services Research.* 17(237). DOI: 10.1186/s12913-017-2183-6.

Bukonda, N. Tavrow, P. Abdallah, H. Hoffner, K. Tembo, J. (2002). Implementing a national hospital accreditation program: the Zambian experience. *International Journal for Quality in Health Care.* Vol. 14 No. 1. 7-16.

Buor, D. (2003). Analysing the primacy of distance in the utilization of health services in the Ahafo-Ano South district, Ghana. *International Journal of Health Planning and Management.* 18. 293-311. DOI: 10.1002/hpm.729.

Burgert, C. Colston, J. Roy, T. Zachary, B. (2013). Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys. *DHS Spatial Analysis Reports No. 7.* Calverton, Maryland, USA: ICF International.

Callaway, B. Sant'Anna, P. (2020). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics.*

Callaway, B. Li, T. Oka, T. (2018). Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics.* Vol. 206. No. 2. 395-413.

Cameron, A. Miller, D. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources.* 50.2. 317–372.

Campbell, O. Graham, W. (2006). Strategies for reducing maternal mortality: getting on with what works. *Lancet.* 368, 1284–1299.

Canay, I. (2011). A Simple Approach to Quantile Regression for Panel Data. *Econometrics Journal.* 14. 368-386.

Carroll, R. Rupert, D. Stefanski, L. (1995). Measurement Error in Nonlinear Models. London: Chapman and Hall.

Cattaneo, M. Crump, R. Farrell, M. Feng, Y. (2019). On Binscatter. DOI: arXiv:1902.09608.

Cattaneo, M. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*. 155(2):138-54.

Chabe-Ferret, S. (2015). Analysis of the bias of Matching and Difference-in-Difference under alternative earnings and selection processes. *Journal of Econometrics.* Vol. 185. No. 1.

Chalkley, M. Mirelman, A. Siciliani, L. Suhrcke, M. (2016). Paying for performance for health care in low- and middle-income countries: an economic perspective. *CHE Working Paper.* No. 140.

Chowdhury, M. Ronsmans, C. Killewo, J. Anwar, I. Gausia, K. Das-Gupta, S. Blum, L. Dieltiens, G. Marshall, T. Saha, S. Borghi, J. (2006). Equity in use of home-based or facility-based skilled obstetric care in rural Bangladesh: an observational study. *Lancet.* 367, 327–332.

Chernozhukov, V. Hansen, C. (2008). Instrumental Variable Quantile Regression: A Robust Inference Approach. *Journal of Econometrics.* 142. 379-398.

Chernozhukov, V. Fernandez-Val, I. Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*. 78. 1093–125.

Chernozhukov, V. Fernandez-Val, I. Kowalski, A. (2015). Quantile regression with censoring and endogeneity. *Journal of Econometrics.* 186. 201–21.

Clarke, D. Matta, B. (2018). Practical Considerations for Questionable IVs. *Stata Journal.* 18:3. 663-691.

Cleveland, E. Dahn, B. Lincoln, T. Safer, M. Podesta, M. Bradley, E. (2011). Introducing health facility accreditation in Liberia. *Global Public Health.* Vol. 6. No. 3. 271-282.

Cohen, J. Easterly, W. (2009). *What Works in Development? Thinking Big and Thinking Small.* Brookings Institution Press.

Conley, T. Hansen, C. Rossi, P. (2012). Plausibly Exogenous. *The Review of Economics and Statistics.* 94:1. 260-272.

Cooper, Z. Gibbons, S. Jones, S. McGuire, A. (2011). Does hospital competition save lives? Evidence from the English NHS patient choice reforms. *The Economic Journal.* 121.

Crépon, B. Duflo, E. Gurgand, M. Rathelot, R. Zamora, P. (2013). Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment. *Quarterly Journal of Economics* 128.2.: 531–80.

Cronin, C. Guilkey, D. Speizer, I. (2017). The effects of health facility access and quality on family planning decisions in urban Senegal. *Health Economics.* DOI: 10.1002/hec.3615.

Cronin, C. Guilkey, D. Speizer, I. (2016). The individual's choice of facility for maternal health and family planning services in a dense urban environment: The case of Senegal. *Working Paper.*

Dal Bó, E. Finan, F. Rossi, M. (2013). Strengthening state capabilities: the role of financial incentives in the call to public service. *Quarterly Journal of Economics.* 128:1169-218.

Das, J. Hammer, J. (2007). Money for nothing: The dire straits of medical practice in Delhi, India. *Journal of Development Economics.* 83. 1. 1–36.

Das, J., and J. Hammer. (2007). Location, Location, Location: Residence, Wealth and the Quality of Medical Care in Delhi, India. *Health Affairs* 26 (3): 338–51.

Das, J. Holla, A. Mohpal, A. Muralidharan, K. (2016). Quality and Accountability in Health Care Delivery: Audit-Study Evidence from Primary Care in India. *American Economic Review.* Vol. 106. No. 12. 3765-3799.

Das, J. Woskie, L. Raibhandari, R. Abbasi, Jha, A. (2018). Rethinking assumptions about delivery of healthcare: implications for universal health coverage. *BMJ.* 361.

Dammert, A. (2009). Heterogenous Impacts of Conditional Cash Transfers: Evidence from Nicaragua. *Economic Development and Cultural Change***.** Vol. 58. No. 1. 53-83.

Davezies, L. D'Haultfoeuille, X. Fougère, D. (2009), Identification of peer effects using group size variation. *Econometrics Journal.* 12. 397–413.

Daw, J. Hatfield, L. (2018). Matching and Regression to the Mean in Difference-in-Differences Analysis. *Health Services Research.* Vol. 53. No. 6. 4138-4156.

Daw, J. Hatfield, L. (2018). Matching in Difference-in-Differences: between a Rock and a Hard Place. *Health Services Research.* Vol. 53. No. 6. 4111-4117.

de Chaisemartin, C. D'Haultfoeuille, X. (2020). Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review.*

de Geyndt, W. (1995). Managing the Quality of Health Care in Developing Countries. *World Bank Technical Paper.* No. 258.

de Savigny, D. Mayombana, C. Mwageni, E. Masanja, H. Minhaj, A. Mkilindi, Y. Mbuya, C. Kasale, H. Reid, G. (2004). Care-seeking Patterns for Fatal Malaria in Tanzania. *Malaria Journal.* 3:27.

Deaton, A. Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine.*

Delfgauuw, J. Dur, R. (2008). Incentives and workers' motivation in the public sector. *Economic Journal.* 118. 171 – 191.

Deserranno E. (2019). Financial incentives as signals: experimental evidence from the recruitment of village promoters in Uganda. *American Economic Journal: Applied Economics.* 11:277-317.

Ding, P. Li, F. (2019). A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment. *Political Analysis.* Vol. 27. No. 4.

Djebbari, H. Smith, J. (2008). Heterogeneous Impacts of PROGRESA. *Journal of Econometrics*. Vol. 145. 64-80.

Donabedian, A. (1966). Evaluating the Quality of Medical Care. *The Milibank Memorial Fund Quarterly*. Vol. 44. No. 2.

Donabedian, A. (1988). The quality of care. How can it be assessed? *JAMA*. 260. 12. 1743–1748.

Donato, K. Miller, G. Mohanan, M. Truskinovsky, Y. Vera-Hernandez, M. (2017). Personality Traits and Performance Contracts: Evidence from a Field Experiment among Maternity Care Providers in India. *American Economic Review*.

Duke, T. Yano, E. Hutchinson, A. (2015). Large-scale data reporting of paediatric morbidity and mortality in developing countries: it can be done. *Arch Dis Child* 2015; 101: 392–7.

Duong, D. Binns, C. Lee, A. (2005). Utilisation of delivery services at the primary health care level in rural Vietnam. *Social Science and Medicine*. 59(12). 2585-2595. DOI: 10.1016/j.socscimed.2004.04.007.

DuMouchel, W. Duncan, G. (1983). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. *Journal of the American Statistical Association*. 78. 383. 535-543.

Egger, P. Ehrlich, M. (2013). Generalized propensity scores for multiple continuous treatment variables. *Economics Letters*. 119. 32-34. DOI:10.1016/j.econlet.2013.01.006.

Eijkenaar, F. (2013). Key Issues in the Design of Pay for Performance Programs. *European Journal of Health Economics*. Vol. 14. 117-131.

Elkies, N. Fink, G. Barnighausen, T. (2015). "Scrambling" geo-referenced data to protect privacy induces bias in distance estimation. *Population and Environment*. 37(1). DOI: 10.1007/s11111-014-0225-0.

Falk, A. Ihino, A. (2006). Clean Evidence on Peer Effects. *Journal of Labour Economics*. 24:1. 39-58.

Fan, J. Gijbels, I. Hu, T-C. Huang, L-S. (1996). A Study of Variable Bandwidth Selection for Local Polynomial Regression. *Statistica Sinica*. 6(1). 113-27.

Ferraz, C. Finan, F. (2008). Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes. *Quarterly Journal of Economics*. 123. 703-7.

Filmer, D. Hammer, J. Prichett, L. (2002). Weak Links in the Chain II: A Prescription for Health Policy in Poor Countries. *The World Bank Research Observer*. 17.1. 47-66.

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

Flores, C. (2007). Estimation of Dose-Response Functions and Optimal Doses with a Continuous Treatment. *Working Papers 0707. University of Miami, Department of Economics*.

Flores, C. Flores-Lagunes, A. Gonzalez, A. Neumann, T. (2012). Estimating the Effects of Length of Exposure to Instruction in a Training Program: The Case of Job Corps. *The Review of Economics and Statistics*. 94(1). 153-171. DOI: 10.1162/REST a 00177.

Flores, C. Mitnik, O. (2013). Comparing Treatments Across Labor Markets: An Assessment of Non-experimental Multiple-Treatment Strategies. *The Review of Economics and Statistics*. 95(5). 1691-1707.

Fuchs, V. (1968). The growing demand for medical care. *The New England Journal of Medicine*. 279. 190-195. DOI: 10.1056/NEJM196807252790405.

Gaarder, M. Glassman, A. Todd, J. (2010). Conditional cash transfers and health: unpacking the causal chain. *Journal of Development Effectiveness*. 2.1. 6-50.

Gabrysch, S. Campbell, OM. (2009). Still Too Far To Walk: Literature Review of the Determinants of Delivery Service Use. *BMC Pregnancy and Childbirth*. 9(34). DOI:10.1186/1471-2393-9-34.

Galizzi, M. Tammi, T. Godager, G. Linnosmaa, I. Wiesen, D. (2015). Provider Altruism in Health Economics. Discussion Paper 4.

Gage, A. Carner, F. Blossom, J. Aluvaala, J. Amatya, A. Mahat, K. Malata, A, Roder-DeWan, S. Twum-Danso, N. Yahya, T. Kruk, M. (2019). In Low- and Middle-Income Countries, Is Delivery In High-Quality Obstetric Facilities Geographically Feasible? *Health Affairs*.

Gaynor, M. Moreno-Serra, R. Propper, C. (2013). Death by market power: Reform, competition, and patient outcomes in the National Health Service. A*merican Economic Journal: Economic Policy*. 5, 134–166.

Gelman, A. Hill, J. (2007). Data Analysis using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Gertler, P. van der Gaag, J. (1990). The willingness to pay for medical care : evidence from two developing countries. Baltimore, MD : The Johns Hopkins University Press.

Gertler, P. Vermeersch, C. (2012). Using Performance Incentives to Improve Health Outcomes. Washington, DC: World Bank, Report No.: Policy Research Working Paper 6100.

Gibbons, C. Suarez Serrato, J. Michael, U. (2019). Broken or Fixed Effects. *Journal of Econometric Methods*. Vol. 8. No. 1. 1-12.

Giesselmann, M. Schmidt-Catran, A. (2018). Interactions in Fixed Effects Regression Models. *DIW Berlin Discussion Papers*.

Gneezy, U. Meier, S. Rey-Biel, P. When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives*. 25.4. 191-210.

Godager, G. Hennig-Schmidt, H. Iversen, T. (2016). Does performance disclosure influence physicians' medical decisions? An experimental study. *Journal of Economic Behaviour & Organisation*. 131. 36-46.

Goodchild, M. Haining, P. Wise, S. (1992). Integrating GIS and spatial analysis: problems and possibilities. *International Journal of Geographical Information Systems* 6: 407–23.

Goodman-Bacon, A. (2021). Difference-in-Differences with Variation in Treatment Timing. *Journal of Econometrics.*

Government of South Africa. (2003). National Health Act, No. 61 of 2003. Government Gazette No. 26595.

Government of South Africa. (2010). Re-engineering primary health care in South Africa: discussion document. Pretoria: National Department of Health.

Gravelle, H. Santos, R. Siciliani, L. (2014). Does a hospital's quality depend on the quality of other hospitals? A spatial econometrics approach. *Regional Science and Urban Economics.* 49.

Grepin, K. Habyarimana, J. Jack, W. (2019). Cash on delivery: results of a randomised experiment to promote maternal health care in Kenya. *Journal of Health Economics.* 65. 15-30.

Griffiths, P. Stephenson, R. (2001). Understanding users' perspectives of barriers to maternal health care use in Maharashtra, India. *Journal of Biosocial Sciences.* 33. 339-359.

Griliches, Z. Hausman, J. (1986). Errors in variables in panel data. *Journal of Econometrics.* 31(1). 93-118. DOI: 10.1016/0304-4076(86)90058-8.

Grossman, M. (1972). The Demand for Health: A Theoretical and Empirical Investigation. *NBER.* DOI: 10.7312/gros17900.

Gruber, J. Hendren, N. Townsend, R. (2014). The Great Equalizer: Health Care Access and Infant Mortality in Thailand. *American Economic Journal: Applied Economics.* 6.1. 91-107.

Guagliardo, M. (2004). Spatial accessibility of primary care: concepts, methods and challenges. *International Journal of Health Geographics.* 3(3).

Guccio, C. Lisi, D. (2016). Thus do all. Social interactions in inappropriate behaviour for childbirth services in a highly decentralized healthcare system. *Regional Science and Urban Economics.* 61. 1-17.

Gwatkin, D. Rustein, S. Johnson, K. Pande, R. Wagstaff, A. (2003). Initial country-level information about socioeconomic differentials in health, nutrition and population. Washington DC: World Bank, Health, Population and Nutrition Group.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica.* 50. 1029–1054.

Hainmueller, J. Mummolo, J. Xu, Y. (2019). How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practise. *Political Analysis.* Vol. 27. 163-192.

Hausman, J. (1978). Specification Tests in Econometrics. *Econometrica.* 46(6). 1251-71.

Health Systems Trust. (2013). National Health Care Facilities Baseline Audit: national summary report.

Heckman, J. Ichimura, H. Todd, P. (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*. Vol. 64. No. 4. 605-654.

Heckman, J. Smith, J. Clements, N. (1997). Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts. The *Review of Economic Studies*. 64. 487-535.

Heckman, J. Urzua, S. Vytlacil, E. (2006). Understanding Instrumental Variables in Models with Essential Heterogeneity. *The Review of Economics and Statistics*. Vol LXXXVIII. No. 3.

Hirano, K. Imbens, G. (2004). 'The Propensity Score with Continuous Treatments', in Gelman, A. and Meng, X-L. (1st ed). *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. John Wiley & Sons Ltd.

Hjortsberg, C. (2003). Why do the sick not utilise health care? The case of Zambia. *Health Economics*. 12. 755-770. DOI: 10.1002/hec.839.

Ho , D. Imai, K. King, G. Stuart, E. (2007). Matching as nonparametric pre-processing for reducing model dependence in parametric causal inference. *Political analysis*. 15(3):199-236.

Hug, L. Alexander, M. You, D. (2019). National, Regional and Global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis. *Lancet Global Health*. 7. 6.

Hunter, J. Asmall, S. Ravhengani, N. Chandran, T. Tucker, J. Mokgalagadi, Y. (2017). The ideal clinic in South Africa: Progress and challenges in implementation. In A. Padarath, & P. Barron (Eds.), *South African health review 2017*. Durban, South Africa: Health Systems Trust.

Ibnat, D. Leonard, K. Bawo, L. Mohammed-Roberts, R. (2019). The Three-Gap Model of Health Worker Performance. *World Bank Policy Research Working Paper*. 8782.

Imbens, G. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *Review of Economics and Statistics*. 86. 1. 4–29.

Imai, K. Ratkovic, M. (2013). Estimating Treatment Effect Heterogeneity in Randomised Programme Evaluation. *The Annals of Applied Statistics*. Vol. 7. No. 1. 443-470.

Imai, K. Strauss, A. (2011). Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign. *Political Analysis*. 19:1. 1-19.

International Labour Organisation. (2016). Discussion Note for Targeting in Malawi and Implications for the Future of the Social Cash Transfer.

Ioannides, Y. (2013). From Neighbours to Nations: the Economics of Social Interactions. Princeton University Press.

Johnson, I. Roger Moon, H. (2021). Estimation of Peer Effects in Endogenous Social Networks: Control Function Approach. *The Review of Economics and Statistics*. 103(2).

Kabane, S. (2016). Ideal Clinic Scale Up Plan 2016/17 [Slideshow]. Available at: https://www.idealhealthfacility.org.za/App/Document/Download/120

Kairies, N. Krieger, M. (2013). How do Non-Monetary Performance Incentives for Physicians Affect the Quality of Medical Care? A Laboratory Experiment. *RUHR Economic Papers.*

Karaca-Mandic, P. Norton, E. Dowd, B. (2011). Interaction Terms in Nonlinear Models. *Health Services Research.* 47.1. 255-274.

Karra, M. Fink, G. Canning, D. (2017). Facility distance and child mortality: A multi-country study of health facility access, service utilization, and child health outcomes. *International Journal of Epidemiology.* 46(3).

Kelejian, H. Prucha, I. (1998). A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. *Journal of Real Estate Finance and Economics.* 17. 99-121.

Kim, E. Singh, K. Speizer, I. Angeles, G. Weiss, W. (2019). Availability of health facilities and utilization of maternal and newborn postnatal care in rural Malawi. *BMC Pregnancy and Childbirth.* 19. 503.

Koenker, R. (2004). Quantile Regression for Longitudinal Data. *Journal of Multivariate Analysis.* 91. 74-89.

Kolstad, J. (2013). Information and quality when motivation is intrinsic: evidence from surgeon report cards. *American Economic Review.* 103:2875-910.

Kondylis, F. Manacorda. M. (2012). School Proximity and Child Labor: Evidence from Rural Tanzania. *The Journal of Human Resources.* 47(1) 32-63.

Kleibergen, F. Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics.* 133. 97–126.

Kluve, J. Schneider, H. Uhlendorff, A. Zhao, Z. (2012). Evaluating continuous training programmes by using the generalized propensity score. *Journal of the Royal Statistical Society.* 175. 587-617. DOI: 10.1111/j.1467-985X.2011.01000.x.

Kruk, M. et al. (2018). High-quality health systems in the Sustainable Development Goals era: time for a revolution. *Lancet Global Health.* 6:e1196-252.

Kruk, M. Chukwuma, A. Mbaruku, G. Leslie, H. (2017). Variation in quality of primary-care services in Kenya, Malawi, Namibia, Rwanda, Senegal, Uganda and the United Republic of Tanzania. *Bulletin of the World Health Organisation.* 95. 408-418.

Kruk, M. Gage, A. Arsenault, C. Jordan, K. Leslie, H. Roder-DeWan, S. Adevi, O. Barker, P. Daelmans, B. Dauboya, S. English, M. Garcia Elorrio, E. Guanais, F. Gureje, O. Hirschhorn, L. Jiang, L. Kelley, E. Tekle Lemango, E. Liliestrand, J. Malata, A. Reddy, T. Rowe, A. Salomon, J. Thapa, G. Twum-Danso, N. Pate, M. (2018). High-quality health systems in the Sustainable Development Goal era: time for a revolution. *The Lancet Global Health.* Vol. 6.

Kruk, M. Gage, A. Joseph, N. Danaei, G. Garcia-Saiso, S. Salomon, J. (2018). Mortality due to low quality health systems in the Universal Health Coverage era: a systematic analysis of amenable deaths in 137 countries. *The Lancet*. 392. 10160. 2203-2212.

Kumar, S. Dansereau, E. Murray, C. (2014). Does distance matter for institutional delivery in rural India? *Applied Economics*. 46(33). DOI: 10.1080/00036846.2014.950836.

Kusuma, D. Cohen, J. McConnell, M. Berman, P. (2016). Can cash transfers improve determinants of maternal mortality? Evidence from the household and community programs in Indonesia. *Social Science and Medicine*. 163. 10-20.

Kuuire, V. Kangmennaang, J Atuoye, K. Antabe, R. Boamah, S. Vercillo, S. Amoyaw, J. Luginaah, I. (2017). Timing and utilisation of antenatal care service in Nigeria and Malawi, *Global Public Health*. 12:6, 711-727.

Lagarde, M. Blaauw, D. (2014). Pro-social preferences and self-selection into jobs: Evidence from South African nurses. *Journal of Economic Behaviour Organisation*. 107. 136-52.

Lagarde, M. Haines, A. Palmer, N. (2007). Conditional cash transfers for improving uptake of health interventions in low- and middle-income countries: a systematic review. *JAMA*. 298. 1900–1910.

Lagarde, M. Huicho, L. Papanicolas, I. (2019). Motivating provision of high quality care: it is not all about the money. *BMJ*. 366.

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

Lagarde, M. Haines, A. Palmer, N. (2009). The Impact of Conditional Cash Transfers on Health Outcomes and Use of Health Services in Low and Middle Income Countries. *Cochrane database of systematic reviews*.

Lannes, L. Meessen, B. Sourcat, A. Basinga, P. (2016). Can performance-based financing help reaching the poor with maternal and child health services? The experience of rural Rwanda. *International Journal of Health Planning and Management*.

Lavy, V. Strauss, J. Thomas, D. de Vreyer, P. (1996). Quality of Health Care, Survival and Health Outcomes in Ghana. *Journal of Health Economics*. 15. 333-357. DOI: 10.1016/0167-6296.

Lavy, V. (1996). School supply constraints and children's educational outcomes in rural Ghana. *Journal of Development Economics*. 51. 291-314.

Lavy, V. Germain, J-M. (1994). Quality and cost in health care choice in developing countries. *Living standards measurement study (LSMS) working paper; no. LSM 105. Washington, D.C. : The World Bank*.

Lavy, V. Schlosser, A. (2011). Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics*. 3. 1–33.

Lazear, E. (2001). Educational Production. *Quarterly Journal of Economics*. 116.3. 777-803.

Lechner, M. (2010). The Estimation of Causal Effects by Difference-in-Difference Methods. *Foundations and Trends in Econometrics*. Vol. 4. No. 3. 165-224.

Lee, L. F. Liu, X. and Lin, X. (2010). Specification and Estimation of Social Interaction Models with Network Structure, Contextual Factors, Correlation and Fixed Effects. *The Econometrics Journal*. 13: 145–176.

Lee, L. F. (2002). Consistency and Efficiency of Least Squares Estimation for Mixed Regressive, Spatial Autoregressive Models. *Econometric Theory*. 18(2). 252–277.

Lee, S. (2018). Intrinsic Incentives: A Field Experiment on Leveraging Intrinsic Motivation in Public Service Delivery. *SSRN*.

Lee, S. (2007). Endogeneity in quantile regression models: a control function approach. *Journal of Econometrics*. 141. 1131–58.

Leonard, K. Masatu, M. (2017). Changing health care provider performance through measurement: The Long Term Impacts of a Program to Encourage Quality in Outpatient Care. *Social Science and Medicine*. 181. 54-65.

Leonard, K. Masatu, M. (2006). Outpatient process quality evaluation and the Hawthorne Effect. *Social Science and Medicine*. 63:9. 2330-2340.

Leonard, K. Masatu, M. (2010a). Professionalism and the know-do gap: exploring intrinsic motivation among health workers in Tanzania. *Health Economics*. 19:12. 1461-1477.

Leonard, K. Masatu, M. (2010b). Using the Hawthorne effect to examine the gap between a doctor's best possible practice and actual performance. *Journal of Development Economics*. 93:2. 226-234.

Leonard, K. (2014). Active patients in rural African health care: implications for research and policy. *Health Policy and Planning*. 29. 85-95. DOI:10.1093/heapol/czs137.

Leonard, K. Mæstad, O. (2016). 'Analyzing the determinants of health worker performance'. In Scheffler, R. Herbst, C. Lemiere, C. Campbell, J. *Health Labor Market Analysis in Low- and Middle-Income Countries: An Evidence-Based Approach*.

Leonard, K. Masatu, M. Vialou, A. (2007). Getting Doctors to Do Their Best: The Roles of Ability and Motivation in Health Care. *Journal of Human Resources*. 42. 3. 682–700.

Lepine, A. Legarde, M. Nestour, A. (2018). How effective and fair is user fee removal? Evidence from Zambia using a pooled synthetic control. *Health Economics*.

Liu, X. Lee, L. Bollinger, C. (2010). Improved efficient quasi maximum likelihood estimator of spatial autoregressive models. *Journal of Econometrics*. 159, 303–319.

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

Linard C, Gilbert M, Snow RW, Noor AM, Tatem AJ (2012) Population Distribution, Settlement Patterns and Accessibility across Africa in 2010. PLoS ONE 7(2): e31743.

Lin, X. (2010). Identifying Peer Effects in Student Academic Achievement by Spatial Autoregressive Models with Group Unobservables. *Journal of Labour Economics*. 28(4). 825-860.

List, J. Shaikh, A. Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*. 22. 773-793.

Liu, X. Patacchini, E. Rainone, E. (2017). Peer effects in bedtime decisions among adolescents: a social network model with sampled data. *The Econometrics Journal*. 20.

Lohela, T. Campbell, O. Gabrysch, S. (2012). Distance to care, facility delivery and early neonatal mortality in Malawi and Zambia. *PloS one*. 7(12). DOI: 10.1371/journal.pone.0052110.

Long, S. Yemane, A. Stockley, K. (2010). Disentagling the Effects of Health Reform in Massachusetts: How Important Are the Special Provisions for Young Adults? *American Economic Review*. 100. 297-302.

Longo, F. Siciliani, L. Gravelle, H. Santos, R. (2017). Do hospitals respond to rivals' quality and efficiency? A spatial panel econometric analysis. *Health Economics*. 26.

Lori, J. Perosky, J. Munro-Kramer, M. Veliz, P. Musonda, G. Kaunda, J. Boyd, C. Bonawitz, R. Biemba, G. Ngoma, T. Scott, N. (2019). Maternity waiting homes as part of a comprehensive approach to maternal and newborn care: a cross-sectional survey. *BMC Pregnancy and Childbirth*.

Macarayan, E. Gage, A. Doubova, S. Guanais, F. Lemango, E. Ndiaye, Y. Waiswa, P. Kruk, M. (2018). Assessment of quality of primary care with facility surveys: a descriptive analysis in ten low-income and middle-income countries. *The Lancet*. 6(11). DOI: 10.1016/S2214-109X(18)30440-6.

Machado, J. Santos Silva, J. (2019), Quantiles via Moments. *Journal of Econometrics*. 213.1: 145–173.

Mackinnon, J. Webb, M. *(2017)*. Wild Bootstrap Inference for Wildly Different Cluster Sizes. *Journal of Econometrics*. 32.2. 233-254.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. 281-297.

Madara, J. Burkhart, J. (2015). Professionalism, Self-regulation, and Motivation: How Did Health Care Get This So Wrong? *JAMA Network*.

Malqvist, M. Sohel, N. Do, T. Eriksson, L. Persson, L-A. (2010). Distance decay in delivery care utilisation associated with neonatal mortality. A case referent study in northern Vietnam. *BMC Public Health*. 10. 762. DOI: 10.1186/1471-2458-10-762.

Manang, F. Yamauchi, C. (2018). The Impact of Access to Health Facilities on Maternal Care Use, Travel Patterns and Health Status: Evidence from Longitudinal Data from Uganda. *Economic Development and Cultural Change*. Pre-print. DOI: 10.1086/702794.

Manthalu, G. (2019). User fee exemption and maternal health care utilisation at mission health facilities in Malawi: An application of disequilibrium theory of demand and supply. *Health Economics*. DOI: 10.1002/hec.3856.

Manthalu, G. Yi, D. Farrar, S. Nkhoma, D. (2016). The effect of user fee exemption on the utilization of maternal health care at mission health facilities in Malawi. *Health Policy and Planning*. 31(9).

Manski, C. (1990). Nonparametric Bounds on Treatment Effects. *American Economic Review*. 80. 2. 319-323.

Manski, C. (2000). Economic Analysis of Social Interactions. *The Journal of Economic Perspectives*. 14(3), 115–136.

Manski, C. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies*. 60(3), 531–542.

Markovic, A. Ryan, A. (2017). Pay-for-Performance: Disappointing Results or Masked Heterogeneity? *Medical Care Research & Review*. Vol. 74. No. 1. 3-78.

Mate, K. Rooney, A. Supachutikul, A. Gyani, G. (2014). Accreditation as a path to achieving universal quality health coverage. *Globalization and Health*. Vol. 10. No. 68.

Masters, S. Burstein, R. Amofah, G. Abaogye, P. Kumar, S. Hanlon, M. (2013). Travel time to maternity care and its effect on utilization in rural Ghana: A multilevel analysis. *Social Science and Medicine*. 93. 147-154. DOI: 10.1016/j.socscimed.2013.06.012.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behaviour. In: Zarembka, P. Ed. Frontiers in Econometrics, Academic Press, New York. 105-142.

McGuire, F. Kreif, N. Smith, P. Stacey, N. Edoka, I. (Unpublished Manuscript). Do Quality Improvement Policies Work for All? Heterogeneous effects and the Impact of Baseline Quality Levels.

McGuire, F. Kreif, N. Smith, P. C. (2021). The Effect of Distance on Maternal Institutional Delivery Choice: Evidence from Malawi. *Health Economics*. 30(9). 2144-2167.

McIntyre, D. Ataguba, J. (2018). Access to quality health care in South Africa: Is the health sector contributing to addressing the inequality challenge? *DFID Research Paper*.

McLaren, Z. Ardington, C. Leibbrandt, M. (2014). Distance decay and persistent health care disparities in South Africa. *BMC Health Services Research*. 14(541). DOI: 10.1186/s12913-014-0541-1.

McWilliams, M. (2020). Professionalism Revealed: Rethinking Quality Improvement in the Wake of a Pandemic. *NEJMCatalyst*. 1.5.

Melly, B. Santangelo, G. (2015). The changes-in-changes model with covariates. *Working Paper*.

Meyer, B. (1995). Natural and Quasi-Experiments in Economics. *Journal of Business & Economic Statistics*. Vol. 13, No. 2. 151-161.

Miller, S. Belizan, J. (2015). The true cost of maternal death: individual tragedy impacts family, community and nations. *Reproductive Health.* 12. 56.

Ministry of Health. (2004). Programme of Work. (2004-2010). Government of Malawi.

Ministry of Health. (2010). Health Sector Strategic Plan. (2011-2016). Government of Malawi.

Ministry of Health (2017). Health Sector Strategic Plan II. (2017-2022). Government of Malawi.

Ministry of Health. ICF International. (2014). Malawi Service Provision Assessment (MSPA) 2013-14.

Mohanan, M. Vera-Hernández, M. Das, V. (2015). The know-do gap in quality of health care for childhood diarrhea and pneumonia in rural India. *JAMA Paediatrics.* 169. 349-57.

Montiel Olea, J. Pflueger, C. (2013). A Robust Test for Weak Instruments. *Journal of Business and Economic Statistics.* 31. 358-369.

Moscelli, G. Gravelle, H. Siciliani, L. (2021). Hospital competition and quality for non-emergency patients in the English NHS. *RAND Journal of Economics.* 52:2. 382-414.

Moscone, F. and Tosetti, E. (2011). GMM estimation of spatial panels with fixed effects and unknown heteroskedasticity. *Regional Science and Urban Economics*. 41, 487–497.

Mukhopadhyay, A. Sahoo, S. (2016). Does access to secondary education affect primary schooling? Evidence from India. *Economics of Education Review.* 54. 124-142. DOI: 10.1016/j.econedurev.2016.07.003.

Mullen, K. Frank, R. Rosenthal, M. (2010). Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. *RAND Journal of Economics*, *41*, 64-91.

Muller, I. Smith, T. Mellor, S. Rare, L. Genton, B. (1998). The effect of distance from home on attendance at a small rural health centre in Papua New Guinea. *International Journal of Epidemiology.* 27. 878-884.

Muthathi, I. Rispel, L. (2020). Policy context, coherence and disjuncture in the implementation of the Ideal Clinic Realisation and Maintenance programme in the Gauteng and Mpumalanga provinces of South Africa. *Health Research Policy and Systems.*

Muthathi, I. Levin, J. Rispel, L. (2020). Decision space and participation of primary healthcare facility managers in the Ideal Clinic Realisation and Maintenance programme in two South African provinces. *Health Policy and Planning.*

Mwabu, G. Ainsworth, M. Nyamete, A. (1993). Quality of Medical Care and Choice of Medical Treatment in Kenya: An Empirical Analysis. *The Journal of Human Resources.* 28(4). 838-862.

National Academy of Sciences, Engineering and Medicine. (2018). Crossing the Global Quality Chasm: Improving Health Care Worldwide. *The National Academies Collection.* Washington (DC).

National Department of Health. (2019). Primary Health Care Definitions and Classifications, unpublished. South Africa

National Department of Health. (2021). Annual Performance Plan 2021/2022. South Africa

National Statistics Office. (2018). Malawi Population and Housing Census. Government of Malawi.

National statistical office. (2017). *Malawi demographic and Health survey 2015-16*. Zomba, Malawi.

Nesbitt, R. Lohela, T. Soremekun, S. Vesel, L. Manu, A. Okyere, E. Grundy, C. Amenga-Etego, S. Owusu-Agyei, S. Kirkwood, B. Gabrysch, S. (2016). The influence of distance and quality of care on place of delivery in rural Ghana. *Scientific Reports*. 6. DOI: 10.1038/srep30291.

Ngoma, T. Asiimwe, A. Mukasa, J. Binzen, S. Sebanescu, F. Henry, E. Hamer, D. Lori, J. Schmitz, M. Marum, L. Picho, B. Naggayi, A. Musonda, G. Conlon, C. Komakech, P. Kamara, V. Scott, N. (2019). Addressing the Second Delay in Saving Mothers, Giving Life Districts in Uganda and Zambia: Reaching Appropriate Maternal Care in a Timely Manner. *Global Health Science Practise*. 11. 7. 1.

Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica*. Vol. 49. No. 6. 1417-1426.

Noble, M. Zembe, W. Wright, G. Avenell, D. (2013). Multiple deprivation and income poverty at small area level in South Africa in 2011. Cape Town, South Africa: Southern African Social Policy Research Institute.

O'Donnell, O. (2007). Access to Health Care in Developing Countries: Breaking Down Demand Side Barriers. *Cad Saude Publica*. 23(120). 2820-2834.

Olivella, P. Siciliani, L. (2017). Reputational concerns with altruistic providers. *Journal of Health Economics*. 55, 1-13.

Ommeh M, Fenenga CJ, Hesp CJ, Nzorubara D, Rinke De Wit TF. (2019). Using mobile transport vouchers to improve access to skilled delivery. Rural and Remote Health; 19: 4577. https://doi.org/10.22605/RRH4577.

O'Neil, S. Kreif, N. Grieve, R. Sutton, M. Sekhon, J. (2016). Estimating causal effects: considering three alternatives to difference-in-differences estimation. *Health Services & Outcomes Research Methods*. 16. 1-21.

Padarath, A. King, J. English, R. (2015). South African Health Review 2014/15. Durban: Health Systems Trust.

Papanicolas I, McGuire, A. (2015). Do financial incentives trump clinical guidance? Hip replacement in England and Scotland. *Journal of Health Economics*. 44:25-36.

Papanicolas I, Smith, P. (2015). The Role of Practitioner Motivation In Designing Provider Payment Reforms and Other Incentives. International Symposium on Health Care Policy. The Commonwealth Fund.

Partners in Health. (2013). The Role of Maternity Waiting Homes as Part of a Comprehensive Maternal Mortality Reduction Strategy in Lesotho. *PIH Reports.* Vol 1. Issue 1.

Paul, E. Albert, L. Bisala, B. (2018). Performance based financing in low income and middle-income countries: isn't it time for a rethink? *BMJ Global Health.* 3.

Paul, B. Rumsey, D. (2002). Utilisation of health facilities and trained birth attendants for childbirth in rural Bangladesh: An empirical study. *Social Science and Medicine.* 54(12). 1755-1765.

Penn-Kekana, L. Pereira, S. Hussein, J. Bontogon, H. Chersich, M. Muiania, S. Portela, A. (2017). Understanding the implementation of maternity waiting homes in low- and middle-income countries: a qualitative thematic synthesis. *BMC Pregnancy and Childbirth.* 17. 269.

Perez-Heydrich, C. Warren, J. Burgert, C. Emch, M. (2013). Guidelines On the Use of DHS GPS Data. Spatial Analysis Reports No. 8. Claverton, Maryland, USA: ICF International.

Pickett, J. (2016). Primary Care at What Price? The Role of Consumer Information Under Quality Uncertainty. *Publicly Accessible Penn Dissertations.*

Poovan, P. Kifle, F. Kwast, B. (1990). A maternity waiting home reduces obstetric catastrophes. *World Health Forum.* Vol. 11. 440-445.

Powell, D. (2020). Quantile Treatment Effects in the Presence of Covariates. *Review of Economics and Statistics.* Vol. 102. No. 5.

Powell-Jackson, T. Hanson, K. (2012). Financial incentives for maternal health: impact of a national programme in Nepal. *Journal of Health Economics.* 31.1. 271-284.

Powell-Jackson, T. Hanson, K. Whitty, C. Ansah, E. (2014). Who benefits from free healthcare? Evidence from a randomized experiment in Ghana. *Journal of Development Economics.*

Presidential Initiative on Safe Motherhood, 2012. Viewed 7 July 2020, < https://www.health.gov.mw/index.php/directorates/safe-motherhood/presidential-initiative-on-martenal-health-safe-motherhood#:~:text=The%20Presidential%20Initiative%20on%20Maternal,mortality%20rates%20in%20the%20country >

Puhani, P. (2012). The Treatment Effect, the Cross Difference, and the Interaction Term in Nonlinear "Difference-in-Differences" Models. *Economics Letters.* 115.1. 85-87.

QGIS Development Team, (2009). QGIS Geographic Information System. Open Source Geospatial Foundation.

Revelli, F. (2006). Performance rating and yardstick competition in social service provision. *Journal of Public Economics.* 90, 459-475.

Rhodes, W. (2010). Heterogeneous Treatment Effects: What Does a Regression Estimate? *Evaluation Review.* 34(4). 334-361.

Rose, C. (2021). Identification of Peer Effects with Miss-specified Peer Groups: Missing Data and Group Uncertainty. arXiv: 2104.10365.

Rosenbaum, P. Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika.* Vol. 70. No. 1. 41-55.

Rosenbaum, P. (1987). The Role of a Second Control Group in an Observational Study. *Statistical Science.* 2(3). 292-316.

Rosenthal, M. Frank, R. Li, Z. Epstein, A. (2005). Early Experience with Pay-for-Performance: From Concept to Practice. *JAMA.* 294(14). 1788– 93.

Rosenzweig, M. & Wolpin, K. (1986). Evaluating the Effects of Optimally Distributed Public Programs: Child Health and Family Planning Interventions. *The American Economic Review.* 76(3). 470-482.

Rowe, A. Rowe, S. Peters, D. Holloway, K. Cahlker, J. Ross-Degnan, D. (2018). Effectiveness of strategies to improve health-care provider practices in low-income and middle-income countries: a systematic review. *The Lancet Global Health.* Vol. 6. No. 11.

Rowe, S. Peters, D. Holloway, K. Chalker, J. Ross-Degnan, D. Rowe, A. (2019). A systematic review of the effectiveness of strategies to improve health care provider performance in low- and middle-income countries: Methods and descriptive results. *PloS One*, Vol. 14. No. 5.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.

Rubin, D. (1976). Inference and missing data. *Biometrika.* 63. 581-592.

Ryan, A. (2018). Well-Balanced or too Matchy–Matchy? The Controversy over Matching in Difference-in-Differences. *Health Services Research.* Vol. 53. No. 6. 4106-4110.

Saleh, S. Bou Sleinman, J. Dagher, D. Sbeit, H. Natafgi, N. (2013). Accreditation of hospitals in Lebanon: is it a worthy investment? *International Journal for Quality in Health Care.* Vol. 25. No. 3.

Samai, O. Sengeh, P. (1997). Facilitating emergency obstetric care through transportation and communication, Bo, Sierra Leone. *International Journal of Gynecology & Obstetrics.* 59(2). 157-164.

Sant'Anna, P. Zhao, J. (2020). Doubly Robust Difference-in-Difference Estimators. *Journal of Econometrics.*

Sarma, S. (2009). Demand for Outpatient Healthcare Empirical Findings from Rural India. *Applied Health Economics and Health Policy.* 7(4). 265-277. DOI: 10.2165/10899650-000000000-00000.

Schaffer, M. (2010). xtivreg2: Stata module to perform extended IV/2SLS, GMM and AC/HAC, LIML and k-class regression for panel data models.

Schafer, J. (1999), Multiple Imputation: a primer. *Statistical Methods in Medical Research.* 8. 3-15.
Schmidt, J-O. Ensor, T. Hossain, A. Khan, S. (2010). Vouchers as demand-side financing instruments for health care: a review of the Bangladesh maternal voucher scheme. *Health Policy.* 96. 98-107.

Schoeps, A. Gabrysch, S. Niamba, L. Sié, A. Becher, H. (2011). The effect of distance to health care facilities on childhood mortality in rural Burkina Faso. *American Journal of Epidemiology.* 173(5). 492-498. DOI: 10.1093/aje/kwq386.

Schultz, P. (2004). Health economics and applications in developing countries. *Journal of Health Economics.* 23. 637-641. DOI: 10.1016/j.jhealeco.2004.04.002.

Schultz, P. (2005). Effects of Fertility Decline on Family Well-Being: Opportunities for Evaluating Population Programs. *Working Paper, Yale University.*

Seljeskog, L, Sundby, J. Chimango, J. (2007). Factors Influencing Women's Choice of Place of Deliverry in Rural Malawi – An Exploratory Study. *African Journal of Reproductive Health.* 10(3). 66-75.

Shannon, G. Bashshur, R. Metzner, C. (1969). The concept of distance as a factor in accessibility and utilization of healthcare. *Medical Care Review.* 26, 143–161.

Sherry, T. Bauhoff, S. Mohanan, M. (2017). Multi-tasking and Heterogeneous Treatment Effects in Pay-for-Performance in Health Care: Evidence from Rwanda. *American Journal of Health Economics.* Vol. 3. No. 2. 192-226.

Shleifer, A. (1985). A Theory of Yardstick Competition. *RAND Journal of Economics.* 16. 3.

Silan, V. Kant, S. Archana, S. Misra, P. Rizwan, S. (2014). Determinants of underutilisation of free delivery services in an area with high institutional delivery rate: a qualitative study. *North American Journal of Medical Sciences.* 6.7. 315-320.

Skiles, M. Burgert, C. Curtis, S. Spencer, J. (2013). Geographically linking population and facility surveys: methodological considerations.

Smith, P. Mosialos, E. Papanicolas, I. Leatherman, S. (2010). Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects. Cambridge University Press. Cambridge. UK.

Sojourner, A. (2013). Identification of peer effects with missing peer data: Evidence from project star. *The Economic Journal.* 123: 574–605.

Stacey, N. Mirelman, A. Kreif, N. Suhrcke, M. Hofman, K. Edoka, I. (2021). Facility standards and the quality of public sector primary care: Evidence from South Africa's "Ideal Clinics" program. *Health Economics.* 30(7). 1543-1558.

Staiger, D. Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica.* 65. 557–586.

Stock, R. (1983). Distance and the utilization of health facilities in rural Nigeria. *Social Science & Medicine.* 17(9). 563-570. DOI: 1.1016/0277-9536(83)90298-8.

Stokes, J. Kristensen, S. Checkland, K. Cheraghi-Soh, S. Bower, P. (2017). Does the impact of case management vary in different subgroups of multimorbidity? Secondary analysis of a quasi-experiment. *BMC Health Services Research.* Vol. 17.

Strupat, C. (2017). Do Targeted Reproductive Health Services Matter? – The Impact of a Midwife Program in Indonesia. *Health Economics*. 26(12). 1667-1681.

Sun, S. Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*. 225. 2. 175-199.

Singh, K. Speizer, I. Kim, E. Lemani, C. Tang, J. Phoya, A. (2018). Evaluation of a maternity waiting home and community education program in two districts of Malawi. *BMC Pregnancy and Childbirth*.

Scott, N. Kaiser, J. Vian, T. Bonawitz, R. Fong, R. Ngoma, T. Biemba, G. Boyd, C. Lori, J. Hamer, D. Rockers, P. (2018). Impact of maternity waiting homes on facility delivery among remote households in Zambia: protocol for a quasi-experimental, mixed-methods study. *BMJ Open*.

Schmidheiny, K. Siegloch, S. (2020). On Event Studies and Distributed-Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization. *SSRN*.

Solon, G. Haider, S. Wooldridge, J. (2015). What Are We Weighting For?. *Journal of Human Resources*. 50.2. 301-316.

Tanser, F. Gijsbertsen, B. Herbst, K. (2006). Modelling and understanding primary health care accessibility and utilization in rural South Africa: An exploration using a geographical information system. *Social Science & Medicine*. 63. 691-705. DOI:10.1016/j.socscimed.2006.01.015.

Tegegne, T. Chojenta, C. Loxton, D. Smith, R. Kibret, K. (2018). The impact of geographic access on institutional delivery care use in low and middle-income countries: Systematic review and meta-analysis. *Plos One*. 13(8). DOI: 10.1371/journal.pone.0203130.

Terza, J. Basu, A. Rathouz, P. (2007). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modelling. *Journal of Health Economics*. 27(3). 531-543. DOI: 10.1016/j.jhealeco.2007.09.009.

Terza, J. (2018). Two-Stage Residual Inclusion Estimation in Health Services Research and Health Economics. *Health Services Research*. 53(3). DOI: 10.1111/1475-6773.12714.

Thaddeus, S. Maine, D. (1994). Too Far To Walk: Maternal Mortality in Context. *Social Science and Medicine*. 38(8). 1091-1110.

Thomas, D. Lavy, V. Strauss, J. (1996). Public policy and anthropometric outcomes in Cote d'Ivoire. *Journal of Public Economics*. 61:155-92.

Thornton, R. (2008). The Demand for, and Impact of, Learning HIV Status. *American Economic Review*. 98(5). 1829-63. DOI: 10.1257/aer.98.5.1829.

Todd, P. (2007). 'Evaluating Social Programs with Endogeneous Program Placement and Selection of the Treated', in Schultz, P. and Strauss, J. (1st ed). *The Handbook of Development Economics*. 3847-3894. Elsevier.

United Nations, Department of Economic and Social Affairs. (2016). Study on Aging in Sub-Saharan Africa: Sampling Manual. Sampling Manual.

United Nations Inter-agency Group for Child Mortality Estimation (UN IGME). (2020). Levels & Trends in Child Mortality: Report 2020, Estimates developed by the United Nations Inter-agency Group for Child Mortality Estimation. United Nations Children's Fund, New York.

UNICEF. (2019). 2018/19 Health Budget Brief South Africa.

Uny, I. (2017). Weighing the Options for Delivery Care in Rural Malawi: Community Actors' Perceptions of the 2007 Policy Guidelines and Redefined Traditional Birth Attendants Roles. *PhD Thesis.*

Van de Poel, E. Flores, G. Ir, P. O'Donnell, O. Van Doorslaer, E. (2014). Can vouchers deliver? An evaluation of subsidies for maternal health care in Cambodia. *Bulletin of the World Health Organization.* 92.5. 331–339.

Van De Poel, E. Flores, G. Ir, P. O'Donnell, O. (2015). Impact of Performance-Based Financing in a Low-Resource Setting: A Decade of Experience in Cambodia. *Health Economics.*

Van Doorslaer, E. Masseria, C. (2004). Income-Related Inequality in the Use of Medical Care in 21 OECD Countries. *OECD Health Working Paper.*

VanderWeele, T. (2015). Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press.

Van Kippersluis, H. Rietveld, C. (2018). Beyond Plausible Exogeneity. *The Econometrics Journal.* 21:3. 316-331.

Villeval, M-C. (2020). Performance Feedback and Peer Effects: A Review. *Working Paper.*

Von Hinke, S. Leckie, G. Nicoletti, C. (2019). The Use of Instrumental Variables in Peer Effects Models. *Oxford Bulletin of Economics and Statistics.* 81: 5.

Wagstaff, A. van Doorslaer, E. (2000). Income Inequality and Health: What Does the Literature Tell Us?. *Annual Review of Public Health.* 21(1). 543-567. DOI: 10.1146/annurev.publhealth.21.1.543.

Wang, W. Lee, L. F. (2013a). Estimation of spatial autoregressive models with randomly missing data in the dependent variable. *Econometrics Journal.* 16. 73–102.

Wang, W. Lee, L. F. (2013b). Estimation of spatial panel data models with randomly missing data in the dependent variable. *Regional Science and Urban Economics.* 43. 521–38.

Wang, R. & Ware, J. (2013). Detecting Moderator Effects Using Subgroup Analyses. *Prev Sci.* Vol. 14. No. 2. 111-120.

Wild, K. Barclay, L, Kelly, P. Martins, N. (2013). The Tyranny of Distance: Maternity Waiting Homes and Access to Birthing Facilities in Rural Timor-Leste. *Bulletin of the World Health Organisation.*

Witter, S. Somanathan, A. (2012). Demand-side financing for sexual and reproductive health services in low and middle-income countries: a review of the evidence. *Policy Research Working Paper.* World Bank.

Witter, S. Fretheim, A. Kessy, F. Lindahl, A. (2013). Paying for performance to improve the delivery of health interventions in Low- and middle-income countries (review). *Cochrane Database Systematic Reviews*. 2. 1–82.

Wooldridge, J. (2014). Quasi-maximum likelihood estimation and testing for non-linear models with endogenous explanatory variables. *Journal of Econometrics*. 182. 226-234.

Wong, E. Popkin, B. Guilkey, D. Akin, J. (1987). Accessibility, quality of care and prenatal care use in the Philippines. *Social Science and Medicine*. 24(11). 927-944.

World Development Indicators. (2019). Washington, DC: The World Bank. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=MW. [accessed 25 September 2021].

World Development Indicators. (2019). Washington, D.C. The World Bank.

World Health Organisation. (2015). Tracking Universal Health Coverage: First Global Monitoring Report.

World Health Organisation. (2021). Tracking Universal Health Coverage: 2021 Global Monitoring Report.

World Health Organization. (2015). *Tracking Universal health coverage: first global monitoring report*. Joint World Bank Report.

World Health Organisation. (2017). Primary health care systems (PRIMASYS): case study from South Africa. Geneva. Licence: CC BY-NC-SA 3.0 IGO.

World Health Organisation (2019). https://www.who.int/news-room/fact-sheets/detail/maternal-mortality [accessed, 25 September 2021].

World Health Organisation. (2016). Namibia: Maternity waiting homes protect newborns and mothers. https://www.who.int/news-room/feature-stories/detail/namibia-maternity-waiting-homes-protect-newborns-and-mothers. [accessed 25 September 2021].

Yanagisawa, S. Oum, S. Wakai, S. (2006). Determinants of Skilled Birth Attendance in Rural Cambodia. *Tropical Medicine and International Health*. 11(2). 238-251.

Yelowitz, A. (1995). The Medicaid Notch, Labour Supply, and Welfare Participation: Evidence from Eligibility Expansions. *The Quarterly Journal of Economics*. 110(4). 909-939.

Zeldow, B. Hatfield, L. (2021). Confounding and Regression Adjustment in Difference-in-Difference Studies. *Health Services Research*. Vol. 56. No. 5.

Zeng, W. Shepard, D. Nguyen, H. Chansa, C. Das, A. Qamruddinb, J. Friendmand, J. (2018). Cost-effectiveness of results-based financing. *Bulletin of the World Health Organisation*. 96. 760–771.

Zuanna, T. Fonzo, M. Sperotto, M. Resti, C. Tsegaye, A. Azzimonti, G. Manenti, F. Putoto, G. Bertoncello, C. Zanovello, S. (2019). Effects of maternity waiting homes on perinatal deaths in an Ethiopian hospital. A case-control study. *European Journal of Public Health*.