Journal of Development Economics - Registered Report Stage 1: Proposal

Community Health Provision at Scale: Evidence from a Randomized Trial in Uganda

Authors (institutional affiliation): Phyllis Awor (Makerere University); Martina Björkman-Nyqvist (Stockholm School of Economics); Andrea Guariso (Trinity College Dublin, *corresponding author*); Jakob Svensson (Stockholm University, IIES)

Date of original draft: 27/04/2020 *(original submission)*
Date of latest draft: 24/03/2021 *(re-submission)*

Abstract:

We study a scaled-up version of an incentivized Community Health Worker program aimed at improving primary healthcare provision and reducing child mortality in Uganda. An earlier proof-of-concept study (Björkman-Nyqvist et al., 2019) found that program achieved its intended result. In this paper we study the impact of the rolled out program to assess whether the program's effectiveness is retained when the program is operating at scale and how other health actors active in the primary health care sector are affected. We will answer these questions using a large-scale cluster-randomized controlled trial across 500 trial villages and more than 12,500 households. The household survey is complemented with a survey of approximately 1,300 Community Health Workers and 700 drug shops that were active in the trial villages at baseline.

Keywords: primary health care; child mortality; community health workers; social entrepreneurship; Randomized Controlled Trial; Uganda.

JEL codes: O12, I11, I12

Study pre-registration: PACTR (PACTR201609001398349) and AEA (AEARCTR-0002392)

Proposed timeline: The study presented in this submission started in 2015. Respondent listing activities and baseline data collection took place from November 2015 through June 2016. Program implementation started immediately after baseline data collection. A short compliance survey was conducted between October 2017 and May 2018, roughly 18 months after the beginning of program implementation. Endline data collection was originally planned to start in May 2020, roughly 4 years after the beginning of program implementation. However, in light of the current COVID-19 pandemic, the data collection has been postponed to start in December 2020. Endline data collection is expected to take 7 months. The final clean dataset is expected to be available within three months from the end of data collection and we plan to write the research paper within 6 months.

1. Introduction

Despite improvements in under-five child mortality in recent decades, an estimated 5.3 million children still die from preventable diseases worldwide every year (WHO, 2019). A majority of these deaths occur in the poorest countries of the world, in areas within countries of underserved populations with inadequate access to basic preventative and curative services.

A common approach to improve primary healthcare and reduce child mortality in low and middle-income countries is Community Health Worker (CHW) programs (Singh and Sachs, 2015). Although global statistics are quite poor and incomplete, the latest WHO figures record more than 1.8 million CHWs across 57 countries (WHO, 2018). The design of the CHW programs, including the selection, training, tasks definition, and remuneration (or lack thereof) of the CHWs, varies from setting to setting and the existing evidence provides mixed results on their overall effectiveness (Sloan et al, 2008; Arifeen et al, 2009; Darmstadt et al, 2010; Bhutta et al, 2011; Bhandari et al, 2012; Kirkwood et al, 2013; Amouzou et al, 2016; Boone et al, 2016). Evaluations of various proof-of-principle programs, however, indicate that well-designed CHW programs can lead to dramatic reductions in child mortality (Kidane and Morrow, 2000; Kumar et al, 2008; Baqui et al, 2008; Björkman Nyqvist et al, 2019). Typically, such proof-of-principle trials are run in selected areas, often with intensive supervision and support. Thus, whether such programs can retain their effectiveness when expanded at scale remains an open question.

With just a few notable exceptions (e.g. Banerjee et al, 2017), there are few attempts to experimentally study whether positive results from pilot studies carry through once the program runs at scale. At the same time, there are several widely recognized challenges in using results from pilot projects to draw conclusions about policies implemented at scale, including market equilibrium effects, context dependence, site-selection bias, and piloting bias. In this project, we study a scaled-up CHW program, which has previously been shown to be highly effective in reducing child mortality in a proof-of-concept study implemented in Uganda (Björkman Nyqvist et al, 2019). To assess its impact and study some of the potential threats to external validity, we design a large-scale randomized controlled trial, which will enable us to understand whether the program retains its effectiveness when implemented across a more heterogeneous population and under improved general health conditions, as well as whether it affects the local market for health provision - which are all dimensions that may potentially dilute its overall impact.

The community health promoters (CHP) program that we study in this project is implemented in Uganda by two Non-Governmental Organizations – Living Goods (LG) and BRAC. The details of the program are provided below, but one of the key innovations is the introduction of a set of financial incentives for the community health workers. While WHO recommends remunerating CHWs, it also acknowledges that the existing evidence on the benefit of remuneration is of very

low quality and that many programs across the globe still rely on volunteers (WHO, 2018b). In the CHP program, instead, the community health workers benefit from two different categories of financial incentives. First, they make profits by selling a range of health-related products to community members while carrying out their standard activities as community health workers. As it will be explained better below, the supply chain for these products is directly managed by the NGOs. Second, they receive additional performance-based remuneration based on a set of key health activities that they perform, which include sick child assessment, registration and support of pregnant women, and visits to newborns in the first week of life. The program was first introduced in Uganda in 2009 with roughly 500 CHPs. A randomized trial across 214 villages started in 2010 and found that the program reduced under-5 mortality by 27% over three years (Björkman Nyqvist et al, 2019). Since then, the CHP program has been massively scaled up across Uganda: by 2016 there were more than 5,500 CHPs, serving a population of about 4.5 million people, and the stated target is to reach 13,000 CHPs by the end of 2021.[1] The new study presented here exploits this expansion to study the impact of the program when it is running at scale. The randomized trial we implement is on a large scale: covering 500 villages and spanning 13 districts of Uganda. Baseline data collection in 2015-2016 surveyed 12,615 households, 18,424 children below 5 years of age, 1,363 Community Health Workers, and 754 drug shops (more details below). A short compliance survey was administered to the households between October 2017 and May 2018, while endline data collection is scheduled for winter 2020.[2] The rich dataset will enable us to address two main questions.

1) *Can the reduction in child mortality observed in the "proof-of-concept" study be sustained when the program operates at scale?*

As discussed and shown in Banerjee et al (2016, 2017), there are many reasons why results might not carry through once programs run at scale. Factors such as market equilibrium effects, context dependence, site-selection bias, and piloting bias, might impact the program when at scale. In addition, the overall health context and primary health needs evolve over time. Over the past decade basic healthcare has steadily improved in many countries, including Uganda, where child mortality dropped nationwide from 77.4 deaths per 1000 live-births in 2010 to 46.6 deaths per 1000 live-births in 2018 (World Bank, 2019). Understanding whether a program that was effective in the context of Uganda in 2010 retains its effectiveness when operating at scale and within a significantly improved health context, is a highly policy relevant question.

[1] The program is moreover being exported to Kenya, Myanmar, and Sierra Leone, although with some local adaptations from the original Ugandan model.

[2] The COVID-19 pandemic forced us to postpone the endline survey, which was originally planned to start in May 2020, until the emergency is over. This delay is not affecting in any way the study design, as the program keeps running in all treatment areas.

There are relatively few studies investigating the impact of effective proof-of-concept studies as they are scaled-up.[3] To the best of our knowledge, this will be the first study to assess the impact of the scaling up of a community healthcare program. While we will not be able to exactly isolate the individual channels that might lead to differential results in the scaled-up assessment, with the rich data at our disposal we will be able to discuss and provide suggestive evidence for many of them. In particular, the design will allow us to speak directly to one of the potentially more serious challenges: market equilibrium effects.

2) *What is the impact of introducing a new incentivized community health worker program on other existing health service providers?*

One challenge that is particularly relevant for any CHW program is the risk of market equilibrium effects. That is, the new health service providers (community health workers in our case) may impact the motivation and behavior of other health agents operating in the same health market. These reactions might enhance or hamper the effectiveness of the new program, with ambiguous consequences on the overall quantity and quality of healthcare provided to the communities. And although in the first pilot study we observed a positive impact of the program on household health outcomes, we could not with any certainty assess the impact on other health agents in the market because of lack of data. This new study will instead allow us to understand whether and how the CHP program affects the supply-side of primary health care provision in the target communities. We will focus on the two main health actors whose primary catchment area overlaps with the CHPs: other CHWs active in the community and drug stores operating in (or nearby) the trial villages. On the former, we want to understand how other CHWs that are operating as part of either governmental or other non-governmental programs react to the introduction of the new CHPs. The new CHP program is expected to increase the number of active CHWs as well as their activities, ultimately ensuring better primary healthcare provision to the local community. Existing evidence, however, suggests that the introduction of a remunerated community health program in a context where most workers are volunteers might also create tensions and ultimately lead to an overall reduction in both quantity and quality of service provision (Deserranno et al, 2020). We will investigate these potential channels using experimental data derived from a survey administered to the entire universe of all types of CHWs operating in the study villages at baseline and endline. We will measure their level of knowledge, activity, and motivation. Our design, and the data we will collect, will also allow us to study changes in the type and number of health workers as well as associated changes in the type and quality of health service provision in the community.

[3] Banerjee et al. (2017) is probably the best-known example. Banerjee, Duflo, and Kremer (2016) discuss the success of Development Innovation Ventures (DIV) – a USAID program to bring demonstrated pilot studies to scale. The Yale Research Initiative on Innovation and Scale (Y-RISE) advances research on the effects of policy interventions when delivered at scale.

The focus on drug stores is motivated by the fact that CHPs sell a variety of health products, which include high quality drugs, at competitive prices (typically below market price), thus increasing the competition in the local market for drugs. Also in this case it is a priori unclear how drug stores might react and what might be the consequence for the community. Increased competition from a high-quality actor might push some stores out of business, thus decreasing the availability of heath products for the community, but it might also induce them to increase the quality (and variety) of products they offer, thus improving the overall drugs available to the community (as found in Björkman Nyqvist et al, 2018). Also in this case we will survey the entire universe of drug stores operating in the villages at baseline and endline. We will monitor entry and exit of drugs stores from the market and we will measure the quality of the drugs, focusing on the treatments for two of the main killer diseases for children under-5 in the study region: malaria (Artemisinin-based combination therapy or ACT) and pneumonia (antibiotics).

Background and Intervention

The first CHW program in Uganda was set up by the government in 2001 – the so-called Village Health Team (VHT) program, which is still operating today across the country – with the aim of making basic health care and health education accessible to everyone. The VHT program resembles a standard CHW program: VHT members are volunteers, selected among community members, who receive an initial basic training and are then expected to conduct home visits, educate households on essential health behaviors, provide basic medical advice, and refer the more severe cases to the closest health center. Although some non-experimental evidence suggests that the program can promote healthcare and improve health outcomes (Brenner et al, 2011; Kalyango et al, 2012; Kayemba Nalwadda et al, 2013; Ministry of Health Uganda, 2015), multiple challenges have been highlighted in relation to motivation, remuneration, training, performances, and retention of the health workers (Kimbugwe et al, 2014; Turinawe et al, 2015; Mays et al, 2017) and in an in-depth assessment of CHW programs across the world the WHO only assigned a score of 8 points out of 36 to the VHT program (Bhutta et al, 2010).

Against this background, in 2007 Living Goods, an NGO active in Uganda, in collaboration with BRAC Uganda started developing a new community health delivery model. Living Goods and BRAC use data-driven performance management, regular in-service training, and supportive supervision to enable CHWs to deliver high-quality primary health care services. Importantly, unlike the VHT and many other volunteer-based programs, the Community Health Promoter (CHP) program provides a rich set of financial incentives. First, the program harnesses the power of franchised direct selling to provide CHPs with incentives to increase access to low-cost, high-impact health products and basic newborn and child health services. In addition to that, the program also provides performance-based incentives to encourage core health activities, such as household visiting, sick child assessment, registration and support of pregnant women, and visits

to newborns in the first week of life. Both the lists of health products and of remunerated activities are regularly revised and re-targeted according to evolving needs. The program is organized into geographically based branches, each managed by a branch managers that is supervised by either Living Goods or BRAC, depending on the location. Each CHP gets assigned to a specific cluster, which in most cases corresponds to a village.

The CHPs are selected through a competitive process among female community members aged 18 to 45 who applied for the position in each village and who possessed basic writing and math skills. Eligible candidates receive 3 weeks of health and business training. At the end of the training, candidates need to pass a skills test in order to be equipped as an active CHP. The NGOs provide an initial set of products to all newly recruited CHPs, together with a uniform, a mobile phone, and a set of training materials and visual aids to use during household visits. CHPs also attend a one-day training each month to review and refresh key health and business topics.

The CHPs tasks mirror the standard CHWs tasks (conduct home visits, educate households on essential health behaviors, provide basic medical advice, referring the more severe cases to the closest health center), but on top of that, as mentioned above, they also sell preventive and curative health products. The product line they have at disposal includes prevention goods (e.g. insecticide treated bednets, water purification tablets, and vitamins), curative treatments (e.g., oral rehydration salts, zinc, and ACTs), as well as other health-related commodities (e.g. diapers, hand soap, fortified food) and durables with health benefits (e.g. improved cook stoves, solar lights, and water filters). These products are sold by the CHP at a discount. The retail price is determined by the NGOs head office with a target of keeping prices for preventive and curative products about 20% lower than the prevailing local market prices. The CHPs in turn purchase these products directly from Living Goods or BRAC branches at wholesale prices between 30-50% below market prices and therefore earn an income on each product sold. Thus, the CHPs operated as micro-entrepreneurs with financial incentives to meet household demand. The broad product mix has three potential benefits: (i) driving up total sales and income for the CHPs; (ii) enabling the NGOs to cross-subsidize prices (dropping prices on essential health products and increasing the margins on other products); (iii) motivating CHPs to be out visiting households regularly by including high-velocity items (such as soap and fortified foods) in the product mix. The business training received by the CHPs stresses the importance of building up a customer-base by providing free services like health education, referrals, and newborn visits. As described above, the income deriving from the micro-entrepreneurial activity is then increased through performance-based incentives, designed by the NGOs to further encourage key health activities such as household visiting, sick child assessment, registration and support of pregnant women, and visits to newborns in the first week of life. Since 2013, Living Goods and BRAC also equip the CHPs with smartphones that includes a rich mobile health application. The application helps guide the CHW through workflows, keep track of their stock, serve as a client management system, and prioritize certain

activities based on timeliness (e.g. pregnancy follow-up) or household risk. Overall, this allows monitoring the CHPs' activity, while collecting real-time health data from the field.

A first evaluation of the impact of the CHP program began in 2010 (Björkman Nyqvist et al, 2019) when the NGOs were still operating in just few hundred villages. The evaluation was based on a cluster-randomized controlled trial that involved 214 villages in 10 districts across Uganda. The villages were stratified by geographical zones and 115 villages were randomly assigned to the treatment group, where the CHP program started operating in January 2011, while 99 villages were assigned to the control group. The evaluation was based on an endline survey collected at the end of 2013, which covered 7,018 households and 11,563 children under-5 that lived in the same village throughout the trial. The study found that over the three years the CHP program reduced under-5 mortality rate by 27% (adjusted rate ratio 0.73, 95% CI 0.58-0.93) in the treatment compared to the control arm. The effects were of similar order of magnitude for infant mortality (adjusted rate ratio 0.67, 95% CI 0.51-0.87) and neonatal mortality (adjusted rate ratio 0.73, 95% CI 0.55-0.98).

Following the first study, the program has been massively scaled up across Uganda, as well as implemented in other countries, such as the neighboring Kenya. The study presented in this submission takes advantage of this scaling up to investigate the two key questions outlined above. This new study involves the same main actors of the first one: program implementers (Living Goods and BRAC), data collection agency (Innovations for Poverty Action), and funding agency (Children's Investment Fund Foundation). This helps ensuring that the design, the management, and the implementation of the research program remains the same as in the first study. There are, however, some important differences: the new study focuses on a program that is running at a scale that is more than tenfold the size it had at the time of the first evaluation, it studies its impact in a changed overall environment and measures it over a longer time period[4], it relies on a much larger sample (500 villages and more than 12,500 households), it exploits a much richer set of data, which includes new surveys from other health service providers in the community, and it relies on a panel of households identified at baseline, rather than on a cross-section.

2. Research Design

Methodology

In order to answer the two research questions outlined above, we rely on a large-scale cluster randomized controlled trial embedded in the scale up of the CHP program across the country. The

[4] We initially expected the evaluation to cover 4 years, but the delays caused by the COVID-19 pandemic led us to extend it by another 9 months.

study covers a sample of 500 villages (LC1), organized into 15 zones that span 13 districts across all four regions of Uganda. The sample is stratified by zone so that within each of the 15 zones the villages are randomly divided into a treatment group and a control group, using balanced randomization[5]. Following randomization, at least one CHP were assigned to each one of the 250 villages in the treatment group.[6] No CHP was assigned to the 250 control villages.

The empirical analysis will be based on two rounds of data collection - baseline and endline - each one including multiple components: a household survey, administered to a subset of households living in the study villages; a CHW survey, administered to the entire universe of CHWs (government, NGO, and private) operating in the study villages; and a drug shops survey, to tests the type and quality of drugs sold in the entire universe of drug shops in the study villages. Baseline data was collected in early 2016. Endline data will be collected as soon as the conditions will allow fieldwork to be carried out without any risks for public health.[7] The rich data at our disposal will provide a comprehensive picture of basic health outcomes as well as key features of both the demand and supply of primary healthcare in the study villages and will allow us to measure the causal effect of the program by comparing outcomes in the treatment and control areas after the program has been running for more than 4 years.

Data

Sample Size

The sample size was designed to detect a reduction in *under-5 mortality* (primary outcome of interest), defined as number of under-5 deaths per 1,000 child-years of exposure to the risk of death under the age of 5. We used data from the control group in the proof-of-concept study conducted by the research team in similar settings (Björkman Nyqvist et al, 2019) to obtain the relevant inputs for the computation. A total sample of 500 clusters (250 per study arm) and 25 households per cluster at baseline (12,500 households in total) allows us to detect a reduction in child mortality of 20% or larger, at the 5% significance level with 80% power, assuming between-cluster coefficient of variation equal to 0.43 and attrition rate of 16% (or 4 households per cluster). Under the same assumptions, this design and sample size also has 80% power to detect a 21% reduction in infant mortality and a 25% reduction in neonatal mortality at the 0.05 significant level. We report the details on the power calculation and the data used in Appendix A.

[5] The randomization was performed by the research team using Stata software.
[6] In line with the standard protocol followed by the NGOs running the program, the number of CHPs assigned to each village depended on the village size.
[7] Endline data collection was expected to start in May 2020, but has been postponed due to the COVID-19 pandemic. According to the latest information from the field, data collection should be allowed to begin by the end of the year.

Variations from the intended sample size

Attrition poses a threat to the ability to detect the expected effect size. The survey implementer (IPA Uganda) has extensive experience in tracking respondents and has built up a rich protocol to reduce attrition as much as possible. At baseline we recorded a comprehensive set of information to allow us tracking back respondents. This included names and nicknames of the respondents and other household members, multiple phone numbers, GPS coordinates, and description of landmarks near the household. About 18 months after the beginning of program implementation we conducted a short compliance survey that covered all households in our sample with the aim to study whether the program was implemented in the treatment villages and not in control areas and update the contact details. For the compliance survey we managed to survey 90.2% of the baseline sample. At endline we will try in any case to track back also the remaining 9.8% of the respondents.[8] We plan to use the rich contact details to track participants even if they move outside the baseline village.[9] For endline we also set up a protocol to replace households that cannot be surveyed, despite all best efforts. The replacement strategy – described below in details – will provide us with a subset of households without baseline information.

Clearly, attrition poses a threat not only to the power of the study, but also to its integrity. Indeed, attrition bias may be introduced whenever people join or drop out of the study in a way that is systematically related to the program provided. We do not expect our program to directly affect migration decisions, in line with the results from the first study conducted by the research team (Björkman Nyqvist et al, 2019). One might still suspect that the program, by improving the quality of primary health care in the village, could reduce attrition in treatment area compared to control areas. However, during the compliance survey we observed, if anything, somewhat higher attrition in treatment (10.4%) as opposed to control (9.3%) areas (p-value = 0.09). Nevertheless, in the empirical analysis section we will describe how we will formally test for attrition and take it into account in our analysis.

Contamination poses a risk to an experiment by diminishing the experimental contrast between the treatment and control groups. In the case of the CHP program, by design randomization takes place at the village level, which reduces the threat of contamination as opposed to settings in which randomization happens at the individual level. At the same time, there is a risk of contamination deriving from CHPs interacting with and providing their services to households located outside of their village. In the previous study we observed that household interaction with CHPs dramatically dropped for households located further than 500mt from the house of the CHP. In order to limit contamination, we therefore selected villages with a minimum distance of 1km from each other. While the design does not (and does not want to) prevent households in control areas to interact

[8] The power calculation described above and in Appendix A assumes endline attrition equal to 16%.
[9] Due to budgetary constraints we expect to be able to only track households that moved within the same district.

with CHPs in treatment villages, we believe this risk will remain low.[10] Finally, in order to reduce contamination at the stage of data collection enumerators and team leaders will always be blinded on the treatment status of villages and respondents.

Partial- or non-compliance with the assigned treatment poses another threat to the power of the experiment. If a unit randomly assigned to receive a certain program in fact does not receive it, the experimental contrast of the study will be diminished, and so will the power of the study. In the case of the programs studied here, the intervention takes place at the village level and we can identify two levels of compliance. The first is at the village level: whether all and only the treatment villages received the CHP program. The second one is at the individual/household level: whether a household has been directly or indirectly exposed to the CHP program. The latter is going to be a key dimension that we are going to measure, analyze, and discuss in our study – namely whether and how households interact with CHPs and benefit from their services. For the former, we believe that within our setting it is relatively easy for the implementing partners to have full control over the program implementation and to adhere to the agreed protocol. We nevertheless worked on minimizing the risk of non-compliance in two ways. First, the research team and the implementing partners – who have been collaborating on different research projects for more than 9 years – have regular contacts to ensure smooth running of the study and to promptly tackle any potential issue or question. Among other things, the NGOs granted the research team access to their administrative data that record details on programs and activities village by village. Secondly, 18 months after the beginning of program implementation we administered a short compliance survey in all study villages to check program implementation in the field. The compliance survey showed that indeed randomization significantly increased the likelihood that a household report interacting with a CHPs in the 30 days preceding the survey.

In any case, in order to preserve the integrity of the study, in the analysis we will always assign a respondent to the treatment or control group depending on the location at baseline, and irrespectively of whether the respondent remained in the same location throughout the study and/or interacted with a CHP. Hence, our final estimates will be interpreted as the impact of introducing the CHP program in a village, irrespectively of whether the respondent interacted with the CHP or not.

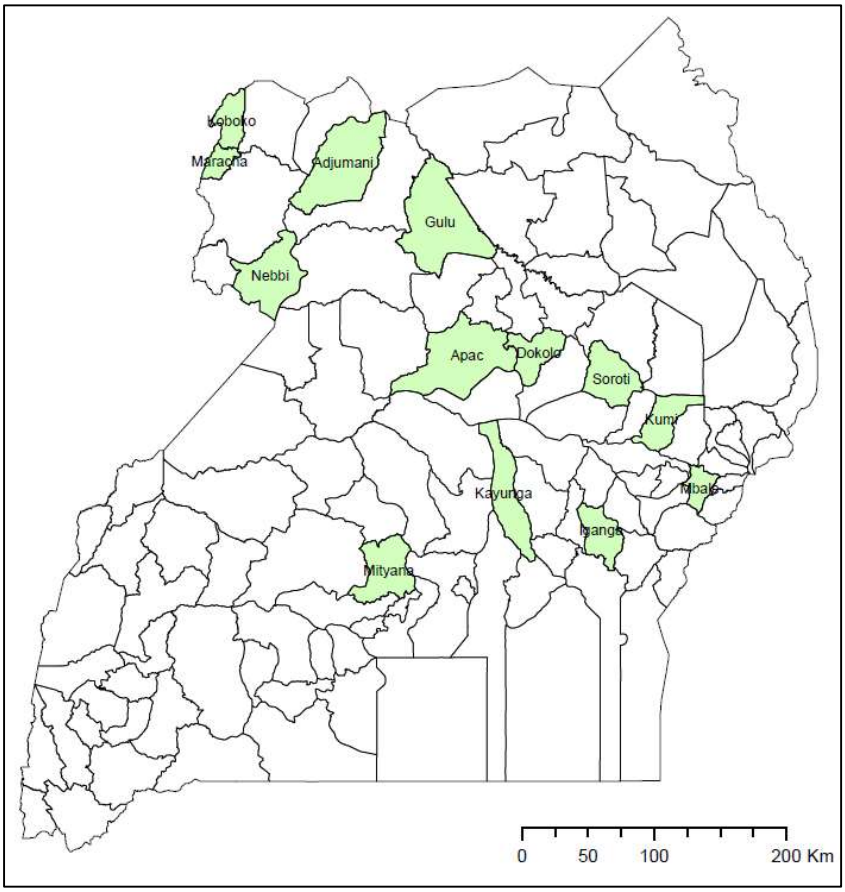
Villages Selection

The village selection proceeded in steps. First, the NGOs followed their standard procedures to identify and map villages that they considered eligible for expanding the CHP program, within the

[10] Moreover, such contamination would if anything lead us to underestimate the impact of the program. And even in the first study the estimated 27% reduction in child mortality happened despite the fact that 5.4% of the households in control areas reported interacting with a CHP in the 30 days preceding the survey.

areas surrounding their branch offices.[11] This resulted in a list of 810 villages. From this list, we selected the 500 study villages in a way that maximized distance between villages, so to reduce the risk of contamination between treatment and control arms.[12] The villages are equally split between the two NGOs: 250 villages fall within the reach of a BRAC branch (11 branches in total) and 250 villages fall within the reach of a Living Goods branch (4 branches in total). Figure 1 show a map of Uganda where the 13 districts that are part of the study are colored in green.

Figure 1. Map of Uganda with Study Districts



[11] The two organizations collected information on a number of different variables per village including: distance to branch office, number of households in the village, estimated population density, area economic status, distance to nearest health facility or clinic, MTN phone network, and presence of other NGOs with ICCM programs. The NGOs combined all these dimensions in a weighting index, which they then used to identify eligible villages, where the CHP program could take place.

[12] Concretely, we calculated the pairwise distance between villages, by district, and we dropped the villages closest to their next neighboring village. We then made cross-district adjustments to maintain even numbers of villages in each district, to enable us to have an equal amount of villages in the treatment and control arms of the RCT in each district. Distance was computed based on the location of the local chairman's house. We also verified the village location against administrative maps of Ugandan districts provided by UBOS, updated in 2012, using QGIS (v2.12.0).

Baseline

Survey Components

At baseline we administered three different surveys within all study villages. First, we designed a household survey to collect information on health knowledge, attitudes, and behaviors. The survey was answered by a female member of the household (see below for eligibility criteria). During the household visits we also collected anthropometric data on all children below five years of age living in the household. Second, we designed a CHW survey to learn about activities, motivation, and knowledge of all CHWs (government, NGO, and private health workers) active in the study villages. Third, we designed a drug quality survey to test the type and quality of medicine for pediatric malaria and pneumonia sold in the drug stores of the villages. We describe here in more details how we identified the sample for each survey component and how we administered baseline surveys. We will then describe the plan for endline data collection.

A. Household Survey

Eligibility

Based on the above power calculation, we targeted 25 households within each village. Our goal at baseline was to identify households with the highest likelihood of having (at least) one child born during the study period (i.e. within 4 years following the baseline survey). This selection was meant to increase the power of the study, compared to a purely random sample, given that the primary health outcomes we are interested in relate to maternal and child health. On the basis of conversations with local informants and pre-testing, four simple criteria were identified as best predictors of whether a woman in the study areas was likely to deliver during the evaluation timeline: 1) currently pregnant, 2) aged 16-35 years old, 3) with a young child less than three years old, 4) married (formally or informally).

We therefore defined an eligible household as a household with a female permanent resident aged 16-35 years old, and either currently pregnant, and/or with a young child, and/or married. A preference was placed on pregnant women, then women with a young child, then married women. The woman was then identified as the primary respondent of the baseline household survey. When multiple women in one household met the same criteria, we selected the woman most likely to remain in the household (i.e. household head or wife of household head), and if we still needed to choose, we prioritized the youngest eligible woman.

In practice, to identify the 25 study households we proceeded as follows. First, we performed a village census listing activity to collect a comprehensive list of households within each village. During this exercise the field officer sat with a Listing Committee, typically composed of the Local

Chairman of the village, members of the Village Health Team (VHT), and other knowledgeable village members that the chairman deemed helpful for this exercise, and they provided a list of all households in the village. Second, for each household in the list, the Listing Committee reported the eligibility criteria. Third, we digitized the lists and used Stata software to rank the households within each village, following the priority of our criteria. Within the same category, we randomized the order of the households in the list. In this way we identified the first 25 households of each list as our sample households and noted additional eligible households as potential replacements. Replacement households were selected following the (random) order of appearance on the list.

Questionnaire

The household survey was administered to the primary female respondent and had 8 sections:

1) household member rosters;
2) basic medical knowledge;
3) incidence and treatment of malaria, diarrhea, and pneumonia in children under 5 years old;
4) nutrition;
5) pregnancy-related knowledge, attitudes, and behaviors;
6) childbirth and neonatal knowledge, attitudes, and behaviors;
7) interactions with health workers;
8) socioeconomic status

Sections 5 and 6 were administered to all women in the household who had recently been pregnant or gave birth, respectively. When one of these women was not present (3% of the cases), the main respondent was asked to answer questions about this woman's pregnancy and/or childbirth to the best of her knowledge. All questions were read to the respondent in her preferred language[13] and she was given time to answer. The respondent was asked questions alone, where no other people could listen to her responses, and only after granting consent. The household survey took on average around 90 minutes to complete.

Once the household survey was completed, an especially trained field officer took anthropometric measurements of all children 59 months old and younger living in the household. If children were not present, the field officer returned up to two more times to seek anthropometric measurements. In 404 cases (2% of the total) anthropometric measures could not be recorded because the child was never at home. The measurements recorded for each child were weight (measured three times in case difference between first two measurements was greater than 0.5 kg); height (measured twice); mid-upper arm circumference (MUAC) (measured twice); a visual check for swelling in the extremities (edema); and a physical check for pitted edema.

[13] The household survey was translated into nine local languages to ensure respondent comprehension of the questions: Luganda, Lusoga, Lugiso, Ateso, Kumam, Madi, Alur, Kakwa, and Luo.

B. CHW Survey

Eligibility

At baseline we surveyed the entire universe of active CHWs in the village. In practice, this included all CHWs that had conducted any community health work in the preceding six months independently on whether they were government, NGO, or private health workers. CHWs were first identified during the village census activity thanks to the help of the Listing Committee. To ensure full coverage of active CHWs in the village, the list was checked and updated when the survey team was in the village to administer surveys, by asking village members about any CHW operating in the village. Finally, all surveyed CHWs were asked about any additional CHWs that might have been operating in the village.

Questionnaire

The CHW survey had 6 sections:

1) demographics;
2) CHW organization/entity information;
3) CHW motivations and organizational support;
4) basic medical knowledge;
5) recent activities;
6) recent activities for children < 5 years.

The survey was administered in the respondent's preferred language and took on average about 45 minutes to complete.

C. Drug Quality Survey

Eligibility

At baseline we collected sample drugs from the entire universe of drug stores operating in the study villages. In order to identify the stores, a field officer asked a few village members (at least three) all places where they could purchase medicine within the village. The field officer then visited all locations mentioned by the village members. These included shops, pharmacies, medical clinics, and informal locations such as personal stores or households. In villages without any drug outlets, the field officer visited the closest drug shop that village members typically attended for purchasing medicines, even if outside of the village boundaries.

Survey

The medicines were collected from the shop using a covert shopper methodology. Field officers were trained to act as typical locals in order to get unbiased service from drug shop providers. These drug purchasers followed strict protocols, to maximize consistency and appear as much as

possible as regular clients of the shop. They entered each village on foot and, once they reached the shops, they explained that they had a 3-year-old child with malaria and a 4-year-old child with pneumonia symptoms and asked for the appropriate treatments.[14] Once outside the store and away from view from the drug shop owner, the field officer placed the drugs in an envelope with an ID number on it to allow identifying the sample. The field officer also completed a short survey that recorded: GPS location, name of drug shop and directions to reach it, drug name, cost, dosage, packaging, and any additional note about the purchase. Drugs were subsequently sent to Kampala where they quality tested using Raman spectroscopy. Appendix B provides the details on the procedure followed for testing the drugs.

Compliance Survey

The compliance survey was administered between October 2017 and ended in May 2018, roughly 18 months since program implementation.

The compliance survey only consisted of a short household survey to both the control and treatment areas (less than 30 mins) and it served two main purposes: 1) track back and update the contacts of households identified at baseline, and 2) study compliance of program implementation using information provided by the households.[15] While doing so, we also collecting information on a small set of outcomes 18 months since the start of program implementation. The survey included the following short sections:

1) Contact details;
2) Interactions with CHWs;
3) Child mortality and disease prevalence;
4) Antenatal and postnatal care.

We tracked all the households that were surveyed at baseline, irrespectively of whether they are still living in the same location or not. We used a combination of strategies in order to find back the households. First, we used of the multiple mobile phone contacts that were recorded at baseline. Second, we used GPS location to find back the exact household location. Finally, we relied on the help of neighbors and village chairpersons. By resorting to this strategy we managed to track 90.2% of the original households. This includes both households that remained in the same location (94.3% of the surveyed households) and households that had migrated to a new location but could still be reached (5.7% of the surveyed households).

Within each household, the target respondent was the same female respondent that was interviewed at baseline. When the original respondent was not available, we interviewed the next

[14] We focused on malaria and pneumonia because they are prevalent diseases in the study areas and CHPs sell drugs to treat them.
[15] We did not survey the health workers and did not carry out any direct monitoring of their presence and activities.

knowledgeable female person in the household. Overall, 87.8% of the surveys were answered by the original respondents.

Endline

The endline data collection is going to mirror the baseline process and components, but it will rely on an even richer set of tools.

A. Household Survey

At endline we plan to track back and survey all baseline households. We will follow a similar approach as the one adopted for compliance survey to track back respondents. For households that could be tracked back at compliance survey, we will rely on their updated contacts, but we will also attempt to track back again households that could not be found at the compliance survey stage, using their baseline contacts. The target respondents will be the baseline respondents, or the next most knowledgeable woman in the household whenever the original respondent is not available.

The household survey will mirror the structure and content of the baseline, but will include a richer section on interaction with CHWs. The sections on pregnancy and childbirth will be again administered to all women in the household who had been recently pregnant or gave birth (over the previous 24 month), respectively. Following the same procedure as baseline, especially trained enumerators will also collect anthropometric measures from all children under-5 living in the households.

In order to preserve sample size, at endline we will replace baseline households that despite our best efforts cannot be tracked back. Although for the replacement households we will not have baseline data, we plan to use this additional sample to enrich our final analysis. The replacement strategy will be as follows. First, we will reach the point where the baseline HH was supposed/expected to be found based on available information (which includes GPS coordinates). Considering as staring point the location where the baseline HH was expected to be found, we will check the first household on its right-hand side for eligibility. Eligibility will be based on the following 3 conditions:

1. The household is NOT already part of our survey;
2. The household has been living in the village for at least 4 years;
3. The household had a child born over the past 5 years (irrespectively of whether the child died or is still alive);

If the household is not eligible, we will move to the household on the right-hand side of it and check eligibility, until we find a valid replacement, or until we reach the end of the street (or

village). In the latter case, we will go back to the original starting point and start checking households on the left-hand side, following the same procedure.[16]

B. CHW Survey

The endline CHW survey will take two slightly different forms. The main survey will be administered to all CHWs active within the village at the time of the endline data collection. Similarly to the procedure followed at baseline, this will include all CHWs that conducted any community health work in the preceding six months. The procedure to identify the CHWs as well as the structure of the survey will mirror baseline.

A second, shorter, version of the survey will be administered to CHWs that were surveyed at baseline but that are no longer active as CHWs within the community. This survey will not have any question on recent activity and performances, but will include some additional questions on their reasons for dropping out.

As we plan to survey the entire universe of active CHWs, there is no replacement strategy for the CHW survey component.

C. Drug Quality Survey

Also in this case we will follow the same procedure as baseline to collect a sample of drugs from all drug stores in the village. We will survey all drug stores that are active at endline. This means that we will include in our sample also all new drug stores that became active during the study period. We will also attempt to record any information on drug store surveyed at baseline that are not operating any more by the time of endline, by asking community members when and why they closed down. Appendix B provides the details on how the drug sample collection and testing are performed.

D. LC1 Survey

At endline we will administer also a short survey to the LC1 chairperson. The aim of this survey will be first of all to record any relevant village-level event that took place during the study period. This will include major investments, NGO interactions, natural disasters, and government funding. The second objective of the survey will be to double check with the chairperson the list of drug stores as well as CHWs operating within the village.

Data collection and data management

All data collection rounds are managed by IPA. Survey data is always collected in digital form, using tablets and the SurveyCTO data collection platform. At baseline all surveys were administered on Samsung Galaxy Tab4 tablets using SurveyCTO Collect. Each completed survey

[16] In case that is still not enough, we will move to check households on the opposite side of the street.

was then uploaded to the project server each day and data was downloaded and aggregated. Throughout the data collection process, the data remained encrypted with Boxcryptor. The survey was translated into nine local languages to ensure respondents comprehension of the questions: Luganda, Lusoga, Lugiso, Ateso, Kumam, Madi, Alur, Kakwa, and Luo. Enumerators and team leaders are always blinded on the treatment status of villages and respondents.

IPA ensures data quality through two processes: survey audits and high frequency checks of recorded data. High frequency checks are performed on a daily basis during data collection, on incoming data, using Stata software. A code is written to look for data outliers, logical inconsistencies, and missing data. In this way issues are immediately identified and actions can be taken to revise, clarify, or correct the wrong answers with the enumerators and, whenever needed, with the help of the original respondent. Audit surveys, instead, repeat a small subset of questions from the original surveys; auditors locates respondents few days after the survey has been completed and administer a (much) shorter version of it, probing whenever the answer they receive differ from the one originally recorded by the enumerator, in order to assess whether there were any issues in survey administration, comprehension, or completion. Audit surveys are conducted with 10% of the household and CHW respondents. Whenever discrepancies arise, enumerators and auditors meet to understand the origin of the discrepancy and verify the correct response, in limited cases verifying answers directly with the household respondent. The verified answers are then corrected in the database.

Stopping Rule and Withdrawal

Risks associated with this study are expected to be minimal. The only real risk associated with the survey is the possibility of conjuring up painful memories in parents who have experienced the death of a child. Though it is highly unlikely, if a subject experiences significant, emotional distress he or she will be referred to a local public health or medical professional. All participants will always be able to interrupt their participation in the study at any point in time at no cost. The informed consent will stress this possibility and will contain the relevant IPA contact details that can be used to request withdrawal even after the survey has been completed. A copy of the form will be left with the participants. In the event of a problem related to the study, the researchers will immediately contact the IRB and send a report detailing the adverse advent.

Outcomes

Primary Outcomes

To assess the impact of the scaled-up program, in relation to the *first* research question, the primary outcome of interest is *under-5 mortality*. We will compute mortality at the cluster level using information contained in the household survey. The survey records: 1) detailed birth information

on all children under five living in the households at the time of the survey; 2) detailed birth and death information on all children that died under the age of five during the study period.

At endline, for each child, we will define the number of month of exposure to the risk of death during the trial period, defined as the difference between the birth date of the child, or the start date of the trial if the child was born before that date, and the date that the child turned five years if that occurred during the trial period, or the date of the endline household survey if the child was less than five years old at that time, or the date of the death of the child. Under-five mortality will then be calculated as number of under-five deaths over the trial period per 1,000 child-years of exposure to the risk of dying under the age of five. We will also compute infant mortality as number of deaths during the trial period arising within the first year of life per 1,000 infant-years of exposure, with infant-years of exposure calculated in a similar way as the child-years of exposure described above. Finally, we will compute neonatal mortality as the number of deaths during the trial period within the first month of life per 1,000 births. We will also collect data on miscarriages and stillbirths (in-utero deaths) during the trial period. All mortality measures will be defined at the village level. [17]

To study how other health actors react to the CHP program, in relation to the *second* research question, we will look at both the extensive and the intensive margins. We will start by studying the impact on the extensive margin. Here, we will focus on the number of drug shops and the number of active CHWs operating in the study villages at endline. These outcomes will be defined at the village level. We will identify as active CHW any CHW that carried out any CHW-related activity over the six months preceding the survey. Next, we will look at the intensive margin. Here, we will focus on the quality of the drugs sold in the drug shops, as well as the level of interaction and type of activities carried out by the CHWs operating in the study villages. In this case the outcomes will be defined at drug sample and CHW level, respectively.

Secondary outcomes

By relying on the different survey tools mentioned above, we will collect a range of additional outcomes that will allow us to dig deeper into the mechanisms behind the main result.

[17] International organizations such as UN and WHO typically express mortality in terms of deaths per 1,000 live-births. Such organizations use data collected over long periods of time and rely on a life-table approach to compute mortality as a probability. Given that our evaluation lasts only for four years, the most appropriate approach in our case is to compute mortality as a ratio, following the steps described above, and to express it in terms of years of exposure. For completeness and in other to facilitate comparisons with other estimates, we will in any case also report results obtained using a life-table approach. Finally, and for completeness, we will report child mortality measures simply expressed in terms of number of child deaths in the village during the study period. In order to make the procedure clearer, the attached do file includes the codes that will be used to compute the different child mortality measures.

Concerning the first research question, the secondary outcomes will serve to investigate the following secondary hypotheses:

1.1) The program increased the chances that a household interacts with and benefits from services provided by the CHPs;
1.2) The program increased the overall amount and quality of health services received by households;
1.3) On top of the impact on child mortality, the program improved additional health outcomes, related to family planning, pregnancy, newborn and child health;
1.4) The program improved the basic health knowledge of the households;
1.5) The program improved the health behavior (both preventive and curative) of the households;

Concerning the second research question, the secondary outcomes will serve to investigate the following secondary hypotheses:

2.1) The program increased the (average) satisfaction, motivation, and confidence of the CHWs operating in the village;
2.2) The program lowered the overall turnover of the CHWs in the village, increased the number of CHWs that work for multiple health organizations, and reduced the total amount of time CHWs dedicate specifically to the government VHT program in the village;
2.3) The program increased the (average) health knowledge of the CHWs operating in the village;
2.4) The program increased the amount of (self-reported) activities of the CHWs operating in the village;
2.5) The program impacted the supply of drugs in the community, by reducing the number of drug stores operating in the local markets, raising the quality of their service, and lowering the price of the drugs;

Appendix C reports the complete list of variables that we will investigate in our analysis, arranged by category.

3. Analysis

Empirical Model

Our primary specification is straightforward and will entail the regression of the outcome variables on a dummy for the treatment status of the village,

$$Y_{i,v,b} = \beta T_{v,b} + \tau_b + \varepsilon_{i,v,b} \qquad (1)$$

where Y is the outcome for individual i (depending on the outcome, it might be a child, a woman, or a community health worker), living in village v, in the catchment area of branch b. In some cases the outcomes will be defined at the village level (e.g. child mortality[18] or drugs quality). Given that we considered the NGOs' branches as blocking variable when performing the initial randomization, all specifications will include branch fixed effects τ_b. Standard errors will be clustered at the village level. The coefficient of interest β will capture the impact of the program on outcome Y.

We will also run two augmented versions of our main regressions, where we will include a control for the baseline value of the outcome variable in order to increase the precision of the estimates.

Correction for Multiple Hypothesis Testing

Given the number of outcomes in our study, multiple testing is a concern. We will therefore follow Kling, Liebman and Katz (2007) and calculate standardized effects for each family of outcomes (see Appendix C for the detailed list of variables). We will also report both robust standard errors as well as the p-values of tests of the null that treatment has no effect computed using randomization inference. [19] The do file attached to this submission includes the specific commands that will be used to run the analysis.

Sample

Households

The main analysis will include the full sample of households that we have identified at baseline and that we have been able to track till endline, plus the replacement households. As described above, at endline we will track back and survey baseline households even if they moved outside the study village (as long as we will be able to track them) and these households will always be included in our analysis. Whenever we will not be able to track back a baseline respondent, in order to preserve power, we will replace the household, following the procedure described above.

We will show the robustness of all our results by excluding the replacement households from the analysis as well as by including baseline controls.

CHW and Drugs stores

For CHWs and drug stores we will have two repeated cross-sections covering all active health workers and stores, which will be included in the analysis. Outcomes in this case will be mostly defined at the village level (see details in Appendix C).

[18] See previous section as well as attached do file for details on how mortality rates will be computed.
[19] We will construct these p-values using 1,000 randomly selected permutations of the randomization allocation. The p-value is then constructed based on the proportion of test statistic values (squared of the estimated coefficients) that are greater than the actual test statistic value.

Data Checks

Balance checks

Following standard practice, in our analysis we will report balance checks performed on the variables collected at baseline. As we already have baseline data, Appendix D reports 10 tables with the full set of checks performed on the baseline variables. The tables show 150 different comparisons between treatment and control group. Out of these we observe 12 variables (8%) in which the difference is significant at 10% level (p-value<0.1), 8 variables (5.3%) in which it is significant at 5% level (p-value<0.05), and 3 variables (2%) in which it is significant at 1% level (p-value<0.01). Overall, the tables clearly show no systematic difference between the treatment and the control groups at the onset of the study. These checks allow us to credibly attribute to the program any difference that we will observe between the groups at endline.

Attrition

We will check that any attrition caused by households that moved to a different district (or that we are not able to track in any way) is non-systematic. In practice, we will run the empirical model mentioned above, using baseline data and replacing the dependent variable Y with an indicator for whether the households could be tracked till endline or not. Non-systematic attrition would imply the coefficient β to be not statistically distinguishable from zero.

On top of comparing the level of attrition across study arms, we will also assess whether the composition of the households lost at follow-up varied across the two study arms. We will start by simply checking the characteristics of the households that we lost at follow-up as and compare them to those that we could track back[20]. We will then interact these characteristics with the treatment indicators, to assess whether certain specific types of households were more or less likely to drop out from one study group.

Because of all the precautions taken during survey work, we do not expect to observe differential attrition. However, in case the analysis will suggest otherwise, we will present bounds of treatment effects, using the approach of Lee (2009).

Missing values and Outliers

As described above, in the data collection section, we plan to identify unusual missing values and outliers straight away during data collection, through the high frequency checks that we will perform on a daily basis on incoming data. This will give us the opportunity to double check and revise any missing entry or outlier due to errors that might have taken place during data collection. We will therefore consider the final dataset to contain only "true" missing values and unusual

[20] We will consider baseline household wealth, household composition, presence of a pregnant woman, and basic health indicators.

values. We therefore do not plan to introduce any correction in our main results. As a robustness check, we will in any case generate our main results by geographic zone, to check that our results are not driven by any one specific zone.

Discussion and Heterogeneous effects

There are several reasons why a scaled-up program may have differential impact from the pilot program. In the setting we are considering, we believe the most important reasons relate to market equilibrium effects, context dependence, site-selection bias, and piloting bias. Below we briefly describe why we think these challenges may be important and how we intend to study them.

- *Market equilibrium effects.* The introduction of the CHP program is likely to impact the motivation and behavior of other health agents operating in the same health market. These reactions might enhance or hamper the effectiveness of the new program, with ambiguous consequences on the overall quantity and quality of healthcare provided to the communities. As discussed above, assessing these potential market equilibrium effects is a core focus of our second research question. Specifically, we will zoom in on the two main health actors whose primary catchment area overlaps with the CHPs: Community Health Workers (as part of either governmental or other non-governmental programs) and drug stores operating in the trial villages. At baseline we collected comprehensive data on these actors by surveying all active CHWs as well as by collecting data from all drug stores operating in (or nearby) the trial villages. In this way we will be able to measure how their behavior changes, both at the intensive and extensive margin, as the new community health program is introduced.

- *Site-selection bias.* By studying market equilibrium effects we will also indirectly deal with site-selection biases. In principle, market equilibrium effects could have been present also in the first pilot study. However, that trial was run mostly in villages where other community health workers were (at least de facto) not active; implying that whether they changed their behavior or not likely played a minor role for the estimated effects (which raises site-selection concerns). In this study, instead, the baseline level of activity of the other health actors appear significantly larger, and through our design we will be able to quantify any market equilibrium effects.[21] In sum, if we find differential effects, we will be able to determine whether market equilibrium effects/ site-selection concerns can explain these differences.

[21] According to our baseline data (Table D.2), about 18% of the households were visited by a health care provider in the 30 days preceding the survey. This share was just about 4% in the pilot study.

- *Context dependence* and *piloting bias*. As also discussed above, an important difference between the pilot study and the current one is that the CHP program has been massively scaled-up. Studying the program at scale allows us to speak to the important challenges of both context dependence and piloting bias. Specifically, the program is now implemented in a larger and more heterogeneous population and we can, through heterogeneity analysis, explore whether there are differential effects depending on the setting. The overall health context has also significantly improved compared to the first study, with child mortality dropping from 77.4 deaths per 1000 live-births in 2010 to 46.6 deaths per 1000 live-births in 2018 (World Bank, 2019). The new RCT is implemented at a scale that is essentially twice as large in terms of sample size compared to the pilot study, which enables us to detect smaller effect size when it comes to child mortality, for example. Here too we plan to study heterogeneity, in this case by initial health conditions in the village. On top of this, we plan to test whether across the two trials we observe changes in the mediating factors - i.e. in the various tasks performed by the CHPs (the inputs in the health production function) - and/or we observe changes in the impact these factors have on the core health outcomes. While such a decomposing will not provide an exact answer, if we observe lower inputs (i.e. less effort or activity by the CHPs), this would be consistent with the idea that the program, when running at scale, cannot sustain the level of engagement we observed in the pilot program (piloting bias concern). If, on the other hand, we observe changes in the impact of inputs on child mortality, one explanation would be that for the same level of engagement, less is achieved in terms of mortality reduction (context dependence).

- *Other channels.* Two other potential channels relevant for interpreting our results are *different time frame* and *political reactions*. On the former, the pilot program was ran for a period of three years while this program is evaluated after more than four years. For some key outcomes, such as child mortality, we will be able to study if this matters, since we will be able to construct mortality measures over the entire trial period. Regarding the latter constraint, we do not expect this dimension to play a major role, as there has been no major changes in the political environment over the past years.

Building up on the above points, we plan to run a set of heterogeneity checks, to gain additional insights on the impact of the program. In particular, we plan to examine heterogeneous effects of the treatment with respect to:

1) *village health conditions*: baseline child mortality rate and baseline average health conditions.

2) *village characteristics*: BRAC vs Living Goods area; baseline average health knowledge of the CHWs operating in the village; baseline average wealth in the village; baseline size of the village and density of CHWs; baseline distance from main road and nearest hospital.
3) *household characteristics*: baseline wealth[22]; baseline distance from the CHW and CHP houses.

Any potential additional heterogeneity check that we might include in the final paper will be explicitly described as exploratory and "not pre-specified".

[22] Wealth will be constructed using principal component analysis and combining all the asset-ownership and household-related questions included in the survey.

4. Bibliography

Arifeen, SE, Hoque DM, Akter T, Rahman M, Hoque ME, Begum K, Chowdhury EK, et al. 2009. Effect of the Integrated Management of Childhood Illness Strategy on Childhood Mortality and Nutrition in a Rural Area in Bangladesh: A Cluster Randomised Trial. *Lancet* 374 (9687): 393–403.

Amouzou A, Hazel E, Shaw B, Miller NP, Tafesse M, Mekonnen Y, Moulton LH, et al. 2016. Effects of the Integrated Community Case Management of Childhood Illness Strategy on Child Mortality in Ethiopia: A Cluster Randomized Trial. *American Journal of Tropical Medicine and Hygiene* 94 (3): 596–604.

Baqui, AH, El-Arifeen S, Darmstadt GL, Ahmed S, Williams EK, Seraji HR, Mannan I, et al. 2008. Effect of Community-Based Newborn-Care Intervention Package Implemented through Two Service-Delivery Strategies in Sylhet District, Bangladesh: A Cluster-Randomised Controlled Trial. *Lancet* 371 (9628): 1936–44.

Banerjee A, Banerji R, Berry J, Duflo E, Kannan H, Mukherji S, Shotland M and M Walton. 2016. Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India. *NBER Working Paper* No. 22746.

Banerjee A, Banerji R, Berry J, Duflo E, Kannan H, Mukherji S, Shotland M and M Walton. 2017. From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *Journal of Economic Perspectives* 31(4): 73-102.

Banerjee A, Esther D, and M Kremer. 2016. *The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy*. The State of Economics, the State of The World Conference at the World Bank.

Bhandari N, Mazumder S, Taneja S, Sommerfelt H, and TA Strand. 2012. Effect of Implementation of Integrated Management of Neonatal and Childhood Illness (IMNCI) Programme on Neonatal and Infant Mortality: Cluster Randomised Controlled Trial. *BMJ* 344: Article e1634.

Bhutta ZA, Zohra SL, Pariyo G, and L Huicho. 2010. *Global Experience of Community Health Workers for Delivery of Health Related Millennium Development Goals: A Systematic Review, Country Case Studies, and Recommendations for Integration into National Health Systems*, WHO Geneva.

Bhutta ZA, Soofi S, Cousens S, Mohammad S, Memon ZA, Ali I, Feroze A, et al. 2011. Improvement of Perinatal and Newborn Care in Rural Pakistan through Community-Based Strategies: A Cluster-Randomised Effectiveness Trial. *Lancet* 377 (9763): 403–12.

Björkman Nyqvist, M., Guariso A, Svensson J, and D Yanagizawa-Drott. 2019. Reducing Child Mortality in the Last Mile: Experimental Evidence on Community Health Promoters in Uganda. *American Economic Journal: Applied Economics*, 11(3): 155-92.

Björkman Nyqvist M, Svensson J, and D Yanagizawa-Drott. 2018. Can Competition Reduce Lemons? A Randomized Intervention in the Antimalarial Medicine Market in Uganda. *Mimeo*.

Boone P, Elbourne D, Fazzio I, Fernandes S, Frost C, Jayanty C, King R, et al. 2016. Effects of Community Health Interventions on under-5 Mortality in Rural Guinea-Bissau (EPICS): A Cluster-Randomised Controlled Trial. *Lancet Global Health* 4 (5): 328–35.

Brenner JL, Kabakyenga J, Kyomuhangi T, Wotton KA, Pim C, Ntaro M, Bagenda FN, Gad NR, Godel J, Kayizzi J. 2011. Can volunteer community health workers decrease child morbidity and mortality in southwestern Uganda? An impact evaluation. *PLoS One*, 6(12): e27997.

Darmstadt GL, Choi Y, Arifeen SE, BariS, Rahman SM, Mannan I, Seraji HR, et al. 2010. Evaluation of a Cluster-Randomized Controlled Trial of a Package of Community-Based Maternal and Newborn Interventions in Mirzapur, Bangladesh. *PLOS One*: e9696.

Deserranno E, Nansamba A, and N Qian, 2020. Aid Crowd-Out: The Effect of NGOs on Government-Provided Public Services. *Mimeo*.

Kalyango JN, Rutebemberwa E, Alfven T, Ssali S, Peterson S, Karamagi C. 2012. Performance of community health workers under integrated community case management of childhood illnesses in eastern Uganda. *Malar J*. 11(1):282.

Kayemba Nalwadda C, Guwatudde D, Waiswa P, Kiguli J, Namazzi G, Namutumba S, Tomson G, Peterson S. 2013. Community health workers–a resource for identification and referral of sick newborns in rural Uganda. *Tropical Med Int Health* 18(7):898–906.

Kidane G, and RH Morrow. 2000. Teaching Mothers to Provide Home Treatment of Malaria in Tigray, Ethiopia: A Randomised Trial. *Lancet* 356 (9229): 550–55.

Kimbugwe G, Mshilla M, Oluka D, et al. 2014. Challenges Faced by Village Health Teams (VHTs) in Amuru, Gulu and Pader Districts in Northern Uganda. *Open J Prev Med* 4:740–50.

Kirkwood BR, Manu A, ten Asbroek AHA, Soremekun S, Weobong B, Gyan T, Danso S, et al. 2013. Effect of the Newhints Home-Visits Intervention on Neonatal Mortality Rate and Care Practices in Ghana: A Cluster Randomised Controlled Trial. *Lancet* 381 (9884): 2184–92.

Kumar V, Mohanty S, Kumar A, Misra RP, Santosham M, Awasthi S, Baqui AH, et al. 2008. Effect of Community-Based Behaviour Change Management on Neonatal Mortality in Shivgarh, Uttar Pradesh, India: A Cluster-Randomised Controlled Trial. *Lancet* 372 (9655): 1151–62.

Lee, D. S. 2009. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economic Studies*, 76: 1071–1102.

Mays DC, O'Neil EJ, Mworozi EA, et al. 2017. Supporting and retaining Village Health Teams: an assessment of a community health worker program in two Ugandan districts. *Int J Equity Health*, 16:129.

Ministry of Health Uganda. 2015. *National Village Health Teams (VHT) assessment in Uganda*. Kampala: Ministry of Health Uganda.

Singh P, and JD Sachs. 2013. 1 million community health workers in sub-Saharan Africa by 2015. *Lancet*, 382: 363-365

Sloan NL, Salahuddin A, Satindra NM, Nuzhat C, Mushtaque C, Ubaider R, and B Winikoff . 2008. Community-Based Kangaroo Mother Care to Prevent Neonatal and Infant Mortality: A Randomized, Controlled Cluster Trial. *Pediatrics* 121 (5): Article 3.

Turinawe EB, Rwemisisi JT, Musinguzi LK, de Groot M, Muhangi D, de Vries DH, Mafigiri DK, and R Pool. 2015. Selection and performance of village health teams (VHTs) in Uganda: lessons from the natural helper model of health promotion. *Hum Resour Health*, 13:73

WHO, 2018. *The 2018 update, Global Health Workforce Statistics*, World Health Organization, Geneva. Available at: https://www.who.int/hrh/statistics/hwfstats/en/

WHO, 2018b. *WHO guideline on health policy and system support to optimize community health worker programmes*, World Health Organization, Geneva. Available at: https://www.who.int/hrh/statistics/hwfstats/en/

WHO, 2019. *Children: reducing mortality.* Online report updated on 19 September 2019, available at: https://www.who.int/news-room/fact-sheets/detail/children-reducing-mortality

World Bank, 2019. *World Bank DataBank* - Mortality rate, under-5 (per 1,000 live births) in Uganda. Available at: https://data.worldbank.org/indicator/SH.DYN.MORT?locations=UG

5. Appendices

Appendix A – Power Computation

The sample size was designed to detect a reduction in under-5 mortality, expressed as number of under-5 deaths per 1,000 child-years of exposure to the risk of death under the age of 5. We used data from the control group in the proof-of-concept study conducted by the research team in the same setting (Björkman Nyqvist et al, 2019) to estimate the parameters needed for the computation (Table A.1).

The outcomes are expressed as rates. The number of clusters required per arm (c) is

$$c = \left(t_{\alpha/2} + t_{\beta}\right)^2 \left(\frac{\frac{(\lambda_C + \lambda_T)}{y} + k^2(\lambda_C^2 + \lambda_T^2)}{(\lambda_C - \lambda_T)^2} \right)$$

where k is the common value of the between-cluster coefficient of variation, λ_C and λ_T are the mortality rates in the control and treatment group, respectively, and y is the number of observed child-years per cluster.

The between-cluster coefficient of variation k can be estimated as

$$\hat{k} = \frac{\hat{\sigma}}{r}$$

where r is the overall mortality rate computed from all clusters and

$$\hat{\sigma}^2 = s^2 - \frac{r}{\bar{y}_H}$$

with s^2 being the empirical variance of the observed rates and $\bar{y}_H$ being the harmonic mean of the y_j. Using the data from Björkman Nyqvist et al (2019) we estimate $\hat{k}$ to range between 0.30 (for neonatal mortality) and 0.43 (for under-5 mortality). Although we expected the new study design to make the clusters more similar, we considered a conservative $\hat{k}$=0.43 in our power computation.

Table A.2 reports the details of the computation. Using the inputs from Björkman Nyqvist et al, (2019), transforming them to fit the new 48-months trial period, and assuming 16% attrition, a sample size of 25 households per cluster at baseline and 250 clusters per arm (500 in total) will have 80% power to detect a 20% reduction in under-5 mortality at the 0.05 significant level. It also has 80% power to detect a 21% reduction in infant mortality at the 0.05 significant level, and 80% power to detect a 25% reduction in neonatal mortality at the 0.05 significant level.

Table A.1 Inputs for power computation

Neonatal Deaths:	106
Infant (Under-1) Deaths:	160
Under-5 Deaths:	206
Live Births:	2978
Years of Exposure to risk of death Under-1:	3015
Years of Exposure to risk of death Under-5:	10731
# of months-HHs:	105421

Note: Data from the control areas in Björkman Nyqvist et al (2019).
The reference period is about 36 months (2011-2013).

Table A.2 Computation details

	U5	***U1***	***Neonatal***
alpha	0.05	0.05	0.05
beta (power)	0.8	0.8	0.8
t_a/2	1.96	1.96	1.96
t_b	0.842	0.842	0.842
Mortality C	0.019	0.053	0.036
Reduction mortality	**20%**	**21%**	**25%**
Mortality T	0.015	0.042	0.027
Study-Months	48	48	48
k	0.43	0.43	0.43
# of HHs at Baseline	**25**	**25**	**25**
Attrition	0.16	0.16	0.16
# of HHs at Endline	21	21	21
Total Years of Exposure / cluster	102.6	28.8	28.5
Clusters/arm	250	250	250
Total clusters	**500**	**500**	**500**

Appendix B – Drugs testing

This appendix provides the details of the procedure adopted for testing drug quality at baseline. The same procedure will be followed at endline.

As explained in the main text, the medicines were collected using a covert shopper methodology. Field officers were trained to act as typical locals in order to get unbiased service from drug shop providers. They entered each village on foot and asked a few village members (at least three) all places where they could purchase medicine within the village. The field officer then visited all locations mentioned by the village members. Once they entered the shop, they explained that they had a 3-year-old child with malaria and a 4-year-old child with pneumonia symptoms and asked for the appropriate treatment. Once outside the store and away from view from the drug shop owner, the field officer placed the drugs in an envelope with an ID number on it to allow identifying the sample. The field officer also completed a short survey that recorded: GPS location, name of drug shop and directions to reach it, drug name, cost, dosage, packaging, and any additional notes about the purchase. Drugs were subsequently sent to Kampala.

In Kampala the drugs were coded as antimalarials, antibiotics for pneumonia, or other types of drugs. All drugs were additionally coded as appropriate or inappropriate drugs for the condition, based on the lists of medically-appropriate antimalarials and antibiotics provided by the National Drug Register of Uganda (updated September 2015), Essential Medicines and Health Supplies List for Uganda (latest version: 2012), the Uganda National Clinical Guidelines for Management of Common Conditions (latest version: 2012), as well as input from child survival specialists and pediatricians. Only drugs on these lists were tested for drug veracity. Painkillers such as paracetamol were not included in the list and, thus, were not tested.

Drugs were quality tested using Raman spectroscopy with a TruScan handheld scanner (software version 1.3; manufactured by Ahura Scientific Inc.) An important advantage of Raman spectroscopy compared to laboratory methods is speed. Another important advantage is that compared to laboratory testing, which requires a fairly large set of pills to test, and thus would require multiple purchases or purchase of more than one dose of tablets, the TruScan method provides a quality indicator per tested tablet. Methods comparing Raman spectroscopy to traditional laboratory methods have found a high degree of consistency across methods, and the Raman method is therefore viewed as suitable when conducting field studies[23] (Bate et al. 2009[24]).

[23] Nine out of the ten largest pharmaceutical companies worldwide rely on Raman spectroscopy technology to authenticate inputs. Moreover, a growing number of national drug enforcement agencies, for example the National Agency for food and Drug Administration and Control in Nigeria (NAFDAC), use the TruScan Raman Spectrometer to test for counterfeit and substandard medicines.

[24] Bate, R., R. Tren, K. Hess, L. Mooney, and K. Porter, 2009, Pilot Study Comparing Technologies to test for Substandard Drugs in Field Settings, *African Journal of Pharmacy and Pharmacology*, 3(4):165-170.

A. Building a Reference Library

For each drug sample to be tested, a pharmaceutical industry expert purchased authentic drug samples from the official importer/ manufacturer licensed by the Ugandan National Drug Authority (NDA).

Chemical signatures were developed for all authentic drug samples[25]. The signatures were saved in a Reference Library. Ten trial tests were done on each authentic drug, and all had to Pass to confirm its robustness.

B. Testing Drug Samples

As tablets and capsules were always sold in batches, three different units from each purchased sample were tested for authenticity. In practice, to perform the test the individual drug sample is placed in the holding vessel of the TruScan machine and analyzed against its corresponding authentic sample chemical signature via spectrometry. The TruScan scanner illuminates the sample (pill) with a laser beam and measures the reflecting Raman spectra. The Raman spectra provides a fingerprint by which the molecule composition of the sample can be identified. The fingerprint is then tested against the authentic reference sample and if they are sufficiently similar, as given by a probabilistic algorithm, the sample passes the test and is considered authentic. The machine then gives the sample a *Pass* or *Fail* report and the corresponding P-value.

The manufacturing details, dosage, expiration date, brand name, ingredient names and potency were also recorded alongside the test results.

D. Baseline results

In total, at baseline we collected 1,435 drug samples from 585 drug shops across our study sample. We coded the drugs into the following categories according to active ingredients: antibiotics, ACT antimalarial medicine, non-ACT antimalarial medicine, asthma medicine, allergy medicine, painkillers, and vitamins[26]. Table B.1 details the total number of each type of drug collected.

[25] There was just one exception, Erythromycin, whose florescence does not permit signature building and therefore had to be excluded from the sample.

[26] Here is the list of drug categories by active ingredients:
Antibiotics: Amoxicillin Trihydrate, Ampicillin + Cloxacillin, Ampicillin Trihydrate, Azythromycin, Cephalexin, Chloramphenicol, Ciprofloxacin, Cloxacillin, Doxycycline, Erythromycin Stearate, Metronidazole, Nitrofurantoin, Phenoxymethylpenicillin, Phenoxymethylpenicillin potassium, Sulfamethoxazole + Trimethoprim, Tetracycline.
ACT antimalarial dugs: Artemether + Lumefantrine, Artesunate Amodiaquine.
Non-ACT antimalarial drugs: Chloroquine Phosphate, Quinine, Septrine, Artesunate.
Drugs for asthma: Aminophylline, Prednisolone, Salbutamol Sulphate.
Allergy drugs: Cetirizine Hydrochloride, Chlorphenamine, Chlorpheniramine Maleate, Dexamethasone.
Painkillers: Acetaminophen- Chlorpheniramine Maleate, Aspirin, Diclofenac, Ibuprofen, Paracetemol, Piroxicam.
Vitamins: Ascorbic Acid, Calcium Folinate, Calcium lactate, Folic acid, Magnesium, Multivitamin, Pyridoxine hydrochloride (B6), Vitamin A, Vitamin B.

Table B.1. Types of drugs supplied at drug shops, per illness described

	Malaria	Pneumonia
	(n = 563)	(n = 872)
Antibiotics	2.70%	59.10%
ACT antimalarial drugs	92.70%	0.90%
Non-ACT antimalarial drugs	1.10%	0.60%
Drugs for asthma	0.00%	20.20%
Allergy drugs	0.90%	10.20%
Painkillers	0.50%	3.30%
Vitamins	0.40%	1.30%
Other/Unknown	1.80%	4.50%

Of the 585 drug outlets where any treatment was purchased or received, 89.2% of vendors provided the covert shopper with at least one ACT antimalarial drug, the recommended treatment for malaria. In regards to pneumonia, only 39.8% of visited shops provided the covert shopper with amoxicillin, the recommended treatment for pneumonia (Kabra, Lodha et al. 2010), whereas 36.1% of visited shops provided our shopper with other kinds of antibiotics.

We aimed to test all collected drugs that were categorized as antibiotics (n = 530) and ACT antimalarial medicines (n = 530), as these are the drugs that CHP will sell in the treatment village. We were unable to test 106 samples for the following reasons: a) the TruScan could not create a chemical signature for the sample due to high fluorescence (n = 64); b) we could not locate a brand sample from the manufacturer to build the chemical signature reference (n = 18); c) there were fewer than 3 pills in the sample (n = 8); or d) other problems with testing (n = 16). In total, we excluded 15 ACT antimalarial samples from our analysis of malaria drugs and 91 antibiotic samples from our analysis of pneumonia drugs.

In total, we therefore tested 954 drug samples (both capsules and tablets) for quality. A drug sample was considered substandard if at least 2 of the 3 samples failed to sufficiently match the chemical signature of the reference sample from the manufacturer.

Overall, 8.7% of the tested drugs did not pass the TruScan test and are therefore to be considered substandard quality drugs. Table B.2 presents the results of the testing for drug veracity. Looking specifically at the drugs that will be introduced to the local market through the CHP in each village, we found 11.1% of ACT antimalarial drugs were substandard and 7.1% of amoxicillin samples were substandard.

Table B.2. Percent of substandard drugs from collected samples

	Overall (n=954)	ACT antimalarials (n = 515)	Antibiotics (n = 439)	Amoxicillin (n = 240)
Percent of substandard drugs	8.70%	11.10%	5.90%	7.10%

Appendix D shows that the sample was balanced across treatment and control groups as was the share of substandard drugs identified (Table D.8).

Appendix C –Variables of Interest

As described in the main text, in this project we will collect a rich set of variables that will allow us to investigate a range of secondary hypotheses. We report them here, as they will be referenced in the table below.

In relation to the first research question, we will investigate the following secondary hypotheses:

1.1) The program increased the chances that a household interacts with and benefits from services provided by the CHPs;
1.2) The program increased the overall amount and quality of health services received by households;
1.3) On top of the impact on child mortality, the program improved additional health outcomes, related to family planning, pregnancy, newborn and child health;
1.4) The program improved the basic health knowledge of the households;
1.5) The program improved the health behavior (both preventive and curative) of the households;

In relation to the second research question, we will investigate the following secondary hypotheses:

2.1) The program increased the (average) satisfaction, motivation, and confidence of the CHWs operating in the village;
2.2) The program lowered the overall turnover of CHWs in the village, increased the number of CHWs that work for multiple health organizations, and reduced the total amount of time CHWs dedicate specifically to the government VHT program in the village;
2.3) The program increased the (average) health knowledge of the CHWs operating in the village;
2.4) The program increased the amount of (self-reported) activities of the CHWs operating in the village;
2.5) The program impacted the supply of drugs in the community, by reducing the number of drug stores operating in the local markets, raising the quality of their service, and lowering the price of the drugs;

Below we report the full list of the variables we plan to investigate in our analysis, arranged by category. The first column reports a short description of the variable as well as the specific questions from the endline survey that we plan to use to generate it. The questions should be considered only indicative, as we are going to revise and edit the survey during piloting phase. The second column of the table reports the source used to generate the variable, while the third and last column indicates to which one of the secondary hypotheses listed above it refers.

Variable Description	Source	Hypothesis
1. Households interactions with CHWs		
a. Household interactions with CHWs in general and CHPs in particular - *HH visited by any CHWs/CHPs over the previous 30 days* - *HH received any health service from the CHWs/CHPs (health products / education / diagnosis / referral / maternal care / follow-up visit)* - *HH knows how to contact the CHWs/CHPs in the village*	HH survey	1.1
2. Health services		
a. Household received follow-up health visits by any health staff… - *…following health-related problems with children under-5 (malaria, diarrhea, pneumonia) to specifically find out about child's recovery* - *…during pregnancy to monitor pregnancy* - *…after delivery to check the mother and child health* o *If so, was the visit performed during the first week of life?*	HH survey	1.2
b. Household received referrals to a health facility due to health-related problems with children under-5, or pregnancy - *…distinguishing between source of the referral (CHP, VHT, other CHW…)*	HH survey	1.2
c. Pregnant woman received counselling and health recommendations… - *…on where to deliver* - *…on medicines to take (Folic Acid, Iron and/or Vitamins / Malaria Prophylaxis / Deworming medicine)* - *…on newborn feeding practices*	HH survey	1.2
3. Health Outcomes		
a. Anthropometric measures for children under-5 - *Height-for-age (expressed in z-scores)* - *Weight-for-height (expressed in z-scores)* - *MUAC-for-age (expressed in z-scores)*	Anthropometric survey	1.3
b. Malaria, diarrhea and pneumonia prevalence among children under-5 - *Child fell sick with malaria in the previous 3 months* - *Child fell sick with diarrhea in the previous 3 months* - *Child fell sick with pneumonia in the previous 3 months*	HH survey	1.3
c. Share of miscarriages and stillbirths during the study period	HH survey	1.3

d. Unmet need for family planning and unwanted pregnancies	HH survey	1.3
4. Health Knowledge		
a. Respondent knowledge concerning causes and treatment of malaria, diarrhea, and pneumonia - *Respondent believes mosquito bites are the only cause of malaria* - *Respondent believes one can make environmental changes to prevent malaria* - *Respondent knows Zinc can be used to treat diarrhea* - *Respondent knows diarrhea can be transmitted by drinking un boiled/untreated water?*	HH survey	1.4
b. Respondent knowledge concerning nutrition and breastfeeding practices - *Respondent knows about vitamins & added nutrients* - *Respondent knows colostrum is healthy* - *Share of correct answers on a short case study presented to the respondent, which compares health evolution of two children treated differently in terms of nutrition and breastfeeding.*	HH survey	1.4
c. Respondent knowledge concerning family planning - Respondent knows about family planning methods (share)	HH survey	1.4
5. Household Health Behavior		
a. Household standard prevention and treatment practices for diarrhea, malaria, and pneumonia - *Respondent washes hands with soap most of the time* - *HH treats malaria with ACT drugs* - *HH treats pneumonia with antibiotic* - *HH treats diarrhea with ORS and Zinc*	HH survey	1.5
b. Household food consumption habits - *Child has varied diet (based on number of different food categories consumed the previous date, obtained from a detailed food consumption section)*	HH survey	1.5
c. Ante-natal and post-natal care practices, including breast-feeding practices - *Women sought ANC at least 4 times* - *Woman fed newborn within 1hr of birth* - *Woman fed baby non-breastmilk fluids after 6 months*	HH survey	1.5

- *Woman during pregnancy took Folic Acid / Iron and/or Vitamins / Malaria Prophylaxis / Deworming medicine* - *Woman took Vitamin A and/or folic acid in first two months after delivery* - *Woman gave birth outside a health facility* - *Woman devised a birth plan*		
6. Community Health Workers knowledge and activity		
a. Level of satisfaction and confidence of health workers - *First principal component from on a set of questions related to satisfaction (e.g. "I am satisfied with the community thanks and recognition I receive for my work")* - *Self-reported level of confidence that the CHW provides correct advise and/or treatment services for the community* - *Revenues as health worker*	CHW survey	2.1
b. Level of CHW turnover (*village level variable*) - *Measure constructed using information provided by CHWs on when they joined the organization and whether / how often the organization changed workers* c. CHW activities across different organizations - CHW works (at least part time) for the VHT government program o Time spent working as VHT member, in a regular week - CHW works for more than one health organization o Total time spend working as CHW across all health organizations, in a regular week	CHW survey	2.2
d. Knowledge of health workers concerning malaria, diarrhea, and nutrition - *CHW believes mosquito bites are the only cause of malaria* - *CHW believes one can make environmental changes to prevent malaria* - *CHW believes combination ORS & Zinc can treat diarrhea* - CHW knows correct signs/symptoms of Pneumonia and Malaria - *Share of correct answers on a short case study presented to the respondent, which compares health evolution of two children treated differently in terms of nutrition and breastfeeding.* - *CHW knows which food contains more protein value than others* - *CHW knows danger signs during pregnancy*	CHW survey	2.3
e. Level of self-reported activity of the health workers - *Days worked as CHW if last 30 days*	CHW survey	2.4

- *Number of Activities in last 30 days (Pregnant women visited; Newborn babies visited; Children < 5 years visited; Family planning visits; People tested for malaria; People treated for malaria; Patients referred to health center)* - *Number of Health forums / health education campaigns held*		
7. Drugs availability		
a. Number of drug stores that opened (closed down) during the study period (*village level variable*)	Drug quality survey	2.5
b. Drug store provided the appropriate medicine to treat the disease (malaria and pneumonia)	Drug quality survey	2.5
c. Price of the drugs sold by the store	Drug quality survey	2.5

Appendix D – Balance Checks

This appendix reports 10 tables containing the full set of balance checks performed on the variables collected at baseline.

Table D. 1: Baseline Balance Checks – Households I

	Mean (SE)		Difference	Obs.
Variable	Control	Treatment	(p-value)	
Age	23.364	23.396	0.033	12,615
	(3.659)	(3.711)	(0.670)	
Able to read	0.625	0.633	0.008	12,612
	(0.484)	(0.482)	(0.539)	
Work in health sector	0.012	0.010	-0.002	12,612
	(0.107)	(0.098)	(0.349)	
# adult (18+) HH members	2.429	2.427	-0.003	12,615
	(1.041)	(1.045)	(0.882)	
# HH members of age 5-17	2.033	2.057	0.018	12,615
	(2.366)	(2.373)	(0.697)	
# children under-5	1.453	1.468	0.015	12,615
	(0.805)	(0.816)	(0.359)	
Currently pregnant	0.202	0.197	-0.005	12,615
	(0.401)	(0.397)	(0.467)	
# years in the village	6.866	6.814	-0.057	12,573
	(6.793)	(6.607)	(0.682)	
Malaria (only) caused by mosquito	0.091	0.093	0.002	12,605
	(0.287)	(0.290)	(0.800)	
Knows malaria prevention measures	0.744	0.758	0.014	11,337
	(0.436)	(0.428)	(0.150)	
Zinc to treat diarrhea	0.327	0.331	0.004	12,510
	(0.469)	(0.471)	(0.725)	
Diarrhea from untreated water	0.479	0.480	0.001	12,439
	(0.500)	(0.500)	(0.889)	
Initial breastmilk healthy	0.752	0.764	0.011	11,669
	(0.432)	(0.425)	(0.237)	
Knows food w/ added nutrients	0.476	0.470	-0.007	12,600
	(0.499)	(0.499)	(0.495)	
Ever used family planning	0.597	0.609	0.012	12,611
	(0.491)	(0.488)	(0.251)	
Currently using family planning	0.333	0.333	-0.000	11,634
	(0.471)	(0.471)	(0.971)	
Desired # months to next pregnancy	35.110	34.598	-0.518	9,623
	(21.367)	(21.427)	(0.343)	
Discuss plans for child w/ husband	0.541	0.540	-0.001	12,562
	(0.498)	(0.498)	(0.902)	

Notes: The table shows the mean and standard error of the mean (in parenthesis) for different variables. The sample consists of all households surveyed at baseline. The full sample includes 12,614 households. Columns 3 show the difference in means and the p-value of the test for means equality (H_0: =mean is equal across treatment and control). The test for equality takes into account district fixed effects and clustering at the village level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table D. 2: Baseline Balance Checks – Households II

	Mean (SE)		Difference	Obs.
Variable	Control	Treatment	(p-value)	
Health care provider visited HH, past 30d	0.179	0.174	-0.004	12,563
	(0.383)	(0.379)	(0.657)	
VHT visited HH, past 30d	0.089	0.093	0.005	12,563
	(0.285)	(0.291)	(0.490)	
Health care personnel provided any service	0.139	0.137	-0.002	12,563
	(0.346)	(0.344)	(0.836)	
Knows any CHW in the village	0.551	0.563	0.013	12,595
	(0.497)	(0.496)	(0.430)	
# health service centers within 1h	1.977	2.052	0.076	12,615
	(1.194)	(1.243)	(0.105)	
memeber of any voluntary org.	0.509	0.510	0.001	12,615
	(0.500)	(0.500)	(0.951)	
Asset Index (PCA)	-0.015	0.015	0.035	12,061
	(1.819)	(1.842)	(0.629)	
Wealth Index (PCA)	-0.021	0.021	0.047	11,837
	(1.914)	(1.940)	(0.548)	

Notes: The table shows the mean and standard error of the mean (in parenthesis) for different variables. The sample consists of all households surveyed at baseline. The full sample includes 12,614 households. Columns 3 show the difference in means and the p-value of the test for means equality (H_0=mean is equal across treatment and control). The test for equality takes into account district fixed effects and clustering at the village level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table D. 3: Baseline Balance Checks – Children U5

	Mean (SE)		Difference	Obs.
Variable	Control	Treatment	(p-value)	
Age in months	26.119	26.002	-0.122	18,396
	(16.807)	(16.596)	(0.590)	
Girl	0.494	0.499	0.005	18,395
	(0.500)	(0.500)	(0.499)	
Has treated bednet	0.090	0.084	-0.005	18,068
	(0.286)	(0.277)	(0.407)	
Slept under bednet yesterday	0.034	0.033	-0.001	18,306
	(0.181)	(0.179)	(0.818)	
Fell sick with malaria, past 3m	0.607	0.612	0.004	18,345
	(0.488)	(0.487)	(0.666)	
Treated for malaria	0.967	0.964	-0.003	11,162
	(0.178)	(0.186)	(0.406)	
Treated with ACT for malaria	0.806	0.815	0.007	11,175
	(0.395)	(0.389)	(0.403)	
Received visit from health care provider for malaria	0.042	0.042	0.000	10,760
	(0.200)	(0.200)	(0.923)	
Received referral for malaria	0.130	0.132	0.002	11,155
	(0.336)	(0.339)	(0.803)	
Fell sick with pneumonia, past 3m	0.383	0.390	0.007	18,358
	(0.486)	(0.488)	(0.510)	
Treated for pneumonia	0.882	0.887	0.004	7,080
	(0.322)	(0.316)	(0.656)	
Treated with antibiotics for pneumonia	0.480	0.510	0.032	7,099
	(0.500)	(0.500)	(0.029)**	
Received visit from health care provider for pneumonia	0.032	0.031	-0.000	6,256
	(0.177)	(0.173)	(0.931)	
Received referral for pneumonia	0.117	0.120	0.003	7,084
	(0.321)	(0.325)	(0.778)	
Fell sick with diarrhea, past 3m	0.328	0.346	0.017	18,359
	(0.469)	(0.476)	(0.046)**	
Treated for diarrhea	0.817	0.824	0.002	6,173
	(0.387)	(0.381)	(0.847)	
Treated with Zinc/ORS for diarrhea	0.428	0.447	0.007	6,181
	(0.495)	(0.497)	(0.649)	
Received visit from health care provider for diarrhea	0.029	0.042	0.012	5,055
	(0.167)	(0.202)	(0.069)*	
Received referral for diarrhea	0.110	0.105	-0.004	6,172
	(0.313)	(0.307)	(0.718)	
Illness prevented child from eating, past 24h	0.197	0.199	0.002	15,886
	(0.398)	(0.400)	(0.751)	

Notes: The table shows the mean and standard error of the mean (in parenthesis) for different variables. The sample consists of all children under-5 living in the households surveyed at baseline. The full sample includes 18,396 children living in 12,614 households. The variables are based on answers provided by the primary respondent to the household survey. Columns 3 show the difference in means and the p-value of the test for means equality (H_0:=mean is equal across treatment and control). The test for equality takes into account district fixed effects and clustering at the village level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table D. 4: Baseline Balance Checks – Pregnancies and Prenatal care

	Mean (SE)		Difference	Obs.
Variable	Control	Treatment	(p-value)	
Pregnancy in the HH in past 3y	0.916	0.917	0.001	12,614
	(0.277)	(0.276)	(0.878)	
Current pregnancy in the HH	0.216	0.208	-0.008	12,614
	(0.412)	(0.406)	(0.300)	
# months pregnant	5.299	5.316	0.022	2,525
	(2.345)	(2.421)	(0.819)	
Current pregnancy was not planned	0.529	0.521	-0.009	2,505
	(0.499)	(0.500)	(0.651)	
Seeking antenatal care	0.618	0.625	0.007	2,537
	(0.486)	(0.484)	(0.747)	
Referred to antenatal care	0.334	0.323	-0.009	2,530
	(0.472)	(0.468)	(0.662)	
Visited by health worker	0.035	0.034	-0.001	2,533
	(0.184)	(0.182)	(0.915)	
Recommended vitamins/iron/folic acid	0.424	0.444	0.018	2,521
	(0.494)	(0.497)	(0.383)	
Took Iron during pregnancy	0.353	0.360	0.008	2,538
	(0.478)	(0.480)	(0.683)	
Took Vitamins during pregnancy	0.224	0.235	0.013	2,538
	(0.417)	(0.424)	(0.490)	
Took Folic Acid during pregnancy	0.356	0.389	0.033	2,538
	(0.479)	(0.488)	(0.099)*	
Recommended malaria prophylaxis	0.371	0.394	0.022	2,517
	(0.483)	(0.489)	(0.298)	
Took something against malaria	0.418	0.434	0.016	2,515
	(0.493)	(0.496)	(0.410)	
Recommended drug against worms	0.222	0.217	-0.005	2,510
	(0.416)	(0.412)	(0.793)	
Took something against worms	0.237	0.248	0.012	2,511
	(0.425)	(0.432)	(0.477)	
Changed diet during pregnancy	0.601	0.586	-0.014	2,529
	(0.490)	(0.493)	(0.478)	
Fell sick during pregnancy (malaria)	0.431	0.458	0.026	2,538
	(0.495)	(0.498)	(0.209)	
Fell sick during pregnancy (not malaria)	0.357	0.330	-0.025	2,538
	(0.479)	(0.471)	(0.218)	

Notes: The table shows the mean and standard error of the mean (in parenthesis) for different variables. Most variables in the table focus on the restricted sample of women living in the sample household that were pregnant at the time of the survey. The full sample includes 2,538 pregnant women living in 12,614 households. Columns 3 show the difference in means and the p-value of the test for means equality (H_0:=mean is equal across treatment and control). The test for equality takes into account district fixed effects and clustering at the village level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table D. 5: Baseline Balance Checks – Deliveries and postnatal care

	Mean (SE)		Difference	Obs.
Variable	Control	Treatment	(p-value)	
Delivery in the HH in past 3y	0.819	0.826	0.007	12,614
	(0.385)	(0.379)	(0.371)	
Someone in HH had a stillbirth in past 3y	0.027	0.020	-0.007	12,614
	(0.162)	(0.140)	(0.009)***	
Someone in HH had a miscarriage in past 3y	0.088	0.087	-0.001	12,614
	(0.283)	(0.281)	(0.830)	
Gave birth in a health facility	0.805	0.787	-0.019	10,867
	(0.396)	(0.409)	(0.148)	
Recommended to deliver with medical personnel	0.699	0.703	0.003	10,823
	(0.459)	(0.457)	(0.804)	
Had prepared a birth plan	0.548	0.551	0.002	10,876
	(0.498)	(0.497)	(0.891)	
Saught antenatal care	0.982	0.982	-0.000	10,868
	(0.131)	(0.132)	(0.930)	
Took Iron during pregnancy	0.666	0.678	0.012	10,857
	(0.472)	(0.467)	(0.352)	
Took Vitamins during pregnancy	0.749	0.755	0.005	10,857
	(0.433)	(0.430)	(0.626)	
Took Folic Acid during pregnancy	0.485	0.480	-0.008	10,857
	(0.500)	(0.500)	(0.579)	
Took something against malaria	0.803	0.808	0.004	10,733
	(0.398)	(0.394)	(0.653)	
Took something against worms	0.616	0.601	-0.016	10,630
	(0.486)	(0.490)	(0.136)	
Fell sick during last pregnancy (malaria)	0.530	0.533	0.002	10,850
	(0.499)	(0.499)	(0.869)	
Fell sick during last pregnancy (other)	0.329	0.329	0.001	10,850
	(0.470)	(0.470)	(0.950)	
Took vitamin/folic acid after delivery	0.355	0.352	-0.002	10,864
	(0.478)	(0.478)	(0.844)	
Fed colostrum	0.872	0.880	0.009	10,407
	(0.335)	(0.325)	(0.248)	
Gave non-milk liquids in first 3d	0.274	0.267	-0.005	10,529
	(0.446)	(0.443)	(0.600)	
Gave food/liquid in first 6m	0.621	0.627	0.007	4,633
	(0.485)	(0.484)	(0.678)	
Gave solid food in first 6m	0.395	0.405	0.008	10,575
	(0.489)	(0.491)	(0.468)	
Visited by health worker in first week	0.076	0.080	0.004	10,561
	(0.265)	(0.272)	(0.539)	
Baby weight (kg)	3.233	3.246	0.016	7,923
	(0.849)	(0.808)	(0.429)	

Notes: The table shows the mean and standard error of the mean (in parenthesis) for different variables. Most variables in the table focus on the restricted sample of women living in the sample household that delivered a baby within the three years before the survey. The full sample includes 10,876 women living in 12,614 households. Columns 3 show the difference in means and the p-value of the test for means equality (H_0=mean is equal across treatment and control). The test for equality takes into account district fixed effects and clustering at the village level. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Table D. 6: Baseline Balance Checks – Mortality (previous 3 years)

	Mean (SE)		Difference	Obs.
Variable	Control	Treatment	(p-value)	
HH experienced child death in past 3y	0.075	0.068	-0.007	12,564
	(0.263)	(0.251)	(0.112)	
# under-5 deaths	1.688	1.580	-0.108	500
	(1.373)	(1.421)	(0.384)	
# infant deaths	1.324	1.236	-0.088	500
	(1.250)	(1.204)	(0.423)	
# neonatal deaths	1.060	0.916	-0.144	500
	(1.172)	(1.040)	(0.147)	
U5 mortality rate per 1,000 births	64.370	59.548	-4.823	500
	(55.201)	(55.514)	(0.329)	
Infant mortality rate per 1,000 births	45.019	42.298	-2.721	500
	(41.609)	(39.651)	(0.453)	
Neonatal mortality rate per 1,000 births	35.613	31.051	-4.562	500
	(39.259)	(35.292)	(0.170)	
U5 mortality rate per 1,000 yrs of exposure	23.412	21.511	-1.901	500
	(20.585)	(19.816)	(0.289)	
Infant mortality rate per 1,000 yrs of exposure	51.884	47.481	-4.403	500
	(51.418)	(47.227)	(0.317)	

Notes: The table shows the mean and standard error of the mean (in parenthesis) for different mortality measures over the three years preceding the survey, computed at the village level (with the exception of the first variable, defined at the household level). Columns 3 show the difference in means and the p-value from testing whether the mean is equal (H_0:=mean is equal across the two groups). The test for equality takes into account district fixed effects and clustering at the village level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table D. 7: Baseline Balance Checks – Anthropometric measures for children under-5

	Mean (SE)		Difference	Obs.
Variable	Control	Treatment	(p-value)	
Age in months	26.176	26.006	-0.176	17,803
	(16.758)	(16.542)	(0.431)	
Girl	0.486	0.492	0.005	17,803
	(0.500)	(0.500)	(0.462)	
Height-for-age (z-score)	-0.521	-0.496	0.022	17,803
	(1.505)	(1.505)	(0.544)	
Hfa z-score < -2	0.146	0.137	-0.008	17,803
	(0.353)	(0.344)	(0.276)	
Weight-for-height (z-score)	-0.553	-0.578	-0.023	17,298
	(1.167)	(1.139)	(0.370)	
Wfh z-score < -2	0.095	0.096	0.000	17,298
	(0.293)	(0.294)	(0.927)	
Weight-for-age (z-score)	-0.663	-0.669	-0.005	17,341
	(1.058)	(1.081)	(0.811)	
Wfa z-score < -2	0.091	0.099	0.008	17,341
	(0.287)	(0.298)	(0.102)	
MUAC	14.697	14.740	0.043	17,796
	(1.424)	(1.443)	(0.162)	
MUAC-for-age (z-score)	-0.285	-0.238	0.046	16,819
	(0.950)	(0.952)	(0.051)*	

Notes: The table shows the mean and standard error of the mean (in parenthesis) for different variables. The sample consists of all children under-5 whose anthropometric measures were captured at baseline. The full sample includes 17,803 children. Columns 3 show the difference in means and the p-value from testing whether the mean is equal (H_0:=mean is equal across the two groups). The test for equality takes into account district fixed effects and clustering at the village level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table D. 8: Baseline Balance Checks – Drugs

	Mean (SE)		Difference	Obs.
Variable	Control	Treatment	(p-value)	
Village has at least one drug shop	0.972	0.972	0.000	500
	(0.165)	(0.165)	(1.000)	
# drug shops in the village	1.540	1.472	-0.068	500
	(1.412)	(1.216)	(0.548)	
Village has at least one open drug shop	0.768	0.761	-0.006	500
	(0.390)	(0.391)	(0.851)	
# open drug shops in the village	1.340	1.252	-0.088	500
	(1.380)	(1.032)	(0.404)	
Malaria drug purchased	0.836	0.812	-0.028	648
	(0.371)	(0.392)	(0.356)	
Price paid for malaria drug	2246.072	2137.795	-78.984	534
	(1044.624)	(973.274)	(0.330)	
Pneumonia drug purchased	0.737	0.751	0.019	648
	(0.441)	(0.433)	(0.606)	
Price paid for pneumonia drug	2452.834	2395.536	-35.979	482
	(1452.462)	(1386.247)	(0.799)	
Store gave ACT to treat malaria	0.914	0.863	-0.058	587
	(0.281)	(0.345)	(0.028)**	
Store gave antibiotic to treat pneumonia	0.746	0.777	0.043	585
	(0.436)	(0.417)	(0.217)	
Substandard drug	0.096	0.077	-0.021	954
	(0.295)	(0.267)	(0.239)	
Substandard ACT	0.116	0.104	-0.016	515
	(0.321)	(0.306)	(0.581)	
Substandard antibiotic	0.071	0.047	-0.026	439
	(0.258)	(0.211)	(0.266)	

Notes: The table shows the mean and standard error of the mean (in parenthesis) for different variables. The sample consists of all drugs purchased at baseline. The full sample includes 500 villages, 648 open drug shops, and 954 tested drug samples. Columns 3 show the difference in means and the p-value from testing whether the mean is equal (H_0:=mean is equal across the two groups). When testing equality we take into account district fixed effects (15 districts) and clustering at the village level (500 villages). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table D. 9: Baseline Balance Checks – CHWs I

	Mean (SE)		Difference	Obs.
Variable	Control	Treatment	(p-value)	
VHT member	0.933	0.960	0.028	1,359
	(0.249)	(0.197)	(0.034)**	
Age	41.175	42.049	0.978	1,361
	(10.963)	(10.748)	(0.093)*	
Female	0.460	0.485	0.023	1,363
	(0.499)	(0.500)	(0.322)	
Years lived in the village	28.279	29.182	0.970	1,359
	(13.760)	(14.695)	(0.201)	
Years as CHW	7.446	7.654	0.196	1,112
	(5.670)	(5.842)	(0.579)	
Worked as CHW before	0.249	0.256	0.005	1,363
	(0.433)	(0.437)	(0.838)	
# VHTs in the village	6.616	3.395	-3.240	1,273
	(79.448)	(1.630)	(0.281)	
Last training < 6	0.584	0.571	-0.022	1,283
	(0.493)	(0.495)	(0.466)	
Last meeting w/ supervisor < 3	0.656	0.697	0.033	1,089
	(0.475)	(0.460)	(0.267)	
Satisfaction level (0-10)	6.763	7.160	0.345	1,357
	(2.268)	(2.332)	(0.012)**	
Confidence level (0-10)	7.735	8.216	0.451	1,361
	(2.010)	(1.813)	(0.000)***	
Earn anything as CHW	0.369	0.445	0.075	1,363
	(0.483)	(0.497)	(0.005)***	
Earning as CHW (000)	13.087	15.173	2.035	1,359
	(41.299)	(33.644)	(0.314)	
Use cell phone as CHW	0.562	0.554	-0.010	1,362
	(0.496)	(0.497)	(0.724)	
Any product or medicine provided	0.530	0.583	0.048	1,363
	(0.499)	(0.493)	(0.119)	
Malaria (only) caused by mosquito	0.314	0.322	0.006	1,355
	(0.464)	(0.468)	(0.824)	
Malaria treated with ACT	0.860	0.873	0.018	1,362
	(0.348)	(0.333)	(0.361)	
Loose stool symptom of diarrhea	0.767	0.795	0.029	1,362
	(0.423)	(0.404)	(0.228)	
Diarrhea treated with ORS	0.145	0.160	0.010	1,359
	(0.352)	(0.367)	(0.624)	
Received any supply last 3m	0.274	0.269	-0.014	1,363
	(0.446)	(0.444)	(0.554)	

Notes: The table shows the mean and standard error of the mean (in parenthesis) for different variables. The sample consists of all Community Health Workers (CHWs) surveyed at baseline. The full sample includes 1,363 CHWs. Columns 3 show the difference in means and the p-value from testing whether the mean is equal (H_0:=mean is equal across the two groups). When testing equality we take into account district fixed effects (15 districts) and clustering at the village level (500 villages). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table D. 10: Baseline Balance Checks – CHWs II

	Mean (SE)		Difference	Obs.
Variable	Control	Treatment	(p-value)	
# days worked last week	1.547	1.747	0.164	1,361
	(1.850)	(1.838)	(0.103)	
# days visited HHs last week	1.314	1.483	0.144	1,361
	(1.594)	(1.591)	(0.101)	
# HHs visited yesterday	2.253	2.262	-0.005	1,363
	(6.372)	(6.855)	(0.989)	
# days worked last month	5.583	6.069	0.362	1,358
	(6.779)	(6.879)	(0.341)	
Any visit last month	0.806	0.823	0.013	1,361
	(0.396)	(0.382)	(0.576)	
# visits last month	14.412	17.538	2.414	1,361
	(32.790)	(44.129)	(0.286)	
# people treated for malaria 1ast month	1.756	2.041	0.205	1,335
	(7.716)	(9.840)	(0.620)	
# referrals to health centers last month	2.287	3.267	0.885	1,348
	(4.416)	(12.722)	(0.176)	
# pregnant women visited last month	1.756	2.012	0.229	1,352
	(2.605)	(2.935)	(0.133)	
# newborn visited last month	1.205	1.358	0.132	1,354
	(1.828)	(2.330)	(0.233)	
# health meetings last month	0.760	0.624	-0.155	1,361
	(2.306)	(1.231)	(0.141)	
# child died last month	0.170	0.235	0.057	1,361
	(0.530)	(0.616)	(0.124)	

Notes: The table shows the mean and standard error of the mean (in parenthesis) for different variables. The sample consists of all Community Health Workers (CHWs) surveyed at baseline. The full sample includes 1,363 CHWs. Columns 3 show the difference in means and the p-value from testing whether the mean is equal (H_0:=mean is equal across the two groups). When testing equality we take into account district fixed effects (15 districts) and clustering at the village level (500 villages). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

6. Administrative information

Funding: This work was supported by the Children's Investment Fund Foundation, London, UK [grant name "Evaluation of Living Goods/BRAC entrepreneurial CHW model in Uganda - Phase II", July 2015].

Institutional Review Board (ethics approval): The trial was approved by the Uganda National Council for Science and Technology (UNCST) (SS3938), the Innovations for Poverty Action IRB (13774), and by the Mildmay Uganda Research and Ethics Committee (MUREC) in Uganda (0109-2015). The trial was registered in the Pan African Clinical Trials Registry (PACTR201609001398349) and in the American Economic Association's registry for randomized controlled trials (AEARCTR-0002392).

Declaration of interest: none

Acknowledgements: We thank Charles Ntale and the staff at IPA-Uganda, specifically Carin Mirowitz, Paola Elice, Laura Schmucker, Eva Noble, and Agne Pupienyte, for precious help during different phases of the study preparation. We thank Living Goods and the BRAC-Uganda staff members, especially Molly Christiansen, Chuck Slaughter, Justin Loiseau, Sharmin Sharif, Munshi Sulaiman, and Inshallah Franco for insightful discussions over the years about the CHP program, and Mihretab Salasibew and his colleagues at the Children Investment Fund Foundation for continuous support throughout the study.