Nigeria

Subsidy Reinvestment and Empowerment Programme

(SURE-P)

Maternal and Child Health Initiative

Impact Evaluation Concept Note

Final Version

12 March, 2013

Contents

Acknowledgements

This impact evaluation is a collaboration between the SURE-P Maternal and Child Health Initiative, National Primary Health Care Development Agency, Federal Ministry of Health, the University of Sussex, University College London, and the Development Impact Evaluation Initiative (DIME) of the World Bank. The team gratefully acknowledges the invaluable support of Dr. Muhammad Ali Pate, Honorable Minister of State for Health, Dr. Ado J.G, Muhammad, Executive Director of the National Primary Health Care Development Agency, Marie Francoise Marie-Nelly (World Bank Country Director for Nigeria), and Dr. Ugo Okoli, SURE-P Project Director. We are also grateful to our peer reviewers (Ayodeji Oluwole Odutolu, Christel Vermeersch, Edit V. Velenyi, David Evans, Jed Friedman, Nana Twum-Danso, Simon Brooker, and Victor Orozco), and the several individuals listed in the section on staffing below. This impact evaluation is made possible through generous funding from the Bill and Melinda Gates Foundation, the Strategic Impact Evaluation Fund, and the Bank-Netherlands Partnership Program.

List of acronyms

CCT – Conditional Cash Transfer

CIC – Changes-in-changes

CPS – Country Partnership Strategy

CHEW – Community Health Extension Worker

DID – Difference-in-differences

FMOH – Federal Ministry of Health

ICC – Intra-cluster correlation

MCH – Maternal and child health

MDG – Millennium Development Goal

MSS – Midwives Service Scheme

NPHCDA – National Primary Health Care Development Agency

PHC – Primary Healthcare Centre

RHAP – Reproductive Health Action Plan

SURE-P – Subsidy Reinvestment and Empowerment Programme

VHW – Village Health Worker

WDC – Ward Development Committee

1. Background

1.1 Context

Maternal and child health in Nigeria have steadily improved over the last two decades. Maternal mortality has been estimated at 545 per 100,000 live births in 2008, down from 1,100 in 1990, and neonatal mortality has fallen to 40 per 1,000 live births, down from 45 in 1990[1]. Although significant, these improvements have been slower than in most countries in the region and are insufficient for Nigeria to fulfil the Millennium Development Goals (MDGs) related to child and maternal health by 2015. For maternal mortality, the target is set at 250 per 100,000 live births, hence at less than one half of its current value.

A key contributing factor to these poor maternal and neonatal outcomes is the very low coverage of antenatal care and skilled birth attendance. In 2008, only about 58 per cent of pregnant women attended one or more antenatal visits and just 39 per cent of childbirths took place under the supervision of a skilled birth attendant. Low service coverage is concentrated amongst the poorest and affects especially the populations of Northern Nigeria. This problem reflects both supply-side and demand-side constraints. On the supply side, issues include a chronic short supply of qualified midwives and other health workers, substandard healthcare infrastructures (with some lacking running water and sanitation), and persistent under-provision of essential supplies, such as drugs and midwifery kits. On the demand side, factors explaining low utilisation include lack of economic resources to meet user fees and transportation costs, perceived irregular service availability, high waiting times, lack of information and individual undervaluation of the benefits of antenatal and obstetric care[2].

To address these challenges, the Federal Ministry of Health (FMOH) is introducing important reforms and is committed to learning which of these are working and worth scaling up. These have

[1] Source: *Launching the SURE-P Maternal and Child Health Initiative – Discussion document,* NPHCDA, February 2012.

[2] It is possible that, in face of irregular service availability and high waiting times, low demand for antenatal care may actually reflect an accurate valuation of its expected benefits. Our survey design has the potential for shedding light on this issue by: a) asking pregnant women about their use of antenatal care and skilled birth attendance and respective maternal health outcomes; b) asking pregnant women about the leading deficiencies of the quality care received; c) interviewing primary health centres' managers and eliciting the likelihood of severe insufficiencies in the quality of the services delivered in each healthcare facility.

included the earmarking and securing of budget appropriations for immunization and large-scale maternal and child health (MCH) initiatives, such as the Midwives Service Scheme (MSS[3]). Evidence on the causal impact of these programs is, however, lacking, hence the scope for using previous experience to inform policy design has been limited.

To begin addressing this lack of evidence, this Concept Note describes a rigorous impact evaluation of the *SURE-P Maternal and Child Health Initiative* (SURE-P thereinafter), an ambitious follow-up program to MSS, implemented by the Nigeria National Primary Health Care Development Agency (NPHCDA), that provides demand and supply side incentives, community monitoring, and increased human resources to improve the rates and quality of antenatal care and skilled birth attendance in Nigeria.

1.2 Links with World Bank strategy and policy implications

The impact evaluation described in this concept note is highly consistent with the World Bank's new Country Partnership Strategy (CPS) for Nigeria for the period FY14-17. In particular, through providing evidence on effective mechanisms through which to provide basic health services, thereby contributing to the reduction of inequality and social exclusion, the impact evaluation will contribute to Pillar 3 of the CPS – Fostering Social Inclusion and Reducing Vulnerability. This activity will support the CPS focus on building accountable and inclusive health systems by, among other things, focusing on health metrics for evidence-based decision making and improving outcomes for mothers and children.

In addition, IE can serve as a powerful tool with which to address weaknesses in existing governance capacity and accountability, and thus the workshop will contribute to Pillar 1 of the new CPS – Improving Governance. This applies to both the proposed midwives incentives and community monitoring interventions, to be experimentally tested through the impact evaluation (see

[3] The Midwives Service Scheme (MSS) was initiated in 2009 to mobilise newly qualified, retired, and unemployed midwives for deployment to primary health care facilities in underserved communities in order to increase skilled birth attendance and, consequently, a reduction in maternal, newborn, and child mortality in Nigeria. This was to be achieved through providing 24 hour midwife coverage at these health facilities, along with essential commodities. The program was initially implemented in 163 clusters, each comprising four primary healthcare facilities and a referral hospital, starting January 2010. Geographically, the program was spread across Nigeria's 36 states and the Federal Capital Territory. The SURE-P build on lessons learned from the MSS.

below). The impact evaluation is also aligned with the Bank's Reproductive Health Action Plan (RHAP) for 2010-2015, which consistently emphasizes the need for evidence-based interventions and accountability for results. RHAP notes the need to use context/country-specific learning to draw overall policy lessons. Impact evaluation supports the production and dissemination of high-quality, context specific evidence and improves our understanding of how to make development work to deliver results on the ground

This impact evaluation is also serving as an example for the Bank's new approach to the science of service delivery. This is built on agile implementation, evidence-based design, and structured learning to (i) improve the effectiveness and operational efficiency of development investment, (ii) promote evidence as a key input in the policy and operational design process, and (iii) build locally targeted and globally relevant knowledge.

This impact evaluation will be carried out under the Bank's collaboration with the FMOH and the Bill & Melinda Gates Foundation to proactively use impact evaluation to improve the effectiveness of priority FMOH programs, with a focus on the Saving One Million Lives Initiative.[4] As such, intermediate and final analyses carried out during the course of this work will be immediately available to policymakers and program managers to guide planning, implementation, and policymaking. In particular, it is expected that the SURE-P impact evaluation will contribute to the following policy areas: (i) compensation / motivation for frontline primary-level health workers; (ii) bottom-up approaches to improving accountability in the delivery of health services; (iii) demand promotion, in particular whether investment in the conditional cash transfer yields sufficient returns; and (iv) the effectiveness of the SURE-P as a whole.

2. SURE-P: motivation, structure and goals

2.1 Motivation and existing evidence

A large body of international evidence indicates that high maternal and neonatal mortality rates are strongly associated with shortages of qualified health staff, insufficient supply of drugs and

[4] This collaboration also includes work on health service delivery indicators and resource tracking.

equipment and inadequate infrastructures (World Bank, 2009). Moreover, this evidence suggests that interventions focused on training and posting midwives in needy communities, as well as on improving primary care infrastructure and the supply chain of essential supplies, have prompted significant increases in the use of antenatal care and, subsequently, reduced maternal and neonatal mortality (see Hatt et al. 2007, World Bank, 2009 and references therein). In Nigeria, a large-scale intervention with these characteristics, the Midwives Service Scheme (MSS), was initiated in 2009 to mobilise newly qualified, retired, and unemployed midwives for deployment to primary health care facilities in underserved communities. Although the programme is associated with sizable increases in utilisation rates within two years of implementation, the non-experimental impact evaluation of the MSS has not yet been completed, whence it is difficult to quantify the share of this increase that is attributable to the programme.

International evidence also shows that increasing the availability of healthcare services alone may not be sufficient to boost coverage rates if, in addition to the supply-side constraints, there are also important constraints to the demand, such as those in the Nigerian case (Glassman et al., 2007). Over the last decade, conditional cash transfers (CCTs) have been used worldwide as a means of relaxing these constraints. In a CCT, money is transferred to beneficiaries conditional on them taking one or more predefined actions; it thus works by simultaneously relaxing financial barriers and encouraging behavioural change[5]. CCTs have been shown to successfully reduce poverty and encourage parental investments in the health and education of their children (for a literature review see Fiszbein, Schady et al., 2009). In the context of health care, Lagarde et al. (2007) shows that, CCTs can be effective in increasing preventive health care use[6]. This finding is corroborated using data from African countries by Evans et al. (2012), in the case of Tanzania, and Akresh et al. (2012), in the case of Burkina Faso. Moreover, CCTs have previously been used to boost antenatal care use and skilled birth attendance. The most cited example of this is India's Janani Suraksha Yojana (JSY), which used a one-off cash transfer to encourage pregnant women to give birth with skilled attendance. Although Lim et al. (2010) suggest that the programme had a positive impact on skilled

[5] One of the microeconomic foundations for the use of CCTs relates to individual suboptimal valuation of the returns to investments in human capital. First, households may perceive the payoffs from their health investments, such as antenatal visits, as too low (either because they do not internalize the relevant externalities or because they have limited exposure to health services which might bias their perceived individual returns downwards). Second, deprived households are also likely to over-discount the future health benefits of their decisions and hence under-invest in the present. Both mechanisms are likely to affect individual choices on whether to seek antenatal care and skilled childbirth.

[6] For evidence indicating that CCTs can also improve infant mortality and morbidity see Barham (2005), Hernández et al. (2005) and Gertler (2004).

birth attendance and neonatal mortality, a more sophisticated quantitative analysis fails to find evidence of the latter (Powell-Jackson et al. 2012).

2.2 Essential structure

The design of SURE-P draws on this pool of evidence and innovates by combining a supply-side intervention with a demand-side CCT[7]. On the supply-side, SURE-P aims to recruit, train and deploy a total of 5,400 midwives and 14,100 community health extension workers (CHEWs), as well as to upgrade essential infrastructures and guarantee the adequate provision of supplies and equipment to primary health centres between the end of 2012 and 2015. In addition SURE-P will, during the same period of time, hire and train a total of 38,700 village health workers (VHW), who are expected to establish the connection between the primary healthcare centres (PHCs) and pregnant women in each village.

On the demand-side, SURE-P introduces a CCT, whereby all pregnant women will be given a total cash payout of 5,000 Naira (about USD 32), conditional on attending antenatal care (ANC), skilled birth attendance and postnatal care. This payment will be provided in 4 tranches: N1,000 upon registration in a PHC, N1,000 after completing the set of standard antenatal visits, N2,000 upon institutional delivery and N1,000 after zero-dose immunisation is given to the newborn baby. Pregnant women will register at PHCs and spot-verification of actual uptake of services will be carried-out by State focal personnel and local healthcare staff. It is expected that, on average, these payments will more than offset the charges that pregnant women need to pay for institutional deliveries and ante-natal visits (a fee for service and the price of the midwifery kit – or "Mama kit" – used in the delivery). Payments will be made directly to the pregnant women through mobile banking and the conventional banking system; the strengths and weaknesses of these modes of payment were assessed in a dedicated CCT pre-pilot study[8]. In addition, information dissemination

[7] Originally, the *Programa de Asignacion Familial* (PRAF) in Honduras also combined a supply-side health intervention with a CCT; however, the implementation of the supply-side component was particularly poor, hence the evaluation of this programme has very serious limitations (Morris, Flores, et al. 2004).

[8] This was carried-out in April and May 2012 with the goal of testing the operational mechanics of the CCT: data collection and validation, payment methods and other key areas of programme administration. It involved a substantial number of interviews with pregnant women, and focus group discussions with women, health workers and WDC members.

activities will be put in place targeting all women of reproductive age to encourage them to register with their nearest PHC.

Due to the logistical complexity of the programme, the implementation of SURE-P will be staggered in two phases. In Phase I, the supply-side of the intervention will be implemented in 500 PHCs, chosen according to perceived need of the populations and capacity of the existing facilities.[9] In Phase II, the CCT component will be added and the supply side will be expanded to 1,300 facilities by 2015.

2.3 Lessons from the SURE-P predecessors: the problems of attrition and commodity diversion from primary healthcare facilities

SURE-P builds on the experience of the MSS, initiated in 2009, which recruited, trained and deployed midwives to primary health centres in needy communities. Although the causal impact of this programme has not yet been evaluated, a number of important lessons have been learnt from its implementation.

First, attrition amongst the deployed midwives has been particularly high, posing a vital challenge to the sustainability of MSS. This is a well-known problem of this type of intervention (see for example Hatt et al. 2007, on the expansion of midwifery services in Indonesia) that is likely to also affect SURE-P[10]. In order to tackle this issue, the impact evaluation will investigate a system of performance incentives to retain midwives[11], encompassing economic and purely motivational incentives; a detailed account of this is given in the next section.

Second, the experience of MSS has shown that, although adequate quantities of essential drugs and equipment have been regularly supplied to PHCs, these supplies appear to be subsequently diverted and sold in secondary markets, thereby failing to reach the pregnant women. Systematic inventory

[9] The 500 facilities have been pre-selected based on a defined set of criteria.

[10] In the case of MSS, inadequate social amenities, language barriers, harsh working conditions and the fact that 45% of the recruited midwives were young newly graduated midwives deployed to areas which were geographically and socio-culturally distant from their homes are believed to be important contributing factors for the high attrition rates. All these challenges will also be present in the case of SURE-P.

[11] It should be noted that, although maternal health responds to various dimensions of midwives' performance (not just midwifery service availability) the reduction in attrition is a fundamental pre-requisite for health outcomes improvement. Although our questionnaire design will provide information of midwifery care quality, this impact evaluation will focus primarily on attrition rates.

checks and physical inspections at the PHC level are beyond SURE-P's scope and budget. Instead, SURE-P proposes to put in place a community-based scheme to monitor stockouts and increase local accountability. This scheme, described in more detail in the next section, will be the subject of a separate experimental impact evaluation. An overview of the essential structure of SURE-P can be found in Table 1.

Table 1: SURE-P (2012-2015) overview

<table>
<tr><td colspan="2">SURE-P: 2012- 2015
- Scope: nation-wide (36 States and Federal Capital Territory, Abuja);
- 500 PHCs on first stage; scale-up to 1,300 by 2015</td></tr>
<tr><td>Supply-side</td><td>Demand-side</td></tr>
<tr><td>- Recruitment, training and deployment of 5,400 midwives, 14,100 CHEWs and 38,700 VHWs in needy communities
- Provision of essential supplies, commodities and refurbishment of PHCs infrastructures
- Provision of incentives to tackle health workers' attrition (both monetary and non-monetary)
- Monitoring the availability of supplies at the PHC level</td><td>- CCT: pregnant women given a total cash payout of N5,000 conditional on attending antenatal care, skilled birth attendance, and postnatal care
- Informational outreach for awareness-creation and demand promotion</td></tr>
</table>

3. Structure of the impact evaluation

SURE-P is a complex intervention that combines distinct components both on the supply and on the demand side. Thus, although this impact evaluation will measure the overall impact of the programme as a package, it is focused on disentangling the separate contributions made by each of its components as a means of identifying the ones that work and are therefore worth scaling up, and those that need to be redesigned in future interventions. The overall structure of the impact evaluation is schematically described in Figure 1, and a detailed account of each of its components is given in the following sections.

In **Phase I**, SURE-P clusters will be randomly assigned to each of the study arms under both (i) midwives incentives and (ii) community stock monitoring

In **Phase II**, SURE-P clusters will be randomly assigned to each of the study arms under both (i) midwives incentives and

Phase II will also include a non-experimental evaluation of the SURE-P package as a whole

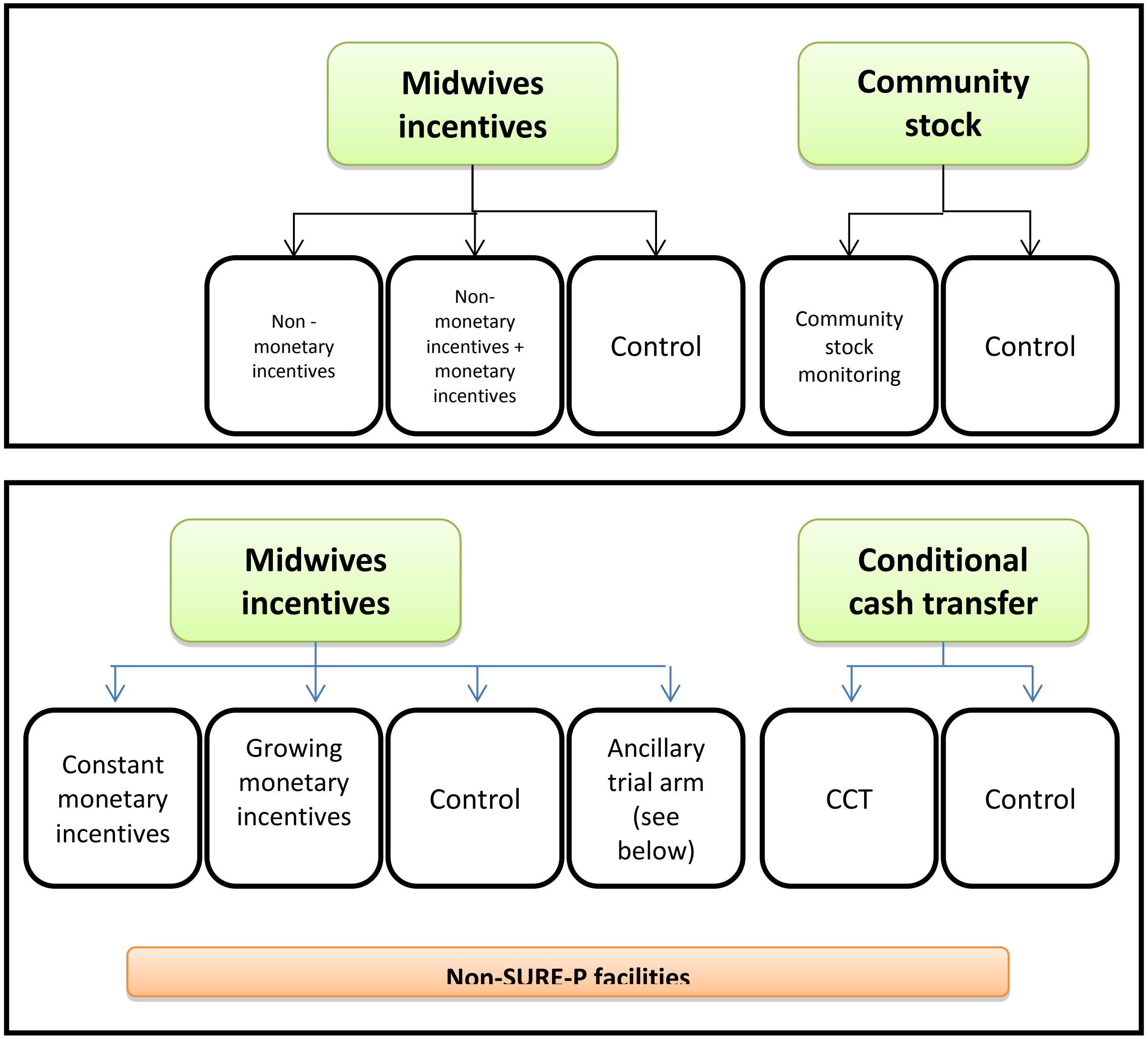


Figure 1: SURE-P impact evaluation

Midwives incentives
Community stock
Non - monetary incentives
Non- monetary incentives + monetary incentives
Control
Community stock monitoring
Control
Midwives incentives
Conditional cash transfer
Constant monetary incentives
Growing monetary incentives
Control
Ancillary trial arm (see below)
CCT
Control
Non-SURE-P facilities

3.1 Experimental components

3.1.1 Incentives to midwives

Midwives attrition has been found to be a leading challenge to the success of the maternal and child health interventions that preceded SURE-P. After extensive discussion with FMOH and SURE-P staff, the lack of adequate incentives to health workers has been indicated as a likely cause of this problem and different incentive structures have been suggested.

There is vast empirical evidence indicating that the performance of healthcare providers responds positively to adequate monetary incentives (Glassman et al., 2007; Basinga et al. 2010). Nonetheless, a more recent strand of research (see Ashraf, Bandiera and Jack, 2012 and references therein), has brought to the fore the role of non-monetary incentives (i.e. incentives other than economic ones) in boosting provider performance through leveraging intrinsic motivation. It is plausible that this type of incentive may be important in the context of SURE-P, for pro-social preferences are believed to be one of the reasons informing midwives' decision of enrolling the programme

The SURE-P impact evaluation will consider the effectiveness of monetary and non-monetary incentives in reducing midwives' attrition: in Phase I of the programme implementation, the impact evaluation will focus on the effect of non-monetary incentives and of monetary incentives over and above non-monetary [12] in Phase II it will, in addition, determine the impact of different structures of monetary incentives.

This experimental design may beg the question of why one should focus on this comparison rather than on a simpler evaluation of the effect of monetary *vs* non-monetary incentives. It should, however, be noted that while there is extensive evidence suggesting that monetary incentives can crowd-out intrinsic motivation, thereby reducing performance, the introduction of non-monetary

[12] It will clearly be desirable to minimize attrition far beyond the twelve-month period considered here. The SURE-P team will consider keeping these incentives in place for a longer time period if this impact evaluation shows that they are effective in the shorter run.

incentives has been considered relatively safe[13]. They are also relatively less costly to implement, hence their use has been recommended without significant reservations. Thus, in this setting, the most relevant policy question is whether non-monetary incentives should be supplemented with monetary ones and, if so, what is their appropriate level. We thus seek to determine the incremental value of monetary incentives (over and above non-monetary ones). This question is also novel in the literature, since little work has been devoted to the combined effect of monetary and non-monetary incentives (Ariely, Bracha and Meier 2009, AER: 544-555 is one of the few recent papers on this issue).

In Phase I, two alternative incentive structures are compared: A) non-monetary incentives, consisting of quarterly tokens of appreciation (SURE-P t-shirt, mug, etc.), recognition letters sent directly from the FMOH indicating the incentives to be given in the following quarter,[14] and a laudatory ceremony for the midwives who complete 12 months of service; B) quarterly payments (equal across quarters) made to each midwife's account, in tandem with quarterly letters sent directly from the FMOH, quantifying the accumulated financial incentives received to date and highlighting future incentives payable to those who stay through the following quarter.

In this phase, the experimental evaluation will have three arms: non-monetary incentives only (I.A), non-monetary plus monetary incentives (I.B), and a control arm (I.C). This research design allows determining the impact of non-monetary incentives, and the extent to which this impact increases when these are combined with monetary incentives; this is novel in the economic literature.

In Phase II, armed with an increased sample size (800 further PHCs) and with the impact evaluation results from Phase I it will be possible to test additional combinations of incentives. An experimental design with three arms will be particularly informative: an arm of quarterly monetary incentives which are constant over time and of the same size as in Phase I (II.A); a second arm of quarterly payments that increase over time but where the total amount received over the year is the same as in the constant payments case (II.B); and a control arm (II.C). The rationale for increasing

[13] Li (2012) suggests that non-monetary incentives can also, in special cases, crowd-out intrinsic motivation. Nevertheless, the bulk of the international evidence on performance-worsening effects of incentives concerns monetary incentives.

[14] Focus group meetings will identify locally appropriate tokens as well as the messages embedded in them. An example of a such tokens can be a t-shirt displaying the message "I am a midwife who stays!".

the value of the tranches paid throughout the year is to increase the opportunity cost of attrition at a given moment; this can be key in Nigeria, where widespread mistrust on policy makers' commitment often lead individuals to heavily discount future financial benefits. In contrast, the scheme of constant tranches pays relatively more in the beginning of the year, thereby lending credibility to the scheme[15].

Under the assumption that there are no time-treatment arms interaction effects (hence that differences in retention rates between Phase I and Phase II can be controlled for by using differences in the retention rates of the control arms of the two phases), the combined data of Phase I and II will allow us to compare non-monetary with monetary incentives (both constant or increasing) of a given size (the one used in II.A and II.B). While this comparison is informative, it will be useful to generalise this analysis further, since the relative effect of monetary versus non-monetary incentives may depend on the size of these monetary incentives. Because the available sample size will not allow a non-parametric comparison between several levels of monetary incentives and non-monetary ones, a fully structural model will thus be specified; this will use all the available data, and therefore allow the comparison between non-monetary incentives and any level (within reason) of monetary incentives.[16] It is known however, that the identification of these structural models will be greatly facilitated by the specification of an additional (ancillary) treatment arm with significantly higher, or lower, levels of monetary incentives than the other treatment arms (II.D).[17] In terms of power, the feasibility of this ancillary trial arm can only be assessed after the data from Phase I becomes available.

[15] It would be desirable to minimize attrition beyond the twelve months threshold, which would require incentives to stay in place for a longer time period. This possibility will be considered if the proposed incentives structure proves to be effective in the shorter run.

[16] For examples of such models see for instance Vera-Hernández (2003), Todd and Wolpin (2006), Attanasio et al (2005).

[17] Note that the purpose of this fourth trial arm is purely ancillary in the sense that its only role is to facilitate the identification of the marginal utility of income in the structural model.

Table 2: Incentive schemes to midwives in SURE-P

<table>
<tr><td colspan="6"><i>Phase I</i></td></tr>
<tr><td colspan="2">I.A: Non-monetary</td><td colspan="2">I.B: Monetary and Non-Monetary</td><td colspan="2">I.C: Control</td></tr>
<tr><td colspan="6"><i>Phase II</i></td></tr>
<tr><td>Group II.A: Monetary (constant over time) incentive</td><td colspan="2">Group II.B: Monetary (increasing over time)</td><td colspan="2">Group II.C: Control</td><td>Group II.D (ancillary trial arm): Either significantly smaller <i>or</i> larger monetary incentive than the other treatment arms</td></tr>
</table>

3.1.2 Community monitoring of PHC stocks of drugs and essential supplies

Recent evidence indicates that incorporating reputational concerns in health policy design can be a powerful means of promoting health care providers' performance. For example, Bevan and Hood (2006) and Bevan and Hamblin (2009) show that publicly disseminating information on low performers, or prevaricators, has significant deterrent effects, provided that effective monitoring is simultaneously implemented to stop them from gaming the system.

SURE-P will use mobile phone communications as an inexpensive means of simultaneously achieving the required level of community monitoring and of providing health care professionals with incentives conducive to the reduction of stockouts. The mobile phone numbers of PHC users in the select target communities will be collected at the time of baseline survey, and 80 of them will be called every three months to check whether they were denied drugs and essential supplies by healthcare workers due to reported stockouts.[18] In the treatment group, objective information on stockouts will then be disseminated in the communities: the stockout rate in their PHC as well as several other appropriate benchmarks (i.e. the average of their district, state, the national one, as well as the government benchmark and very importantly the community entitlements). It should be noted that this way of "naming and shaming" poor practice and of publically praising good practice will never single out individual health workers and is exclusively aimed at institutions.

[18] We will suggest to SURE-P that a small top up credit will be provided to ensure responsiveness to the calls. SMS messages will also include information on pregnant women's' healthcare entitlement.

The SURE-P communications team is working to develop culturally appropriate messages and analysing alternative channels for communicating healthcare entitlements and information dissemination: these may include local media, outdoor posters, involvement of community leaders and direct dissemination of information to members of the community through SMS messages. This intervention is expected to ease one of key barriers to public officials accountability in low income settings: lack of information at the community level on performance and entitlements (Bjorkman and Svensson 2012).

3.1.3 Conditional cash transfers

A randomised approach will be used to evaluate the contribution of the CCT component to the overall impact of SURE-P. This will identify the effect of the CCT, over and above the supply-side interventions, and will take advantage of the staggered implementation of this component of the programme, that is dictated by logistical constraints. Thus, in Phase II, SURE-P PHCs will be randomly assigned to either a CCT group or a control group.

3.2 Overall programme evaluation: the impact of SURE-P as a package

The selection of the healthcare facilities that will receive SURE-P in each of its implementation phases is not amenable to randomisation. These have been pre-selected according to perceived need and physical capacity for receiving the intervention. Thus, the impact evaluation of SURE-P as a package will use a quasi-experimental design, with treatment facilities being matched to control facilities based on similar trends in key indicators.[19] This overall evaluation will determine the combined effect of the supply side and the demand sides (CCT) of the intervention, i.e. the impact of SURE-P as a package on skilled birth attendance and uptake of antenatal care.[20] It will thus focus on the second phase of the programme implementation, in order to allow for the rollout of the CCT component and to allow enough time for the initial problems associated with the implementation of a large scale programme to be resolved.

[19] The matching will be based on facility level data, provided by the SURE-P team, 2006 Census data, and data from the 2011 General Household Survey.

[20] It would have been very interesting to include neonatal mortality among our primary outcomes. However, we do not have a surveillance system that would allow us to collect data on neonatal mortality with enough quality as to postulate it to be a primary outcome. However, we will try to adapt the sisterhood method of measuring maternal mortality to the measurement of neonatal mortality.

Table 3: Summary of Main Research Questions

1. Overall programme

What is the overall impact of SURE-P on skilled birth attendance and use of antenatal care?

2. Supply-side intervention: midwives' incentives and community-based stock monitoring

2.1 What is the impact of non-economic incentives on midwives' attrition?

2.2 What is the impact of monetary incentives on midwives' attrition?

2.3 What is the additional impact on midwives' attrition of the provision of monetary incentives, over and above non-economic ones?

2.4 Regarding monetary incentives: provided that the yearly amount paid is the same, does it matter (in terms of midwives' attrition) whether the tranches paid increase or are constant over time?

2.5 What is the level of monetary incentives that produces the same effect on attrition as non-economic ones?

2.6 What is the effect on drug stockouts of a combined intervention that: informs communities of the available antenatal and obstetric health services; subsequently elicits stockout rates at the local health centres from telephone surveys of pregnant women; finally disseminates this information on stockout rates in the communities benchmarking such information to stockout rates in the facilities available to other communities.

3. Demand-side intervention: CCT

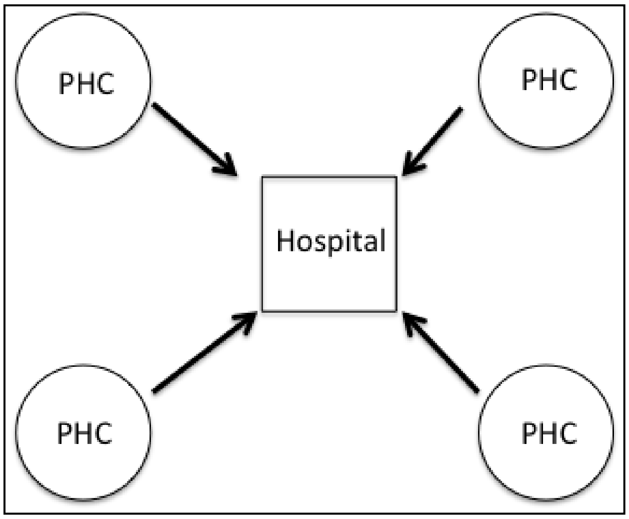
Over and above the effect of the supply side interventions, what is the additional impact on the uptake of skilled birth attendance and antenantal care of a CCT that give mothers a cash transfer conditional on these outcomes?

4. Empirical strategy

4.1 Experimental evaluations: midwives incentives, community-based stock monitoring and CCT

The implementation unit of SURE-P is a healthcare cluster comprising four PHCs and a referral hospital, as shown schematically in Figure 2. This structure reflects the interface between primary and secondary care in Nigeria, and so the four PHCs in each cluster are likely to be exposed to common institutional influences and to face identical logistic and operational constraints. In addition, these PHCs serve communities that are located in relative proximity, therefore sharing many epidemiologic, socioeconomic and cultural features. Very importantly for the SURE-P IE, the midwives of different PHCs are likely to interact within the hospital where their PHC is adhered to. For these reasons, the randomisation unit in all components of the impact evaluation will be the implementation cluster, not the PHC. Also, to minimise the possibility of contamination from affecting our impact estimates, we will refrain from collecting data at the borders of the clusters (a stricter criteria to define the buffer zone is being agreed with the SURE-P team).

Figure 2: SURE-P implementation cluster



Although the randomisation will be done at the cluster level, the outcomes of interest are measured at the individual level. Because the outcomes of individuals within the same cluster are likely to be more correlated than those of individuals in different clusters, realistic intra-cluster correlation coefficients (ICC) are needed for the calculation of sample sizes used in the evaluation. Pagel et al. (2011) estimates ICCs for perinatal outcomes from cluster-randomized control trials in low income countries. The point estimates for ICCs relative to skilled birth attendance, antenatal visits and maternal and neonatal mortality range, in the vast majority of cases, between zero and 0.15. In light of this evidence, which is largely corroborated in the literature, an ICC of 0.15 is used in the following sample size calculations (unless stated otherwise). Likewise, power is specified at the conventional level of 0.8 and statistical significance at 0.05[21]. Required sample sizes will, however, be different for each separate component of the impact evaluation. Table 4 (p. 28) summarises the assumptions made for determining the sample sizes required in each component of the evaluation.

4.1.1 Sample size for midwives incentives (Phase I)

As shown in Table 2, this component of the impact evaluation contrasts two treatment groups – non-monetary incentives (only) and monetary and non-monetary incentives (simultaneously) – and a control group. The main outcome of interest is, in this case, the yearly midwives retention rate, i.e. the proportion of midwives that stay in the location they were assigned to, in the framework of SURE-P, for at least 12 months.

There are no official numbers for midwives attrition but the SURE-P team believes these are especially high: at each moment in time, about 20% of midwife positions in the MSS programme are vacant, despite a sustained inflow of new hires. In light of these concerns, we based our estimations on a baseline retention rate of 0.3 (in the absence of the incentives, 30% of the midwives that start at the beginning of the year stay until the end). As a working hypothesis we further assume that appropriate non-monetary incentives can increase retention rates by 15 percentage points and that a package of simultaneous monetary and non-monetary incentives is able to increase these rates by an additional 15 percentage points. This does not seem implausible in light of the powerful effects of financial incentives on providers' behaviour, reported by Basinga et al. (2010). Although the papers of position of the SURE-P programme do not make explicit a quantitative objective for midwives attrition, the SURE-P team would consider that differences in retention rates of approximately 15 percentage points correspond to the minimum effect size that is relevant for policy.

We set out the power analysis to test three null hypotheses: (1) the retention rates in control and non-monetary incentives (only) arms are the same; (2) the retention rates in the non-monetary incentives arm and in the non-monetary *and* monetary incentives (simultaneously) arm are the same; (3) the retention rates in the control and in the non-monetary incentives *and* monetary incentives (simultaneously) arms are the same. Given that the effect size is expected to be smaller in (1) and (2)

[21] Given the policy significance of SURE-P, power is increased to 0.85 in the evaluation of the overall programme impact.

than in (3), we maximize power with an unbalanced design that increases the number of clusters in the non-monetary arm. In particular, clusters will be allocated as follows: control group – 35 clusters; non-monetary (only) – 51 clusters; monetary and non-monetary incentives – 39 clusters. For a significance level of 0.05 and power of 0.8, we need to interview at least 10 midwives per cluster which is entirely feasible because there are 16 midwives per cluster.[22] Hence, we set out to interview the 16 midwives per cluster to allow for some sample loss; the incremental cost of these additional interviews is relatively small.

In order to minimize sample size loss between the baseline and the follow-up survey, it is important to maintain updated records of midwives' mobile phone numbers, as a means of maintaining the communication between them and the SURE-P team. Small incentives, in the form of free call time, will be given to midwives to incentivise them to keep their contact details up to date[23].

4.1.2 Sample size for monitoring PHC stocks of drugs and essential supplies (Phase I)

There are no reliable estimates of current stockout rates at the PHC level. However, these are believed to be significant and widespread. A current average stockout rate (probability that a pregnant women using a PHC has been denied drugs or essential supplies) of 0.5 is assumed and it is postulated that this component of the programme can reduce it by 15 percentage points, which according to the SURE-P team is a policy-relevant impact size, considered sufficient to potentially extend this component of the intervention to other health care facilities in the future. Furthermore, since the root causes of the diversion of essential supplies to secondary markets are likely to involve a network of complicities that enmeshes the entire cluster, an exceptionally high ICC of 0.3 is assumed.

[22] Sample size calculations were performed using *Stata 11* commands *sampsi* and *samplcus*. Corrections are made to take into account the unbalanced design according to harmonic mean equivalence (see Torgerson and Togerson, 2008 – p. 108-13).

[23] The plan includes collecting the dates of birthdays of select relatives and friends of each midwife. On these dates midwives are entitled to free phone calls to this select group of contacts, conditional on previously contacting the SURE-P team, thereby keeping their personal contact details up to date. Moreover, both the midwives and their select contacts with be sent plastic cards with the contact details of the SURE-P team, in order maintain the possibility of communication even in case of mobile phone loss.

By allocating an equal number of clusters to the treatment and control groups (62 clusters to each group), the minimum number of observations per cluster is 19 telephone interviews at baseline and follow up.

Similarly to the case of midwives attrition, incentives will be put in place to minimise sample size loss, due to individuals changing their mobile phone numbers. Moreover, all individuals will also receive plastic protected cards with the telephone numbers of the SURE-P team, so that communication can be re-established in case of mobile phone loss or malfunction.

4.1.3 Sample size for midwives incentives (Phase II)

The sample size for midwives incentives in Phase II can be calculated much more precisely after we have collected the Phase I data. Based on the sample size calculations of Phase I, it seems realistic to assume that we would need 185 clusters for Phase II: 35 for control, 50 for constant monetary incentives, 50 for increasing monetary incentives, and 50 for the ancillary arm. As in Phase I, we will try to interview the 16 midwives of each cluster. We plan to carry out more precise calculations once we have the data from Phase I.

4.1.4 Sample size for the CCT component (Phase II)

CCTs have been shown to increase the demand for antenatal care by as much as 10 percentage points over the baseline figures (see for example Gertler et al., 2004). This evaluation also makes the assumption of a 10 percentage points positive effect on the proportion of women attending four or more antenatal visits; this is considered the minimum impact size for the intervention to be potentially scaled up in the future.

As for the other experimental interventions, the randomisation unit for this component of the programme will be the cluster. For the assumed parameter values mentioned above, and allocating 93 clusters to the CCT arm and 92 clusters to the No-CCT arm, the minimum number of interviews of new mothers needed to attain 80% power is 12.

4.1.5 Detailed description of the randomisation procedure in Phase I

First, the clusters will be randomly allocated to one of three trial arms: the arm receiving non-monetary incentives to midwives, the arm receiving both types of incentives and the control arm. Second, the clusters in these arms will be randomly assigned to the community-based monitoring scheme arm, or the respective control group.

In order to improve power, the randomisation must do its best to balance the outcome variable at baseline. However, because in 4.1.1 the number of clusters in each arm is different across the trial arms, the most commonly used procedures for pre-balancing (i.e. randomising within a set) are no longer appropriate. The following randomisation procedure will instead be implemented: although it is not be straightforward to demonstrate that this is, in general, optimal, it clearly improves power when compared with the alternative of simple randomization. The precise steps followed are the following:

First: amongst the 125 clusters, 16 clusters will be randomly allocated to the trial arm receiving non-monetary midwives incentives and 4 clusters will be allocated to the trial arm receiving both types of incentives.

Second: The remaining 105 clusters will be grouped in sets of 3 clusters with similar values of baseline midwife attrition rate.

Third: Within each set of three clusters, one cluster is randomly allocated to a different arm of the trial: one cluster is assigned to non-monetary incentives arm; another cluster is assigned to the trial arm receiving both types of incentives; and the third cluster is assigned to the control arm.

Fourth: this procedure leads to the allocation of 51 clusters to the non-monetary incentive arm, 39 clusters to the arm receiving both incentives and the remaining 35 clusters to the arm receiving no incentives. Finally, within each set, clusters will be grouped in pairs according to the stockout rates at baseline; within each pair, one cluster will be randomly allocated to the community-based stock monitoring intervention and the other to the control group.

4.1.6 Econometric analysis of the experimental data

The main hypothesis of the experimental components will be analyzed using Logit regressions at the individual level with standard errors adjusted for clustering. For Phase I, the two main regressions will be one in which the dependent variable is whether a midwife has stayed in her post for 12 months, and another regression in which the dependent variable is whether the individual was denied drugs in the PHC. In both regressions, the right hand side variables will include the outcome variable at baseline and three binary variables (one for individuals living in a non-monetary incentive arm cluster, one for individuals in the non-monetary plus monetary incentive arm, and a third one for the community monitoring treatment arm) and any other exogenous variables believed to be important ex-ante.

For Phase II, the two main regressions will be one in which the dependent variable is an indicator variable taking the value 1 if a midwife has stayed in her post for 12 months, and another regression in which the dependent variable is whether the woman attended prenatal care. In both regressions, the right hand side variables will include four binary variables (one for individuals living in constant monetary incentive arm clusters, another one for the increasing monetary incentive arm, another one for the ancillary arm, and a fourth one for the CCT arm). We also plan to test for interaction effects between the experimental treatment arms.

4.2 Overall SURE-P evaluation

4.2.1 Quasi-experimental design

Two complementary approaches will be used. The first is a classic difference-in-difference (DID) approach (Ashenfelter and Card, 1985; Heckman and Robb, 1985; Blundell et al., 2004). This method is implemented by comparing the difference in average outcomes before and after the programme for the treatment group with the before and after contrast for the control group[24]. In principle, DID allows the estimation of the average program effect on those exposed to it, under two identifying assumptions: common time effects across groups and no composition changes within each group.

[24] Appropriate methods for adjusting for exogenous covariates, in particular through matching, are reviewed in Blundell et al. (2004); strategies for the computation of standard errors of DID treatment effects in a variety of scenarios are presented in Bertrand, Duflo and Mullainathan (2004).

In the short-run, the outcomes of interest for SURE-P will be the proportion of pregnant women who attend antenatal care ('at least one antenatal visit' and 'four or more antenatal visits' will be examined separately) and the proportion of births under skilled birth attendance[25]. Careful inspection of the pre-existing long trends in these variables will inform the choice of control areas that are plausibly comparable with SURE-P areas, under the assumptions of parallel trends and constant group composition. Dedicated baseline and endline household surveys will then be conducted in both treated and control areas, covering all outcomes of interest and a wide range of socioeconomic variables that are used as covariates in the analysis.

Although DID is a widely used impact evaluation tool, the assumption that, in the absence of the programme, the time trends in the treatment and control groups are the same is ultimately untestable. In order to make the evaluation of SURE-P robust to this issue, DID analysis will be complemented with a second approach that relaxes the common time-effects assumption: the changes-in-changes methodology (CIC) proposed in Athey and Imbens (2006). This approach relaxes the common time-trend assumption allowing the outcomes of interest to be a general function of unobservables. Outcomes are assumed not to depend on treatment assignment, conditional on the value of these unobservables and on the time period in question. In addition, unobservables may vary between groups and are only restricted to be constant over time within each group. The use of the two complementary approaches ensures greater robustness of the overall impact evaluation with respect to particular assumptions and practical limitations of each of the methods used.

4.2.2 Sample size for the overall programme evaluation

International evidence from interventions that are similar to the supply-side component of SURE-P (Hatt et al., 2007; Basinga et al., 2010) have been associated with average increases of about 8 percentage points in the demand for antenatal care and skilled birth attendance. Despite the limitations described above, the MSS programme in Nigeria has also been associated with an average yearly increase in the proportion of women receiving antenatal care of about 10 percentage points. Since SURE-P includes an improved supply-side intervention, incentives to midwives, and a CCT,

[25] Longer-term key outcomes will include maternal and neonatal mortality rates.

it seems sensible to assume a combined positive impact of at least 14 percentage points on the proportion of women receiving antenatal care (up from a baseline proportion of 36%). This is also in line with the SURE-P goals of reaching antenatal care coverage of 75%.

The evaluation of SURE-P, as a package, will focus on the second phase of implementation, which extends the programme to an additional 800 PHCs (200 clusters). Specifically, the overall evaluation of the SURE-P will take as its treatment group those clusters which have been randomly assigned to the CCT intervention and to some form of midwife incentive intervention. There will be 75 such clusters; the treatment group therefore receives (i) supply strengthening, (ii) CCT, and (iii) some form of midwife incentive). The control group will comprise 35 matched clusters. It is assumed that the full SURE-P package will increase the rate of antenatal care from 0.36 in the control group/at baseline to 0.50 in the treatment group. To achieve power of 0.85, we will survey 20 women per cluster at baseline and at follow-up.

Table 4: Sample sizes for impact evaluation

Midwives incentives (Phase I)	Community monitoring (Phase I)	CCT (Phase II)	Overall IE (Phase II)
Baseline ret. rate (control): 0.3 Endline ret. rate: 0.45 (0.60 in the case of non-monetary and monetary) Obs. per cluster: 10 midwives Power: 0.8; sig. level: 0.05; ICC: 0.15	Baseline stockout rate : 0.5 Endline stockout rate: 0.35 Obs. per cluster: 19 phone calls Power: 0.8; sig. level: 0.05; ICC: 0.3	Baseline antenatal: 0.40 Endline antenatal: 0.50 Obs. per cluster: 12 women Power: 0.8; sig. level: 0.05; ICC: 0.15	Baseline antenatal: 0.36 Endline antenatal: 0.50 Obs. per cluster: 20 women Power: 0.85; sig. level: 0.05; ICC: 0.15

Notes:

- The sample size for midwives incentives for Phase II has been omitted for being the most tentative.
- Midwive retention rate = proportion of midwives who are still in their post after 12 months among those who were in their post at baseline
- Stockout rate = proportion of women who could not get one of a set of essential drugs (including iron and folic acid supplements, IPT for malaria, and tetanus vaccine) in the Primary Health Facility in the previous three months to the interview
- Antenatal = proportion of pregnant women who completed at least four antenatal care visits among those who were due to have a baby within the previous 3 months to the interview

Table 5: SURE-P impact evaluation design – Phase 1

Midwives Incentives → **Community Monitoring** ↓	**1A.** Non-monetary incentives	**1B.** Non-monetary + monetary incentives	**1C.** Control	*Total*
2A. Community monitoring intervention	26 clusters	19 clusters	18 clusters	*63 clusters*
2B. Community monitoring control	25 clusters	20 clusters	17 clusters	*62 clusters*
Total	*51 clusters*	*39 clusters*	*35 clusters*	*125 clusters*

Table 6: SURE-P impact evaluation design – Phase 2

Midwives Incentives → **Conditional Cash Transfer** ↓	**3A.** Monetary incentives (constant over time)	**3B.** Monetary + incentives (increasing over time	**3C.** Control	**3D.** Ancillary trial arm	*Total*
4A. Conditional cash transfer treatment	25 clusters	25 clusters	18 clusters	25 clusters	*93 clusters*
4B. Conditional cash transfer control	25 clusters	25 clusters	17 clusters	25 clusters	*92 clusters*
Total	*50 clusters*	*50 clusters*	*35 clusters*	*50 clusters*	*185 clusters*

Table 7: SURE-P IE survey sample sizes

Survey	Baseline Phase 1	Follow-up Phase 1	Baseline Phase 2	Follow-up Phase 2
Pregnant women/households	125 clusters * 20 women/cluster = 2,500	125 clusters * 20 women/cluster = 2,500	• Overall SURE-P IE: 110 clusters * 20 women/cluster = 2,200 • Remaining CCT control clusters: 75 clusters * 12 women/cluster = 900 • Total: 3,100	• Overall SURE-P IE: 110 clusters * 20 women/cluster = 2,200 • Remaining CCT control clusters: 75 clusters * 12 women/cluster = 900 • Total: 3,100
Midwives	125 clusters * 16 midwives/cluster = 2,000	125 clusters * 16 midwives/cluster = 2,000	185 clusters * 16 midwives/cluster = 2,960	185 clusters * 16 midwives/cluster = 2,960
Facility Managers	125 clusters * 4 facilities/cluster = 500	125 clusters * 4 facilities/cluster = 500	185 clusters * 4 facilities/cluster = 740	185 clusters * 4 facilities/cluster = 740
WDC chairmen	125 clusters * 4 chairmen/cluster = 500	125 clusters * 4 chairmen/cluster = 500	185 clusters * 4 chairmen/cluster = 740	185 clusters * 4 chairmen/cluster = 740
Community leaders	125 clusters * 4 chairmen/cluster = 500	125 clusters * 4 chairmen/cluster = 500	185 clusters * 4 chairmen/cluster = 740	185 clusters * 4 chairmen/cluster = 740

5. Cost Analysis

Cost is an important concern for the NPHCDA and SURE-P teams, and the IE will therefore include cost and cost-effectiveness analysis. This analysis will consider both budget and actual expenditure patterns across the respective interventions and the program as a whole. To the extent possible, data on cost-effectiveness will be benchmarked to comparable analyses for national and international programmes.

The cost analysis will include the following: (i) describing the cost of the results (achieved and expected) of the project, including a categorization of costs (facilities, renovations, equipment, commodities, salaries, incentives, overhead, etc.); (ii) comparing the cost effectiveness of different parts of the programme across various geographic settings; (iii) exploring alternative options which would produce better cost effectiveness; (iv) sensitivity analysis, including alternative discount rates and assumptions concerning key cost drivers over time; and (v) assessing financial and opportunity costs to beneficiary households and to their communities.

Data for costing will be derived from SURE-P budget and planning documents, SURE-P project staff, and dedicated IE data collection including surveys of households and individuals, midwives, public health facility managers, and community leaders (including Ward Development Committee chairmen).

6. Evaluation limitations and their mitigation

The implementation of SURE-P is likely to face a number of operational challenges that may also affect the impact evaluation of the programme. This section lists the most pressing of them and proposes strategies to either circumvent them, or mitigate their effect.

6.1 Incentives to health workers

The proposed incentive schemes face potential limitations. First, they are designed to reduce staff attrition and improve the performance of midwives, but not of CHEWs. In general, CHEWs are part of the local community, hence unlikely to attrite. Nonetheless, their motivation is important, for they are responsible for the interface between PHCs and the community. It would thus be interesting to evaluate the combined effect of midwives incentives and different incentives regimes for CHEWs. However, SURE-P does not include explicit incentives for CHEWs; also, interacting the three incentives regimes for midwives with CHEWs' incentives would increase the number of cells in Table 2, thereby requiring unfeasibly large sample sizes[26].

Second, as shown by Gneezy and Rustichini (2000), extrinsic incentives need to be substantial enough in order to compensate for a possible crowding-out of intrinsic motivation. In the case of SURE-P, although the FMOH believes that the proposed extrinsic incentives are sufficiently high, this belief is not backed by hard evidence. In the literature, this issue has been accounted for in the design of impact evaluations: Ashraf, Bandiera and Jack (2012) split the extrinsic motivation group

[26] This would be true even in the simplest case of two incentives regimes for CHEWs: incentives *vs.* no incentives.

in two distinct treatment groups; while one is given an admittedly low-powered financial incentive, the other is paid an amount that is presumably higher than the minimum necessary to prompt behavioural change. However, in the case of SURE-P, it is not possible to follow this approach in the first phase of implementation, as it is infeasible to partition the evaluation sample (124 clusters) in more than two treatment groups at the conventional levels of power and statistical significance.

Third, resentful demoralisation of the midwives allocated to the control group could bias the impact evaluation. Extensive consultation with FMOH and SURE-P has indicated that, while this type of problem is likely to occur if incentives differ within the same cluster, it is not expected to be a problem when incentives differ only between clusters. Thus, by implementing the randomisation at the cluster level, the potential for this type of bias is minimized.

6.2 CCT component

Extensive consultation with FMOH and SURE-P staff suggested that political support for the CCT component might be at risk if it is perceived as unfairly distributed across senatorial districts. Concurrently, logistic constraints, such as the capacity to verify pregnant women's place of residence and availability of payment modalities, may dictate an earlier introduction of CCT in some states than in others. Both issues may restrict the number of clusters effectively available for randomisation during phase 1. This problem is expected to become less relevant as the implementation progresses and generalises to a larger number of clusters.

It could also be argued that the CCT component may provide an incentive for increased fertility rates. This is however unlikely for, as made clear in Fiszbein et al. (2009), this type of effects generally is very small, even in the cases of CCTs that impact on household finances far more substantially than SURE-P.

6.3 Monitoring PHC stocks of drugs and essential supplies

Although access to mobile phones in Nigeria is widespread,[27] utilisation is highest amongst the most affluent. Thus, by relying on mobile phone communications as a means of monitoring PHC performance, the most needy households may be missed-out, leading to an underestimation of the number of pregnant women affected by stockouts. This imbalance can, however, be minimised by re-weighing the sample of mobile users in order to make it representative of the cluster population.

In addition, although information on the 'named and shamed' PHCs will not be made public outside the catchment areas of the treated clusters, it is possible that it eventually reaches control group PHCs. We are minimizing this possibility as PHC workers report to a given hospital and we are not allowing treatment to differ intra-cluster.

6.4 Overall programme evaluation

In principle, it is possible that pregnant women who reside in areas adjacent to the treated ones may self-select into treatment. We see the likelihood of this as very limited, given that no informational campaigns will be run in the untreated areas and that, even in treated areas, the expected demand increases are relatively modest. Nonetheless, this type of crossover might potentially drain SURE-P clusters of human and material resources, therefore diluting the effect of the programme. A number of complementary approaches are likely to mitigate this problem. First, geographic buffer zones can be used to prevent control group contamination. Second, even if careful geographic delimitation of buffer zones fails to successfully isolate treatment and control areas, the impact of the program can still be estimated excluding the limited number of areas exposed to control group contamination. Third, in an unlikely scenario of systematic crossovers, these could be exploited as an opportunity to estimate the impact of different degrees of SURE-P intensity. Within this treatment intensity framework, exposure to the SURE-P can be expected to be maximal at the intended treatment areas and will decrease in intensity with the distance to these areas. This approach has been used in a variety of cases, such as in Frolich and Lechner (2010).

[27] According to the General Household Survey – National Bureau of Statistics, 2011 – about 64% of the Nigerian population have access to a mobile phone. Average access rates range from 84% in urban areas down to 54% in rural ones.

6.5 Operational challenges to the evaluation

A number of operational challenges are likely to emerge throughout the implementation of SURE-P. Some of these need to be tackled in a timely manner in order not to compromise the programme evaluation. Namely, it will be critical to ensure that salaries are paid regularly and that incentives, both monetary and non-monetary, are delivered on time, in urban centres and in hard-to-reach areas. Our questionnaire design will feature questions aimed at assessing the strength of the implementation of these components of the programme.

Another aspect that will be vital is the effective monitoring of the CCT component: money needs to reach pregnant women and spot-verification of actual uptake of services needs to be accurate. The SURE-P team is piloting the key operational aspects of the CCT in order to identify the most effective modalities for cash transfers. Finally, our field coordinator will work closely with the SURE-P team in Abuja to monitor these issues and minimize their effects.

7. Data

The SURE-P impact evaluation will use a combination of existing survey and administrative data, and purposively collected data. Existing data sources, to be used for matching SURE-P clusters to comparison clusters in the overall evaluation of the programme and to verify pre-treatment balance between treatment and control groups for the experimental evaluations, include: (i) NPHCDA health facility data, (ii) the 2006 census, (iii) the 2008 Demographic and Health Survey, (iv) the 2011 General Household Survey; and (v) a recently completed GIS mapping of health facilities conducted by the FMOH Department of Public Health (HIV/AIDS) division. Available administrative data will also be used for these purposes, contingent on quality.

Furthermore, several dedicated surveys will be fielded as part of the impact evaluation. These include:

- Q2/2013: baseline survey with modules for households and eligible women (women that have had a baby in the past 12 months), midwives, primary health facility managers, ward development committee chairmen, and community leaders

- Q2-3/2014: first round of follow-up data collection
- Q2-3/2015: second and final round of follow-up data collection.

Additionally, focus groups will be carried out in Q1/2 of 2013 to inform the design of non-monetary incentives for midwives.

8. Budget and Timeline

8.1. Budget

The estimated SURE-P IE data collection budget is given in Table 5.

Table 8: Estimated SURE-P IE data collection budget (in USD)

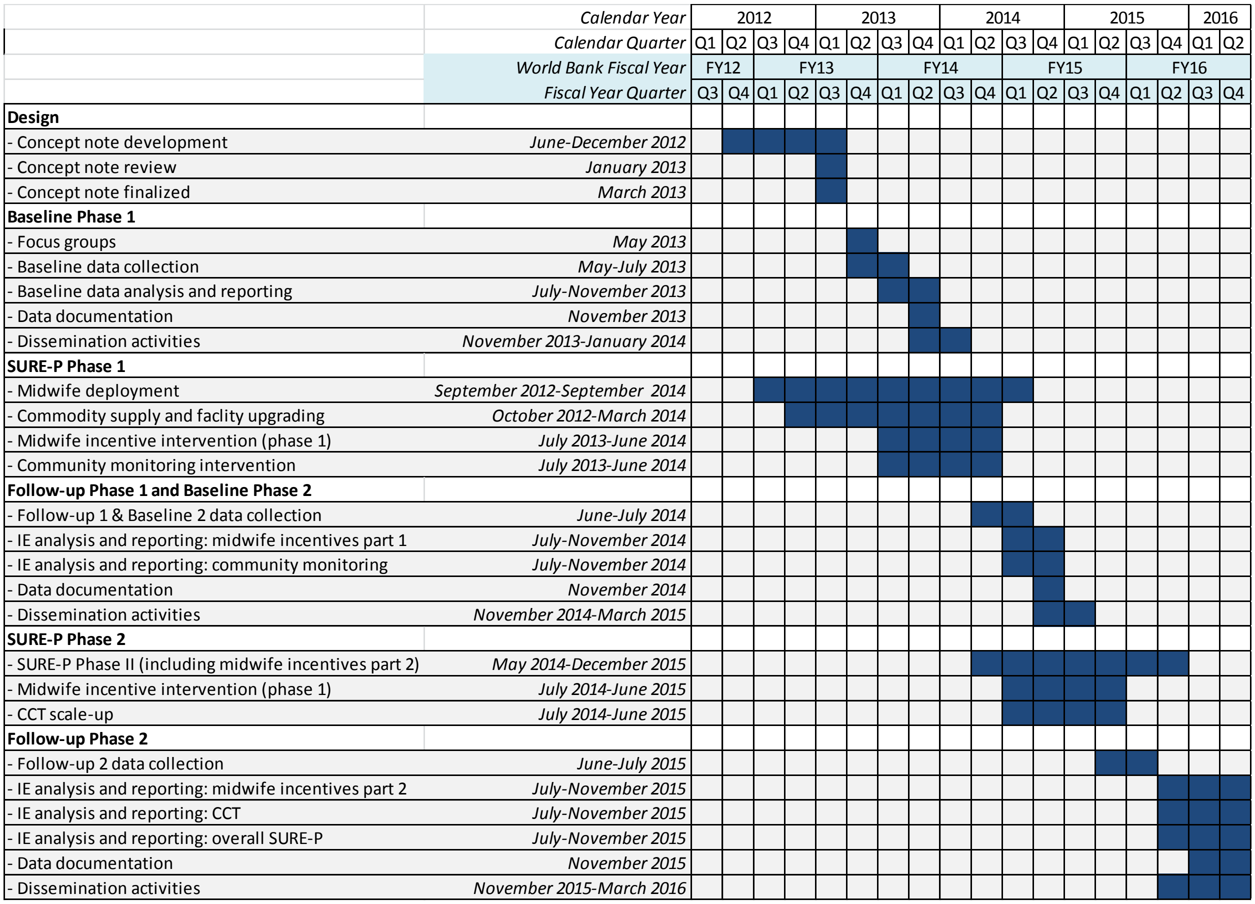
Summary of data collection costs	**Baseline 1**	**Follow-up 1**	**Baseline 2**	**Follow-up 2**	**Total**
Preparatory focus groups	$18,000	0	0	0	$18,000
Impact evaluation surveys	$435,000	$435,000	$595,800	$595,800	$2,061,600
Tracking of midwives	$9,532	0	$14,108	0	$23,640
Total	$462,532	$435,000	$609,908	$595,800	**$2,103,240**

Impact evaluation research, analytical, and coordination costs are estimated at USD 110,000 per year for 3.5 years (June 2012-January 2016), for a total of USD 385,000. Total budget for the impact evaluation is therefore $2,488,240.

8.2. Timeline

The SURE-P IE timeline is shown in Figure 2.

Figure 3: SURE-P IE timeline



	Calendar Year	2012				2013				2014				2015				2016	
	Calendar Quarter	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2
	World Bank Fiscal Year	FY12		FY13				FY14				FY15				FY16			
	Fiscal Year Quarter	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Design																			
- Concept note development	*June-December 2012*		■	■	■	■													
- Concept note review	*January 2013*					■													
- Concept note finalized	*March 2013*					■													
Baseline Phase 1																			
- Focus groups	*May 2013*						■												
- Baseline data collection	*May-July 2013*						■	■											
- Baseline data analysis and reporting	*July-November 2013*							■	■										
- Data documentation	*November 2013*								■										
- Dissemination activities	*November 2013-January 2014*								■	■									
SURE-P Phase 1																			
- Midwife deployment	*September 2012-September 2014*			■	■	■	■	■	■	■	■	■							
- Commodity supply and faclity upgrading	*October 2012-March 2014*				■	■	■	■	■	■	■								
- Midwife incentive intervention (phase 1)	*July 2013-June 2014*							■	■	■	■								
- Community monitoring intervention	*July 2013-June 2014*							■	■	■	■								
Follow-up Phase 1 and Baseline Phase 2																			
- Follow-up 1 & Baseline 2 data collection	*June-July 2014*										■	■							
- IE analysis and reporting: midwife incentives part 1	*July-November 2014*											■	■						
- IE analysis and reporting: community monitoring	*July-November 2014*											■	■						
- Data documentation	*November 2014*												■						
- Dissemination activities	*November 2014-March 2015*												■	■					
SURE-P Phase 2																			
- SURE-P Phase II (including midwife incentives part 2)	*May 2014-December 2015*										■	■	■	■	■	■	■		
- Midwife incentive intervention (phase 1)	*July 2014-June 2015*											■	■	■	■				
- CCT scale-up	*July 2014-June 2015*											■	■	■	■				
Follow-up Phase 2																			
- Follow-up 2 data collection	*June-July 2015*														■	■			
- IE analysis and reporting: midwife incentives part 2	*July-November 2015*																■	■	■
- IE analysis and reporting: CCT	*July-November 2015*																■	■	■
- IE analysis and reporting: overall SURE-P	*July-November 2015*																■	■	■
- Data documentation	*November 2015*																	■	■
- Dissemination activities	*November 2015-March 2016*																■	■	■

9. Staffing

The SURE-P impact evaluation is led by lead investigators Pedro Rosa Dias (Lecturer in Economics, University of Sussex) and Marcos Vera-Hernández (Senior Lecturer in Economics, University College London), co-investigator Marcus Holmlund (IE Coordinator, DIME), and Vincenzo Di Maro (Economist, DIME; IE TTL). Research, analytical, and coordination support is provided by Bright Orji (SURE-P IE Project Advisor), and Olufemi Adegoke and Felipe Dunsch (DIME Coordinating Team).

Oversight and strategic guidance is provided by Dr. Ugo Okoli (SURE-P Project Director), Marie Francoise Marie-Nelly (World Bank Country Director for Nigeria), Arianna Legovini (Manager, DIME), Trina Haque (Sector Manager for Health, Nutrition, and Population, West and Central Africa), Benjamin Loevinsohn (Lead Public Health Specialist, Health, Nutrition, and Population, West and Central Africa), Dan Kress (Deputy Director and Chief Economist, Policy Analysis and Financing, Bill & Melinda Gates Foundation), Hong Wang (Senior Program Officer, Bill & Melinda Gates Foundation), and Mara Hansen (Associate Program Officer, Bill & Melinda Gates Foundation).

The SURE-P Programme Team will be involved at all stages of this research, and a SURE-P IE cluster has been created as the primary link for this work. This includes Dr. Sidi Ali Mohammed (Head, Health Workforce/Supply), Amina Muhtar (Head, Conditional Cash Transfer Planning and Evaluation Unit), Jamila Bello-Malabu (Human Resources for Health Officer), and Ejeckam Chukwuebuka Chukwukadibia (Team Leader, Monitoring and Evaluation).

The research team is working closely with colleagues in the World Bank Nigeria Country Office including Dinesh Nair (Senior Health Specialist), Oluwole Odutolu (Senior Health Specialist), and Shunsuke Mabuchi (Health Specialist).

The team is supported by Melanie Melindji (Program Assistant, World Bank, Washington DC), Janet Adebo (Team Assistant, World Bank, Nigeria Country Office), and (Ugonne Eze (Team Assistant, World Bank, Nigeria Country Office).

References

Akresh, R., De Walque, D. and Kazianga, H. 2012. Alternative cash transfer delivery mechanisms: impacts on routine preventative health clinic visits in Burkina Faso, NBER working paper 17785.

Ariely, D., Bracha, A. and Meier, S. 2009. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. American Economic Review 99(1), 544-55.

Ashenfelter, O. and Card, D. 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. The Review of Economics and Statistics, 67, 648-660.

Ashraf, N. Bandiera, O. and Jack, K. 2012. No Margin, No mission? A field experiment on incentives for pro-social tasks. EOPP working paper 035.

Athey, S. and Imbens, G. W. 2006 Identification and inference in nonlinear difference-in-differences models, Econometrica, 74, 431-497.

Attanasio, O., Gómez, L., Heredia, P. and Vera-Hernández, M. 2005. The Short-Term Impact of a Conditional Cash Subsidy on Child Health and Nutrition in Colombia. Centre for the Evaluation of Development Policies Report Summary, Institute for Fiscal Studies, London.

Attanasio, Orazio p, Costas Meghir, and Ana Santiago. 2012. Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA. The Review of Economic Studies (August).

Barham, Tania. 2005. Providing a healthier start to life: the impact of conditional cash transfers on infant mortality. Department of Agriculture and Resource Economics, University of California at Berkeley, CA.

Basinga, P., Gertler, P., Binagwaho, A., Soucat, A. Sturdy, J. and Vermeersch, C. 2010. Paying primary health care centers for performance in Rwanda. World Bank policy research working paper 5190.

Bertrand, M., Duflo, E., and Mullainathan, S. 2004. How much should we trust differences-in-differences estimates?” Quarterly Journal of Economics, 119, 249–275.

Bevan, G. and Hamblin, R. 2009. Hitting and missing targets by ambulance services for emergency calls: effects of different systems of performance measurement within the UK. Journal of the Royal Statistical Society Series A, 172: 161-190.

Beven, G. and Hood, C. 2006. What’s measured is what matters: targets and gaming in the English public health care system. Public Administration 84:517–538.

Blundell, R. Costa Dias, M., Meghir, C. and Van Reenen, J. 2004. Evaluating the employment impact of a mandatory job search program," Journal of the European Economic Association 2(4): 569-606, 06.

Björkman, M. and J. Svensson. 2009. Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda. The Quarterly Journal of Economics 124 (2):735–769.

Fiszbein, A., Schady, N., Ferreira, F., Grosh, M., Kelleher, N., Olinto, P., Skoufias, E. 2009. Conditional cash transfers: reducing present and future poverty. Policy Research Report, The World Bank - Washington DC.

Frolich, M. and Lechner, M. 2010. Exploiting Regional Treatment Intensity for the Evaluation of Labor Market Policies. Journal of the American Statistical Association, American Statistical Association. 105(491): 1014-1029.

Gertler, P. 2004. Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment. American Economic Review, 94(2): 336–341

Glassman, A., Todd, J., and Gaarder, M. 2007. Performance-based incentives for health: conditional cash transfer programs in Latin America and the Caribbean. Working Paper 120. Washington, DC: Center for Global Development.

Gneezy, U. and Rustichini, A. 2000. Pay enough or don't pay at all. Quarterly Journal of Economics, 791-810.

Hatt, L., Stanton, C. Makowiecka, K., Adisasmita, A., Achadi, E. and Ronsmans, C. 2007. Did the strategy of skilled attendance at birth reach the poor in Indonesia?" Bulletin of the World Health Organization 85(10), 774–82.

Heckman, J. and Robb, R. 1985. Alternative Methods for Evaluating the Impact of Interventions: An Overview Journal of Econometrics, 1985, 30(1-2), pp. 239-67.

Hernández, B., Ramírez, D., Moreno, H., and Laird, N. 2005. Evaluación del Impacto de Oportunidades en la Mortalidad Materna e Infantil. In External evaluation of the impact of the human development program Oportunidades 2004, ed. Hernández-Prado, B. and Hernández-Ávila, M. 73–95. Mexico: National Institute of Public Health.

Lagarde, M., A. Haines, and N. Palmer. 2007. Conditional Cash Transfers for Improving Uptake of Health Interventions in Low- and Middle-Income Countries: A Systematic Review. The Journal of the American Medical Association 298 (16),1900–1910.

Levy, D. and Ohls, J. 2007. Evaluation of Jamaica's PATH Program: Final Report. Mathematica Policy Research, Washington, DC.

Li, X. 2012. Time is Money: a natural field experiment on how a non-monetary timely feedback system motivates volunteers. National University of Singapore Working Paper.

Lim, S. Dandona, L., Hoisington, J.A., James S.L., Hogan, M.C., and 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

List, J., Sadoff, S. and Wagner, M. 2010. So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. NBER working paper 15701.

Morris, S., Flores, R., Olinto, P. and Medina, J.M. 2004. Monetary Incentives in Primary Health Care and Effects on Use and Coverage of Preventive Health Care Interventions in Rural Honduras: Cluster Randomised Trial. The Lancet 364: 2030–37.

Pagel, C., Prost, A. Lewycka, S., Das, S. Colbourn, T., Mahapatra, R. Azad, K., Costello, A. and Osrin, D. 2011. Intracluster correlation coefficients and coefficients of variation for perinatal outcomes from five cluster-randomised controlled trials in low and middle-income countries: results and methodological implications. Trials 12/151: 1-12.

Powell-Jackson, T., Mazumdar, S. and Mills A. 2012. Financial Incentives in Health: New Evidence from India's Janani Suraksha Yojana. London School of Hygiene and Tropical Medicine. Mimeo.

Todd, P. E. and Wolpin, K. I. 2006. Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility, American Economic Review 96(5), 1384–1417.

Torgerson, D. and Torgerson, C. 2008. Designing and running randomised trials in health, education and the social sciences. Basingstoke, Palgrave Macmillan.

Vera-Hernández, Marcos. 2003. Structural Estimation of a Principal-Agent Model: Moral Hazard in Medical Insurance. The RAND Journal of Economics 34 (4) (December 1): 670–693.

World Bank. 2009. Reducing maternal mortality: strengthening the World Bank response. Population and reproductive health policy report - Washington DC

World Health Organization. 2007. Maternal mortality in 2005: estimates developed by WHO, UNICEF, UNFPA and The World Bank. Geneva