IDB WORKING PAPER SERIES Nº IDB-WP-974

In-kind Incentives and Health Worker Performance: Experimental Evidence from El Salvador

Pedro Bernal
Sebastian Martínez

Inter-American Development Bank
Social Protection and Health Division

December 2018

IDB
Inter-American
Development Bank

In-kind Incentives and Health Worker Performance: Experimental Evidence from El Salvador

Pedro Bernal
Sebastian Martínez

December 2018

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library
Bernal, Pedro.
In-kind incentives and health worker performance: Experimental evidence from El Salvador / Pedro Bernal, Sebastian Martínez.
p. cm. — (IDB Working Paper Series ; 974)
Includes bibliographic references.
1. Medical personnel-Administration-El Salvador. 2. Medical care-Evaluation-El Salvador. 3. Medical care surveys-El Salvador. 4. Public health-El Salvador. 5. Incentives in industry-El Salvador. I. Martínez, Sebastián. II. Inter-American Development Bank. Social Protection and Health Division. III. Inter-American Development Bank. Strategic Development Effectiveness Division. IV. Title. V. Series.
IDB-WP-974

http://www.iadb.org

scl-sph@iadb.org

https://www.iadb.org/es/proteccionsocial

In-Kind Incentives and Health Worker Performance: Experimental Evidence from El Salvador

Pedro Bernal
Sebastian Martinez*

Abstract

Maintaining high standards of care from doctors, nurses and other health professionals is of critical importance for an effective and efficient health system. Yet deficient levels of health worker performance, including low effort, absenteeism, and lack of compliance with clinical guidelines, have been documented across numerous countries and contexts. In response, various pay-for-performance interventions that reward providers based on measures of quality of care and health outcomes have been tested, with mixed results. This study experimentally evaluates the effects of team in-kind incentives on health worker performance in El Salvador. Thirty-eight out of 75 community health teams were randomly assigned to receive in-kind incentives linked to performance over a 12-month period. All 75 teams received monitoring, performance feedback and recognition for their achievements, allowing us to isolate the impacts of the incentive. While both treatment and control groups exhibit improvements in performance measures over time, the in-kind incentives generated significant improvements in community outreach, quality of care, timeliness of care, and utilization of services after 12 months. Contrary to conventional knowledge, gains are largest for health teams at the bottom and top of the baseline performance distribution. These results suggest that even small in-kind incentives can be a powerful tool to improve health worker performance and may be a viable alternative to monetary incentives in certain contexts.

Keywords: Pay for Performance, Performance Incentives, In-Kind, Team Incentives, Health Services, El Salvador
JEL codes: I12, I15, I18, C93

* *Acknowledgements:* We thank the Ministry of Health in El Salvador and the Salud Mesoamerica Initiative for assistance and funding for this study. Ferdinando Regalia, Emma Iriarte, Luis Tejerina and Maria Deni Sanchez were instrumental for the realization of this study. Álvaro González, Diego Ríos, Mauricio Dinarte, Ariel Habed, Jennifer Nelson and Jose Carlos Gutiérrez provided valuable inputs for the design of the incentive scheme. We thank Álvaro Canales, and Tamara Arnold who helped in the design of the sampling scheme and data collection instruments and Cinzia Innocenti who oversaw data collection. We thank Elizabeth Bastias-Butler and Christina Memmott for outstanding research assistance. We thank Jessica Goldberg and Nicolás Ajzenman for valuable comments on earlier drafts. We received comments from seminar participants at the Inter-American Development Bank and the LACEA Impact Evaluation Network. All opinions in this paper are those of the authors and do not necessarily represent the views of the Government of El Salvador, or the Inter-American Development Bank, its Executive Directors, or the governments they represent.
Authors: Bernal (pbernallara@iadb.org, corresponding author), Martinez (smartinez@iadb.org). Inter-American Development Bank, 1300 New York Avenue, NW, Washington DC 20577.

I. Introduction

Deficient health provider performance has been documented in multiple middle and low-income countries, including failure to meet coverage targets, absenteeism from post of duty, non-compliance with clinical guidelines and even malpractice (Berendes et al., 2011, Das and Hammer, 2004, 2007; Das et al., 2013, Banerjee et al., 2004; Barber et al, 2007; Planas et al. 2015). While the causes of poor performance are complex, provider effort is thought to play an important role as often providers fail to do what is within their knowledge and means. This is often illustrated with lack-of-compliance with clinical guidelines even when providers have adequate knowledge and resources in multiple settings (Das, Hammer and Leonard, 2008). Effort is particularly important for tasks such as community outreach and patient treatment compliance, which are harder to monitor and are outside the direct control of providers. As a response, public and private health administrators have implemented incentive schemes that reward providers for accomplishing determined health outputs or outcomes (Miller and Babiarz, 2014; Basinga et al., 2011; Bonfrer et al., 2013; Gertler et al., 2014).

We study the effects of in-kind group incentives for community health teams in El Salvador's public sector. Teams were composed of doctors, nurses and community health workers providing primary care services in catchment areas of approximately 3000 individuals in rural areas and 9000 individuals in urban areas. The incentive model was based on eleven maternal and child health targets, measured every six months using a combination of medical record reviews and household surveys. Teams were awarded points for each target met, and points were redeemable for in-kind incentives such as computers, microwaves, air-conditioners and other workplace assets to be shared by the team members.

We identify the incentive effects using a randomized controlled trial. Thirty-eight out of 75 community health teams were randomly assigned to receive incentives during the first twelve months of the pilot. The control group became eligible for incentives thereafter. All health teams received the same performance feedback, supervision, and recognition for their achievements, allowing us to isolate the effect of the in-kind incentives. Furthermore, the experiment was conducted with existing health teams already operating within the organizational structure of the Ministry of Health, allowing us to overcome potential selection of health workers into the incentive model or selection of individuals into teams.

During the year-long experiment, in-kind incentives led to improved performance on a variety of dimensions, including community outreach, quality of care, timeliness of care, and utilization of services. Average standardized effects ranged from 0.10 to 0.17 standard deviations, with the largest effects on community outreach (0.17 standard deviations) and quality of care (0.14 standard deviations). Substantial gains were achieved by health teams with the lowest baseline performance, whose gains centered on activities requiring less team member coordination such as quality of care which is influenced mostly by physicians, and community outreach, which is mostly the responsibility of community health workers. Even larger gains were observed for teams with the highest baseline performance, including for more complex outcomes that require changing patient behaviors such as timeliness of prenatal and postnatal care and utilization of health services including family planning, deworming of children and MMR vaccination. Teams with intermediate performance at baseline that required little effort to meet the performance goals exhibit no response to the incentive scheme. Our results are robust to several specifications and we find no evidence of them being driven by changes in reporting or by shifting effort away from non-contracted outcomes. Finally, 18 months after the start of the experiment control teams were

incorporated into the incentive scheme and show improvements in performance, however treatment teams continue performing significantly higher on several domains.

Our contribution to the pay-for-performance in health literature is fourfold.[2] First, pay-for-performance interventions typically bundle incentives with audits, feedback and information, and even public recognition thus obscuring the effect of incentives alone (Mendelson et al, 2017; Ivers et al, 2012). We conduct one of the first experiments to isolate the effects of incentives from information and public recognition[3]. Second, we contribute to the literature on team incentives and show that they can be effective particularly in settings were teams are small, members have clearly defined roles and their activities are interdependent. Previous evidence on the effect of team-level incentives has been concentrated primarily on private firms (Friebel et al, 2017; Boning, Ichniowski, and Shaw, 2007), and less is known for the public and health sectors.

Third, we show that modest in-kind incentives can produce substantial effects on outreach, quality and coverage of public health services. Compared to monetary incentives, in-kind incentives may be less prone to controversy and expectations of permanency as part of remuneration packages. As such, they may be easier to negotiate in settings of strong unions and with administrators hesitant to assume permanent financial commitments. While traditional economic theory suggests that in-kind incentives are no better than an equivalent monetary amount[4] (Waldfogel, 1996), findings from the behavioral economics and social psychology literature suggest that in some settings in-kind incentives could elicit greater effort (Kube, Maréchal and Puppe, 2012; Heyman and Ariely, 2004) and be less prone to backfiring by crowding out intrinsic motivation (Lacetera and Macis, 2010).[5] A common explanation for this finding is that monetary incentives might be perceived as a payment in a market interaction, whereas in-kind incentives might be perceived as a gift-giving act that elicits loyalty and a response independent of the gift value (Heyman and Ariely, 2004; Bareket-Bojmel, Hochman and Ariely, 2017). Moreover, recipients of team in-kind incentives do not only derive utility from the material value of the goods (i.e using a microwave they earned), but also symbolic and social utility from it, as it can be a salient reminder of the team's performance from which team-members can derive pride (Lactera, Macis, and Slonim, 2012; Scott, 2013). In our knowledge, there is no previous evidence of the effect of team-based in-kind incentives in a field experiment.[6]

Finally, we show heterogeneous responses to a non-linear incentive scheme with targets on performance. While the literature discusses some common concerns of non-linear schemes, such a potential failure to motivate those further away from performance targets (Miller and Babiarz, 2014), to our knowledge there is little empirical evidence on this. Our results suggest that the non-linear scheme implemented in El Salvador was effective at improving performance for those well below the incentive threshold as well as for those substantially above this threshold at baseline.

---

[2] Pay-for-performance experiments with credible attribution include Basinga et al, 2011; Bonfrer et al, 2013; de Walque et al, 2015; Gertler et al, 2014; Shen et al, 2017; Celhay et al, 2018. The Health Results Innovation Trust Fund is undertaking impact evaluations of performance-based financing in health interventions in over 20 countries, primarily concentrated in Sub-Saharan Africa (http://www.rbfhealth.org/).

[3] Our findings also relate to the literature of awards, which are usually bundled with some form of material benefit or prize (Kosfeld and Neckermann, 2011) and to the literature on social recognition for public workers (Gauri et al 2018). In our setting we are able to identify the added-value of providing in-kind incentives for teams in addition to social recognition.

[4] This is more so if incentives are for teams, as different team members might have different preferences for specific in-kind incentives while monetary incentives are fungible and could be easily divided among team members.

[5] The backfiring of monetary incentives might be more relevant in pro-social tasks liking blood donations (Lacetera and Macis, 2010). However, in some type of government work, like that in public health, tasks might be perceived similarly as pro-social if they have wider social benefits beyond those enjoyed by employers and employees (Ashraf, Bandiera and Jack, 2014).

[6] Ashraf, Bandiera and Jack (2014) estimate the effect of pure non-monetary incentives (a star stamp for each sale) and compare it to that of financial rewards based on sales, but they do not test the effect of an in-kind incentive.

The scheme was only ineffective at motivating those at or slightly above the performance threshold who could have achieved the reward with a small marginal effort.

The paper is organized as follows. The next section describes El Salvador's health system and the Mesoamerica Health Initiative under which the pay for performance experiment took place. The third section describes the incentive scheme, how performance was measured, and feedback was provided. The fourth section discusses data sources and the fifth section presents our identification strategy. Results are presented in section six. In section seven we discuss robustness checks and potential alternative explanations for our results. Section eight discusses the results and concludes.

II. Background

A. The Health System in El Salvador

In 2010, El Salvador's Ministry of Health (MoH) started a reform of its public health system with a primary objective of improving access to primary care services in the country.[7] The reform was centered on the creation of government funded and managed community health teams.[8] These teams were designed to serve the primary care needs of the population within a pre-defined catchment area of approximately 3000 individuals in rural areas and 9000 individuals in urban areas. Each rural team was composed of seven members: a physician, a professional nurse, an auxiliary nurse, three community health workers (CHWs, one for every 200 families) and a multi-purpose worker. Urban teams had the same composition but increased the number of CHWs to six (one for every 300 families). Teams worked in primary health care units, which were organized in a network of basic, intermediate and specialized care, each providing more specialized services.

Community health teams were responsible for delivery of primary care services in a designated geographical catchment area. Teams had a clearly defined portfolio of approximately 300 primary health care services which included health education and promotion (age-appropriate nutrition, sexual and reproductive education, breastfeeding, etc.) preventive care (vaccinations, preventive care for low-birthweight, malnutrition, overweight, chronic conditions; reproductive health, etc.) curative care (infectious, non-communicable, and sexually transmitted diseases, etc.), and community-based rehabilitation.

To provide these services, health teams conducted a census of their catchment areas to obtain health and demographic data used to generate a health-risk profile of families and individuals. This risk profile determined the services required by patients according to established guidelines (MINSAL 2011). For instance, for children 24 to 60 months old with no observed health risks, CHWs should provide home-visits twice per year and refer to well-visits[9] with a physician or professional nurse in the health unit twice a year. Children younger than 24 months should receive home-visits and well-visits in the unit every other month up to 12 months of age and every three months up to 24 months of age. The guidelines contained job-descriptions for each member of

[7] This section relies on MINSAL (2010), which provides a comprehensive account of the Health Reform in El Salvador and in MINSAL (2011), which details the guidelines of care of community health teams.

[8] The term for these teams in Spanish is *Equipos Comunitarios de Salud Familiar*, or *ECOS-F.*

[9] Well-visits are preventive visits scheduled periodically. For children this are more frequent when they are younger and become less frequent as they age. A usual child well-visit includes a complete physical exam, weight and height measurements to track growth and some screening tests. Nutritional advice is also usually provided during these visits.

the team and their role in providing the established services according to the risk profile of the population.

The country's health reform was implemented by stages, prioritizing the poorest municipalities with a rapid expansion of health units between 2010 and 2013. Prior to the reform, the country had 377 primary care units and by 2015 they were 747,[10] each staffed by a community health team. The MoH was the main provider of health services in these areas. For example, in 2014, the MoH provided around 93 percent of postnatal care services in rural areas and close to 75 percent in urban areas.[11]

B. Salud Mesoamerica Initiative

To implement the health system reform the government of El Salvador combined national and donor funds, including funds from the *Salud Mesoamerica Initiative* (SMI). SMI is a public-private partnership[12] that aims to reduce maternal and child health inequalities in the Mesoamerica region[13] by extending coverage and improving the quality of health care for the poor. SMI combines grants equal to 50 percent of the project value with an equivalent percent of country funding. The funding is provided with a Results-Based Aid (RBA) model in which national country governments are offered an incentive equal to 25 percent of the total funding envelope if they achieve 80 percent of pre-established targets at the end of an 18-month funding period (Bernal et al., 2018). Targets are independently verified through a combination of household and facility surveys (Mokdad et al, 2015). The incentive SMI offered to national governments had no restrictions, except being earmarked for the health sector.

In the case of El Salvador, SMI provided funding for 14 of the poorest municipalities (see Table 1 for a list of municipalities), which were mostly served by MoH public health services.[14] SMI financed two 24-month phases,[15] each with its own performance framework of 8 to 11 indicators. The performance framework for the first phase focused mainly on the supply of inputs at the primary level for family planning, prenatal and child preventive care as well as the establishment of community health teams. The performance framework for the second phase focused on improving coverage of family planning, maternal and child services, as well as the timeliness and quality of maternal care. By the end of the first phase, the country reached 80% of targets and was granted the 25% performance tranche which was used for the construction and remodeling of health units and budget support for the payroll of personnel in SMI municipalities. The experiment of in-kind incentives for health workers was introduced as part of SMI's second phase, in October 2015, in an effort to motivate health teams and improve the quality of services. The experiment is described in detail in the next section.

[10] The initial number of primary care units comes from MINSAL (2010), whereas the most recent one was provided by the Human Resources Department of the Ministry of Health. The 2015 figure includes basic, intermediate and specialized units. Units had the teams necessary according to their catchment area.

[11] Estimates from most recent national health survey in the country, UNICEF's Multiple Indicator Cluster Survey in 2014.

[12] The public-private partnership combined funding from the Bill & Melinda Gates Foundation, the Carlos Slim Foundation and the Government of Spain that was managed by the Inter-American Development Bank.

[13] For the purposes of SMI the Mesoamerican region included Belize, Guatemala, El Salvador, Honduras, Nicaragua, Costa Rica, Panama, and the state of Chiapas in Mexico.

[14] In 2011, prior to the start of SMI, around 85.4 percent of care sought by women was provided by public facilities and by 2016 (IHME, 2011), the figure was close to 92 percent by 2016, of which around 97% was provided at MoH facilities.

[15] While the design originally was set for two phases, a third phase was to be approved by 2018.

III. Description of the Intervention

A. Incentive Scheme

The intervention that we study here consisted of an in-kind group incentive to health teams linked to 11 key maternal and child health indicators. The indicators mapped closely to SMI's performance framework at a national level, covering outcomes related to family planning, prenatal, postnatal and child care. The indicators were designed to promote community outreach, increase utilization, and improve the timeliness and quality of care.

Each indicator was assigned a performance target. If a community health team met the target, it received points redeemable for in-kind incentives. Targets were set based on SMI's baseline information (Mokdad et al, 2015), the feasibility of reaching the target over an 18-month period, and their alignment to SMI's performance framework. A description of indicators, targets and their points can be found in Table 2. Points assigned to each indicator were weighted based on the expected level of effort to achieve the target and ranged from 5 to 15 points. For example, utilization indicators related to micronutrients or use of modern family planning methods were assigned the highest number of points (15) whereas process indicators such as references to institutional delivery in a mother's birth plan received the fewest points (5).

The value of the award was determined on a sliding scale (Table 3), based on the number of points accumulated by the health team every six-month cycle. Teams had to accumulate a minimum of 60 points to receive the minimum incentive of $650 USD worth of goods and were rewarded up to a maximum of $1,000 USD in goods if they reached 90 or more points in that 6-month cycle. The experiment was implemented for a total of three six-month cycles.[16] Teams chose their awards from a list of in-kind incentives created by the Ministry of Health based on focus groups with health team representatives. Awards included laptops, air conditioners, microwaves, and other assets that could be used by the team to improve productivity and comfort in their place of work. The awards were assigned to the unit, not team members, as they were property of the MoH.

B. Performance measurement

All indicators were measured independently every six months using household surveys and medical record reviews. Details of the data collection can be found in Section IV. By design, only a small sample of observations was collected for each team to determine whether there was statistical evidence that the community health team had achieved the target of each indicator using exact binomial tests.[17] If there was statistical evidence that a team reached or surpassed the desired threshold for an indicator, the team obtained all points corresponding to that indicator. Otherwise it obtained no points. The use of exact hypothesis tests on a binary outcome of reaching each target was motivated by operational considerations, since establishing sufficiently precise point-estimates to reward teams based on partial achievement of the target would have required prohibitively costly sample sizes for each catchment area.

[16] The total incentive amount was initially set at $1000 dollars for each of 4 verification cycles excluding baseline. However, due to time constraints in the financing of the pilot, only 3 verification cycles (plus baseline) were possible. Incentives for the final two cycles were combined, and teams were granted up to $2000 in incentives in the third and final cycle.

[17] The tests were performed to assess the null hypothesis that the team had achieved the target established for the indicator. Due to the small sample sizes for each team, the test was performed based on the binomial distribution rather than a continuous distribution approximation. These tests were useful to identify those teams that were performing well below the target.

The data source for verification of each target was selected based on feasibility of obtaining a sample in each verification cycle. For most indicators related to utilization (i.e. current use of family planning methods) and community outreach (i.e., women received information on family planning methods), household surveys were required. On the other hand, medical record reviews were preferred for most quality of care indicators (i.e., clinical quality of prenatal care) and timeliness of care indicators (i.e., timely antenatal care). Some exceptions were indicators related to narrow population groups (for instance institutional delivery or children between 12 and 23 months that received the measles, mumps and rubella (MMR) vaccination) which due to sample size considerations[18], required a proxy measure from medical records. Table 2 includes the source of verification used for each of the eleven performance indicators.

C. Performance Feedback

The results of each independent verification cycle were shared with teams in an individually tailored report. The report included the points obtained for each indicator and the details of the measurement of each indicator, including its definition and data source. Reports were provided for each of the 4 verification cycles to all 75 teams, independent of treatment status. Their design was identical, except for the fact that teams assigned to the treatment group had a legend explaining the incentive amount obtained during that cycle. Teams assigned to control had only a description of the points obtained overall, with no mention of incentives. An illustration of the differences between the reports is presented in Figure 1. A sample report in full is included in Appendix 1.

The reports were provided to all teams at an event held at the end of each verification cycle, with the presence of representatives of each team, as well as municipal, regional and central level MoH authorities. During those events, aggregated results at the department or municipality level were presented and discussed. In addition, teams that achieved 60% of points or higher[19] received a certificate and public recognition from central MoH authorities. All certificates were designed identically, but certificates for treatment group teams included the incentive amount. An example of the certificates provided by treatment status is shown in Figure 2.

There were four verification cycles in total. The first was conducted at baseline, prior to random assignment of teams to treatment and control groups. No certificates were provided at the end of the baseline (only the reports). The other three cycles were performed every six months. A timeline with the dates of each data collection period and performance event can be found in Figure 3.

IV. Data

Household surveys and medical record reviews were collected by independent surveyors at the end of each verification cycle. The survey group was blinded as to the treatment status of each

[18] For timing and budgetary constraints sampling for household surveys was done based on a stratified random sample of 30 dwellings within the catchment area of the community health team. Hence, narrow population groups such as children between 12 and 23 months, could be underrepresented or absent from the selected sample particularly in catchment areas where there is a smaller share of children. To guarantee adequate sample sizes for indicators measured in this type of population groups a proxy indicator that could be measured from medical records was selected. The sampling strategy is discussed in more detail in Section IV.

[19] The percentage was obtained by adding the total points obtained in a verification cycle and dividing them by the total points from indicators that could be measured in that cycle. This was done, since in some cycles, for reasons beyond the control of the teams, there was not enough sample to estimate compliance with the threshold, and hence the points from this indicator were not counted towards the total. One of the most common reasons for no data availability was the security situation in certain areas, where the data collection team could not enter.

community health team. The surveys were collected every six months starting in September 2015 (Baseline). The three remaining cycles took place in March 2016 (6th-month follow-up), September 2016 (12th-month follow-up), and March 2017 (18th-month follow-up). Each cycle consisted of a household survey and medical record reviews in the catchment areas of each of the 75 participating community health teams. Independent samples were obtained for each cycle, so in effect data consist of a panel of repeated cross-sections for each catchment area.

The household survey captured information on sociodemographic characteristics, family planning, and women and children's health care utilization and behaviors. Surveys were applied to a stratified random sample of 30 dwellings within the catchment area of each community health team. Catchment areas are subdivided in zones that are served by different community health workers. The sample within each team was stratified by zone to guarantee that all zones were represented and to reduce sampling variation. For each selected dwelling, up to two women ages 15 to 49 years old[20] were selected at random using a modified Kish table to respond the women and child health survey modules (the latter, when they had children under five). Data on maternal health (prenatal care, delivery and postnatal care) were collected on the most recent birth of each woman.

The sample of 30 dwellings per catchment area was designed so that, in expectation, two thirds of dwellings included women between the ages of 15 to 49, and about one-third of dwellings had children under five. This was done to achieve the minimum necessary sample sizes required to distinguish whether a target had been met or not for the population sub-groups of interest, such as use of family planning methods for reproductive age women or consumption of deworming medication for children. On average, 2308 dwellings were surveyed for each verification cycle of which 1447 had reproductive age women and 781 had children under five.

Medical record reviews were used to measure targets for population sub-groups that represented only a small percentage of the population such as pregnant women or children under two-years old. Reviews were conducted for medical records of patients with a recent delivery, pregnant women, and children 6 to 23 months old. The target sample for each catchment area was 17 records of deliveries, 17 of pregnant women and 18 of children 6 to 23 months old.[21] In total, an independent random sample of 806 records was obtained for women with recent deliveries, 1159 records for pregnant women, and 1336 records for children 6 to 23 months old.

Medical record reviews for women with recent deliveries captured data on prenatal care, delivery and postnatal care. Prenatal care data included dates of each visit, the clinical examinations performed during each visit (blood pressure, weight, fundal height, fetal heart rate) and lab tests (urine, blood, HIV). These data allowed for construction of quality of prenatal care measures, which consisted of adherence to the country's clinical guidelines on examinations that should be performed during pregnancy, by gestational age. Data were also collected on the birth plan, which documents a pregnant woman's intentions for the upcoming birth, including the desired place of delivery and transportation arrangements to the selected place. Finally, data were collected on the date and place of the birth and first postnatal check-up.

[20] All cycles of data collection followed this procedure of selecting at random at most two women 15 to 49 in case there were more than two or selecting all of them when two or less lived in the dwelling, except for baseline in which data collection was done with only one woman selected at random.

[21] If fewer records were located in the reference period, then all records were selected. Since deliveries are a rare event, the actual average sample per catchment area was well below the target.

In the case of clinical records for pregnant women at the time of data collection, data were collected on the dates of prenatal care visits and the date of last menstruation, providing a measure of timeliness of prenatal care (i.e. whether the first prenatal care visit occurred within the first trimester of pregnancy). Independent samples of medical records were drawn for pregnant women and women with deliveries, since each measurement cycle focused on the six-month period immediately preceding measurement and using patients with completed deliveries to measure timeliness of prenatal care would have fallen outside this window.

Finally, the medical record review for children age 6 to 23 months old captured data on MMR (measles, mumps, and rubella) and BCG (tuberculosis) vaccinations, as well as prescription and delivery of micronutrient supplements aimed at reducing the prevalence of anemia.

The data collected through household surveys and medical records was used to construct indicators of health care performance for each team. There were a total 11 indicators used to measure performance including use and information received of modern family planning methods, timely prenatal care (first visit within the first trimester of pregnancy), quality of prenatal care, use of birth plan, institutional delivery, timely postnatal care for women (within one week after delivery), consumption of deworming pills for children under five, consumption of micronutrient powders among children 6 to 23 months old, MMR vaccination for children between 12 and 23 months old, and knowledge among mothers of children under five of treatment of diarrhea with oral rehydration salts (ORS) and zinc. Different sources were used to measure team performance, population-based measures from household surveys were used for indicators on family planning, consumption of deworming pills and knowledge of treatment of diarrhea with ORS and zinc. For all others, medical records were used due to data quality and sample size considerations.[22]

The performance indicators were either direct indicators of a desired outcome or, in the case of four indicators, they were proxies of a desired outcome selected due to operational and sample constraints. The construction of each indicator can be found in Appendix 2. For purposes of our analysis the performance indicators are used (proxies and direct). A discussion and analysis comparing results from proxies to those of direct outcomes is presented in Section VII.

Finally, although the objective was to collect data for all 75 health teams in each verification cycle, the surveyors were unable to collect the full household survey or medical records sample in some catchment areas, primarily due to security concerns with gang violence which is prevalent in the intervention areas. We have complete data (all four cycles for all 11 indicators) for 64 out of 75 catchment areas. We use this as our main analysis sample. Appendix 3 presents a breakout of the sample by source and verification cycle. We find no evidence of a relationship between experimental assignment and attrition as we discuss in Section VII.

V. Identification and Estimation Strategy

A. Identification and Experimental Design

Our objective is to identify the causal effects of the in-kind incentives on provider performance on outcomes related to the delivery of maternal and child care, as well as on population-level health

[22] For the indicators of prenatal care quality and timeliness, medical records were preferred since they were more precise in terms of exact dates and procedures performed. Even though household surveys that rely on maternal recall can be accurate for some specific interventions during delivery (Chen et al, 2014) and for conditions during pregnancy (Krakowiak, 2015), to our knowledge its reliability in terms of specific clinical items during prenatal care is not that well studied. Moreover, differential recall bias related to pregnancy outcomes and women characteristics has long been known to be a concern in the epidemiological literature (Yawn et al 1998).

behaviors including utilization, timeliness of care and health related knowledge. The challenges of identifying these effects empirically are multiple and are not limited to the health pay for performance literature. Health workers could self-select into compensation schemes that include some form of performance incentives (Prendergast, 1999). For example, more productive workers might be attracted to jobs with performance pay rather than a fixed salary. A naïve comparison of individuals with performance incentives to workers on fixed salary could reflect differences in worker productivity rather than the effect of incentives. Selection issues could also occur at the team level, as different workers might prefer team compensation or at the organization level, since different types of organizations could favor providing performance incentives (Boning, Ichniowski, and Shaw 2007). Finally, performance incentives, particularly in the health sector, are usually deployed in conjunction with other interventions that could have an effect on their own such as audits, information feedback, and public reporting,[23] which could confound the effect of incentives (Mendelson et al, 2017).

To overcome these challenges, our sample is composed exclusively of existing MoH community health teams. Since teams were constituted by the MoH prior to the start of the incentive pilot, we avoid selection of individual health workers into teams. Moreover, all teams are subject to the same MoH institutional environment, including compensation rules. Random assignment of teams to a treatment group with incentives and a control group without reduces the potential of selection on unobservables. Finally, all teams received identical measurement, performance feedback, and public recognition for their achievements, allowing us to credibly isolate the effects of the in-kind incentives.

Our sample is composed of 75 community health teams[24] within 8 administrative regions in 14 municipalities. Administrative regions[25] are responsible for the network of primary and secondary units in a pre-defined area including supplying inputs, supervision, and management (all under the command of the central MoH). We implemented a stratified random assignment procedure within three blocks, two rural and one urban. The two rural blocks of about the same size were constructed based on administrative regions using the MoH *Sistemas Básicos de Salud Integral* (SIBASI). All urban teams were grouped into one block. Blocking was done to guarantee balance within each administrative region and to improve precision (List, Sadoff, and Wagner, 2010). A summary of the number of health teams by block is presented in Table 4.

The random assignment was conducted in a public event in October 2015, with representatives of all health teams and national authorities, in which the incentive scheme was presented along with the random assignment procedure. Teams were assigned at random[26] to one of two stages. Stage I teams (treatment group) were eligible for the incentive scheme during the 6th, 12th, and 18th-month verification cycles, whereas stage II teams (control group) would only be eligible for the incentive scheme during the 18th-month verification cycle (see Figure 3). As previously

[23] If members of community health teams are intrinsically motivated, reporting on performance even if it is among their peers could have an effect of its own as they might increase effort to perform better or at the same level as their peers. Kolstad (2013) finds evidence that the intrinsic motivation response to information of performance outweighed that of profit incentives after the introduction of report cards for cardiac surgery in Pennsylvania.

[24] All community health teams participated independently on the randomization, except for the two urban teams of Sensuntepeque, which were considered as one for the purpose of random assignment since the two teams shared the same premises. This was not the case for all other teams.

[25] The MoH divides the country in 17 SIBASIs or administrative regions.

[26] Random assignment was done by introducing into an opaque box an equal number of marked balls with the numbers I and II (and totaling the amount of teams within a block), shuffling the box and then drawing a ball at random for each team within a block. Teams that selected a ball marked with I were assigned to Stage I (treatment) whereas those with a ball marked with II were assigned to Stage II (control). This procedure was repeated for the three blocks.

mentioned, all teams, regardless of their assignment would receive the same performance measurement, reports, feedback and certificates of recognition.

Since teams in the control group were aware of their pending incorporation into the incentive scheme as of the 18th month verification cycle, one threat to identification in this phase-in design is that control teams, in anticipation of future incentives, could start working to improve their performance since the start of the experiment. If this mechanism is at work, then our estimates of the incentive effects would be attenuated. We discuss this and other threats to identification in Section VII.

B. Estimation Strategy

To estimate the effect of incentives we use the following regression framework:

$$(1)\ Y_{ijbt} = \delta T_{jb} + \varphi_b + \beta X_{ijbt_0} + \varepsilon_{ijbt}$$

where Y_{ijbt} is the outcome of interest for individual i, in the catchment area of team j, of block b, in time t; T_{jb} is an indicator equal to one if the team j in block b was assigned to treatment; φ_b are block effects; X_{ijbt_0} represents baseline covariates; and ε_{ijbt} is an error term, which is clustered at the catchment area level. The parameter of interest is δ, which captures the average difference between treatment and control groups within each block. Under standard identification assumptions, that is, if treatment assignment is independent of the outcomes of interest, then, δ, will capture the effect of in-kind incentives. While this assumption is untestable, we provide evidence that randomization was successful in producing balance for a comprehensive set of health team, patient and catchment area characteristics presented in Table 5. The balance test was performed by estimating equation (1), without baseline covariates, on a set of baseline variables. In addition to the balance test presented in Table 5, we also include similar balance tests for all the outcomes analyzed in the results section.

Our preferred regression estimate of δ pools data from the 6th and 12th month follow-up rounds (during which the control group was not eligible for incentives) and includes controls for team-level outcomes of interest at baseline, Y_{jbt_0}. We pool follow-up rounds to increase the sample size and reduce variance (McKenzie, 2012) and we include the baseline outcome as a control to improve statistical power (McConnell and Vera-Hérnandez, 2015) and to account for chance differences between groups. As a robustness check we also compare the estimates of our main outcomes with those using differences-in-differences and present disaggregated effects by follow-up round in Section VII.

Our primary outcomes are the 11 indicators used to measure performance. To summarize this information, we first present results on indices of outcomes grouped into four domains: community outreach, quality of care, timeliness of care, and utilization. We then present individual estimates for each outcome. The grouping of indicators into each of the domains follows Table 2. The indices on each domain are average standardized treatment effects (ASTE), which are defined as follows:

$$(2)\ ASTE_d = \frac{1}{K}\sum_{k\epsilon K}\frac{\hat{\delta}_k}{\hat{\sigma}_k}$$

where $ASTE_d$ is the average standardized treatment effect for domain d, which is composed of K outcomes, $\hat{\delta}_k$ is the treatment effect for outcome k estimated from equation (1) and $\hat{\sigma}_k$ is the

standard deviation of the control on outcome k. Using this approach, which is akin to that used by Kling et al. (2004), Spenkcuch (2012), and Clingingsmith et al (2009) has several advantages. First, since it aggregates the effect of different measures within a domain, it increases statistical power. Second, it reduces the number of hypothesis test performed on the main analysis from 11 to 4, and therefore reduces the likelihood that we find effects just by chance (Type I error). Finally, it allows us to combine effects from outcomes from different sources (medical records, household surveys) and population groups (women 15-49, pregnant women, children under 5, etc.) into a single metric. We follow Kling and Liebman (2004) in estimating the standard errors for ASTE for each domain and use a seemingly unrelated regression framework to account for covariance across estimates.[27]

VI. Results

A. Overall performance score

Figure 4 presents the kernel density plots of health team's performance scores by time period (baseline, 6-month, 12-month and 18-month) and treatment status. The overall performance score is the aggregate weighted measure of targets met by a health team, with weights equal to the points established for each indicator (Table 2). The score determined whether teams in the treatment group received incentives, according to the sliding scale presented in Table 3. It is not a measure of the magnitude of the gains, but rather of compliance with the targets of all indicators and a summary measure that teams received on their performance every cycle. For each period, we perform a Kolmogorov-Smirnov (KS) test of equality of distributions between treatment and control and include the p-values of these tests at the bottom of Figure 4.

Figure 4 reveals several findings. First, there are no statistical differences between the distributions at baseline and six-months, the first period where treatment teams were eligible for incentives. Differences appear 12-months after the scheme was introduced, as seen both in the figure and confirmed with the result of the KS test. Second, treatment and control distributions both shifted right over time, indicating improvement after the scheme was in place. Third, while at baseline both distributions (treatment and control) spanned almost the full range of potential performance scores (0-100) and about half were below the threshold of eligibility for incentives (60 points), only a small share of teams remained below this threshold 12-months after. Additionally, the distribution of performance scores was compressed between 40 and 100. Only 12 percent of all teams remained below the minimum incentive threshold 12-months after baseline. This is evidence that low performance teams at baseline had a marked improvement and suggests a sort or "threshold effect" as teams strived to be above the 60 percent threshold from which incentives and public recognition were given.

The fact that control teams had a marked improvement in performance during the first six months that mirrored treatment could have several explanations. First, all teams received a baseline report with their performance, and this information could have had an effect on its own. There is empirical evidence that among physicians, report-cards on quality of care can improve their performance even if it is unrelated to patient demand (Kolstad, 2013). Second, all teams regardless of their experimental assignment received public recognition if they met 60 percent of more the established targets. Social recognition can be an effective motivator in labor relations

[27] We also estimated bootstrapped standard errors clustered by community health team and stratified by block with 1000 replications to account for an unknown form of correlation within teams and across outcomes. The results are essentially the same. Moreover, the magnitude of our estimate of standard errors using the Kling and Liebman (2004) approach was slightly more conservative than those of the bootstrap for our main results.

(Kosfeld and Neckermann, 2011). Moreover, qualitative evidence suggests that public recognition with the certificates was something that motivated deeply community health teams and those not getting one felt compelled to strive harder (Munar et al, 2018).[28] Third, all teams were part of SMI, which had incentives at the national level. Feedback and support were provided to all teams by municipal and administrative region supervisors to create local plans of improvement aimed at achieving SMI targets. Fourth, it could be explained by an anticipation effect, since control teams knew they would be eligible for incentives at 18-months. We discuss this further in Section VII.

Treatment teams improved their overall performance scores relative to controls 12-months after baseline. An important part of the improved performance is the reduction of teams performing below the performance threshold, suggesting that low performance teams had a marked improvement. Finally, at 18-months, once control teams were eligible for incentives, the performance distributions of treatment and control are again equal and compress around the highest performance scores. While Figure 4 gives a clear picture of team performance, it does not provide an idea of the magnitude of the gains in performance, since it is just a measure of compliance with targets for the 11 indicators. We look in detail for the magnitudes of improvements by domain in the next subsection and at the gains by baseline performance score in subsection D.

B. Main Outcomes

Community Outreach

We study two community outreach outcomes, both measured from household surveys. First, information on modern family planning refers to women 15 to 49 years old[29] that declare to have received information on modern family planning methods[30] by health personnel during the last six months. The second is knowledge by mothers of children under five on treatment of diarrhea with oral rehydration salts (ORS) and zinc. Treatment of children's diarrhea with zinc in addition to ORS is recommended by the WHO, since it reduces the severity and duration of diarrhea episodes particularly among children in low-resource settings[31] (Lazzerini and Wanzira, 2016). The outcome is relevant in El Salvador since only about half of baseline diarrhea cases sought medical treatment. Furthermore, zinc had been introduced recently, in 2014, as part of the clinical guidelines for treatment of diarrhea in the country.

Results are presented in Panel A of Table 6. In-kind incentives increased community outreach by 0.17 standard deviations. Prior to the introduction of the incentive scheme there was adequate balance (column 2). When analyzing the outcomes individually (Panel B of Table 6), we observe that information on family planning increased by 0.058 percentage points, which represents an increase of 11.5 percent relative to baseline. On the other hand, knowledge of treatment of diarrhea with ORS & Zinc increased by 0.079 percentage points, which is about twice the baseline

[28] As stated by one of the participants in the feedback events regarding the performance certificates *"When we go to the meetings, and I love that.... when we are all there, they say, 'We are going to give this diploma to the teams that made a great work.' They call them up, and it looks like they are graduating from something. And my colleagues become really motivated. And maybe you can hear from those who didn't receive anything phrases like 'I will propose something for the next meeting because I want to be there.'"* (Munar et al 2018).

[29] This refers to women in need of contraception which are defined as those age 15 to 49, excluding those sterilized, in menopause, that declare to be virgin or not sexually active, and those pregnant or trying to conceive.

[30] Modern family planning methods include injectable, contraceptive pills, female or male sterilization, intra uterine devices (IUD), implants, emergency contraception, female or male condoms, contraceptive diaphragm and sponges.

[31] This is since in these setting children are likely to be malnourished or zinc deficient.

level (only 7 percent of women declared that they would use Zinc in addition to ORS for a diarrhea episode for their children at baseline).

Quality of Care

For quality of care, we group two outcomes related to the care received by women during pregnancy at the health facility: quality of prenatal care and reference to institutional delivery. Both are measured from medical records. On average, about 95 percent of pregnant women in the catchment area received prenatal care from the community health teams at MoH facilities. Quality of prenatal care is an indicator equal to one if the pregnant women received all the clinical examinations required by national guidelines during all prenatal care visits[32]. Examinations vary by gestational age, but include weight, blood pressure, fundal height (after 14 weeks of gestation), fetal heart rate and fetal movements (the last two required after 20 weeks of gestation). In addition, lab tests of glucose, HIV, hemoglobin and urine should be done at least once during pregnancy according to clinical guidelines. These examinations and tests permit the detection of critical risks during pregnancy including eclampsia, anemia, diabetes, intrauterine growth restriction, and/or HIV, that could lead to increased maternal and fetal morbidity and mortality if undetected and untreated (WHO, 2016; Bhutta et al., 2014). The second outcome of interest is reference to institutional delivery which measures whether the pregnant women received advice on institutional delivery and registered a delivery plan (a plan developed by provider and patient on the place of delivery, transportation and arrangements during pregnancy).

Table 7 presents the effects of the incentive on quality of care measures. The incentives had a statistically significant effect of 0.14 standard deviations (Panel A), driven by quality of prenatal care which exhibits an increase of 8.4 percentage points (a 14 percent increase relative to a baseline of 58.7 percent). Reference to institutional delivery increased by only 2 percentage points and the result is not statistically significant (Panel B). It is worth noting that there is a statistically significant imbalance in quality of prenatal care at baseline. However, this goes in the opposite direction of the treatment effect, implying that treatment teams had a lower level of this indicator than control ones at baseline and we have no evidence to believe that this reflect a systematic difference between treatment and control groups.[33]

Timeliness of Care

Timeliness of care includes two outcomes: timely antenatal care and timely postnatal care. Both are measured from medical records. The first is an indicator equal to one if the first prenatal care visit occurred during the first trimester of pregnancy, as measured from the date of the last menstrual period and the date of the first prenatal care visit, which are both captured in the medical record. Early prenatal care is part of WHO standards of care (WHO, 2007) since it allows to detect and treat conditions that could compromise fetal and maternal health (Carroli et al., 2001). The second one is an indicator equal to one if the first postnatal care visit occurred within a week after delivery as measured from the date of birth and the date of the first postnatal care visit. Early postnatal care is a critical part of the continuum of care for pregnant women since it allows detection of complications arising after delivery (WHO, 2013), and more than a third of maternal deaths in the region occur during this period (Kassebaum et al, 2013).

[32] On average women received around 4 prenatal care visits in MoH facilities in the 14 participating municipalities.

[33] Balance tests are presented in Table 5, and those presented for all of the other outcomes of interest in Tables 6 to 9. Moreover, our preferred model on Column (4) controls for baseline levels, which should account for any chance imbalance, and is more conservative given the direction of the imbalance compared to the difference-in-difference estimate (see Appendix 5).

Results are presented in Table 8. The average treatment effect is of 0.10 standard deviations (Panel A). Panel B shows that the effect is driven by timely post-natal care, with an increase of 8 percentage points (an increase of about 14 percent relative to baseline). On the contrary timely prenatal care is essentially unchanged, a statistically insignificant coefficient of 0.018 percentage points. Timely prenatal care was 72.6 percent at baseline which is comparable to the 74.7% average of the Latin American region, (Moller et al., 2017). Moreover, evidence suggests that this outcome might be difficult to influence as it requires not only clinical knowledge but also developing strategies to identify pregnant women early (Basinga et al, 2011). The marginal cost of rising this indicator may be increasing at higher levels of compliance since it involves identifying those women who are least likely to seek care early in their pregnancy, such as teenagers and single women.

Utilization

We include a total of five utilization indicators. Two of these are measured with household survey data, namely current use of modern family planning methods by women 15 to 49 in need of contraception[34] and consumption of two doses of deworming pills by children 12 to 50 months in the last 12 months. For the other three measures, sampling constraints required constructing proxies from medical records (we compare these proxies to their equivalent household measures in Section VII for a subset of teams were data are available). These utilization proxies include an indicator for institutional delivery from medical records, an indicator for prescription of micronutrients sachets for children 12 to 23 months and an indicator for MMR vaccination according to the medical records of children of the same age.[35]

Overall the team incentives were effective in producing a small improvement in the utilization of health services, of 0.096 standard deviations as presented in Table 9. This effect came about by relatively small gains in all five indicators of between three to five percentage points. While most of these individual effects are statistically insignificant, we gain precision by pooling them together in a single measure.

C. Non-contracted outcomes

A common concern in the design of pay-for-performance schemes is that if individuals have multiple tasks to complete as part of their job, but only a subset are subject to incentives, then in response they could shift effort away from non-contracted outcomes. This unintended effect is commonly referred as multi-tasking in the theoretical literature of contracts (Holstrom and Miligrom, 1991). While most of the literature focuses on multitasking at the individual level[36], teams could also be subject to this unintended effect if they share a common objective function and are able to coordinate.

[34] See the notes on the Community Outreach subsection for the definition of modern family planning methods and women in need of contraception.

[35] For micronutrients and MMR vaccination this was the source of verification for performance during baseline, but was later changed to children with the prescription of micronutrients filled by the team for micronutrients and MMR vaccination according to vaccination booklets, since they were more up to date. We use the original measures since those were those collected in all four verification cycles.

[36] The empirical evidence on the existence of this effect in pay-for-performance schemes in health is mixed, most likely as it is contract-specific and also since it would depend on the degree in which contracted outcomes share inputs with non-contracted ones (Miller and Babiarz, 2014)

We test whether the incentive scheme had this type of unintended consequences by assessing three non-contracted outcomes: detection of diabetes, hypertension, and cervical cancer among women 15 to 49 years old during the last six months. None of these outcomes were included as part of the incentive scheme, but they were activities to be performed by community health teams in their catchment area. All were constructed from the household survey. We present the average standardized treatment effect in Table 10, finding no evidence of shifting of effort from these activities.

D. Heterogeneity by baseline performance

As discussed in section III, the incentive scheme was structured as a step function based on a non-linear sliding scale starting at a threshold of 60 points out of 100 (Table 3). A common concern with this type of design is that if performance is costly, the marginal cost of improving performance could outweigh the benefits for teams far below the threshold. In this case, the incentives might only affect those teams that are slightly below the threshold where the benefits outweigh the costs (Miller and Babiarz, 2014). This type of response would also be expected from the behavioral goal-gradient hypothesis (Kivetz et al. 2006), that states that effort is increased as the distance to the rewards shortens and decreases once the threshold is passed and the reward (or partial reward) obtained.

To our knowledge there is limited empirical evidence on this type of response in the pay-for-performance in health literature.[37] In this subsection we analyze the response of teams to the incentive scheme by grouping teams into three categories according to their baseline performance. The first group includes 30 teams below the incentive threshold (performance score of 0 to 59 inclusive). The second group is comprised of 19 teams with baseline performance in the first bracket of incentives (performance score of 60 to 69). And the third group is comprised of the 15 highest performing teams that scored above the first bracket at baseline (performance score of 70-100).[38] For each of these groups we analyze the magnitude of improvement on contracted outcomes by estimating the ASTE. This is a more relevant measure than the gains in the performance score, since it reflects the magnitude of gains made by teams in each category and not just the overall compliance with the targets.

Fist, we estimate effects on the pooled ASTE for all 11 contracted outcomes as a single summary measure, using our preferred model that pools the 6^{th} and 12^{th} month measurements and controls for baseline level of each outcome. The results, presented in Figure 5, suggest that the lowest performance teams experienced substantial gains of about 0.12 standard deviations on the pooled ASTE. In contrast, those teams in the first bracket of 60-69 points at baseline achieved essentially no gains. Finally, the highest performing teams at baseline had the largest effect, achieving on average an increase of 0.22 standard deviations on the 11 contracted outcomes.

Next, we analyze results by domain to understand how teams in each baseline performance group achieved their improvements. This is particularly illustrative as domains could require different levels of effort. The results, presented in Table 11, suggest that teams with the lowest-performance score at baseline, made substantial progress in community outreach and quality of care, with effects of 0.26 and 0.13 standard deviations respectively (Table 11, second column).

[37] There is some evidence of this individual behavior in marketing from Kivetz et al. 2006 that study the context of consumer reward programs. There is also some evidence on this effect from the literature on non-monetary incentives for prosocial behavior in the study of blood donations (Lacetera and Macis, 2010).

[38] As a robustness check we conducted the heterogeneity analysis by quintiles of baseline performance score, finding essentially the same results discussed in this section, albeit treatment effects had wider confidence intervals with fewer observations per group.

Teams at the threshold of incentives at baseline show no indication of improvement in any domain, except possibly quality of care, although this result is not statistically significant (third column). Teams on the high end of the performance distribution at baseline, had large improvements on most domains. Interestingly, these teams were the only ones that achieved improvements in the timeliness of care and utilization domains (with 0.29 and 0.19 standard deviations respectively). Moreover, their ASTEs on the community outreach and quality of care domains are of about the same magnitude than those obtained by teams with the lowest performance baseline score.

The performance gains amongst low-baseline-performance teams were concentrated on domains that required less team coordination. For example, community outreach was mainly the responsibility of community health workers, whereas quality of care relied on physicians and professional nurses. On the other hand, high-baseline-performance teams also achieved gains on domains that required more team coordination, such as utilization of family planning methods and consumption of micronutrients. These domains require that health teams coordinate closely to change patient behaviors and follow-up with patients. These results may reflect that high-baseline-performance teams were more closely articulated to begin with and were thus able to respond to incentives on the more complex domains. On the other hand, it appears that gains for low-baseline-performance teams were driven by efforts exerted by individual team members. Finally, as in the aggregate analysis, we find no evidence of heterogeneity of shifting efforts away from non-contracted outcomes for any sub-groups, with small and statistically insignificant coefficients on non-contracted outcomes.

Overall, these results suggest that the in-kind incentives were sufficiently attractive to motivate higher performance for low-baseline-performance teams, however these improvements were concentrated in domains that required relatively less team effort. If all teams were purely maximizing their return to effort, and effort is costly, it might be expected that those above the incentive's threshold would do minimal or no additional effort. Our results are consistent with this hypothesis for teams at the threshold of 60-69 points at baseline which show no indication of responding to the incentive scheme. However, the large effects we find on high-performing teams seems at odds with the theoretical prediction. High-performers appear to increase their effort substantially across all domains, including the costliest in terms of team coordination. One potential explanation is that that they were the most productive teams to begin with, so small changes in effort could be translated into large changes in performance. An alternative explanation, is that for these high-achieving teams, reaching the highest incentive amount could have served as a goal, particularly since they had the lowest distance to the highest possible amount. If this was the case, a behavioral response consistent with the goal-gradient hypothesis would be that they increased their effort as they saw this goal within reach, even if it was costly, as they were highly motivated to reach the maximum incentive possible. Finally, the U-shaped observed on the response to the incentive-scheme is similar to that observed in response to rank-order feedback in lab settings and has been described as "first-place loving" and "last-place loathing" (Gill et al, 2018). While rank-order feedback was provided to all teams[39], it could have been that in-kind incentive effect could have been additive to that of rank-order feedback and hence the results we observe.

[39] Rank-order feedback was provided in public at the time performance certificates were awarded. Those not earning a certificate, that is below the 60 points threshold, did not receive their specific rank, but by not receiving the certificate they were informed effectively that they were the lowest performers.

E. Post-experimental treatment effects

The experiment with differential exposure to in-kind incentives lasted 12 months. By the third follow-up, at 18-months, both treatment and control teams were eligible for incentives, allowing us to study whether control teams catch-up to treatment teams once they became eligible for incentives. We estimate the treatment effect in each period by domain using our preferred model. Results in the fourth column of Table 12 show that the effects on the domains of quality of care and utilization remain significant higher for the treatment group at 18-months. ASTEs on the domains of community outreach and timeliness of care are positive but imprecisely estimated. It is important to note that as shown in Figure 4, both treatment and control teams show improvements in their performance score over time and the control group appears to narrow the performance gap after 18 months.

VII. Robustness checks and competing explanations

A. Differential Attrition

We were unable to obtain complete data for 11 out of 75 teams (6 control and 5 treatment). The primary reason was security concerns in certain catchment areas due to increased gang violence. We test for differential attrition in two ways. First, we create a dummy variable equal to one if a team had data for all key outcomes[40] on all waves and zero otherwise and regress this against treatment assignment and block effects. Second, we, perform baseline balance tests on the analysis sub-sample of 64 teams with complete data, and the full sample of 75 teams. The results of both exercises are presented on Tables 4A and 4B of the Appendix. We find no evidence of systematic differences between treatment assignment and attrition as there are no statistically significant differences in the share of teams with data on all outcomes between treatment and control. Moreover, the baseline balance tests on the full sample are very similar to the analysis sample both in magnitude and statistical significance, suggesting that attrition was unrelated to treatment assignment.

B. Unobserved differences between teams

While the random assignment appears to have generated adequate baseline balance in covariates, as a robustness check we run a separate set of analysis using difference-in-differences (DID) which additionally controls for any time-invariant unobserved team characteristics. The downside of the DID specification is reduced precision. We test for differences between estimates from the DID specification and our preferred model presented in the results section using clustered bootstrap at the team level and stratified by block with 1000 replications to estimate the standard errors. The results, presented in Appendix 5, indicate that in general there are no significant differences between both sets of estimates, and effect sizes are of a similar magnitude. Two exceptions are quality of care, in which the effect using DID is actually larger (0.24 Vs. 0.14 standard deviations) and timeliness of care in which the effect is substantially smaller (0.02 Vs. 0.10 standard deviations). However overall, the results from the DID analysis substantiate the results presented in section VI.

[40] The key outcomes refer to those used in the main analysis and included in Tables 6 to 10 (11 performance outcomes and three non-contracted outcomes).

C. The effect of goods on productivity

The in-kind performance incentives consisted of goods that might affect productivity in subsequent rounds. For example, laptops could be used by teams to increase the efficiency of registration of key information or reduce the time to analyze information. If the in-kind incentives received in the first period affected productivity directly, this would confound the incentive effects in subsequent periods, as differences in productivity in the treatment group could be due to the additional goods obtained as well as incentives. While theoretically possible, we do not expect this bias to be large for two reasons. First, most goods selected by teams were used to increase comfort in the work place, including fans, microwaves, and water dispensers.[41] While these could affect work satisfaction, they are unlikely to have a direct effect on activities conducted in the field, such as community outreach or patient interactions required to improve performance indicators. Second, there were substantial delays in delivering the goods after the first six-month period since the acquisition process involved the purchase of small quantities of a diverse set of items using national procurement systems.

An indirect test of the potential effects of the goods on performance is given by the 6th month follow-up results presented in Table 12, since no team had received any good before the first evaluation period. At the 6th month follow-up, the quality of care and utilization domains already exhibit statistically significant effects. The effect on quality is actually slightly larger at 6 months compared to the 12th months follow-up, and the effect on utilization remains at about the same magnitude in both rounds. The two domains for which there is no statistically significant effects at the 6th month follow-up but large increases at the 12th month follow-up are timeliness of care and community outreach. However, these are precisely the domains requiring effort on behalf of the health team in the field, where the goods offered as incentives are least likely to contribute directly to productivity.[42]

D. Cheating

Incentive schemes are fraught with potential unintended effects, particularly for complex tasks in which there are only imperfect measures of effort. A common concern is that individuals might focus on low-effort activities to obtain the incentive without influencing the underlying outcome of interest. For instance, incentivizing teachers' performance based on their students' test scores could have unintended consequences such as inflating test scores (Koretz, 2002), exempting the worst students from taking the test (Cullen and Reback, 2006) and even changing students' answers (Jacob and Levitt, 2003), rather than improving students' learning and achievement. In the health context, vaccinations could be over-reported if incentives are conditioned on the number of children vaccinated, without actually improving the coverage of immunity among the population (Lim et al, 2008).

The performance scheme we study here uses two different sources to measure performance: a household survey and a medical record review. We have no reason to believe that the household survey could be easily manipulated, and medical records are legal documents that if tampered can lead to legal action. However, we may still be concerned that teams tampered with medical

[41] The list of goods included air conditioners, refridgerators, microwaves, coffee makers, chairs, tables, laptops, printers, projectors, and digital cameras.

[42] For the case of community outreach, we speculate that the larger effect at the 12-month period might be explained by the fact that these activities are mostly performed by community health workers and the incentive scheme was presented first to representatives of teams which were mostly physicians and/or nurses. It might have been easier to mobilize these workers once the team achieved the first diploma and report with the recognition of their work and the amount received.

records to obtain short-terms gains, particularly since medical records were used to measure performance on seven out of the eleven performance indicators. If this were the case, then part of our effects could be due to changes in reporting rather than improvement in the underlying intended outcomes among the population.

To assess this possibility, we compare treatment effects for a subset of outcomes in which we could build comparable indicators from medical records and the household survey. Due to the design of the sampling scheme described in Section IV, we have data from both sources for four outcomes in 33 catchment areas, with comparable cohorts of children and pregnant women.[43] These outcomes are timely prenatal and postnatal care, institutional delivery, and MMR vaccination. The samples of medical records and household surveys were independent, so they do not reflect concordance of data by patient, but rather an estimate of the outcome of interest among the population of the catchment area of the team.[44]

In Table 13 we present results using outcomes from each source, and test whether there are statistically significant differences between them. The results of these tests, which are presented on column (5) indicate that there are no statistically significant differences between data sources, and more importantly, the size of the effects is quite similar for most outcomes. The similar magnitudes are relevant, since we have relatively low power for these tests given the limited sample of catchment areas. The only outcome in which the effect from the medical records proxy is larger than that of the household surveys is postnatal care (0.12 Vs. 0.056 percentage points). However, the baseline level of this indicator is also much lower in medical records than in the household survey (61% Vs. 92%). We believe this occurred since the sampling frame for medical records was based on the actual census of pregnant women maintained by the community health teams, and the clinical record for selected women in the sample was reviewed to find the clinical forms containing the data on the date of delivery and the post-partum date. A common finding was that for a substantial number of records (14 percent at baseline) there was no evidence of the post-partum care in the record since it only included actions taken by physicians and nurses and not those by community health workers. Community health workers were allowed by the countries guidelines to perform the early postnatal checkup to women, since usually women stay at home after delivery. Since the population-based coverage of early postnatal care was high at baseline (92 percent), and we find a significant effect of around 6 percentage points in this measure, the higher effect on the medical record is most likely a catch-up of reporting of services provided to the population. Similar issues occurred for other outcomes in which the medical record measures were substantially lower than the population-based measures (columns 1 and 3 of Table 13). [45]

Taken together, the evidence suggests that there were no systematic differences between data sources, supporting the claim that teams did not tamper with records. Two aspects of the incentive scheme may have helped limit tampering. First, teams were aware that performance would be measured using both medical records and independent household surveys. The latter may have been perceived as an audit on work reported in medical records. Second, the MoH conducted

[43] For this subset of teams, we have adequate balance and we have no evidence of differential attrition by treatment assignment as presented in Appendix 4, which are similar tests to those presented in Section VII A.

[44] The outcomes presented in Table 13 are related to maternal care and vaccination, all of which the main provider in the catchment area were the community health teams. For instance, prenatal care was sought in MoH facilities by 95 percent of women with a pregnancy on average over all 75 community health teams.

[45] For institutional delivery, the issue was that there was a gap in registration that occurred since deliveries were not performed by teams, but by hospitals which had to send the information on deliveries back to teams and this caused a delay in registration. For vaccinations, it was a similar issue to that of postnatal care.

close supervision from the administrative regions and the central level for all teams. Every three months they met to discuss results and reports of performance were analyzed jointly to understand the barriers to achieve targets and determine improvement plans. While this is standard practice at a national level, it may have been particularly salient for the 75 catchment areas studied here given their relevance to the national results-based aid model under SMI.

E. Anticipation

A final threat to identification of incentive effects in our context is anticipation effects from community health teams in the control group, since they were aware of their eligibility for incentives after the initial 12-month experimental period. If control teams started to improve performance in anticipation of becoming eligible for incentives, this would underestimate our treatment effects. As observed in Figure 4, there is a clear improvement in control team performance relative to baseline during over the first 12 months, however this could be explained by the effects of information and recognition for which they were eligible. Unfortunately, we are unable to separate potential anticipation effects from information and recognition effects. If anticipation did play a role in boosting performance for the control group, our treatment estimates would be attenuated downward. Since we do find positive and statistically significant effects for all domains analyzed, we interpret our estimates as a lower-bound of the effect of in-kind incentives.

VIII. Discussion

Improving the quality of health care is a global challenge that is particularly salient in the poorest and hardest to reach areas in low and middle-income countries. Pay for performance is a promising tool to align health provider's incentives more closely with health outcomes. Yet despite growing interest, causal evidence is mixed, and only a few studies are able to isolate the effect of incentives separate from information, feedback and/or recognition. Moreover, the design of incentive schemes, such as whether to pay cash or in-kind incentives and individual versus group incentives likely plays a key role in their effectiveness, unintended consequences, and sustainability, yet we know little about the plusses and minuses of different models. In the health sector in particular, providers may be at least partially motivated by intrinsic factors, and payment of monetary incentives can face institutional barriers that make them untenable in the public sector. As such, how to structure a pay for performance incentive schemes is context specific and remains largely an open question (Miller and Babiarz, 2014).

This study presents experimental evidence on the effect of performance-based in-kind incentives for community health teams, and to our knowledge is the first evidence for this specific incentive model. Conventional economic theory suggests that in-kind and group incentives may be relatively low-powered compared to cash or individual incentives, since benefits are shared by team members and may not align with individual worker preferences, and there is a risk of free-riding among team members. Yet our results suggest in-kind incentives for teams led to substantial improvements in performance across multiple domains including community outreach and quality. The incentives even led to changes in patients' behaviors, which are not under the full control of providers, such as timeliness of care and the utilization of health services. Even though all teams (treatment and control) improved their performance over time, those assigned randomly to receive incentives first had a faster rate of improvement, with these gains concentrated amongst the lowest and highest performance teams at baseline.

While outside the scope of this paper, a key pending topic for further research is understanding the mechanisms driving the changes in performance. Qualitative evidence suggests that while the performance reports and certificates created a sense of competition between teams regardless of their experimental assignment, those eligible for the in-kind incentives were able to motivate all team members to go the extra mile and perform additional home visits, work on weekends, and put in additional effort to achieve their targets and develop creative ways to influence patients' behaviors. For instance, some treatment teams came up with ideas such as informing mothers of the market price of micronutrients sachets (which are provided for free by teams) to improve mothers' valuation of them. Moreover, it seems that in-kind incentives were valued by teams more like an award and recognition (particularly since they were announced in front of their peers) rather than for their pure material value, which might elicit greater effort (Kosfeld and Neckerman, 2011). As a team member put it when discussing the in-kind incentives: "*That's like a bonus for what we were already doing. So, let's try harder. Since they are recognizing our work, and they want to recognize it, well, then it was like a motivation, an extra incentive to strive more, right? To work*" (Munar et al, 2018).

We find no evidence that increased performance on incentivized outcomes came at the cost of shifting away effort from non-contracted outcomes[46] or from gaming the system. The latter despite ample room to simply improve reporting since some proxy measures in medical records tended to be under-reported as compared to those from the household survey. Qualitative evidence suggests that a key element to avoid these unintended consequences was that monitoring and supervision by the MoH focused on the full portfolio of services provided by teams and feedback on the performance reports centered on how to change the underlying outcomes rather than the proxies used for measuring performance. In addition, the fact that every verification cycle collected data on both household surveys and medical records might have reduced the temptation of tampering medical records.

All teams regardless of their experimental assignment were part of the target area of Salud Mesoamerica Initiative, which provides results-based funding with incentives for national governments. While there is evidence that this type of results-based funding is effective in improving the production of health care services (Bernal et al, 2018), our results show that adding relatively small in-kind incentives to local providers could further accelerate change on multiple domains. Overall, our results suggest that these types of incentives could be a powerful tool to improve health worker performance and can be a viable alternative to monetary incentives in certain contexts.

[46] At least from those we measure from women in fertile age related to the prevention of chronic conditions.

IX. References

Ashraf, N., Bandiera, O., & Jack, B. K. (2014). No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics*, *120*, 1–17.

Banerjee, A., Deaton, A., & Duflo, E. (2004). Health, health care, and economic development: wealth, health, and health services in rural Rajasthan. *American Economic Review*, 94(2), 326–330.

Barber SL, Bertozzi SM, Gertler PJ (2007) Variations in prenatal care quality for the rural poor in Mexico. *Health Affairs (Millwood)*. May-Jun;26(3): w310-23.

Bareket-Bojmel, L., Hochman, G., & Ariely, D. (2017). It's (Not) All About the Jacksons: Testing Different Types of Short-Term Bonuses in the Field. *Journal of Management*, *43*(2), 534–554.

Basinga, P., Gertler, P. J., Binagwaho, A., Soucat, A. L., Sturdy, J., & Vermeersch, C. M. (2011). Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: An impact evaluation. *The Lancet*, *377*(9775), 1421–1428.

Berendes, S., Heywood, P., Oliver, S., & Garner, P. (2011). Quality of private and public ambulatory health care in low- and middle-income countries: systematic review of comparative studies. *PLoS medicine*.

Bernal, P., Celhay, P., and Martinez, S. (2018). Is Results-Based Aid More Effective than Conventional Aid? Evidence from the Health Sector in El Salvador. Social Protection and Health Division. *IDB Working Paper* No IDB-WP-859.

Bhutta, Z. A., Das, J. K., Bahl, R., Lawn, J. E., Salam, R. A., Paul, V. K., … Walker, N. (2014). Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost? *The Lancet*, *384*(9940), 347–370.

Bonfrer, Igna, Robert Soeters, Ellen van de Poel, Olivier Basenya, Gashubije Longin, Frank van de Looij, and Eddy van Doorslaer. 2013. The effects of performance-based financing on the use and quality of health care in Burundi: an impact evaluation. *The Lancet*, 381: S19.

Boning, B., C. Ichniowski, and K. Shaw (2007). Opportunity Counts: Teams and the Effectiveness of Production Incentives. *Journal of Labor Economics*, 25(4): 613-50

Carroli, G., Villar, J., Piaggio, G., Khan-Neelofur, D., Gülmezoglu, M., Mugford, M., … Bersgjø, P. (2001). WHO systematic review of randomised controlled trials of routine antenatal care. *The Lancet*, *357*(9268), 1565–1570.

Celhay, P., Gertler, P., Giovagnoli, P., Vermeersch, C., (2018). Long-Run Effects of Temporary Incentives on Medical Care Productivity. *American Economic Journal: Applied Economics* (forthcoming).

Chen, C., McElrath, L., Pierce, C. B., Blomquist, J. L., & Handa, V. L. (2014). Concordance Between Hospital Records and Maternal Recall 5–10 Years After Childbirth. *Obstetrics & Gynecology*, *123*.

Clingingsmith, D., Khwaja, A. I., & Kremer, M. (2009). Estimating the Impact of the Hajj: Religion and Tolerance in Islam's Global Gathering [*]. *Quarterly Journal of Economics*, *124*(3), 1133–1170.

Cullen, J. B., & Reback,R.(2006). Tinkering toward Accolades: School Gaming under a Performance Accountability System, in Timothy J. Gronberg and Dennis W. Jansen (Eds.), *Advances in Applied Microeconomics* 14 (2006), 1–34.

Das, J., & Hammer, J. (2004). Strained Mercy: The Quality of Medical Care in Delhi. *Economic And Political Weekly*, 3228(39), 951–965.

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

Das, J., Hammer, J., & Leonard, K. (2008). The quality of medical advice in low-income countries. *The Journal of Economic Perspectives*, 22(2), 93-114.

Das, J., Holla, A., Das, V., Mohanan, M., Chan, B., & Tabak, D. (2012). In Urban And Rural India, A Standardized Patient Study Showed Low Levels Of Provider Training And Huge Quality Gaps. *Health Affairs*, 31(12), 2774-2784.

de Walque, D., Gertler, P., Bautista-Arredondo, S., and Kwan, A., and Vermeersch, C., de Dieu J., Binagwaho, A., and Condo, J (2015). Using Provider Performance Incentives to Increase HIV Testing and Counseling Services in Rwanda. *Journal of Health Economics*. Volume 40, Pages 1-9

Eijkenaar, F., Emmert, M., Scheppach, M., & Schöffski, O. (2013). Effects of pay for performance in health care: A systematic review of systematic reviews. *Health Policy*, *110*(2–3), 115–130.

Friebel, G., Heinz, M., Krueger, M., Zubanov, N., Bandiera, O., Barankay, I., … Wuest, S. (2017). Team Incentives and Performance: Evidence from a Retail Chain. *American Economic Review*, *107*(8), 2168–2203

Gauri, V., Jamison, J. C., Mazar, N., Ozier, O., Raha, S., & Saleh, K. (2018). Motivating Bureaucrats through Social Recognition Evidence from Simultaneous Field Experiments. *Policy Research Working Paper Series, World Bank*, (8473).

Gill, D., Kissová, Z., Lee, J., & Prowse, V. (2018). First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision. *Management Science*, 29(07).

Heyman, J., & Ariely, D. (2004). Effort for Payment. A Tale of Two Markets. *Psychological Science*, *15*(11), 787–793.

Holmstrom, B., & Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, and Organization*, *7*(0), 24–52.

IHME (2011). SM2015-El Salvador Baseline Household Census & Survey Report. *Institute of Health Metrics and Evaluation*.

Ivers, N., Jamtvedt, G., Flottorp, S., Jm, Y., Sd, F., Ma, O. B., … Brien, M. A. O. (2012). Audit and feedback : effects on professional practice and healthcare outcomes (Review). *Cochrane Database of Systematic Reviews (Online)*, (6), CD000259.

Jacob, B. A., & Levitt, S. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating, *Quarterly Journal of Economics* 118(3), 843–877.

Kassebaum, N. J., Bertozzi-Villa, A., Coggeshall, M. S., Shackelford, K. A., Steiner, C., Heuton, K. R., … Lozano, R. (2014). Global, regional, and national levels and causes of maternal mortality during 1990-2013: A systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, *384*(9947), 980–1004.

Kivetz, R., Urminsky, O., & Zheng, Y. (2006). The Goal-Gradient Hypothesis Resurrected: Purchase Acceleration, Illusionary Goal Progress, and Customer Retention. *Journal of Marketing Research*, *43*(1), 39–58.

Kling, J. R., Liebman, J. B., Katz, L. F., & Sanbonmatsu, L. (2004). Moving to Opportunity and Tranquility: Neighborhood Effects on Adult Economic Self-Sufficiency and Health from a Randomized Housing Voucher Experiment. *Working Paper*, (August).

Krakowiak, P., Walker, C., Tancredi, D., & Hertz-Picciotto, I. (2015). Maternal recall versus medical records of metabolic conditions from the prenatal period: A validation study. *Maternal Child Health Journal*, *19*(9), 1922–2013.

Kolstad, J. T. (2013). Information and Quality When Motivation Is Intrinsic. *The American Economic Review*, *103*(7), 2875–2910.

Koretz, D. M. (2002). Limitations in the Use of Achievement Tests as Measures of Educators' Productivity. *Journal of Human Resources* 37(4), 752–777.

Kosfeld, M., & Neckermann, S. (2011). Getting More Work for Nothing? Symbolic Awards and Worker Performance. *American Economic Journal: Microeconomics*, *3*(August), 86–99

Kube, B. S., Maréchal, M. A., & Puppe, C. (2012). The Currency of Reciprocity : Gift Exchange in the Workplace. *American Economic Review*, *102*, 1644–1662.

Lacetera, B. N., Macis, M., & Slonim, R. (2012). Will There Be Blood? Incentives and Displacement Effects in Pro-Social Behavior. *American Economic Journal: Economic Policy*,

4(1), 186–223.

McConnell, B., & Vera-Hérnandez, M. (2015). Going beyond simple sample size calculations: a practitioner's guide. *Institute of Fiscal Studies,* Working Paper No. W15/17.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, *99*(2), 210–221.

Mendelson, A., Kondo, K., Damberg, C., Low, A., Motuapuaka, M., Freeman, M., … Kansagara, D. (2017). The effects of pay-for-performance programs on health, health care use, and processes of care: A systematic review. *Annals of Internal Medicine*, *166*(5), 341–353.

Miller, G., & Babiarz, K. (2014). Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. *Encyclopedia of Health Economics*, 457–466.

MINSAL (2010). *La Reforma de Salud en El Salvador*. San Salvador, El Salvador.

MINSAL (2011). *Lineamientos Operativos para el Desarrollo de Actividades en los ECOS Familiares y ECOS Especializados*. San Salvador, El Salvador.

Moller, A. B., Petzold, M., Chou, D., & Say, L. (2017). Early antenatal care visit: a systematic analysis of regional and global levels and trends of coverage from 1990 to 2013. *The Lancet Global Health*, *5*(10), e977–e983.

Mokdad, A. H., Colson, K. E., Zúñiga-Brenes, P., Ríos-Zertuche, D., Palmisano, E. B., Alfaro-Porras, E., … Regalia, F. (2015). Salud Mesoamérica 2015 Initiative: design, implementation, and baseline findings. *Population Health Metrics*, *13*(1), 3.

Munar, W., S. Wahid, S., Mookherji, S., Innocenti, C., & Curry, L. (2018). Team- and individual-level motivation in complex primary care system change: A realist evaluation of the Salud Mesoamerica Initiative in El Salvador. *Gates Open Research* 2: 55.

Lacetera, N., & Macis, M. (2010). Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior and Organization*, *76*(2), 225–237.

Lazzerini, M., & Wanzira, H. (2016). Oral zinc for treating diarrhoea in children (Review). *Cochrane Database of Systematic Reviews*, (12), 1–133.

Lim, S. S., Stein, D. B., Charrow, A., & Murray, C. J. L. (2008). Tracking progress towards universal childhood immunisation and the impact of global initiatives: a systematic analysis of three-dose diphtheria, tetanus, and pertussis immunisation coverage. *The Lancet*, *372*(9655), 2031–2046.

List, J. A., Sadoff, S., & Wagner, M. (2010). So you want to run an experiment, now what? Some Simple Rules of Thumb for Optimal Experimental Design. *National Bureau of Economic Research Working Paper Series*, *No. 15701*.

Prendergast, C. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature*, *37*(1), 7–63.

Scott, J. (2004.). The Benefits of Tangible Non-Monetary Incentives. *University of Chicago, Graduate School of Business*.

Scott, A., Sivey, P., Ait Ouakrim, D., Willenberg, L., Naccarella, L., Furler, J., & Young, D. (2011). The effect of financial incentives on the quality of health care provided by primary care physicians. In A. Scott (Ed.), *Cochrane Database of Systematic Reviews* (p. CD008451). Chichester, UK: John Wiley & Sons, Ltd.

Spenkuch, J. L. (2012). Moral hazard and selection among the poor: Evidence from a randomized experiment. Journal of Health Economics, 31(1), 72–85.

Shen, G. C., Nguyen, H. T. H., Das, A., Sachingongu, N., Chansa, C., Qamruddin, J., & Friedman, J. (2017). Incentives to change: effects of performance-based financing on health workers in Zambia. *Human Resources for Health*, 15, 20.

Yawn, B. P., Suman, V. J., & Jacobsen, S. J. (1998). Maternal recall of distant pregnancy events. *Journal of Clinical Epidemiology*, *51*(5), 399–405.

Waldfogel, J. (1996). The Deadweight Loss of Christmas: Reply. *The American Economic Review*, *86*(5), 1306–1308.

WHO. (2007). *Standards for Maternal and Neonatal Care*. Geneva, Switerzland: WHO.

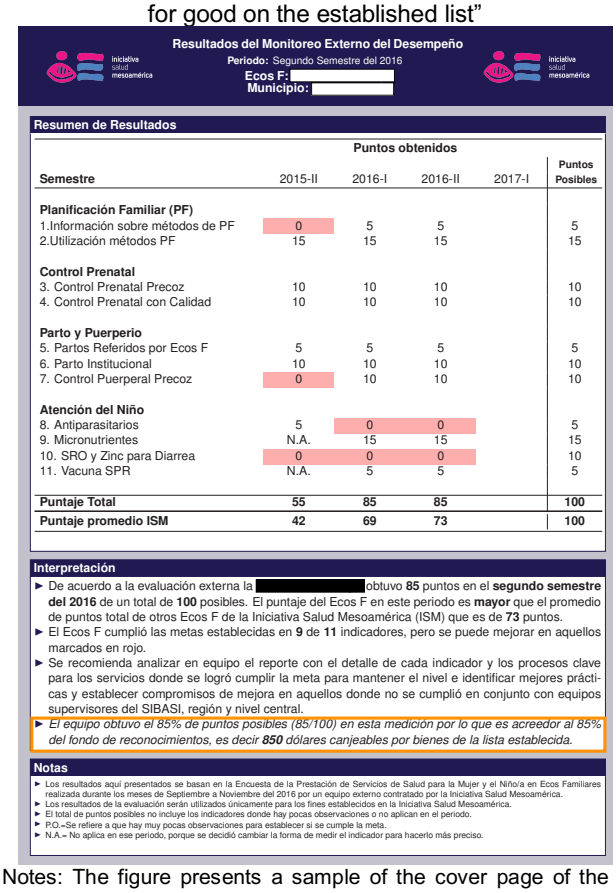
WHO. (2013). *Postnatal care of the mother and newborn 2013*. WHO. Geneva, Switerzland: WHO.

WHO. (2016). *WHO Recommendation on Antenatal care for positive pregnancy experience*. Geneva, Switerzland: WHO.

Figure 1. Cover page of team performance report by treatment status

Treatment

The highlighted section reads:
"The team obtained 85% of the possible points (85/100) in this cycle and has therefore obtained 85% of the performance fund, that is 850 dollars interchangeable for good on the established list"



Resultados del Monitoreo Externo del Desempeño

Periodo: Segundo Semestre del 2016

Ecos F: [redacted]

Municipio: [redacted]

Resumen de Resultados

Semestre	Puntos obtenidos 2015-II	2016-I	2016-II	2017-I	Puntos Posibles
Planificación Familiar (PF)					
1.Información sobre métodos de PF	0	5	5		5
2.Utilización métodos PF	15	15	15		15
Control Prenatal					
3. Control Prenatal Precoz	10	10	10		10
4. Control Prenatal con Calidad	10	10	10		10
Parto y Puerperio					
5. Partos Referidos por Ecos F	5	5	5		5
6. Parto Institucional	10	10	10		10
7. Control Puerperal Precoz	0	10	10		10
Atención del Niño					
8. Antiparasitarios	5	0	0		5
9. Micronutrientes	N.A.	15	15		15
10. SRO y Zinc para Diarrea	0	0	0		10
11. Vacuna SPR	N.A.	5	5		5
Puntaje Total	**55**	**85**	**85**		**100**
Puntaje promedio ISM	**42**	**69**	**73**		**100**

Interpretación

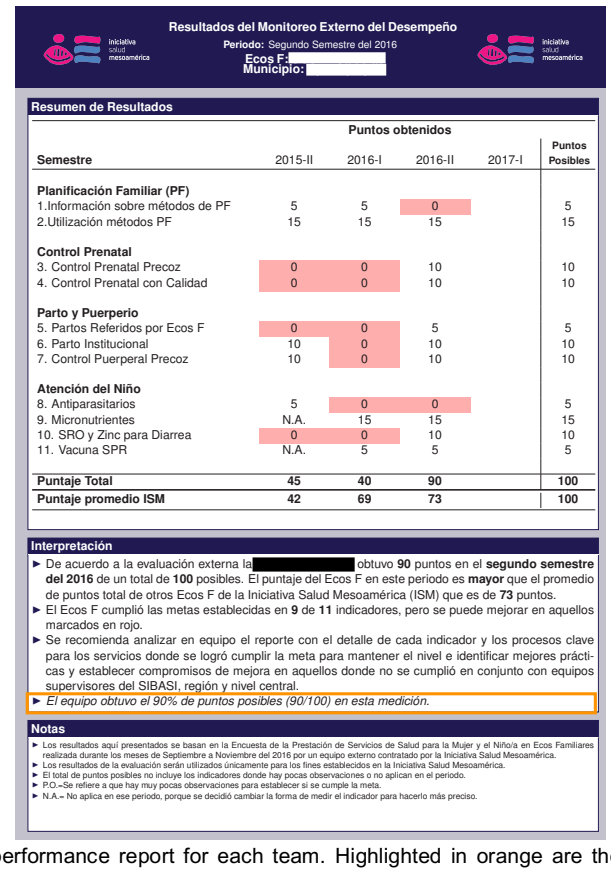
- De acuerdo a la evaluación externa la [redacted] obtuvo **85** puntos en el **segundo semestre del 2016** de un total de **100** posibles. El puntaje del Ecos F en este periodo es **mayor** que el promedio de puntos total de otros Ecos F de la Iniciativa Salud Mesoamérica (ISM) que es de **73** puntos.
- El Ecos F cumplió las metas establecidas en **9** de **11** indicadores, pero se puede mejorar en aquellos marcados en rojo.
- Se recomienda analizar en equipo el reporte con el detalle de cada indicador y los procesos clave para los servicios donde se logró cumplir la meta para mantener el nivel e identificar mejores prácticas y establecer compromisos de mejora en aquellos donde no se cumplió en conjunto con equipos supervisores del SIBASI, región y nivel central.
- *El equipo obtuvo el 85% de puntos posibles (85/100) en esta medición por lo que es acreedor al 85% del fondo de reconocimientos, es decir **850** dólares canjeables por bienes de la lista establecida.*

Notas

- Los resultados aquí presentados se basan en la Encuesta de la Prestación de Servicios de Salud para la Mujer y el Niño/a en Ecos Familiares realizada durante los meses de Septiembre a Noviembre del 2016 por un equipo externo contratado por la Iniciativa Salud Mesoamérica.
- Los resultados de la evaluación serán utilizados únicamente para los fines establecidos en la Iniciativa Salud Mesoamérica.
- El total de puntos posibles no incluye los indicadores donde hay pocas observaciones o no aplican en el periodo.
- P.O.=Se refiere a que hay muy pocas observaciones para establecer si se cumple la meta.
- N.A.= No aplica en ese periodo, porque se decidió cambiar la forma de medir el indicador para hacerlo más preciso.

Control

The highlighted section reads:
"The team obtained 90% of the possible points (90/100) in this cycle"



Resultados del Monitoreo Externo del Desempeño

Periodo: Segundo Semestre del 2016

Ecos F: [redacted]

Municipio: [redacted]

Resumen de Resultados

Semestre	Puntos obtenidos 2015-II	2016-I	2016-II	2017-I	Puntos Posibles
Planificación Familiar (PF)					
1.Información sobre métodos de PF	5	5	0		5
2.Utilización métodos PF	15	15	15		15
Control Prenatal					
3. Control Prenatal Precoz	0	0	10		10
4. Control Prenatal con Calidad	0	0	10		10
Parto y Puerperio					
5. Partos Referidos por Ecos F	0	0	5		5
6. Parto Institucional	10	0	10		10
7. Control Puerperal Precoz	10	0	10		10
Atención del Niño					
8. Antiparasitarios	5	0	0		5
9. Micronutrientes	N.A.	15	15		15
10. SRO y Zinc para Diarrea	0	0	10		10
11. Vacuna SPR	N.A.	5	5		5
Puntaje Total	**45**	**40**	**90**		**100**
Puntaje promedio ISM	**42**	**69**	**73**		**100**

Interpretación

- De acuerdo a la evaluación externa la [redacted] obtuvo **90** puntos en el **segundo semestre del 2016** de un total de **100** posibles. El puntaje del Ecos F en este periodo es **mayor** que el promedio de puntos total de otros Ecos F de la Iniciativa Salud Mesoamérica (ISM) que es de **73** puntos.
- El Ecos F cumplió las metas establecidas en **9** de **11** indicadores, pero se puede mejorar en aquellos marcados en rojo.
- Se recomienda analizar en equipo el reporte con el detalle de cada indicador y los procesos clave para los servicios donde se logró cumplir la meta para mantener el nivel e identificar mejores prácticas y establecer compromisos de mejora en aquellos donde no se cumplió en conjunto con equipos supervisores del SIBASI, región y nivel central.
- *El equipo obtuvo el 90% de puntos posibles (90/100) en esta medición.*

Notas

- Los resultados aquí presentados se basan en la Encuesta de la Prestación de Servicios de Salud para la Mujer y el Niño/a en Ecos Familiares realizada durante los meses de Septiembre a Noviembre del 2016 por un equipo externo contratado por la Iniciativa Salud Mesoamérica.
- Los resultados de la evaluación serán utilizados únicamente para los fines establecidos en la Iniciativa Salud Mesoamérica.
- El total de puntos posibles no incluye los indicadores donde hay pocas observaciones o no aplican en el periodo.
- P.O.=Se refiere a que hay muy pocas observaciones para establecer si se cumple la meta.
- N.A.= No aplica en ese periodo, porque se decidió cambiar la forma de medir el indicador para hacerlo más preciso.

Notes: The figure presents a sample of the cover page of the performance report for each team. Highlighted in orange are the differences between the two. The identity of the teams was masked for confidentiality purposes.

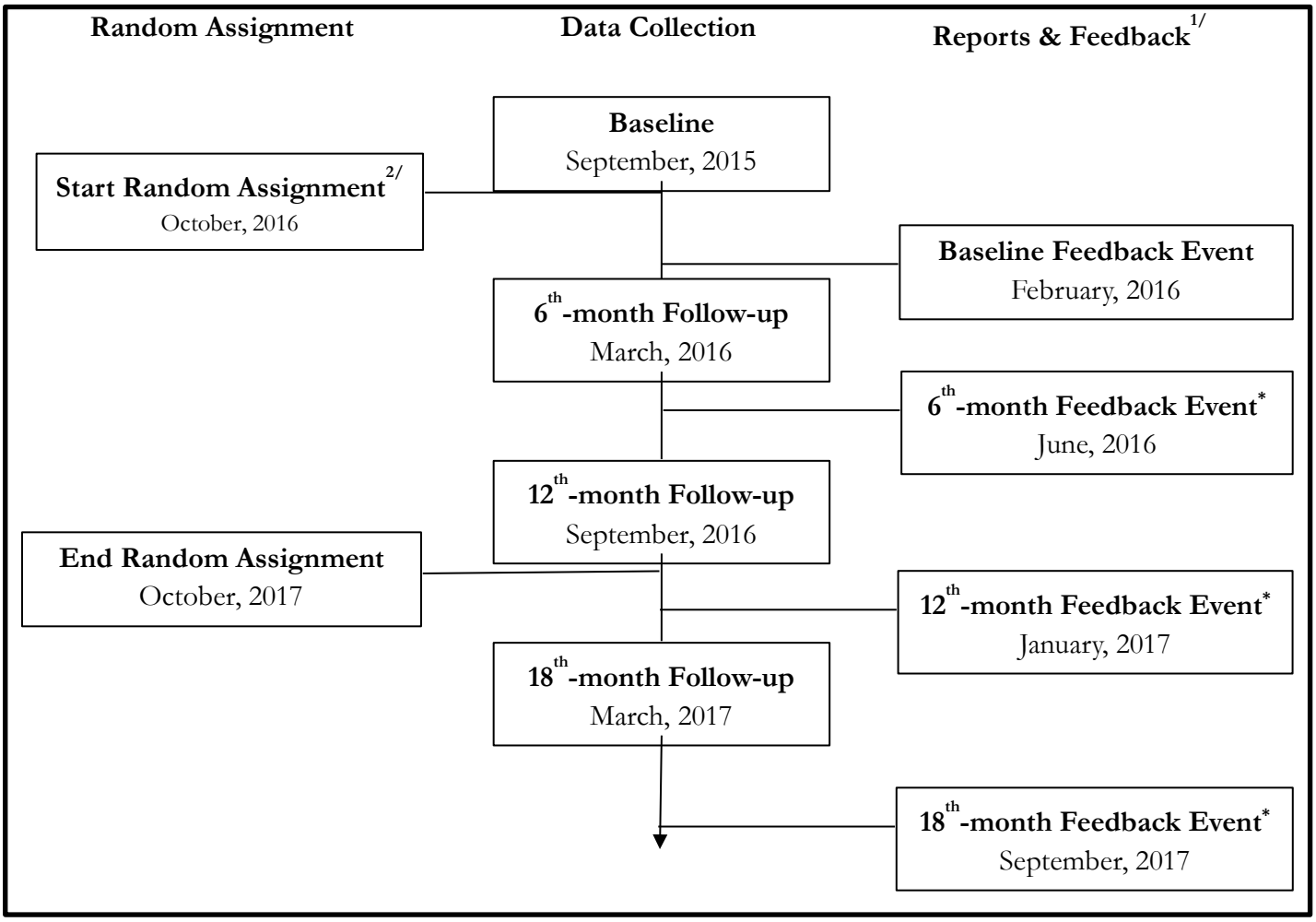
Figure 2. Team performance certificates by treatment status

Treatment

Control

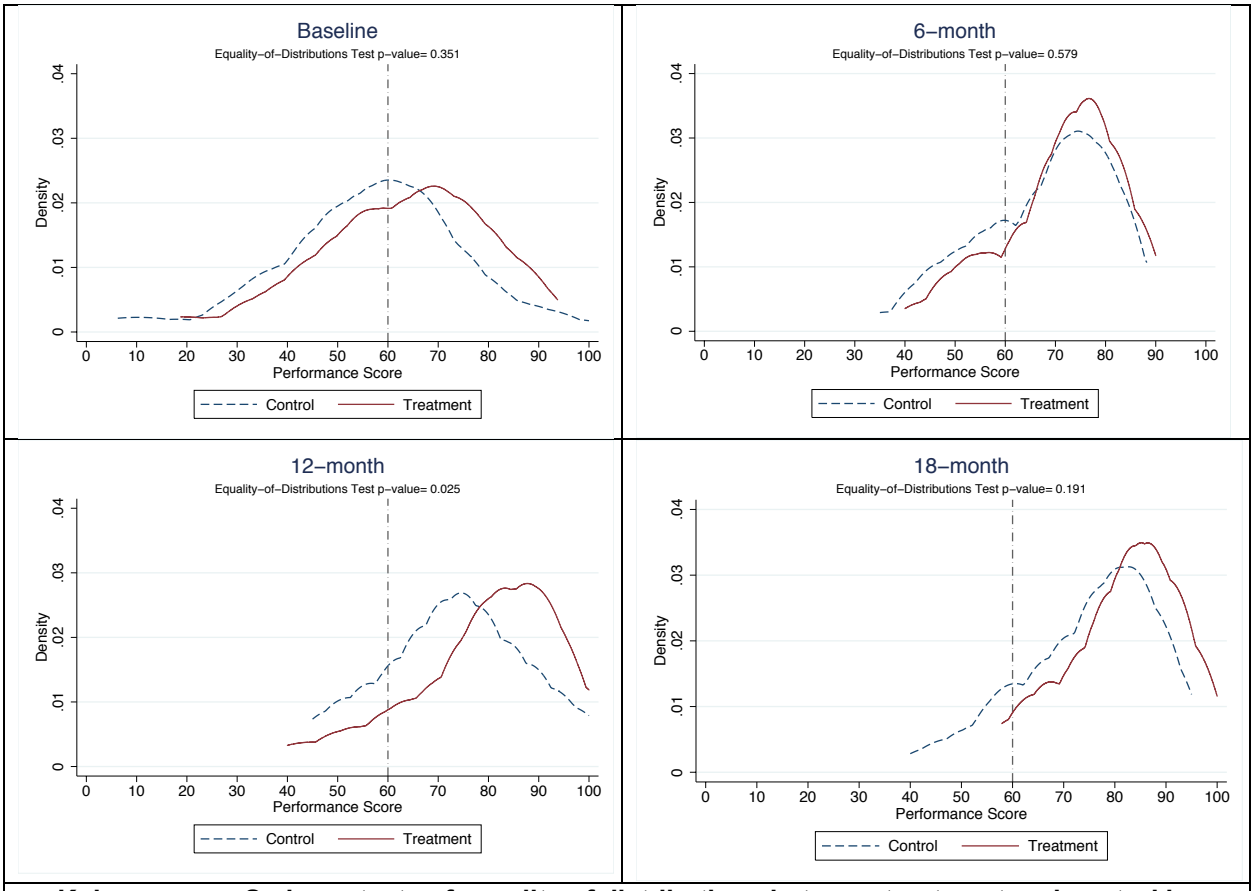
Notes: The figure presents a sample of the certificates of performance that were provided to teams that obtained 60% of more of the total points possible. Highlighted in orange are the differences between the two, which is just the amount obtained in treatment teams and the title of the certificate (a voucher in the case of treatment and just certificate for control). The identity of the teams was masked for confidentiality purposes.

Figure 3. Timeline of data collection and key events



Notes: Dates of data collection refer to the start date. Each data collection lasted about two months and a half. The random assignment of teams into treatment and control ended after the 12th month follow-up was finalized.
1/ Reports and feedback refer to the event in which teams received their performance report, the presentation of overall results as well as their performance certificates. The latter was only starting in the 6th month follow-up.
2/ In the same date of the random assignment the incentive scheme was presented to teams.
* In these events, teams also received performance certificates.

Figure 4. Distribution of team performance scores by treatment status and time period

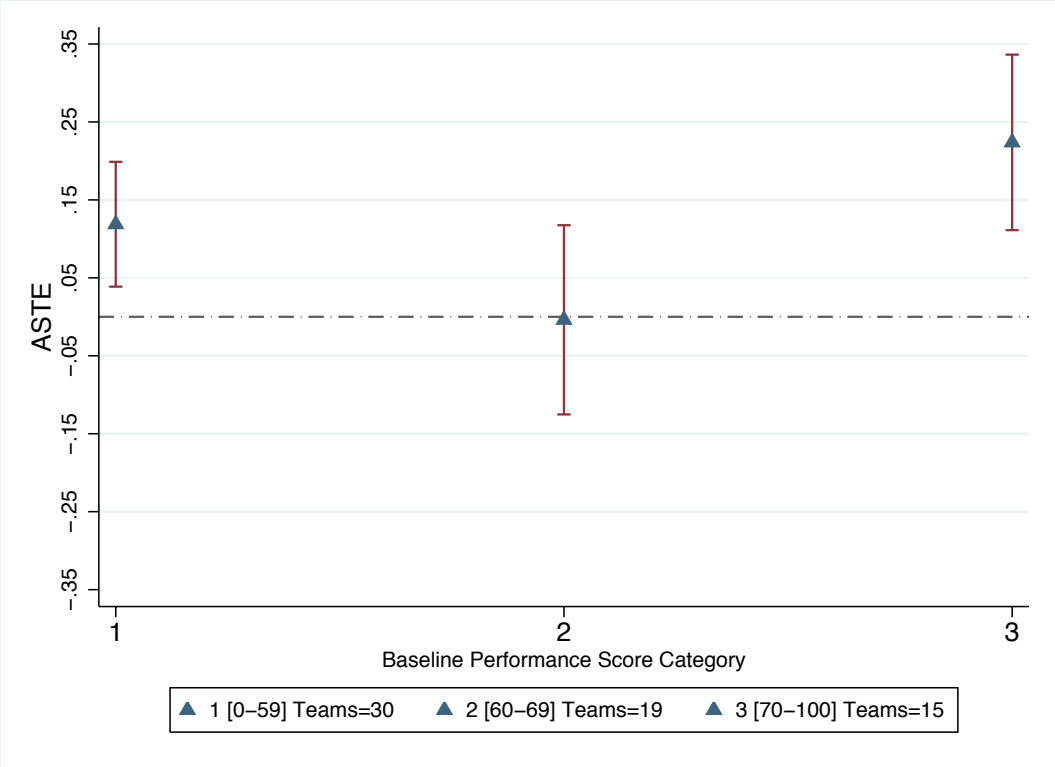


Kolmogorov – Smirnov tests of equality of distributions between treatment and control by period

Period	P-Value
Baseline	0.351
6-month	0.579
12-month	0.025
18-month	0.191

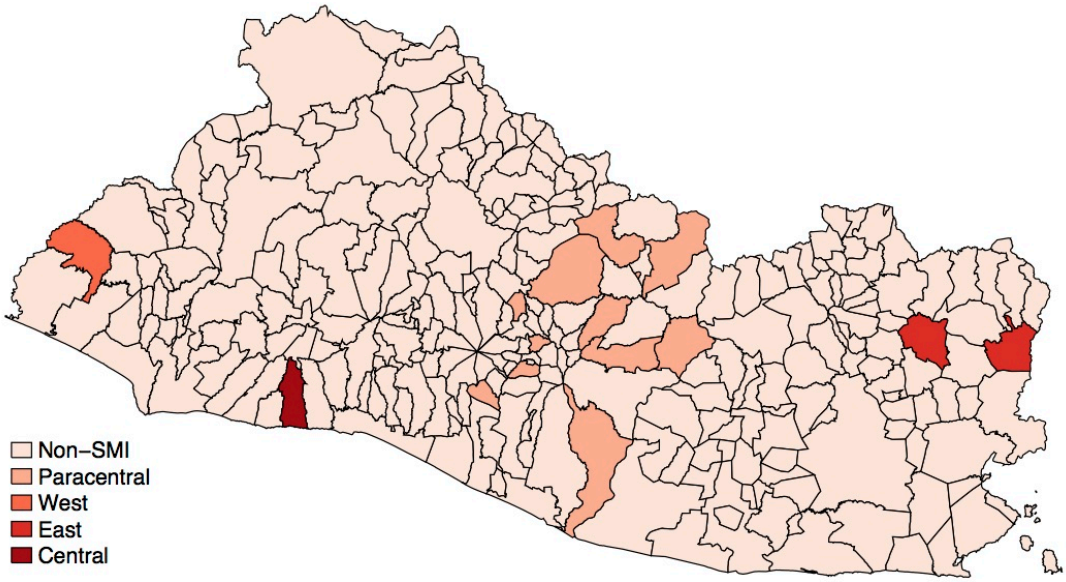
Notes: The graphs present the kernel density of the distribution of the performance score of all teams by treatment assignment. The score is calculated as explained in Section IIIB and was the one informed to teams during the period. The exact p-value of the Kolmogorov-Smirnov test of equality of the treatment and control distributions in each period is presented on the second column. During the 6th and 12th month follow-ups the experimental only treatment teams received incentives. At the 18th month follow-up both treatment and control teams were eligible for incentives.

Figure 5. Average standardized treatment effect of all contracted outcomes by category of baseline performance score



Notes: Categories of baseline performance score are constructed based on the sliding scale for incentives. The first category includes teams with a baseline performance below the threshold for incentives, i.e., 60. The second category includes teams just at or above the incentive threshold, i.e. 60 to 69 inclusive. The third category includes those teams with a baseline performance score of 70 or higher. The number of teams in each category is displayed in the figure label. Average standardized treatment effect of all eleven contracted outcomes, obtained as described in Section V, using the model that pools both the 6th and 12th month follow-ups and with block effects and the baseline value of the outcome of interest in each regression. Confidence interval at 95% constructed from standard errors clustered at the team level, displayed on the red lines around each estimate.

Table 1. Municipalities targeted by Salud Mesoamerica Initiative and number of community health teams per municipality



Geographical Region	SIBASI	Municipality	Community Health Teams in 2015		
			Rural	Urban	Total
Central	La Libertad	Chiltiupán	2	1	3
West	Ahuachapán	Tacuba	9	1	10
East	La Unión	El Sauce	2	1	3
East	Morazán	Sociedad	3	1	4
Paracentral	Cabañas	Ilobasco	12	1	13
Paracentral	Cabañas	Sensuntepeque	7	2	9
Paracentral	Cuscatlán	Monte San Juan	2	1	3
Paracentral	Cuscatlán	San Cristóbal	2	1	3
Paracentral	La Paz	San Antonio Masahuat	1	1	2
Paracentral	La Paz	Santa María Ostuma	2	1	3
Paracentral	San Vicente	Apasteque	6	1	7
Paracentral	San Vicente	San Esteban Catarina	1	1	2
Paracentral	San Vicente	San Ildefonso	2	1	3
Paracentral	San Vicente	Tecoluca	7	3	10
		Total	**58**	**17**	**75**

Notes: The map illustrates all the municipalities of El Salvador and highlights those municipalities targeted by SMI according to their geographical region. SIBASI (*Sistema Básico de Salud Integral*) is the administrative delimitation of the health system that has under its influence the primary and secondary care of a group of municipalities. SIBASIs are independent of each other, but they all depend from the central level. The number of community health teams is that reported by the MoH in 2015.

Table 2. Indicators used to evaluate team performance, targets, and source of verification

Category	Indicator	Target	Points	Source
Outreach	Women 15 to 49 receiving information on modern family planning methods by health personnel in the last six months	80%	5	Household survey
Outreach	Women 15 to 49 with children less than five with knowledge of treatment of diarrhea with oral rehydration salts and zinc at the time of the survey	50%	10	Household survey
Quality	Prenatal care according to national clinical guidelines for women with a delivery in the last four months	80%	10	Medical Records
Quality	Reference to institutional delivery in birth plan for women with a delivery in the last four months	100%	5	Medical Records
Timeliness	First prenatal care visit prior to 12 weeks of gestation for women with pregnancies that reached three months of gestation in the last six months	80%	10	Medical Records
Timeliness	Postpartum care within a week from delivery by health personnel for women with a delivery in the last four months	92%	10	Medical Records
Utilization	Women 15 to 49 in need of family planning using a modern family planning method at the time of the survey	61%	15	Household survey
Utilization	Institutional Delivery for women with a delivery in the last four months	94%	10	Medical Records
Utilization*	Prescription of micronutrients for children 6 to 23 months old in the last six months[1/]	80%	15	Medical Records
Utilization	Consumption of two deworming pills in the last six months by children 18 to 59 months old[2/]	80%	5	Household survey
Utilization	Children 12 to 23 months with vaccination of measles, mumps, and rubella (MMR)	80%	5	Medical Records
	Total points		**100**	

Notes: The target is the one set by the MoH for the indicator in order to receive the points towards the performance score. The points are the ones awarded to the performance score if the target was met. The source describes the verification source for the target. The detailed definition of each indicator can be found in Appendix 2. Prescription of micronutrients is included in the utilization category since the underlying outcome of interest was to improve the consumption of micronutrients sachets, but no feasible indicators could be used other than this.

1/ This definition was changed to micronutrients provided to children 6 to 23 months in the last six months after baseline since it was considered a better proxy. The definition provided in the table was the one used in the analysis since it had baseline data.

2/ At baseline the source of verification was the clinical record of children, but was changed to vaccination records after baseline since it was considered more comprehensive.

Table 3. Sliding scale to determine the amount of the incentive obtained in each cycle

Points obtained[1/]	% of Incentive	Incentive amount (USD)
90-100	100%	$1000
80-89	85%	$850
70-79	75%	$750
60-69	65%	$650
59 or less	0%	$0

[1/] The points obtained are calculated by dividing the total points obtained in each external verification cycle divided by the total number of points possible, i.e. excluding those indicators with no data and multiplying it times 100.

Table 4. Community teams by blocks, administrative regions and random assignment

Adminstrative Region (SIBASI)	Rural Teams Block 1 C	Rural Teams Block 1 T	Rural Teams Block 2 C	Rural Teams Block 2 T	Urban Teams Block 3 C	Urban Teams Block 3 T	Total C	Total T	Total
Ahuachapán	3	6			1		4	6	10
Cabañas			10	9	1	2	11	11	22
Cuscatlán			2	2	1	1	3	3	6
La Libertad	1	1				1	1	2	3
La Paz			1	2	1	1	2	3	5
La Unión			1	1		1	1	2	3
Morazán	2	1				1	2	2	4
San Vicente	9	7			4	2	13	9	22
Total	**15**	**15**	**14**	**14**	**8**	**9**	**37**	**38**	**75**

Notes: The number in each cell is the number of community health teams. Random assignment was done within each block. C=Experimental Control, T=Experimental Treatment.

Table 5. Balance test on catchment area and community health team characteristics

	Control Mean	Mean Difference (Treatment-Control)	Std. Error	P-Value	N
	(1)	(2)	(3)	(4)	(5)
Panel A. Dwelling Characteristics					
Floor of durable material	0.6774	0.0313	(0.052)	[0.551]	1,054
Ceiling of durable material	0.4307	0.0029	(0.077)	[0.970]	1,054
Wall of durable material	0.8226	0.0333	(0.048)	[0.487]	1,054
Electricity	0.7732	-0.0337	(0.040)	[0.401]	1,054
Phone	0.0882	0.0224	(0.017)	[0.197]	1,054
Toilet	0.2448	-0.0753*	(0.044)	[0.093]	1,054
Bono Comunidades Solidarias	0.2922	0.0006	(0.067)	[0.993]	1,054
Dwelling with Children Less Than 2	0.2619	0.0188	(0.027)	[0.485]	1,054
Panel B. Women Characteristics					
Age	30.3	-0.1279	(0.630)	[0.840]	1,095
Single	0.3403	-0.0444	(0.035)	[0.206]	1,052
No Health Insurance	0.9382	-0.0104	(0.017)	[0.549]	1,051
Elementary Education or Less	0.8184	0.0221	(0.031)	[0.475]	1,052
Panel C. Team Characteristics					
Personnel					
Physicians	1.23	-0.1042	(0.194)	[0.594]	64
Professional Nurses	0.98	0.0075	(0.088)	[0.933]	64
Auxiliary Nurses	1.16	-0.3484	(0.269)	[0.200]	64
Community Health Workers	2.97	-0.2079	(0.234)	[0.377]	63
Multi-purpose personnel	0.73	-0.0102	(0.114)	[0.929]	64
Administrative Region (SIBASI)					
Ahuachapán	0.1250	0.0416	(0.080)	[0.606]	64
Cabañas	0.2813	-0.0300	(0.087)	[0.730]	64
Cuscatlán	0.0938	0.0015	(0.073)	[0.984]	64
La Libertad	0.0469	0.0222	(0.052)	[0.671]	64
La Paz	0.0469	0.0323	(0.052)	[0.535]	64
La Unión	0.0469	0.0323	(0.052)	[0.535]	64
Morazán	0.0625	-0.0118	(0.061)	[0.847]	64
San Vicente	0.2969	-0.0881	(0.101)	[0.386]	64

Notes: Sample of analysis is that of 64 community health teams. Column (2) presents the difference between treatment and control groups controlling for blocks. Standard errors clustered at the team level are presented on Column (3). Durable materials include concrete, brick, adobe, concrete blocks, and tiles. Toilet refers to those connected to sewage or a septic tank. *Bono Comunidades Solidarias* is a conditional cash transfer program in El Salvador. Panel A and B present characteristics measured from household surveys at baseline. Panel C presents community health team characteristics at end line (no baseline data are available for these variables).

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6. Treatment Effect on Community Outreach

	Control Mean	Baseline	Post-treatment (No controls)	Post-treatment (Controlling for Baseline)
	(1)	(2)	(3)	(4)
Panel A. ASTE				
ASTE Community Outreach		0.0358	0.1763**	0.1748**
		(0.069)	(0.071)	(0.069)
p-value		[0.603]	[0.013]	[0.011]
Panel B. Treatment effect on individual indicators				
Information on modern family planning (HS)	0.5038	-0.0002	0.0567	0.0580*
		(0.046)	(0.035)	(0.035)
p-value		[0.996]	[0.109]	[0.098]
N		592	1,761	1,761
Knowledge of treatment of diarrhea with ORS & Zinc (HS)	0.0708	0.0185	0.0804**	0.0787**
		(0.030)	(0.037)	(0.035)
p-value		[0.540]	[0.033]	[0.027]
N		552	1,630	1,630

Notes: The sample of analysis is of the 64 teams with data on all waves. Panel A, presents the estimate of the average standardized treatment effect (ASTE) as described in equation (2) using the two individual outcomes presented on Panel (B). Information on modern family planning refers to women age 15 to 49 that received information by health personnel on the last six months on modern family planning methods. Knowledge of treatment of diarrhea with ORS & Zinc refers to women with children less than five that mention Oral rehydration salts and Zinc as the treatments to provide whenever their child has diarrhea. Both outcomes are measured from household surveys (HS). Column (1) presents the control mean at baseline. Column (2) presents the estimate of δ from baseline on the outcome of interest using equation (1), but excluding any baseline covariates. Column (3) presents the estimate of δ using the pooled 6th and 12th month follow-up data excluding baseline covariates. Column (4) presents the same estimate as (3) but includes as baseline covariate the mean team level outcome of interest at baseline. Standard errors are clustered at the team level and are presented in parenthesis. P-values are included in square brackets.
* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 7. Treatment Effect on Quality of Care

	Control Mean	Baseline	Post-treatment (No controls)	Post-treatment (Controlling for Baseline)
	(1)	(2)	(3)	(4)
Panel A. ASTE				
ASTE Quality of Care		-0.1361	0.1349**	0.1427***
		(0.095)	(0.058)	(0.054)
p-value		[0.152]	[0.020]	[0.008]
Panel B. Treatment effect on individual indicators				
Quality prenatal care (MR)	0.5870	-0.1116*	0.0754**	0.0843**
		(0.057)	(0.035)	(0.033)
p-value		[0.057]	[0.034]	[0.013]
N		639	1,264	1,264
Reference to institutional delivery (MR)	0.8727	-0.0153	0.0212	0.0203
		(0.044)	(0.015)	(0.014)
p-value		[0.728]	[0.173]	[0.157]
N		638	1,264	1,264

Notes: The sample of analysis is of the 64 teams with data on all waves. Panel A, presents the estimate of the average standardized treatment effect (ASTE) as described in equation (2) using the two individual outcomes presented on Panel (B). Quality of prenatal care is an indicator equal to one if all the prenatal care visits that a pregnant woman received in the facility included all the clinical measures and lab exams required by national clinical guidelines. These vary by gestational age, but include weight, blood pressure, fundal height (after 14 weeks of gestation), fetal heart rate and fetal movements (the last two after measured after 20 weeks of gestation. In addition, lab tests of glucose, HIV, hemoglobin and urine should be done at least once during pregnancy. Reference to institutional delivery is related to having provided counselling for institutional delivery during prenatal care visits through a delivery plan done in conjunction with the woman. Both outcomes are measured from medical records (MR). Column (1) presents the control mean at baseline. Column (2) presents the estimate of δ from baseline on the outcome of interest using equation (1), but excluding any baseline covariates. Column (3) presents the estimate of δ using the pooled 6th and 12th month follow-up data excluding baseline covariates. Column (4) presents the same estimate as (3) but includes as baseline covariate the mean team level outcome of interest at baseline. Standard errors are clustered at the team level and are presented in parenthesis. P-values are included in square brackets. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 8. Treatment Effect on Timeliness of Care

		Control Mean	Baseline	Post-treatment (No controls)	Post-treatment (Controlling for Baseline)
		(1)	(2)	(3)	(4)
Panel A. ASTE					
ASTE Timeliness of care			0.0954	0.1203**	0.1021*
			(0.089)	(0.057)	(0.052)
	p-value		[0.286]	[0.035]	[0.051]
Panel B. Treatment effect on individual indicators					
Timely Prenatal Care (MR)		0.7264	0.0276	0.0256	0.0187
			(0.036)	(0.028)	(0.025)
	p-value		[0.452]	[0.360]	[0.462]
	N		908	1,849	1,849
Timely Post-natal Care (MR)		0.5807	0.0638	0.0889**	0.0808**
			(0.069)	(0.040)	(0.039)
	p-value		[0.359]	[0.032]	[0.042]
	N		638	1,264	1,264

Notes: The sample of analysis is of the 64 teams with data on all waves. Panel A, presents the estimate of the average standardized treatment effect (ASTE) as described in equation (2) using the two individual outcomes presented on Panel (B). Timely prenatal care is an indicator equal to one if the first prenatal care visit at the facility occurred during the first trimester of pregnancy as measured from the date of the last menstrual period and the date of the first prenatal care visit. Timely post-natal care is an indicator equal to one if the women received postnatal care in the first week after delivery as measured from the date of birth and the date of the first postnatal care visit. Both outcomes are measured from medical records (MR). Column (1) presents the control mean at baseline. Column (2) presents the estimate of δ from baseline on the outcome of interest using equation (1), but excluding any baseline covariates. Column (3) presents the estimate of δ using the pooled 6^{th} and 12^{th} month follow-up data excluding baseline covariates. Column (4) presents the same estimate as (3) but includes as baseline covariate the mean team level outcome of interest at baseline. Standard errors are clustered at the team level and are presented in parenthesis. P-values are included in square brackets.
* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 9. Treatment Effect on Utilization

	Control Mean	Baseline	Post-treatment (No controls)	Post-treatment (Controlling for Baseline)
	(1)	(2)	(3)	(4)
Panel A. ASTE				
ASTE Utilization		-0.0323	0.0897**	0.0958**
		(0.065)	(0.041)	(0.040)
p-value		[0.620]	[0.028]	[0.015]
Panel B. Treatment effect on individual indicators				
Use of modern family planning methods (HS)	0.7444	0.0096	0.0267	0.0340
		(0.048)	(0.030)	(0.025)
p-value		[0.843]	[0.380]	[0.184]
N		592	1,761	1,761
Institutional Delivery (MR)	0.7857	-0.0223	0.0486*	0.0476*
		(0.067)	(0.026)	(0.026)
p-value		[0.741]	[0.067]	[0.068]
N		638	1,264	1,264
Micronutrients for children (MR)	0.8791	-0.0239	0.0299	0.0320
		(0.036)	(0.027)	(0.027)
p-value		[0.512]	[0.279]	[0.240]
N		814	2,279	2,279
Deworming pills consumption (HS)	0.4084	0.0422	0.0535*	0.0450
		(0.056)	(0.032)	(0.031)
p-value		[0.455]	[0.098]	[0.148]
N		431	1,431	1,431
MMR Vaccination (MR)	0.6448	-0.0681	0.0372	0.0518
		(0.060)	(0.035)	(0.035)
p-value		[0.259]	[0.294]	[0.140]
N		814	2,279	2,279

Notes: The sample of analysis is of the 64 teams with data on all waves. Panel A, presents the estimate of the average standardized treatment effect (ASTE) as described in equation (2) using the five individual outcomes presented on Panel (B). HS refers to an outcome that is measured from a household survey. MR refers to an outcome measured from medical records. Use of family planning refers to women 15 to 49 in need of family planning that are using a modern family planning method at the time of the survey. Institutional delivery is an indicator equal to one if the women medical record has evidence of an institutional delivery. Micronutrients for children refers to indication of micronutrients to children 6 to 23 months old in the medical records. Deworming pills consumption refers to children 12 to 59 months old that were administered two doses of deworming in the last 12 months according to their mother. MMR refers to measles, mumps, and rubella vaccination received by children 12 to 23 months old according to their medical record. Both outcomes are measured from medical records. Column (1) presents the control mean at baseline. Column (2) presents the estimate of δ from baseline on the outcome of interest using equation (1), but excluding any baseline covariates. Column (3) presents the estimate of δ using the pooled 6th and 12th month follow-up data excluding baseline covariates. Column (4) presents the same estimate as (3) but includes as baseline covariate the mean team level outcome of interest at baseline. Standard errors are clustered at the team level and are presented in parenthesis. P-values are included in square brackets. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 10. Treatment effect on non-contracted outcomes

		Control Mean	Baseline	Post-treatment (No controls)	Post-treatment (Controlling for Baseline)
		(1)	(2)	(3)	(4)
Panel A. ASTE					
ASTE Non-Contracted Outcomes			0.0937*	0.0379	0.0215
			(0.053)	(0.035)	(0.031)
	p-value		[0.079]	[0.274]	[0.485]
Panel B. Treatment effect on individual indicators					
Diabetes detection (HS)		0.1743	0.0083	0.0115	0.0106
			(0.031)	(0.017)	(0.017)
	p-value		[0.792]	[0.495]	[0.525]
	N		1,096	3,166	3,166
Hypertension Detection (HS)		0.4661	0.0756**	0.0324	0.0143
			(0.033)	(0.025)	(0.024)
	p-value		[0.024]	[0.206]	[0.548]
	N		1,098	3,170	3,170
Cytology performed (HS)		0.2789	0.0485	0.0087	0.0029
			(0.030)	(0.026)	(0.024)
	p-value		[0.107]	[0.740]	[0.905]
	N		1,097	3,174	3,174

Notes: The sample of analysis is of the 64 teams with data on all waves. Panel A, presents the estimate of the average standardized treatment effect (ASTE) as described in equation (2) using the three individual outcomes presented on Panel (B). All outcomes are measured from the household survey for women 15 to 49 years old and refer to whether the woman went to the health facility in the last six months for the detection of the chronic condition (diabetes and hypertension) or for a cytology. Column (1) presents the control mean at baseline. Column (2) presents the estimate of δ from baseline on the outcome of interest using equation (1), but excluding any baseline covariates. Column (3) presents the estimate of δ using the pooled 6^{th} and 12^{th} month follow-up data excluding baseline covariates. Column (4) presents the same estimate as (3) but includes as baseline covariate the mean team level outcome of interest at baseline. Standard errors are clustered at the team level and are presented in parenthesis. P-values are included in square brackets. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 11. Heterogeneity of treatment effects by performance score at baseline

		Post-treatment (Controlling for Baseline)		
		Baseline Performance Score Category		
	Overall	0-59	60-69	70-100
	(1)	(2)	(3)	(4)
ASTE on Contracted Indicators by domain				
ASTE Community Outreach	0.1748**	0.2615***	-0.1132	0.3374**
	(0.069)	(0.096)	(0.105)	(0.141)
p-value	[0.011]	[0.006]	[0.280]	[0.017]
ASTE Quality of Care	0.1427***	0.1330*	0.0706	0.1268
	(0.054)	(0.071)	(0.090)	(0.119)
p-value	[0.008]	[0.061]	[0.431]	[0.288]
ASTE Timeliness of care	0.1021*	0.0859	-0.0328	0.2854***
	(0.052)	(0.058)	(0.139)	(0.093)
p-value	[0.051]	[0.140]	[0.814]	[0.002]
ASTE Utilization	0.0958**	0.0691	0.0217	0.1925***
	(0.040)	(0.070)	(0.067)	(0.049)
p-value	[0.015]	[0.326]	[0.747]	[0.000]
ASTE Non-Contracted Outcomes	0.0215	0.0419	-0.0074	0.0070
	(0.031)	(0.049)	(0.073)	(0.037)
p-value	[0.485]	[0.387]	[0.919]	[0.852]
Community Health Teams	64	30	19	15

Notes: The sample of analysis is of the 64 teams with data on all waves. Column (1) presents the estimates of the average standardized treatment effect (ASTE) post-treatment controlling for team-level baseline outcomes by domain, as presented on Panel A, Column (4) of Tables 6 through 9. Columns (2) presents results for teams below the cutoff for incentives in their performance score at baseline. Column (3) presents results for teams with a baseline performance score of 60 to 69, which is equivalent for the first bracket of the incentive scheme. Column (4) presents results for teams with the highest levels of performance score at baseline (70-100). The performance score is defined as in Section 3. Standard errors are clustered at the team level and are presented in parenthesis. P-values are included in square brackets.
* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 12. Comparison of ASTE by wave and domain

	Baseline	6-Month	12-months	18-months
	(1)	(2)	(3)	(4)
ASTE of Contracted-Indicators by domain				
ASTE Community Outreach	0.0358	0.0970	0.2570***	0.1128
	(0.069)	(0.088)	(0.083)	(0.097)
p-value	[0.603]	[0.271]	[0.002]	[0.244]
ASTE Quality of Care	-0.1361	0.1638***	0.1294*	0.1682***
	(0.095)	(0.061)	(0.071)	(0.064)
p-value	[0.152]	[0.007]	[0.067]	[0.009]
ASTE Timeliness of care	0.0954	0.0399	0.1664**	0.0720
	(0.089)	(0.064)	(0.066)	(0.066)
p-value	[0.286]	[0.536]	[0.012]	[0.275]
ASTE Utilization	-0.0323	0.0915**	0.0964*	0.0932*
	(0.065)	(0.046)	(0.049)	(0.048)
p-value	[0.620]	[0.046]	[0.051]	[0.052]
ASTE Non-Contracted Outcomes	0.0937*	-0.0359	0.0745**	0.0900*
	(0.053)	(0.044)	(0.037)	(0.054)
p-value	[0.079]	[0.415]	[0.045]	[0.096]

Notes: The sample of analysis is of the 64 teams with data on all waves. Column (1) presents the estimates of the average standardized treatment effect (ASTE) post-treatment controlling for team-level baseline outcomes by domain at baseline. Columns (2) to (4) presents the ASTE of teams at each follow-up. Standard errors are clustered at the team level and are presented in parenthesis. P-values are included in square brackets.
* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 13. Comparison of individual outcome effects by source

	Common Sample of Units & Time-Periods				
	Medical Records Proxy		Household Survey		
	Control Mean	Post-treatment (Controlling for Baseline)	Control Mean	Post-treatment (Controlling for Baseline)	P-value of difference between (2) and (4)
	(1)	(2)	(3)	(4)	(5)
Timeliness of Care					
Timely Prenatal Care	0.7109	0.0259	0.5746	-0.0000	
		(0.034)		(0.072)	
p-value		[0.450]		[1.000]	[0.717]
N		927		276	
Units		33		33	
Timely Post-natal Care	0.6080	0.1213**	0.9167	0.0563**	
		(0.051)		(0.021)	
p-value		[0.025]		[0.012]	[0.228]
N		675		275	
Units		33		33	
Coverage					
Institutional Delivery	0.7557	0.0307	0.9167	0.0068	
		(0.037)		(0.031)	
p-value		[0.414]		[0.829]	[0.566]
N		675		276	
Units		33		33	
MMR Vaccination	0.6364	0.0538	0.8611	0.0752	
		(0.051)		(0.062)	
p-value		[0.295]		[0.235]	[0.803]
N		1,105		154	
Units		33		33	

Notes: The sample of analysis is of the 33 teams with data on all waves that had data on both households and medical records for the presented outcomes. All outcomes include common cohorts of patients for medical records and household surveys. Columns (2) and (4) present the estimates of the treatment effect (ASTE) post-treatment controlling for the team-level baseline of the outcome. Standard errors are clustered at the team level and are presented in parenthesis. P-values are included in square brackets.
* p < 0.10, ** p < 0.05, *** p < 0.01.

Appendix

Appendix 1. Sample performance report

iniciativa salud mesoamérica

Resultados del Monitoreo Externo del Desempeño

Periodo: Segundo Semestre del 2016

Ecos F: [redacted]

Municipio: [redacted]

iniciativa salud mesoamérica

Resumen de Resultados

	Puntos obtenidos				
Semestre	2015-II	2016-I	2016-II	2017-I	**Puntos Posibles**
Planificación Familiar (PF)					
1.Información sobre métodos de PF	0	5	5		5
2.Utilización métodos PF	15	15	15		15
Control Prenatal					
3. Control Prenatal Precoz	10	10	10		10
4. Control Prenatal con Calidad	10	10	10		10
Parto y Puerperio					
5. Partos Referidos por Ecos F	5	5	5		5
6. Parto Institucional	10	10	10		10
7. Control Puerperal Precoz	0	10	10		10
Atención del Niño					
8. Antiparasitarios	5	0	0		5
9. Micronutrientes	N.A.	15	15		15
10. SRO y Zinc para Diarrea	0	0	0		10
11. Vacuna SPR	N.A.	5	5		5
Puntaje Total	**55**	**85**	**85**		**100**
Puntaje promedio ISM	**42**	**69**	**73**		**100**

Interpretación

- De acuerdo a la evaluación externa la [redacted] obtuvo **85** puntos en el **segundo semestre del 2016** de un total de **100** posibles. El puntaje del Ecos F en este periodo es **mayor** que el promedio de puntos total de otros Ecos F de la Iniciativa Salud Mesoamérica (ISM) que es de **73** puntos.
- El Ecos F cumplió las metas establecidas en **9** de **11** indicadores, pero se puede mejorar en aquellos marcados en rojo.
- Se recomienda analizar en equipo el reporte con el detalle de cada indicador y los procesos clave para los servicios donde se logró cumplir la meta para mantener el nivel e identificar mejores prácticas y establecer compromisos de mejora en aquellos donde no se cumplió en conjunto con equipos supervisores del SIBASI, región y nivel central.
- *El equipo obtuvo el 85% de puntos posibles (85/100) en esta medición por lo que es acreedor al 85% del fondo de reconocimientos, es decir **850** dólares canjeables por bienes de la lista establecida.*

Notas

- Los resultados aquí presentados se basan en la Encuesta de la Prestación de Servicios de Salud para la Mujer y el Niño/a en Ecos Familiares realizada durante los meses de Septiembre a Noviembre del 2016 por un equipo externo contratado por la Iniciativa Salud Mesoamérica.
- Los resultados de la evaluación serán utilizados únicamente para los fines establecidos en la Iniciativa Salud Mesoamérica.
- El total de puntos posibles no incluye los indicadores donde hay pocas observaciones o no aplican en el periodo.
- P.O.=Se refiere a que hay muy pocas observaciones para establecer si se cumple la meta.
- N.A.= No aplica en ese periodo, porque se decidió cambiar la forma de medir el indicador para hacerlo más preciso.

Note: Sample report for a treatment team. Control reports were identical except for the highlighted part in the cover page that specifies the amount obtained.

iniciativa salud mesoamérica

Resultados del monitoreo externo del desempeño: Planificación Familiar

Periodo: Segundo Semestre del 2016
Ecos F:
Municipio:

iniciativa salud mesoamérica

1. Información sobre métodos de PF

Meta	Estatus	Puntaje Obtenido
80%	**Cumple meta**	**5/5**

- **Descripción** Este indicador mide el porcentaje de mujeres en edad fértil que cumplen criterios para recibir información de planificación familiar que la recibieron por parte del personal de salud.
- **Resultado** De las 11 mujeres en edad fértil que cumplen criterio para recibir información de PF visitadas al azar en el área de influencia del Ecos F, 9 recibieron información sobre métodos de PF modernos por personal de salud en los últimos seis meses.
- Con este resultado se concluye que se cumple la meta considerando el error de muestreo.
- **Recomendación** Se recomienda analizar en equipo el plan local.
- **Fuente** Encuesta de vivienda

2. Utilización de métodos de PF

Meta	Estatus	Puntaje Obtenido
61%	**Cumple meta**	**15/15**

- **Descripción** Este indicador mide el porcentaje de mujeres en edad fértil que cumplen criterios para planificación familiar que utilizan métodos modernos de PF.
- **Resultado** De las 19 mujeres en edad fértil que cumplen con criterio para planificar visitadas al azar en el área de influencia del Ecos F, 18 utilizaban métodos de planificación familiar modernos al momento de la visita.
- Con este resultado se concluye que se cumple la meta considerando el error de muestreo.
- **Recomendación** Se recomienda analizar en equipo el plan local.
- **Fuente** Encuesta de vivienda

Definiciones clave de Planificación Familiar

- **Mujeres que cumplen criterios de planificación familiar**. Para propósitos de la evaluación, éstas son aquellas mujeres en edad reproductiva (de 15 a 49 años de edad) excluyendo las siguientes:
  - mujeres en menopausia
  - con histerectomía
  - vírgenes
  - que no tienen relaciones sexuales y
  - aquellas embarazadas o tratando de quedar embarazadas.
- Las mujeres que cumplen criterios para recibir información de planificación familiar son las mismas que las anteriores pero excluye a las mujeres esterilizadas.
- *Recordar que las mujeres vírgenes o que no tienen relaciones sexuales no se deben dejar fuera de las acciones de promoción de planificación familiar ya que en cualquier momento pueden tener la necesidad de planificar.*
- *Es importante identificar a las mujeres que cumplen criterios para planificar o para recibir información de PF en el área de influencia del Ecos F para poder proveerles información adecuada y satisfacer su necesidades de planificación.*
- **Métodos modernos de planificación familiar :** Los métodos de planificación modernos incluyen:
  - esterilización femenina o masculina
  - dispositivo intra-uterino DIU
  - inyectables
  - implantes
  - píldoras anticonceptivas
  - píldora de emergencia
  - preservativo masculino o femenino
  - diafragma
  - esponja espemicida
- *Al recomendar un método de planificación familiar recuerde valorar a la paciente y conocer sus necesidades para poder identificar el método de planificación más adecuado para ella.*

iniciativa salud mesoamérica

Resultados del monitoreo externo del desempeño: Control Prenatal Parto y Puerperio

Periodo: Segundo Semestre del 2016

Ecos F:

Municipio:

iniciativa salud mesoamérica

3. Control Prenatal Precoz

Meta	Estatus	Puntaje Obtenido
80%	**Cumple meta**	**10/10**

- **Descripción** Este indicador mide el porcentaje de mujeres embarazadas cuyo primer control prenatal ocurrió antes de las 12 semanas de gestación según expediente.
- **Resultado** De los 7 expedientes de mujeres embarazadas seleccionados al azar en el Ecos F, 6 tuvieron su primer control prenatal antes de las 12 semanas de gestación.
- Con este resultado se concluye que se cumple la meta considerando el error de muestreo.
- **Recomendación** Se recomienda analizar en equipo el plan local.
- **Fuente** Encuesta en establecimientos

4. Control Prenatal con Calidad

Meta	Estatus	Puntaje Obtenido
80%	**Cumple meta**	**10/10**

- **Descripción** Este indicador mide el porcentaje de mujeres con parto en los últimos 4 meses (excluyendo el más reciente) que recibieron el control prenatal con calidad de acuerdo a las mejores prácticas en expedientes.
- **Resultado** De los 1 expedientes de mujeres con parto en el periodo evaluado seleccionados al azar en el Ecos F, 1 fue realizada de acuerdo a las mejores prácticas.
- Con este resultado se concluye que se cumple la meta considerando el error de muestreo.
- **Recomendación** Se recomienda analizar en equipo el plan local.
- **Fuente** Encuesta en establecimientos

Resultados por criterio de mejores prácticas en control prenatal

Criterio Clínico	% Cumple	Criterio Clínico[3/]	% cumple	Exámenes de...[4/]	% cumple
Al menos 4 APN	100%	FCF	100%	...Glucosa	100%
Peso[1/]	100%	Movimientos fetales	100%	...VIH	100%
Presión arterial[1/]	100%			...Hemoglobina	100%
Altura uterina[2/]	100%			...Orina	100%

Notas: Los porcentaje son respecto al total de expedientes revisados. Un expediente tiene que cumplir con todos los criterios para ser considerado como de mejores prácticas. Los criterios subrayados son aquellos que están por debajo del 80% [1/] Se mide en todas las visitas. [2/] Se mide a partir de las 14 semanas. [3/] Se mide a partir de las 20 semanas. [4/] Se mide que esté en el expediente al menos una vez.

5. Partos Referidos por Ecos F

Meta	Estatus	Puntaje Obtenido
100%	**Cumple meta**	**5/5**

- **Descripción** Este indicador mide el porcentaje de mujeres con parto en los últimos 4 meses (excluyendo el más reciente) que fueron referidas a una institución de salud como parte del plan de parto en expedientes.
- **Resultado** De los 1 expedientes de mujeres con parto en el periodo evaluado seleccionados al azar en el Ecos F, 1 tenía plan de parto y estaba establecida una institución médica como lugar de parto.
- Con este resultado se concluye que se cumple la meta considerando el error de muestreo.
- **Recomendación** Se recomienda analizar en equipo el plan local.
- **Fuente** Encuesta en establecimientos

6. Parto Institucional

Meta	Estatus	Puntaje Obtenido
94%	**Cumple meta**	**10/10**

- **Descripción** Este indicador mide el porcentaje de mujeres con parto en los últimos 4 meses (excluyendo el más reciente) que tenían registrado el parto institucional en expedientes.
- **Resultado** De los 1 expedientes de mujeres con parto en el periodo evaluado seleccionados al azar en el Ecos F, 1 tenía registrado el parto institucional ya sea con la copia del egreso, la hoja de control postparto o en algún otro lugar del expediente.
- Con este resultado se concluye que se cumple la meta considerando el error de muestreo.
- **Recomendación** Se recomienda analizar en equipo el plan local.
- **Fuente** Encuesta en establecimientos

7. Control Puerperal Precoz

Meta	Estatus	Puntaje Obtenido
92%	**Cumple meta**	**10/10**

- **Descripción** Este indicador mide el porcentaje de mujeres con parto en los últimos 4 meses (excluyendo el más reciente) que recibieron atención postnatal durante la semana posterior al parto según expediente.
- **Resultado** De los 1 expedientes de mujeres con parto en el periodo evaluado seleccionados al azar en el Ecos F, 1 tenía registrada la atención puerperal y esta fue realizada durante la semana posterior al parto.
- Con este resultado se concluye que se cumple la meta considerando el error de muestreo.
- **Recomendación** Se recomienda analizar en equipo el plan local.
- **Fuente** Encuesta en establecimientos

iniciativa salud mesoamérica

Resultados del monitoreo externo del desempeño:
Atención del Niño

Periodo: Segundo Semestre del 2016

Ecos F:
Municipio:

iniciativa salud mesoamérica

8. Antiparasitarios

Meta	Estatus	Puntaje Obtenido
66%	No cumple meta	0/5

- **Descripción** Este indicador mide el porcentaje de niños de 12 a 59 meses que consumieron la dosis adecuada para la edad de antiparasitarios durante los últimos seis meses.
- **Resultado** De los 9 viviendas visitadas al azar con niños de 12 a 59 meses en el área de influencia del Ecos F, 2 consumieron la dosis adecuada para la edad de antiparasitarios en los últimos seis meses.
- Con este resultado se concluye que no se cumple la meta considerando el error de muestreo.
- **Recomendación** Se recomienda analizar en equipo el plan local.
- **Fuente** Encuesta de vivienda

9. Micronutrientes

Meta	Estatus	Puntaje Obtenido
80%	Cumple meta	15/15

- **Descripción** Este indicador mide el porcentaje de niños de 6 a 23 meses que tenían indicación de micronutrientes en expedientes y recibieron la dosis adecuada en farmacia.
- **Resultado** De los 16 expedientes de niños de 6 a 23 meses al momento de la visita seleccionados al azar en el Ecos F, 11 tenían indicación de micronutrientes registrada en el expediente y recibieron la dosis adecuada.
- Con este resultado se concluye que se cumple la meta considerando el error de muestreo.
- **Recomendación** Se recomienda analizar en equipo el plan local.
- **Fuente** Encuesta en establecimientos

10. SRO y Zinc para Diarrea

Meta	Estatus	Puntaje Obtenido
50%	Cumple meta	10/10

- **Descripción** Este indicador mide el porcentaje de madres de niños de 0 a 59 meses que mencionaron que tratarían a sus hijos con sales de rehidratación oral y zinc a sus hijos cuando tuvieran diarrea.
- **Resultado** De las 10 madres de niños de 0 a 59 meses visitadas al azar en el área de influencia del Ecos F, 5 madres mencionaron que tratarían a sus hijos con SRO y Zinc en caso de diarrea o trataron a sus hijos con SRO y Zinc si estos tuvieron un episodio de diarrea en las últimas dos semanas.
- Con este resultado se concluye que se cumple la meta considerando el error de muestreo.
- **Recomendación** Se recomienda analizar en equipo el plan local.
- **Fuente** Encuesta en vivienda

11. Vacuna SPR

Meta	Estatus	Puntaje Obtenido
94%	Cumple meta	5/5

- **Descripción** Este indicador mide el porcentaje de niños de 6 a 23 meses que tenían registro de la vacuna SPR según el libro de vacuna.
- **Resultado** De los 15 niños de 6 a 23 meses al momento de la visita seleccionados al azar en el Ecos F, 13 tenían registro de la vacuna SPR en el libro de vacuna.
- Con este resultado se concluye que se cumple la meta considerando el error de muestreo.
- **Recomendación** Se recomienda analizar en equipo el plan local.
- **Fuente** Encuesta en establecimientos

Appendix 2. Definition of outcomes of interest and proxies

Category	Indicator and Definition	Source
Outreach	**Information on modern family planning*** Share of women 15 to 49 in need of contraception that received information on modern family planning methods by health personnel in the last six months. Women in need of contraception are those age 15 to 49 excluding those sterilized, in menopause, that declare to be virgin or not sexually active, and those pregnant or trying to conceive. Modern family planning methods include any of the following: injectable, contraceptive pills, female or male sterilization, intra uterine devices (IUD), implants, emergency contraception, female or male condoms, contraceptive diaphragm and sponges.	Household survey
Outreach	**Knowledge of treatment of diarrhea with ORS & Zinc*** Share of women 15 to 49 with children less than five with knowledge of treatment of diarrhea with oral rehydration salts and zinc at the time of the survey. All women with these characteristics are asked in the survey about how they would treat their child if he/she felt ill with diarrhea and all options mentioned are recorded by enumerators. Enumerators do not read/provide any options in this question to women.	Household survey
Quality	**Quality prenatal care*** Share of women in the catchment area of community health teams with a delivery in the last four months (excluding the most recent month) that received prenatal care according to national clinical guidelines at the facility. It is equal to one if during all prenatal care visits provided at the facility the women was assessed according to guidelines. The assessment guidelines include: for each visit regardless of gestational age measures of weight and blood pressure, for each visit after 13 weeks of gestation measurement of uterine height, for each visit after 19 weeks of gestation, fetal heart rate. In addition, women should have received blood tests assessing glucose, HIV and hemoglobin, and a urine test at least once during their prenatal care visits.	Medical Records
Quality	**Reference to institutional delivery*** Share of women in the catchment area of community health teams with a delivery in the last four months (excluding the most recent month) with reference to institutional delivery in the birth plan.	Medical Records
Timeliness	**Timely Prenatal Care**	
Timeliness	*Medical Record Proxy** Share of women in the catchment area of community health teams that reached three months of gestation in the last six months for which the first prenatal care visit occurred prior to 12 weeks of gestation. The gestational age at the time of the first visit is obtained by the difference between the date of the first prenatal care visit and the date of last menstrual cycle. If either piece of information is not found in the record it is considered that the woman did not received care, that is zero. When compared to the household survey in Table 13 it is restricted to births occurring from October 2015 to October 2016, the comparable period with the survey.	Medical Records
Timeliness	*Household Survey* Share of live births whose first prenatal care visit occurred prior to the first 12 weeks of gestation and was provided by a physician or a professional nurse. When compared to the medical records proxy in Table 13 it is restricted to births occurring from October 2015 to October 2016, the comparable period with the medical records.	Household Survey
Timeliness	**Timely Post-natal Care**	
Timeliness	*Medical Record Proxy** Share of women in the catchment area of community health teams with a delivery in the last four months (excluding the most recent month) that received postpartum care within a week from delivery by health personnel.	Medical Records

Category	Indicator and Definition	Source
	The time of postpartum care is obtained by the difference between the dates of the first postnatal care visit for the woman and the date of delivery. If either piece of information is not found in the record it is considered that the woman did not received care, that is zero. When compared to the household survey in Table 13 it is restricted to births occurring from October 2015 to October 2016, the comparable period with the survey.	
Timeliness	*Household Survey* Share of live births whose first postpartum care visit occurred within a week after delivery and was provided by either a physician, a nurse or a community health worker either at home or at a health facility. When compared to the medical records proxy in Table 13 it is restricted to births occurring from October 2015 to October 2016, the comparable period with the medical records.	Household Survey
Utilization	**Use of modern family planning methods*** Share of women 15 to 49 in need of contraception using a modern family planning method at the time of the survey Women in need of contraception are those age 15 to 49 excluding those sterilized, in menopause, that declare to be virgin or not sexually active, and those pregnant or trying to conceive. Modern family planning methods include any of the following: injectable, contraceptive pills, female or male sterilization, intra uterine devices (IUD), implants, emergency contraception, female or male condoms, contraceptive diaphragm and sponges.	Household survey
Utilization	**Institutional Delivery**	
Utilization	*Medical Record Proxy** Share of women in the catchment area of community health teams with a delivery in the last four months (excluding the most recent month) that delivered in a health facility according to the records. When compared to the household survey in Table 13 it is restricted to births occurring from October 2015 to October 2016, the comparable period with the survey.	Medical Records
Utilization	*Household Survey* Share of live births delivered in a health facility by skilled provider (physician or professional nurse). When compared to the medical records proxy in Table 13 it is restricted to births occurring from October 2015 to October 2016, the comparable period with the medical records.	Household survey
Utilization	**Micronutrients for children**	
Utilization	*Medical Record Proxy** Share of children age 6 to 23 months at the time of data collection in the catchment area of community health teams that were prescribed micronutrients sachets in the last six months according to the medical record.[1/]	Medical Records
Utilization	*Household Survey* Share of children age 6 to 23 months old that consumed 50 or more micronutrients sachets in the last six months according to maternal recall.	Household survey
Utilization	**Deworming pills consumption*** Share of children age 18 to 59 months old that consumed at least two deworming pills in the last six months according to maternal recall.	Household survey
Utilization	**Measles, Mumps and Rubella (MMR) Vaccination**	
Utilization	*Medical Record Proxy** Share of children age 12 to 23 months old in the catchment area of community health teams immunized with the MMR vaccine according to the medical record. [2/] When compared to the household survey in Table 13 it is restricted to children born from October 2014 to September 2015, the comparable period with the survey.	Medical Records
Utilization	*Household Survey* Share of children age 12 to 23 months old that were immunized with the MMR vaccine according to their vaccination card. When compared to the medical records in Table 13 it is restricted to children born from October 2014 to September 2015, the comparable period with the records.	Household survey

Category	Indicator and Definition	Source
Non-contracted	**Diabetes detection** Share of women 15 to 49 years old that were tested for diabetes in a health facility in the last six months.	Household survey
Non-contracted	**Hypertension Detection** Share of women 15 to 49 years old that were tested for hypertension (blood pressure taken) in a health facility in the last six months.	Household survey
Non-contracted	**Cytology performed** Share of women 15 to 49 years old that were tested for cancer in the uterine cervix (cytology or pap test) in a health facility in the last six months.	Household survey

Notes: *Indicator used to evaluate performance in the incentive scheme.

[1/] This definition was changed to micronutrients provided to children 6 to 23 months in the last six months after baseline since it was considered a better proxy for the incentive scheme. The definition in the table is the only one comparable across all rounds and hence the one used in the analysis.

2/ At baseline the source of verification for the incentive scheme was the clinical record of children but was changed to vaccination records after baseline since it was considered more comprehensive. The definition in the table is the only one comparable across all rounds and hence the one used in the analysis.

Appendix 3. Sample by source and verification cycle

	Baseline	6-month	12-month	18-month	Total	Average per cycle
Panel A. Full Sample by Verification Cycle						
Dwellings in Survey						
Total Surveyed	2,421	2426	2254	2130	9231	2308
With Eligible Women	1574	1673	1694	1519	6460	1615
Completed	1303	1523	1549	1413	5788	1447
With Children Less than 5	553	930	881	761	3125	781
With Children Less than 2	331	482	469	380	1662	416
Medical Records						
Pregnancies	794	884	761	786	3225	806
Deliveries	1127	1148	1152	1207	4634	1159
Children Age 6 to 23 months	1314	1364	1271	1393	5342	1336
Panel B. Main Analysis Sample by Verification Cycle						
Dwellings in Survey						
Total Surveyed	2,056	2,063	1,982	1,815	7,916	1979
With Eligible Women	1,343	1421	1510	1299	5,573	1393
Completed	1098	1,288	1,388	1,213	4987	1247
With Children Less than 5	464	775	798	663	2700	675
With Children Less than 2	280	395	421	341	1437	359
Medical Records						
Pregnancies	733	758	694	666	2851	713
Deliveries	1012	967	1019	1027	4025	1006
Children Age 6 to 23 months	1185	1167	1145	1189	4686	1172

Notes: Panel A presents a summary of the total sample collected in each verification cycle. Panel B restricts the sample to those 64 teams in which data on all outcomes for all 4 verification cycles were collected. The data on the household survey, are the number of dwelling with the specified characteristics. The data on records refer to the total number of medical records reviewed.

Appendix 4. Balance on baseline characteristics and attrition

Table 4A. Relationship of treatment assignment and data availability

	Non-missing main outcomes	Non-missing outcomes for common indicators in medical records and household surveys	P-value of difference
	(1)	(2)	(3)
Treatment	0.0314	0.0141	
	(0.084)	(0.118)	
p-value	[0.710]	[0.905]	[0.880]
Control Mean	0.8378	0.4324	
N	75	75	

Notes: Column (1) presents the result of a regression of a dummy variable equal to one if the community health team had data on all outcomes all four rounds for the incentive scheme on data collection and zero otherwise against an experimental treatment assignment indicator and block effects. Column (2) presents a similar estimate, but the dependent variable is equal to one if the community health team had data on all outcomes used for performance measurement and with their equivalent in household surveys were available in all four round of data collection. Column (3) presents the p-value of the difference between the estimate of Column (1) and (2). Standard errors are clustered at the team level and are presented in parenthesis. P-values are included in square brackets.
* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 4B. Balance on baseline characteristics in different samples

	Sample	Control Mean	Mean Difference (Treatment-Control)	Std. Error	P-Value	N	Community Health Teams
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Dwelling Characteristics							
Floor durable material	1	0.6832	0.0397	(0.045)	[0.384]	1,253	75
	2	0.6774	0.0313	(0.052)	[0.551]	1,054	64
	3	0.6318	0.0902	(0.087)	[0.307]	592	33
Ceiling durable material	1	0.4278	0.0236	(0.067)	[0.725]	1,253	75
	2	0.4307	0.0029	(0.077)	[0.970]	1,054	64
	3	0.4223	0.0858	(0.096)	[0.376]	592	33
Wall durable material	1	0.8220	0.0517	(0.041)	[0.208]	1,253	75
	2	0.8226	0.0333	(0.048)	[0.487]	1,054	64
	3	0.7770	0.0680	(0.077)	[0.386]	592	33
Electricity	1	0.7837	-0.0338	(0.035)	[0.337]	1,253	75
	2	0.7732	-0.0337	(0.040)	[0.401]	1,054	64
	3	0.7449	-0.0446	(0.053)	[0.405]	592	33
Phone	1	0.0958	0.0205	(0.016)	[0.214]	1,253	75
	2	0.0882	0.0224	(0.017)	[0.197]	1,054	64
	3	0.0878	0.0189	(0.022)	[0.397]	592	33
Toilet	1	0.2442	-0.0327	(0.040)	[0.417]	1,253	75
	2	0.2448	-0.0753*	(0.044)	[0.093]	1,054	64
	3	0.2078	-0.0576	(0.061)	[0.354]	592	33
Bono Comunidades Solidarias	1	0.3128	-0.0053	(0.058)	[0.927]	1,253	75
	2	0.2922	0.0006	(0.067)	[0.993]	1,054	64
	3	0.3108	-0.0937	(0.102)	[0.366]	592	33
Dwelling with Children Less Than 2	1	0.2586	0.0199	(0.024)	[0.405]	1,253	75
	2	0.2619	0.0188	(0.027)	[0.485]	1,054	64
	3	0.2905	0.0126	(0.039)	[0.748]	592	33
Panel B. Women Characteristics							
Age	1	30.1778	-0.1377	(0.596)	[0.818]	1,299	75
	2	30.2676	-0.1279	(0.630)	[0.840]	1,095	64
	3	30.3518	0.5379	(0.788)	[0.500]	614	33
Single	1	0.3445	-0.0462	(0.032)	[0.149]	1,251	75
	2	0.3403	-0.0444	(0.035)	[0.206]	1,052	64
	3	0.3176	-0.0450	(0.045)	[0.329]	592	33
No Health Insurance	1	0.9344	-0.0137	(0.017)	[0.418]	1,250	75
	2	0.9382	-0.0104	(0.017)	[0.549]	1,051	64
	3	0.9443	-0.0164	(0.025)	[0.523]	592	33
Elementary Education or Less	1	0.8145	0.0320	(0.028)	[0.250]	1,251	75

	Sample	Control Mean	Mean Difference (Treatment-Control)	Std. Error	P-Value	N	Community Health Teams
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	2	0.8184	0.0221	(0.031)	[0.475]	1,052	64
	3	0.8345	0.0120	(0.041)	[0.773]	592	33
Panel C. Units Characteristics							
Personnel							
Physicians	1	1.2267	-0.1068	(0.159)	[0.505]	75	75
	2	1.2344	-0.1042	(0.194)	[0.594]	64	64
	3	1.2727	-0.2724	(0.417)	[0.519]	33	33
Professional Nurses	1	0.9867	0.0220	(0.077)	[0.777]	75	75
	2	0.9844	0.0075	(0.088)	[0.933]	64	64
	3	0.9697	0.1417*	(0.081)	[0.088]	33	33
Auxiliary Nurses	1	1.1200	-0.3077	(0.217)	[0.160]	75	75
	2	1.1563	-0.3484	(0.269)	[0.200]	64	64
	3	1.2424	-0.6102	(0.629)	[0.340]	33	33
Community Health Workers	1	2.9054	-0.1559	(0.204)	[0.446]	74	74
	2	2.9683	-0.2079	(0.234)	[0.377]	63	63
	3	3.1250	-0.3370	(0.235)	[0.161]	32	32
Multi-purpose personnel	1	0.7467	-0.0173	(0.102)	[0.866]	75	75
	2	0.7344	-0.0102	(0.114)	[0.929]	64	64
	3	0.7273	0.1593	(0.153)	[0.305]	33	33
Administrative Region (SIBASI)							
Ahuachapán	1	0.1333	0.0518	(0.074)	[0.487]	75	75
	2	0.1250	0.0416	(0.080)	[0.606]	64	64
	3	0.1212	0.0201	(0.116)	[0.863]	33	33
Cabañas	1	0.2933	-0.0047	(0.080)	[0.953]	75	75
	2	0.2813	-0.0300	(0.087)	[0.730]	64	64
	3	0.3030	-0.2397*	(0.138)	[0.092]	33	33
Cuscatlán	1	0.0800	-0.0031	(0.063)	[0.960]	75	75
	2	0.0938	0.0015	(0.073)	[0.984]	64	64
	3	0.1515	0.0327	(0.141)	[0.818]	33	33
La Libertad	1	0.0400	0.0251	(0.045)	[0.579]	75	75
	2	0.0469	0.0222	(0.052)	[0.671]	64	64
	3	0.0606	-0.0277	(0.073)	[0.706]	33	33
La Paz	1	0.0667	0.0235	(0.058)	[0.685]	75	75
	2	0.0469	0.0323	(0.052)	[0.535]	64	64
	3	0.0000	0.0000	(0.000)	.	33	33
La Unión	1	0.0400	0.0251	(0.045)	[0.578]	75	75

	Sample	Control Mean	Mean Difference (Treatment-Control)	Std. Error	P-Value	N	Community Health Teams
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	2	0.0469	0.0323	(0.052)	[0.535]	64	64
	3	0.0606	0.1090	(0.082)	[0.192]	33	33
Morazán	1	0.0533	-0.0016	(0.052)	[0.976]	75	75
	2	0.0625	-0.0118	(0.061)	[0.847]	64	64
	3	0.0303	0.0704	(0.071)	[0.326]	33	33
San Vicente	1	0.2933	-0.1162	(0.092)	[0.209]	75	75
	2	0.2969	-0.0881	(0.101)	[0.386]	64	64
	3	0.2727	0.0352	(0.136)	[0.798]	33	33

Notes: Sample 1 refers to the full sample of 75 community health teams. Sample 2 refers to the subset of 64 teams with data on all outcomes for the performance score in all four rounds. Sample 3 refers to the subset of teams with data on all outcomes on all outcomes for the performance score in all four rounds and data on comparable indicators from the household survey. The third column presents the difference between treatment and control groups controlling for blocks. Standard errors clustered at the team level are presented on Column (4). Durable materials include concrete, brick, adobe, concrete blocks, and tiles. Toilet refers to those connected to sewage or a septic tank. *Bono Comunidades Solidarias* is a conditional cash transfer program in El Salvador. Panel A and B present characteristics measured from household surveys at baseline. Panel C presents community health team characteristics at end line (no baseline data are available for these variables).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix 5. Comparison of ASTE using difference-in-difference and controlling for baseline

	Post-treatment (Controlling for Baseline)	DID	Difference (2) Vs (1)
	(1)	(2)	(3)
ASTE of Contracted-Indicators by domain			
ASTE Community Outreach	0.1748**	0.1769*	0.0021
	(0.069)	(0.100)	(0.059)
p-value	[0.011]	[0.078]	
ASTE Quality of Care	0.1427***	0.2436***	0.1010
	(0.054)	(0.093)	(0.089)
p-value	[0.008]	[0.009]	
ASTE Timeliness of care	0.1021*	0.0224	-0.0797
	(0.052)	(0.085)	(0.083)
p-value	[0.051]	[0.791]	
ASTE Utilization	0.0958**	0.1243*	0.0285
	(0.040)	(0.071)	(0.066)
p-value	[0.015]	[0.080]	
ASTE Non-Contracted Indicators	0.0215	-0.0520	-0.0735
	(0.031)	(0.059)	(0.048)
p-value	[0.485]	[0.377]	

Notes: The sample of analysis is of the 64 teams with data on all waves. Column (1) presents the estimates of the average standardized treatment effect (ASTE) post-treatment controlling for team-level baseline outcomes by domain, as presented on Panel A, Column (4) of Tables 6 through 9. Column (2) presents the difference-in-difference estimate with block effects of the pooled post-treatment follow-up at 6^{th} and 12^{th} months relative to baseline. Standard errors in Columns (1) and (2) are clustered at the team level and are presented in parenthesis. Column (3) presents the difference between the estimates of Column (2) and (1). Standard errors of the difference are obtained using bootstrap clustered at the team level. P-values are included in square brackets.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.