Nudging at Scale: Combining Random and Quasi-Random Variation to Evaluate the Scale-Up of Incentives for Immunization in Pakistan

Subhash Chandir [*], Rachel Glennerster[†], Maryiam Haroon[‡], Edward Jee [§], Mubarak Taighoon Shah[¶] & Danya Arif Siddiqi[‖]

September 13, 2024

Abstract

When large government programs are introduced, programmatic imperatives often prevent the ideal randomized evaluation. We leverage detailed administrative data and a combination of both random and quasi random variation to generate a plausible measure of impact despite programmatic constraints. We study the scale up of an incentive for immunization program in Sindh province in Pakistan where seven of the lowest immunization performing districts were chosen to receive the program with the order of rollout randomized. We calculate the probability a district is included in the program by permuting across alternative selection criteria for the included districts and conducting

[*]Interactive Research and Development, Global, Email: subhash.chandir@ird.global

[†]University of Chicago. Email: rglennerster@uchicago.edu, Corresponding author.

[‡]University of Chicago. Email: maryiamharoon@uchicago.edu

[§]University of Chicago. Email: edjee@uchicago.edu

[¶]Interactive Research and Development, Global. Email: mubarak.shah@ird.global.

[‖]Interactive Research and Development, Global. Email: danya.arif@ird.global.

Acknowledgements: We acknowledge the financial support received from GiveWell. We thank the Expanded Programme on Immunization, Sindh, for supporting project implementation for this work. We would also like to thank Sundus Iftikhar and Ghulam Murtaza for their continuous support. We recognize the exceptional research assistance provided by Anbar Aizenman, Aleiba Dahri, Hammad Ezad, Sindhika Jai, Ayman Moazzam, and Yjing Zhang.

a weighted Callaway Sant'Anna estimation which includes data from 5 districts not included in the randomization of the program. This allows us to examine long-term impact of the program even though the rollout lasts only 7 months. Bayesian (or nonparametric) smoothing sharply improves the precision of estimates. Immunization rates for Pentavalent-1 and Measles-1 both rise by almost 15%, almost identical to the previous small scale randomized pilot. Exploiting individual level data we show the incentive increased enrolment of marginal children by 16% although few of these children persisted throughout the schedule while the incentive induced those already enrolled to persist for longer.

JEL codes: D04, I12, C55, H75, I18, C9, C11
Keywords: Childhood immunization, incentives, health behaviour, big administrative data, scale-up.

1 Introduction

The introduction of large scale government programs is rarely done in a way that maximizes the ability of researchers to cleanly evaluate them. While a government might be willing to stagger the phase-in of a program and even randomize the order of phase-in they typically prefer, for administrative ease, to phase-in at the level of a large unit (like a district or states) and to phase-in quickly (given political timelines). When programs are targeted at lagging districts/states governments often prefer to use some rule to determine which receive special support, limiting the scope for randomization. However, it is important to be able to evaluate these programs given the concern that small-scale or nongovernmental run pilots may not generate the same impacts as large government run programs.

In this paper, we evaluate the introduction of a large scale government program that faced precisely these constraints. We use a combination of random and quasi random variation alongside administrative data and minimal assumptions to generate a plausible measure of impact. We conclude that this hybrid approach can be useful for measuring impact when the ideal experiment is not possible but administrative data is available for units outside the randomization frame.

Specifically, we evaluate the impact of a program that provides small incentives to promote childhood immunization introduced by the Government of Sindh, Pakistan to seven low performing districts in 2022. The phase-in was randomized at the level of the district (and then within district by town) and was completed within 7 months. We leverage administrative data available for all 30 districts in the province. We then use both quasi random variation (about which of the 30 potential districts were chosen for the program) alongside random variation in the timing of phase-in (within the 7 targeted districts). The seven districts were chosen following a rule: they had low coverage rates of Pentavalent-3 and Measles-1 vaccines as of the end of August 2021. We assume that the precise selection criteria was one of several equally plausible alternative selection rules: for example low rates of Pentavalent-2 and Measles-1 in July 2021 could have been the criteria. We permute over these different selection criteria to calculate the probability that a district is selected for the program and thus the RCT. We find 5 additional districts with a greater than 40% probability of being in the program. Including administrative data from these districts as additional 'never-treated' controls, whilst conditioning on the probability of inclusion, increases the precision of our estimates over a standard seven district RCT. It also allows us to measure the impact of the program long after the treatment districts have all rolled in. We find a 14.6% and 14.8% increase in Penta-1 and Measles-1 vaccines as a result of the program, almost identical to the results of the earlier small scale pilot of the program (Chandir et al 2022).

Our phase-in design necessitates the use of contemporary difference-in-difference methods designed to address bias. However, these approaches typically reduce potential bias by trading-off efficiency and can result in imprecise impact estimates. Callaway and Sant'Anna (2021) event-study estimates, for example, trace out treatment effect dynamics relative to roll-in and estimate the impact of each post-event time period separately. This means only a fraction of the data is used for a given event-time

estimate, in our case only 2% of available data is used for a given week's estimated effect. To address this we use Bayesian nonparametrics to estimate a latent underlying impact (which moves over time) and shrink estimates close in event-time space towards each other, pooling information across weeks whilst avoiding "forbidden" comparisons highlighted by Goodman-Bacon (2021); Borusyak et al. (2024); Sun and Abraham (2021). This generates smoother and more precise estimates of long run impact (a similar approach is suggested by Faletto (2023)).

Our study utilizes data from Sindh Zindagi Mehfooz (Safe Life) Electronic Immunisation Registry (also known as ZM-EIR or SEIR) which collects real-time information on children's vaccinations administered in all public and private clinics in the province by scanning a QR code on the child's immunization card during vaccine administration.[1][2] The registry is the result of a long term partnership between the Government of Sindh, and the non-governmental organization Interactive Research and Development (IRD).[3] The registry also allows mobile topup payments to be sent automatically to caregivers when their child receives an immunization if the vaccine is given within incentive program districts. Outcome data for this study come from the SEIR registry although independent checks were also performed to ensure the validity of the SEIR data.

During 2017-2020, a pilot version of the incentive program was evaluated with the objective of identifying the most cost-effective size, method, and structure of incentives (Chandir et al 2022). It found, in results similar to the companion evaluation in India (Banerjee et al 2021), that a regular small mobile topup payment was the most cost-effective form of incentive.

The positive outcomes motivated the Government of Sindh to introduce the program to seven low immunization districts. It was rolled out in parts of Karachi (East, West, and Central), Kamber, Hyderabad, Jacobabad and Sujawal, a program area in which there are approximately three million administrations of the six primary vaccines every year. A small mobile top-up of PKR 200 (USD 1.25) is sent for each immunization administered to caregivers when the child is vaccinated within the program district.[4]

For logistical reasons, the program roll-in was staggered: one district adopted the program per month in random order between January and August 2022. Due to COVID-19 restrictions on large gatherings, not all vaccinators from each district could be trained at the same time, leading to a town-by-town phase-in within a district with the order of towns within a district also randomized with a one-week gap. [5]

The decision to prioritize the needs of program roll-in over the needs of the evaluation was crucial for the

[1] If the card is not available vaccinators find the record by inputting characteristics such as phone number, child name, or parent name.

[2] Vaccinations given by non EPI staff (such as teachers) mobilized during specific vaccination campaigns are not captured by SEIR.

[3] Access to the full SEIR registry is restricted to IRD staff who help maintain the registry on behalf of the Sindh government. Non-IRD coauthors were able to submit code to be run on the registry and analyze aggregate data. The study had the approval of the Sindh Government.

[4] The amount was later raised to PKR 275.

[5] The exception was the first district where the order was not randomized.

project's success, but it meant a standard evaluation approach would have been under-powered. The time variation introduced by the randomized roll-in is modest, with only 1 week between towns and 7 months between the first and last roll-in. This variation is particularly modest in comparison to the the high underlying variance in the number of immunizations by week and that the full immunization schedule for one child lasts between one and two years.[6] Many caregivers in a town will not have the chance to respond to the incentive before the next town is rolled in because their child is not yet eligible for their next vaccine, and caregivers may take time to learn about and, therefore, respond to the program.

To address these challenges, we combine variation from the randomized phase-in with quasi-random variation generated from the algorithm for selecting districts into the program. Districts were chosen if, using data available on September 1 2020, the district fell in the lowest 20th percentile on either Pentavalent-3 or Measles-1 vaccines. However, the date the data dump was generated and which vaccines were used to define low vaccination rates are somewhat arbitrary. By calculating which districts would have been chosen if the selection had used slightly different dates or using different vaccines we can estimate the probability a district would have been chosen for the program under a range of possible similar selection criteria. In addition to the 7 districts actually chosen we find 5 additional districts with a nonzero probability of inclusion. We include these 5 in our estimation and reweight our difference-in-difference estimates using the inclusion propensity score. The additional 5 districts substantially improve the precision of our estimates and provide a control group even after the program has fully rolled in allowing us to estimate longer term impacts.

Because towns are rolled in at different calendar times, a standard two-way fixed effect model with heterogeneous treatment effects could be biased and introduce negative weights (Borusyak et al., 2024; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2021), therefore to estimate effects we rely on a modification of Callaway and Sant'Anna (2021)'s estimator where the propensity score enters directly, justified under a design-based approach, rather than estimated as a function of pre-treatment covariates. Event study estimates calculate a separate program effect for each period after a town has been rolled in. The confidence intervals around each individual weekly estimation are large: only 2% of our data are used for any individual weekly estimate. We therefore use Bayesian estimation techniques to shrink our estimates and use data more efficiently. We assume there is a latent impact which moves slowly week to week and this is measured with noise by each individual weekly estimate. Our Bayesian approach is identical to a nonparametric estimation of this latent variable measured with noise. This generates a smoother and more precisely estimated impact over time and for the average post intervention period.

Next, we leverage the rich granularity of the administrative data to relax the parallel trends assumption by only comparing treated children to control children with the same vaccination due date. We find the incentive increased enrolment and early vaccines by 30%, although few of these children persisted to the

[6] If all immunizations are completed in a timely manner the full schedule can be completed in one year but caregivers receive the incentive if they bring their child in their first two years of life.

end of the schedule. The incentive also induced those enrolled prior to the introduction of the incentive to persist and receive more later vaccines.

Pakistan is well suited to study immunization incentives. It has one of the world's highest child mortality rates (74/1000 live births), with 27% of under-5 deaths attributed to vaccine-preventable diseases. Two out of five children miss out on life-saving vaccines, making Pakistan one of the 10 countries with the most unvaccinated children (Centre for Disease Control and Prevention, 2023). In Sindh, a province with an annual birth cohort of 1.7 million, 49% of children aged 12-23 months have received all vaccines classified as essential (National Institute of Population Studies (2019)).

We contribute to the literature in three ways. First we provide additional evidence to the growing literature suggesting small incentives tied to immunization can increase immunization rates. While middle-income countries (MICs) have employed broad income assistance schemes to incentivize immunization through linking cash transfers to vaccination and consistent clinic attendance, the financial burden imposed by these incentives remains unattainable for many low and lower-middle-income countries. However, much smaller transfers that act more like a nudge have been shown, in pilot studies, to increase immunization rates in Rajasthan India (Banerjee et al 2010), Kenya (Gibson et al 2017), Karachi, Pakistan (Chandir et al 2022), and Haryana India (Banerjee et al 2021). While Banerjee et al was tested at relatively large scale (nearly 300,000 children), the program is still being refined to make implementation by the government possible.

Second, we contribute to the literature evaluating how pilot programs fare when scaled up by governments. Muralidharan and Niehause (2017) document that few RCTs are conducted on large populations (Gertler and Boyce 2003, Schultze 2004 are examples that are). Even fewer studies document how impacts change when a program is tested at pilot and then at scale: Akram et al 2017 examine the scale up of a program to support migration in the lean season in Bangladesh; Banerjee et al 2017 discuss the changes necessary to adapt a pilot program to one that can be implemented at larger scale. We test a version of the program very similar to the pilot and find very similar impacts between the individually randomized pilot and the scaled program (evaluated through a clustered RCT). This reflects two opposing forces: the salience of the program was likely higher in the individualized pilot because the program was explained in a one-on-one enrollment interview; because it was a clustered RCT, in the scale-up we are able to capture the impact of the incentive on those who would otherwise never have received a single vaccine, while the pilot was conducted only on those who had received at least one vaccine.

Finally, we provide two methodological contributions to the difference-in-difference literature, (Athey and Imbens, 2022; Arkhangelsky et al., 2023). While inverse probability weighting is a standard part of Callaway Sant'Anna methodology, we propose weighting by the probability of being included in an experiment based on a clear understanding of the selection criteria and quasi random variation in whether a unit was included in the randomization. Whilst Gechter and Meager (2021) combine observational and experimental data across studies to estimate the impacts of conditional cash transfers, we are unaware

of other examples where this combination of randomized and quasi randomized variation has been used to evaluate a specific program in this way. Our second methodological contribution is to show how Bayesian non-parametrics can be used to improve the precision of modern difference-in-difference approaches such as Callaway and Sant'Anna event-time estimates by pooling information across weeks whilst avoiding comparing early to late-treated units.

The remainder of the paper is organized as follows: Section 2 provides some background on study setting and expansion of the incentive program; Section 3 describes the data sources used, presents descriptive statistics and discusses the empirical strategy; Section 4 presents the main findings and heterogeneous treatment effect, Section 6 concludes.

2 Background

2.1 Study Setting: Immunization rates in Pakistan

Despite the cost-effectiveness of early childhood immunization, 25 million children under the age of one failed to receive the recommended childhood vaccines in 2021, leading to over 1.5 million deaths from vaccine-preventable diseases (Centre for Disease Control and Prevention, 2023; World Health Organization, 2023). Eighteen million children received no vaccines at all, with 62% of unvaccinated children residing in 10 low and middle-income countries, including Pakistan (Centre for Disease Control and Prevention, 2023). In Pakistan, only 66% of children between the ages of 12 to 23 months received six primary vaccinations recommended by WHO, and just 51% of children received all age-appropriate vaccinations (National Institute of Population Studies, 2019). Furthermore, a significant regional disparity in vaccination coverage is evident, particularly in Sindh, home to an annual birth cohort of 1.7 million, with only 49% of children aged 12-23 months receiving all essential vaccines.[7] Recently, Pakistan adopted an Electronic Immunization Registry (EIR), generating high-quality, real-time immunization data for children. Sindh was an early adopter of EIRs: with technical support from IRD, the government of Sindh successfully scaled up an EIR across the entire province between October 2017 and March 2018. The program is also being scaled up to other provinces in Pakistan, and other countries are adopting similar registries, providing an opportunity for testing and implementing cost-effective and easily scalable interventions to boost vaccination rates.

2.2 Scale-up of Mobile Incentive Program

In Karachi, between 2017-2020, Chandir et al. (2022) evaluated the impact of small mobile conditional cash transfers (mCCT) of varying amounts, design, schedule, and payment methods on childhood im-

[7] The immunization rates for other provinces are: Punjab (80%), Azad Jammu and Kashmir (75%), Islamabad (68%), Gilgit Baltistan (57%), and Khyber Pakhtunkhwa (55%) (National Institute of Population Studies, 2019)

munization rates in Korangi, Sindh, Pakistan.[8] The study design aimed to identify the most effective and easily scalable program structure for targeting incentives for immunization. The pilot was implemented in all government immunization clinics and a private birthing center in Korangi town. All children visiting treatment centres were screened for eligibility and enrolled in the program if the child was under two years old, visiting for one of the three vaccines, including Bacille Calmette-Guerin (BCG), Pentavalent-1 (Penta-1) or Pentavalent-2 (Penta-2), and the caregiver provided a mobile phone number at the time of vaccine administration. The participants were randomly allocated to the treatment arms through a real-time phone-based application, with 1,600 participants enroled in each arm. The program utilized a study-specific version of the Zindagi Mehfooz electronic immunization registry to trigger payments when a child enrolled in the study received a vaccine. Chandir et al. (2022) find that mobile phone top-up payments are more effective than electronic money, and certain payments are more effective than lottery payments. While higher payments induced more immunizations than lower payments, the difference was relatively modest. The program increased the full immunization coverage rate at 12 months and up-to-date coverage at 18 months at USD 23 per additional fully immunized child (USD 0.8-2.4 per immunization visit) (Chandir et al., 2022).

In response to these promising findings, the Government of Sindh, Pakistan, in collaboration with Interactive Research and Development (IRD), introduced the conditional incentive program at scale[9]. The most cost-effective policy combination was adopted from the pilot study: a certain payment delivered through mobile top-ups, with a flat schedule and a payment slightly above the low payment in Chandir et al. (2022). Specifically, caregivers of children age 0-23 months receive a mobile top-up payment of PKR 200 ($1.25)[10] for any of six targeted chidhood immunizations[11] received at a fixed site (whether government or private clinic) or through vaccinator outreach in the treated district[12]. A child is eligible for an incentive if they receive a vaccine in the program district, regardless of their enrollment district.[13]

The program leverages Sindh's existing system of Electronic Immunization Registry (SEIR, known as ZM-EIR) to send incentive payments. The SEIR records all vaccinations administered in the Province. The vaccines are registered in the registry when the vaccinator scans the QR code on an immunization

[8] The study comprised of seven components, consisting of five conditional cash transfer (mCCT) arms, one SMS reminder arm, and one pure control group. The five mCCT arms had different incentive amounts (ranging from USD 5 to 15 per fully immunized child), payment schedules (flat versus escalating payments over the schedule), structures (certain versus lottery payments), and payment methods (airtime or mobile money).

[9] The scale-up of the mCCT program was funded by GiveWell.

[10] We used 160 PKR = 1 USD exchange rate as of August 1, 2021,

.

[11] The six primary vaccines are BCG, Pentavalent-1, Pentavalent-2, Pentavalent-3, Measles-1 and Measles-2. If the standard vaccine schedule is followed, one of the six incentivized vaccines is administered on each of the scheduled visits so that caregivers get one and only one incentive per visit. If an incentivized vaccine is out of stock, a child may receive other vaccines without receiving an incentive. This is likely to happen in roughly 1 out of 8% of cases. For detailed vaccine schedule, see section A4).

[12] Note: the pilot program was tested across private and government clinics.

[13] We discuss the potential for caregivers from other districts to cross district borders to be eligible for incentives in the sections below.

form with their phone or registers them manually if the caregiver does not have the card (see figures A3 and A4).[14]

[15] Once the vaccination is registered in program districts, an automatic payment in the form of a mobile top-up is triggered to the phone number provided by the caregiver. The program was advertized widely through community members and vaccination staff who informed caregivers about the incentives. Initial communication activities included social media and cable TV advertisements, branded local vehicles, promotional SMS, print advertisement materials, and cards with program information.

The program was introduced in 7 districts, including three in Karachi (East, West, and Central), Kamber, Hyderabad, Jacobabad, and Sujawal. The selected districts were in the bottom 20% of districts for immunization coverage rates for either Pentavalent-3 or Measles-1 vaccines in the 2020 birth cohort, as measured by surviving annual infants in the ZM-EIR as of August 31, 2021 (program and non-program districts presented in Figure (A1)).[16] Figure (1) provides details of the study timeline. For logistical reasons, the introduction was staggered, with one district per month rolled into the program between January and August 2022. For the evaluation of the program, the district order was randomized. However, because of the restrictions imposed due to COVID-19, the training of the vaccinators could not be administered at the district level but had to be administered at the town level. Hence, within each district, each town was enrolled in the program with equal spacing, ensuring that enrollment takes place weekly throughout the month. Again, the order of roll-in at the town level was randomized with University of Chicago researchers prescribing the randomized order and therefore introduction date in each district.[17] The one exception was the first district, Karachi East, where towns were rolled out in a nonrandom order. Figure (A2) provides details of the randomized order and the actual dates of introduction of the program for each town and district. The implementation followed the randomization order closely, with only minor deviations (four out of five towns of district Karachi Central, two out of seven towns of district Kambar, and one out of four towns of district Hyderabad). The roll-in order was maintained, and there was only a 1-3 day difference in launch dates from the originally specified plan.[18]

The scale-up varied from the pilot program evaluated by Chandir et al. (2022) in three main aspects. First, all caregivers in a town became eligible for payments under the program simultaneously, which made it possible to advertise the program in the community. Caregivers were not individually informed

[14] A child can be found in the register by looking up the clinic, caregivers phone number, the child's name, parents name and birthdate).

[15] There is pressure on vaccinators to reach vaccination targets, as measured by the number of children registered or vaccines administered in the system, so most children who come in without a card are put into the registry even though it takes time to find them.

[16] Note: the district where the pilot was carried out was not in the seven scale up districts as it had marginally higher immunization rates but was included in one of the control districts.

[17] Rachel Glennerster oversaw the randomization which was carried out by University of Chicago staff without involvement of IRD coauthors or staff.

[18] The main reason for the deviations was the involvement of health staff (including vaccinators and the respective District Health Officers) in the province-wide polio campaigns, crash immunization activities, and immunization week. They were unavailable for program launch on the specified dates, but the deviations did not impact the original program's design.

of the incentive amount by study staff (as they were in the pilot) as this role was given to vaccinators. Second, the payment amount was not specifically included in text messages sent to caregivers to remind them of their child's immunization due date (for reference, see Figure A5).[19] Finally, due to inflation in Pakistan, the real value of the incentive varied over the course of the program.

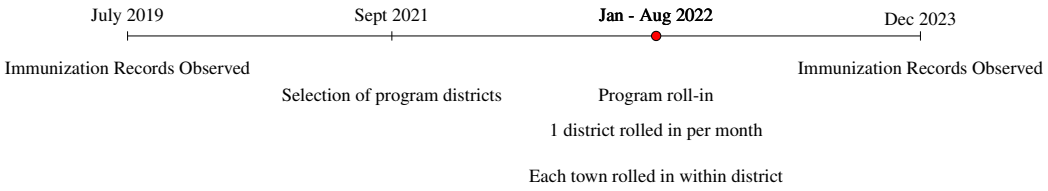


Figure 1: Timeline of Immunization Records and Program Roll-in

3 Empirical strategy

3.1 Data

The analysis in this paper leverages the unique administrative child-level immunisation records from the ZM-EIR registry, which is the same registry used to administer the incentive payments. All vaccines administered in the province at the fixed and outreach, private or government sites are recorded in the ZM-EIR.[20] The ZM-EIR was first deployed in October 2017 and has been scaled up across all 30 districts of Sindh province. As of July 31, 2023, it was used by 3,957 vaccinators working at 1,649 public and 314 private immunisation centres.

We use child-level data from both the seven program districts and five control districts: Karachi (East, West, and Central), Kamber Shahdad Kot, Hyderabad, Jacobabad, Sujawal, Tharparker, Matiari, Sukkur, Korangi and Malir.[21] The analysis was run on 2.9 million children, with 34.5 million immunization events recorded from July 01, 2019, to December 31, 2023.

The registry collects child-level information about the immunization history (detailing the administration dates and sites of first and subsequent antigens). Additionally, it captures information on child and parental characteristics (gender, place and date of birth, maternal education, and language of the caregiver), location (clinic, union council, town and district where the child received its first and all

[19] The text message is sent by Zindage-Mehfooz reminding caregivers that their child is due for a vaccine tomorrow and they should visit the EPI centre to get their child vaccinated. It also provides phone numbers that caregivers can use for further information. The text messages did not specify the incentive amount because of concerns that caregivers might go to clinics in nonprogram districts leading to complaints and accusations of corruption.

[20] Only vaccines not recorded in the ZM-EIR are those administered by by non-regular EPI staff during targeted surge campaigns or relief programs.

[21] In accordance with our data use agreement with IRD, the researchers at University of Chicago's access is limited to data from 12 districts characterized by the lowest immunization coverage rates. These 12 districts exhibited lower vaccination rates depending on the selected antigen and month. For instance, Hyderabad, Kamber, and Sujawal had a low level of Pentavalent-3 and Measles-1 vaccines in the 2020 birth cohort, as documented by surviving annual infants in the Electronic Immunization Registry of the Government of Sindh as of August 31, 2021.

subsequent vaccines) and vaccinator characteristics.[22] The registry only collects information about the children who have been vaccinated; thus, it is not possible to directly calculate immunization rates, which requires additional data on the population. However, the relevant population data is also subject to dispute.

3.2 Outcomes

To avoid p-hacking, we pre-specified the number of children immunized for Pentavalent-1 and Measles-1 as our primary outcomes of interest. We selected an early vaccine (Penta-1) and a later vaccine (Measles-1) in the schedule to capture the effect of the program on an increase in the number of children enrolled and an increase in persistence through the schedule (more vaccines per child).[23] Our secondary outcomes include the number of children receiving other vaccines in the schedule (in logs).[24] Other vaccines in the schedule include (i) Bacille Calmette-Guérin (BCG), (ii) Pentavalent-2, (iii) Pentavalent-3 and (iv) Measles-2 (though Measles-2 is not included in the definition of fully immunized child). While not prespecified, we take the logarithmic transformation of vaccines administered as towns vary greatly in size; thus, the estimated impact has a natural interpretation as the percentage change in vaccines due to the program.[25]

3.3 Data Description and Estimation Challenges

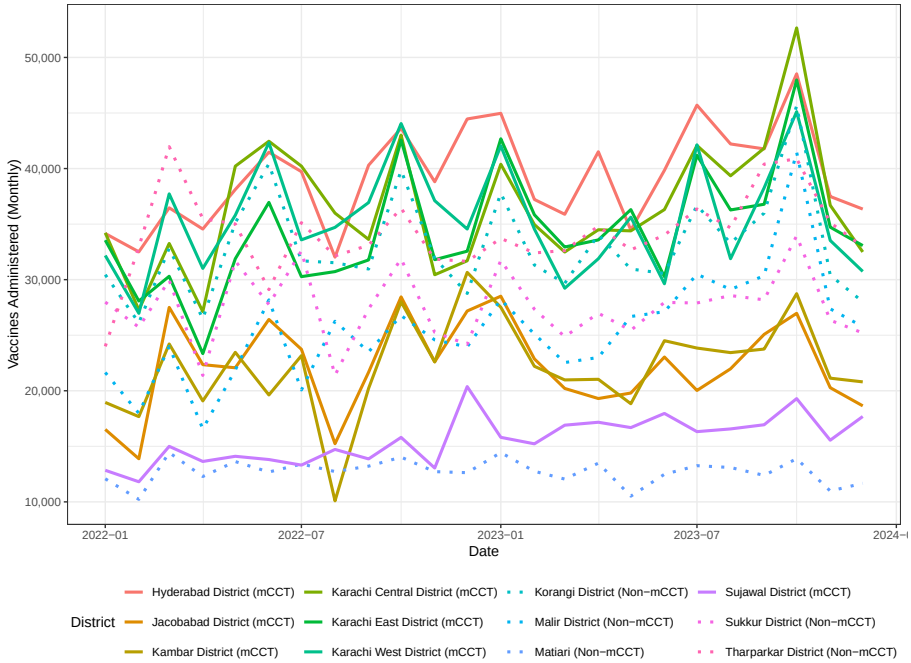
The vaccination data between 2019 and 2022 from our study districts shows huge variation in the number of children immunized on a given day or month (Figure 2). For instance, the number of children immunized in Kambar district in the 3rd week of May 2021 was 3,899, which is 4,700% higher than the preceding week (which was Ramadan). Several large shocks explain variations. Some shocks are common across districts, such as the delta wave in 2021 (before our sample period) and Ramadan. Other shocks impact some districts more than others: for example, the floods of 2022 (which hit as the program was being rolled out) and the vaccinator strike and subsequent outreach campaign in 2021. If we were to use only the seven RCT districts in our estimation, it would be hard to tease out the effects of these shocks from the impact of the program, as 2 of our 7 RCT districts were hit by the floods. However, several control districts were also hit by floods (Chandir et al. 2023). While the data includes 34 million individual immunization observations, there are only seven districts rolled in the program and, on average,

[22] In adherence to our data use agreement with IRD, researchers only have access to data that excludes personal details such as the child's name, address, date of birth, national identity card, and vaccinator name.

[23] The admin data ZM data does not provide comprehensive information on the total number of children between 0-23 months. It only provides information on the immunized children. Thus, we can not calculate immunization rates.

[24] We specified the number of children receiving Pentavalent-1 and Measles-1 vaccination as primary outcomes in the AEA Social Science Registry Glennerster (2022) and present these results in section 4.1. The results for other vaccines are presented in the online appendix.

[25] The decision to roll in by town, not district, only occurred once the program had started, and thus, randomization by the town was not in our pre-analysis plan. Towns vary much more by size than districts, which makes it even more important to use logarithmic transformation.



Note: Each line represents the total number of vaccines administered each month in the 12 study districts. Districts not in the RCT are depicted using dotted lines.

Figure 2: Total vaccines administered per month across districts

three to four towns per district. The entire randomized roll-in takes seven months, and the gap between each town is one week. The qualitative evidence suggests that caregivers take time to learn about the incentive, with many learning when they take their child for immunization. This will delay the response to the program, especially for Measles vaccines, which are due at 9 and 15 months: a child born just after the program starts will not complete their measles vaccines before the last town is rolled in. Similarly, if a caregiver learns about the program in month 2, their child may not become eligible for their next vaccine before the end of the roll-in. Without using the additional non-mCCT districts, there will be no control districts left by the end of August 2022, and we will miss the response of many caregivers. If information about the program diffuses gradually, the full impact will not be seen immediately.

3.4 Estimation Strategy: Combining Random and Quasi Random Variation

In this section, we discuss our estimation strategy for combining both randomized and non-randomized variation in the roll-in of towns into the immunization incentive program. We start by setting out the standard assumptions for estimating a randomized roll-in design (the no anticipation and common trend assumptions) and the rationale that these will hold in our randomized sample. We then discuss the methodology and assumptions needed to identify non-program districts for inclusion in our analysis. Our key assumption is that there were many plausible selection algorithms that IRD could have chosen to select the "seven program districts with the lowest levels of immunization", including focusing on different vaccines or undertaking the analysis at different times. We estimate the probability other districts

could have been included in the program (under a range of plausibly equivalently selection criteria) and include as controls, those with non-zero probability of being included.[26] Using the balancing property of the propensity score (Rosenbaum and Rubin, 1983), we include inverse probability weights in our ATT estimation to reflect the probabilities of being included in the incentive program. Throughout this section, our unit of analysis is town by week, i.e. the log of the number of vaccines administered in a week in a given town. We discuss how our clustering adjusts for two-level randomization (both district and then within the district at the town level). In Section 4.3, we discuss our estimation strategy to exploit individual-level data.

3.4.1 Assumptions for staggered event study

We estimate the effect of the program using Callaway and Sant'Anna (2021) to overcome the well-documented negative weight issue when estimating difference-in-differences with staggered treatment timing. Furthermore, the inverse probability weight approach outlined in Callaway and Sant'Anna (2021) allows us to transparently account for site selection, discussed in Section 3.4.2.

We define the Average Treatment effect on the Treated (ATT) for a town that rolls into treatment in first period g at time t:

$$ATT(g,t) = E\left[Y_t(g) - Y_t(0) | G_g = 1\right]$$

To identify the $ATT(g,t)$ the standard no anticipation assumption must hold:

$$E\left[Y_t(g) | G_g = 1\right] = E\left[Y_t(0) | G_g = 1\right], \quad \forall\, t \leq g$$

In our setting, this means that caregivers can not anticipate the roll-in of the program and thus are unlikely to delay vaccination decisions until the program reaches their town.[27]

Moreover, the incentive payment of $1.25 dollars is sufficiently small that individuals are unlikely to find it worthwhile to invest in finding out the exact timing of the roll-in in their communities and delay vaccination of their child. Finally, qualitative evidence suggests many caregivers learned about the incentives when they visited the centre for the first vaccines after the program rolled in. Thus, the assumption of no anticipation is likely to hold in our case.

[26] Given immunization rates are highly correlated across vaccines and auto-correlated over time, out of 30 possible districts, only 5 could have been included in the RCT sample under alternative permutations of the criteria. Conditional on being one of the 5 additional districts, the lowest probability of inclusion was 40%.

[27] There was no public announcement or schedule of when districts and towns would be rolled in. Also, vaccinators only knew about the program when they were trained for its implementation and it was rolled out immediately after the training.

Additionally, the common trends assumption states that:

$$E\left[Y_t(0) - Y_{t-1}(0)|G_g = 1\right] = E\left[Y_t(0) - Y_{t-1}(0)|D_s = 0, G_g = 0\right]$$

for all roll-in dates, g, and pairs s, t where $t \geq g, t \leq s$ must hold.

The parallel trend assumption is likely to hold for treated units as the treatment timing was randomized at the district and then town level.

3.4.2 Inclusion of non-randomized districts

The assumption of unconditional parallel trends between program and non-program districts is a strong assumption given the 27 non-program districts have mechanically higher Penta-3 and Measles-1 rates than program districts in August 2021 by virtue of being selected for inclusion in the program. If levels are correlated with changes then non-program districts might have systematically different immunization trends than program districts. We address this in two ways: we confine our analysis to only a few non-program districts that could have been selected under plausibly equivalent selection criteria; and we adopt a conditional common trends assumption.

An unconditional common trends assumption assumes that the counterfactual untreated trend for those in the experiment would have been the same for those in the experiment and not in the experiment. Given our knowledge of the selection mechanism, this seems reasonable.

Since selection was based on immunization levels and the common trends assumption allows for differential levels in Y_t across towns, the unconditional common trends assumption above is sufficient to identify the ATT. However, to be conservative and account for the possibility of correlation between levels and trends we introduce the conditional common trends assumption, conditioning on the covariate X:

$$E\left[Y_t(0) - Y_{t-1}(0)|X, G_g = 1\right] = E\left[Y_t(0) - Y_{t-1}(0)|X, D_s = 0, G_g = 0\right]$$

Traditionally, researchers collect and condition on a set of baseline covariates to ensure parallel trends hold. Instead, we exploit our knowledge of the selection mechanism and balancing property of the propensity score to condition directly on $p_{g,t}(X)$, the probability a town is first treated in period g at time t.

IRD's selection mechanism for the 'high-risk districts for immunization' was:

(i). Calculate crude coverage for 2020 birth cohort using ZM-EIR data dump generated on September 1st, 2021 and Government of Sindh data on annual live births/surviving infants.

(ii). Define a district below the 20th percentile of Penta-3 or Measles-1 district coverage rate as 'high-

risk'.

Our identifying assumption is that any of a series of similar selection criteria (type of vaccine and date of data dump) were equally likely to have been chosen. Specifically, that within a given vaccine category (i.e. within BCG, Penta-1, Penta-2 or Penta-3, and within Measles-1 or Measles-2) and within up to 8 months of the actual selection date, the probability of selecting a specific antigen or date to calculate the coverage rate used for the selection criteria is equally likely. We permute through all possible combinations of vaccines and dates and measure how often a given district is included in the experiment. Averaging over all the permutations gives the probability a district is selected for the experiment.[28] Figure 3 shows the probability a given district could have been included in the experiment across vaccine-pairs, averaging over the data dump selection date. Whilst the 7 districts actually included in the mCCT program have higher than average likelihood of inclusion (the 7 mCCT districts are denoted using bold font), there are multiple districts, such as Matiari and Korangi, that are almost as likely as Karachi West, Central, and East to have been selected for the programme but were not. In fact, Hyderabad, which is included in the mCCT program, was the least likely district to enter the experimental sample of the 12 potential districts with positive inclusion probability.

Conditioning on the probability in this way directly accounts for the selection mechanism. Intuitively, $p_{g,t}$ will be greater for districts with higher average pre-program vaccination rates.[29] When one of the 7 districts in the sample that was actually chosen for the RCT has a higher $p_{g,t}$, its treated outcomes will be weighted more and untreated outcomes weighted less and vice versa. With $p_{i,g,t}$ in hand it's possible to calculate the inverse-propensity weighted ATT described by Callaway and Sant'Anna (2021):

$$ATT^{nyt}(i,g,t) = E\left[\left(\frac{G_{i,g}}{E[G_{i,g}]} - \frac{\frac{p_{i,g,t}(1-D_{i,t})(1-G_{i,g})}{1-p_{i,g,t}}}{E\left[\frac{p_{i,g,t}(1-D_{i,t})(1-G_{i,g})}{1-p_{i,g,t}}\right]}\right)(Y_{i,t} - Y_{i,g-1})\right]$$

In our setting, $p_{i,g,t}$ enters directly into the sample analogue estimator, whereas conventionally $p_{i,g,t}(X)$ will be calculated using a logistic regression of treatment status on covariates, X.

3.4.3 Summary parameters

In order to study treatment effect dynamics, we estimate the effect of participating in the treatment with varying lengths of exposure to the program, *e* periods after treatment roll-in where $e = t-g$ i.e. e captures how many weeks have passed since a town first rolled into treatment at g. To ensure compositional shifts in which towns have been treated for e periods do not introduce the illusion of treatment dynamics we

[28] In the appendix we describe in greater detail how propensity scores are calculated using Algorithm 1.

[29] To calculate the propensity a district is treated at a given time period, $p_{g,t}(X)$, we further permute over the inclusion algorithm and our randomization procedure. Since districts have different numbers of towns, this introduces small differences in $p_{g,t}$ across units so we index $p_{i,g,t}$ by town i and calculate the propensity score directly for each town.

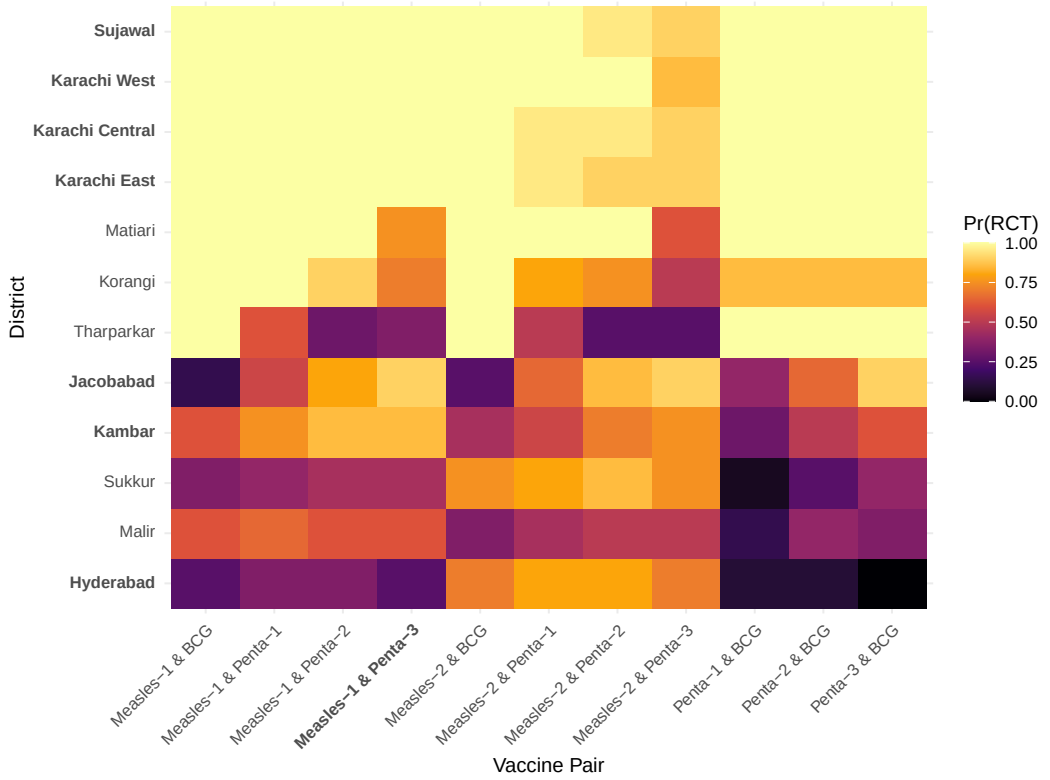


Figure 3: Probability a district is selected into experiment, by vaccine-pair averaging over selection date

Notes: This plot show the probability a district is in the lowest 20th percentile of coverage rates for a given vaccine-pair, averaging over possible data dump generation dates. Bold text denotes the districts and vaccine-pair actually chosen for the roll-out of the program.

balance estimates using the longest possible time period for which all treated towns are observed, which is 67 weeks using the last data dump.[30] Event study estimates by length of exposure to treatment (e) is given by:

Callaway and Sant'Anna (2021) as:

$$\theta_{es}(e) = \sum_{g \in \mathcal{G}} \underbrace{\mathbf{1}\{g + e \leq \mathcal{T}\} P(G = g | G + e \leq \mathcal{T})}_{\omega_{es}} \underbrace{ATT(g, g + e)}_{\text{Effect } e \text{ periods after } g \text{ roll-in}} \tag{1}$$

Furthermore, to summarise the overall treatment effect of participating in the program, we estimate the average effect for a town's overall post-treatment periods given by:

$$\theta_{sel}(\tilde{g}) = \frac{1}{\mathcal{T} - \tilde{g} + 1} \sum_{t=\tilde{g}}^{\mathcal{T}} ATT(\tilde{g}, t) \tag{2}$$

where $\theta_{sel}(\tilde{g})$ is the average effect of participating in the treatment among towns in group $\tilde{g}$ across all post-treatment periods. We find the overall effect by first computing the average effect for each group (across all time periods) and then averaging these together across groups. Thus, the overall effect of participating in the program can be estimated using:

$$\theta_{\text{overall}} = \sum_{g \in \mathcal{G}} \theta_{sel}(g) P(G = g | G \leq \mathcal{T}) \tag{3}$$

3.5 Inference

To calculate standard errors for our two stage randomization procedure, we use a modified multiplier bootstrap which reflects our nested assignment: First, we draw a district-level random variate. Next, we draw random variables at the town level, and finally we draw a correlation parameter that determines how much weight is placed on the district level random shock. Our approach, inspired by Athey and Imbens (2022), ensures the estimator of the variance-covariance matrix, $\mathbb{V}$, fixes the distribution of adoption dates. This design-based approach reflects the fact that, due to implementation requirements, regardless of realised randomization order our experiment must roll-in one town per week. Therefore, we calculate the empirical influence function as described by Callaway and Sant'Anna (2021):

[30] The length of the exposure to the program varies for each town, which could potentially change the composition of groups for each event time and length of exposure. To ascertain that our estimates do not suffer from the issue of compositional changes, we use a balanced panel and aggregate the average treatment effect for towns that are exposed to the treatment for at least some fixed number of time periods. If we do not impose a balanced panel when we compare the magnitude in the coefficient between say week 5 and 10 we are changing both the lag but also which towns are included in the estimation. One could imagine that requiring a balanced panel also means using only a specific set of towns rolled into the program for our treatment estimates. However, in our case all towns have been rolled-in after seven months of implementation phase and, thus, are part of the current estimations. This has been possible given the extensive nature of the data collected by the the Government of Sindh.

$$\widehat{\Psi}(Y_{i,t}, Y_{i,g-1}, G_{gi}) = \left(w_g^{\text{treat}} \cdot (Y_t - Y_{g-1}) - w_g^{\text{treat}} \cdot \overline{w_g^{\text{treat}} \cdot (Y_t - Y_{g-1})}\right) - \left(w_g^{\text{control}} \cdot (Y_t - Y_{g-1}) - w_g^{\text{control}} \cdot \overline{w_g^{\text{control}} \cdot (Y_t - Y_{g-1})}\right)$$

and form a bootstrap draw, $\widehat{ATT}(g,t)^b = \widehat{ATT}(g,t) + V^b$

$\widehat{\Psi}(\cdot)$ where V^b is generated as follows for each draw, b:

(i). For district d in $1, ..., D$ draw $X_d^b \sim N(0,1)$.

(ii). For town t in $1, ..., T$ draw $Y_t^b \sim N(0,1)$.

(iii). Draw $\rho^v \sim U(-1,1)$.

(iv). Generate $\tilde{V}^b = Y_t^b + \rho^b X_d^b$

(v). Calculate $V^b = \frac{\tilde{V}^b - \overline{\tilde{V}}^b}{\sigma_{\tilde{V}^b}}$

By drawing ρ^b each bootstrap draw we allow for arbitrarily strong correlation within each district without taking a stance on the level of correlation across towns within districts.[31]

3.5.1 Threats to Our Identifying Assumption

One objection to our identification assumption is that the choice of selection criteria (from the range of similar potential criteria) was unlikely to be quasi random. It might have been driven by a specific objective either political (the criteria had to be one that included a particular "favored" district) or medical (the vaccines chosen are of particular importance). To be a problem for our identification this unobserved driver of selection would have to be correlated with trends in immunization, not just levels. For example, if a politically favored district regularly gained additional resources this would impact levels and not changes in immunization. To fail our conditional common assumption it would need to be recently favored and suddenly receiving additional resources from both the incentive program and other programs in a way that increased immunization at precisely the time of program roll-in. If the move to favored status happened before the roll-in this would show up in nonparallel trends (see results). We therefore judge this unlikely. However we explore the the possibilities by comparing the allocation of other discretionary (immunization) resources across districts.

There is reason to think that all antigens are not equal and that public health colleagues may have strong reasons to select the vaccines they did. For this reason our plausibly equivalent selection criteria involve permuting within vaccine: for example varying whether the selection criteria is Penta-1 or Penta-3. We draw comfort from the fact that while Penta-3 was chosen by IRD as the key vaccine for the selection

[31] We standardise V^b to ensure the bootstrap estimates are well behaved, see e.g. Shao and Tu (1995)

criteria, Penta-1 was chosen as the primary outcome for this study because both are considered critical vaccines. It is more plausible that Measles-1 had a higher chance of being in the selection criteria than Measles-2 as Measles-2 is not required for the definition of full immunization. However, for our conditional common trends to assumption to be violated it would need to be the case that districts that performed worse on Measles-1 had a systematically different trajectory than those that performed worse on Measles-2.

3.6 Shrinkage of ATTs

The Callaway and Sant'Anna procedure involves estimating multiple ATTs: approximately 1,500 ATTs across 52 time periods and 29 groups. Whilst estimates are then aggregated into summary parameters, which increases precision by averaging over multiple noisy estimates, each component of the summary parameters still utilize a relatively small proportion of available information. For instance, the event study aggregation only averages over 29 total ATTs for each event period e. The estimates for a given week discard any information from surrounding weeks under the assumption that the impact of the program at different time periods could be fully independent from each other. Since each ATT is essentially a difference of means, applying well-known shrinkage estimators such as the James-Stein estimator is relatively straight forward and attractive in this setting given the large number of estimated means to regularise. Instead, we leverage understanding of the program to make additional assumptions to further increase power whilst shrinking estimates. We assume there is a latent program impact measured with noise in any one period and time since the event is the relevant distance measure (i.e. impacts close together in time are more likely to be related than ones further in time) to derive much tighter estimates.

We estimate a hierarchical shrinkage model that pools information across ATTs using a Gaussian Process (GP) prior to embed a notion of similarity across impacts in event time. Specifically, we estimate a hierarchical shrinkage model with the following likelihood over latent ATTs:

$$\widehat{ATT}(g,t) \sim N(ATT(g,t), se(\widehat{ATT}(g,t)))$$
$$ATT(g,t) \sim \mathcal{GP}(0, K_{ATT}\,(g-t, g'-t'))$$

By using a hierarchical model over the latent, unknown ATTs we shrink estimates towards a global mean depending on the pooling metric or signal-to-noise ratio. If $se(\widehat{ATT}(g,t))$ is relatively large compared to the hierarchical variance, then the estimated ATT is shrunk towards the global mean. However, if $se(\widehat{ATT}(g,t))$ is relatively small, then the estimated ATT is left relatively untouched.

We add additional structure to the shrinkage problem by embedding a notion of similarity between ATTs by using the event time as a distance metric. The covariance kernel of the GP treats ATTs with similar event times as more informative about each other than ATTs further away in event-time space. Since both the 'upper', hierarchical likelihood and 'lower' likelihood are jointly estimated, rather than imposing

how much each ATT is shrunk towards a common mean we estimate the hierarchical variance, which measures how heterogeneous effects are across g, t, and use the data to inform the level of shrinkage for a given ATT. Gaussian Processes are sometimes motivated using a 'weight-space' view which shows that a GP with an exponential covariance kernel is akin to an infinite-dimensional basis function expansion – our procedure can be interpreted as a purely non-parametric regularisation exercise. However, instead of choosing a weighting function to aggregate $\widehat{ATT}$s we estimate a weighting function directly from the data, depending on the trade-off between individual $\widehat{ATT}$ sampling variation and how related latent ATTs are in event-time space.[32]

4 Results

4.1 Conditioning on the Probability of Selection in Callaway and Sant' Anna method

4.1.1 Selection of Non-Randomized Control Districts and Propensity Scores

Figure A6 shows the estimated propensity scores for districts at $g = \{3, 10, 25\}$ over 35 weeks. Orange depicts the traditional logistic propensity score, which regresses treatment status on a constant when no pre-treatment covariates are provided. The estimates in blue show the permuted propensity score, accounting for the probability of inclusion into the program. Districts selected for inclusion in the mCCT programme are depicted in bold above plot facets. For districts which had a low chance of being included in the program, e.g. Hyderabad which has a probability of only 47%, the permuted propensity score is lower for earlier gs - the district had a relatively low probability of being treated. In contrast, Sujawal had a very high chance of being treated and therefore has a much higher propensity score on average. Sujawal's propensity score increases rapidly at later time periods because the probability of Sujawal being a never treated district is low (only 6%). Figure 3 reinforces the finding in Figure 2 that while included districts in general have a higher propensity to be in the RCT, there is variation with Hyderabad, Kambar and Jacobabad (included districts) looking similar to many non-included districts and some non-included districts (like Malarni) having similar propensities to many included districts.

4.1.2 Standard Callaway and Sant'Anna

We first show results for the standard Callaway and Sant'Anna difference-in-difference method on the log of the number of vaccines for Pentavalent-1 and Measles-1 as the outcomes (Figures 4 and 5). Our estimation includes five additional (non-treated) similar districts as a control group. We used a balanced panel of 29 treated towns observed for at least 67 weeks post-programme roll-out. For the pre-period, we use 20 weeks before the treatment. In the figures, red dots indicate pre-treatment estimates, and

[32] In Appendix Table A4 we show results robust across an alternative, wider prior specification.

blue represents post-treatment estimates. The orange dots indicate the overall average treatment effect averaging across time.

Figures 4 and 5 show that our estimates are consistent with the parallel trends assumption, as the coefficients before the programme roll-out are close to zero and do not exhibit pretrends. Caregivers are responsive to incentives; taking an average of post treatment effects, an incentive of 1.25 USD (PKR 200) induces a 14.0% (CI: $[3.15, 26.90]$, $p = 0.0128$) increase in the number of Pentavalent-1 vaccines in treated towns as compared to non-treated towns (average treatment effect over time represented by the orange dot presented in figure 4). Similarly, Measles-1 shows a 15.5% increase (CI: $[2.80, 28.1]$, $p = 0.0167$). However, the individual time period estimates are very noisy and variable. Without the additional districts the estimates are substantially noisier although consistent with our main specification (Table A2). We also present results for other vaccines in the online appendix section A7 and find a consistent positive and significant impact of the scale-up of the incentive for immunization program on BCG, Pentavalent-2 and Pentavalent-3 vaccines.

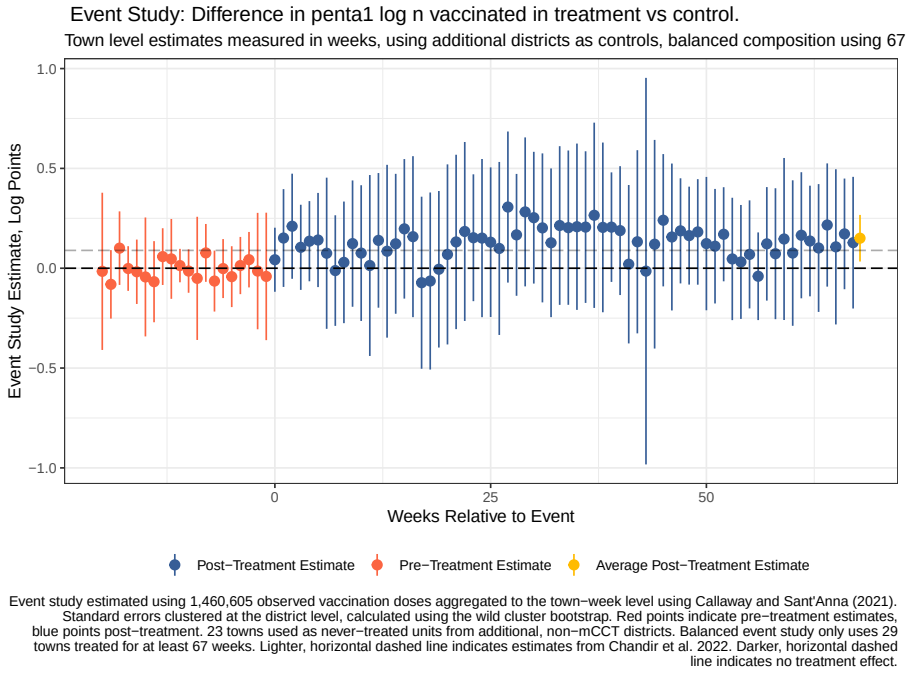


Figure 4: Event Study estimates from Difference in Difference using Callaway and Sant' Anna: Town-level estimates for Pentavalent-1

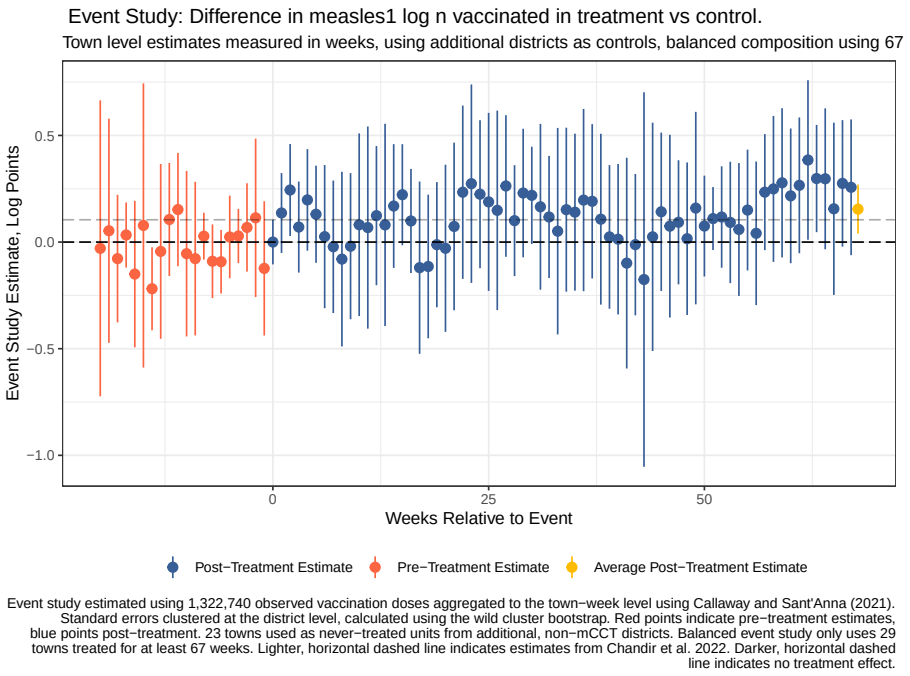


Figure 5: Event Study estimates from Difference in Difference using Callaway and Sant' Anna: Town-level estimates for Measles-1

4.2 Shrinking Callaway Sant'Anna: Hierarchical Bayes Estimation

Our preferred estimation approach uses the hierarchical shrinkage to exploit the bias-variance trade-off and improve precision (Figure 6 and 7). As above, we use a balanced panel with 67 weeks of post-treatment data. Again, we see no pre-trends prior to the roll-in of the immunization incentive. Immediately, following a town being rolled in we see a substantial jump in immunizations. Credibility intervals around each weekly estimate are substantially smaller (averaging a span of 35% for Penta-1 and 36.5% for Measles-1 versus 71.1% and 68.1% respectively for the traditional Callaway and Sant'Anna estimates) and the dynamics are much smoother suggesting a lot of the variation week to week in the non-smoothed was due to noise rather underlying treatment effect heterogeneity. The average post treatment effect is also more precisely estimated although of similar magnitude to the unsmoothed estimate. Specifically we find the incentive increased Penta-1 immunizations by 14.59% (CI: $[10.76, 18.66]$) and Measles-1 immunizations by 14.75% (CI: $[10.57, 19.06]$).

4.3 Leveraging due date by conditioning on those who can respond

In this section, we present our estimates for the impact of the CCTs programme on child vaccination behaviour using extensive individual child-vaccine level data. We make use of immunization records to construct a panel of child-week observations per vaccine for those who are enrolled in the SEIR registry. We leverage the granularity of the data in two ways to improve the precision of our estimates and relax the common trends assumption used previously.

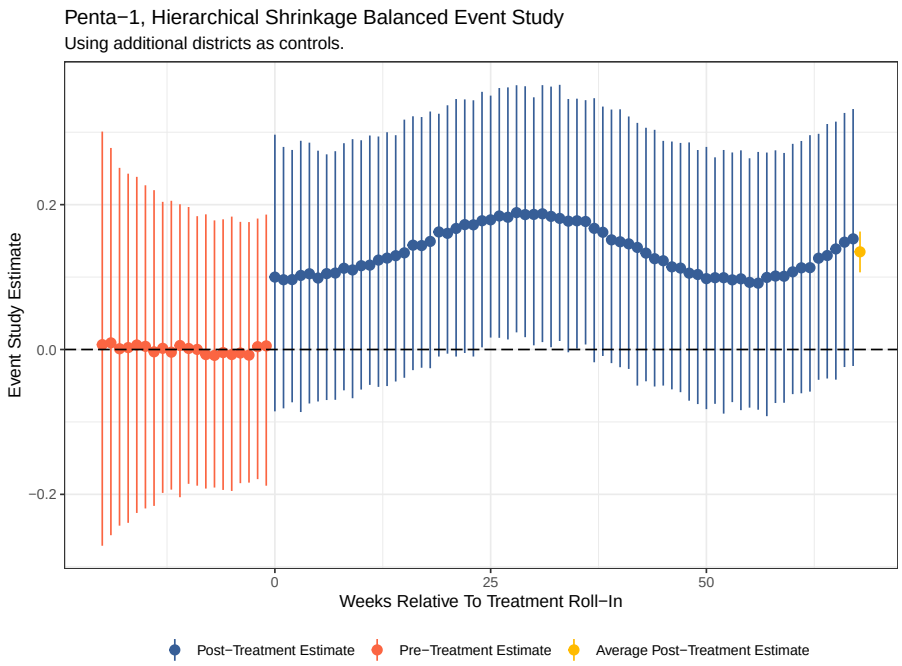


Figure 6: Shrinkage Estimates Gaussian Process Prior: Town-level Pentavalent-1

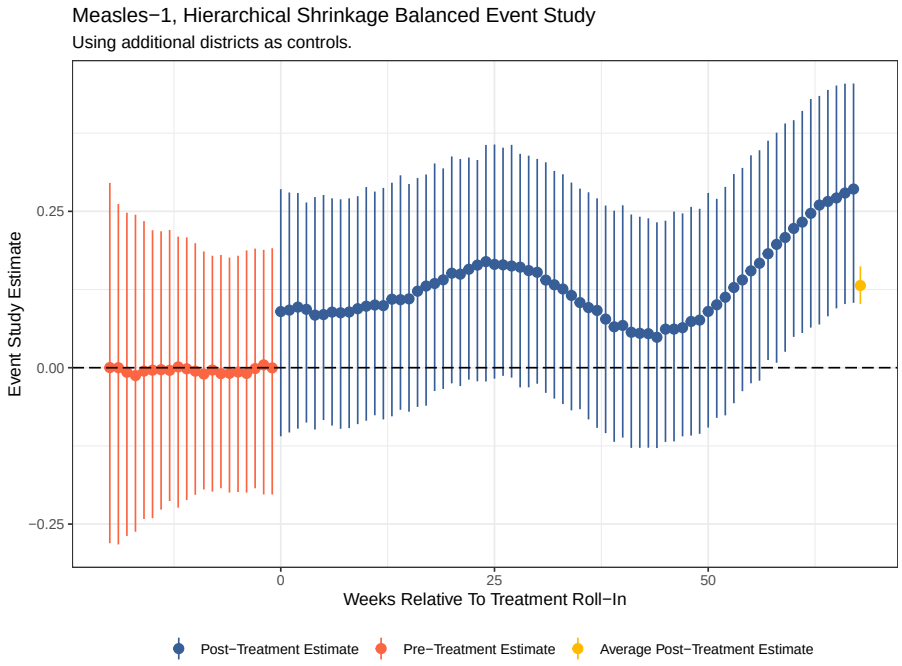


Figure 7: Empirical Bayes Gaussian Process Prior: Town-level estimates for Measles-1

Firstly, we re-define treatment status for a given child as $D_{it} = \mathbb{I}\{t_i \geq g_i \cap b_i \geq g_i\}$ where b_i denotes the week a child is due to be vaccinated – a child is now considered treated if they are both rolled into the program and due to receive the vaccine of interest. Whilst changing the interpration of our estimand, this increases the precision of our estimates. Without adjusting for the due-date of children would lead to an attenuation of our effect since the ITT would average over many children who are technically treated but not yet born, or not yet due to receive a vaccine. Our second innovation conditions on the week a child is due to calculate a conditional ATT.

We use 26 million child-week observations per vaccine to calculate the $ATT(g, t, b)$ – the effect of the program for those first treated at g, at time t, who are due in week b. This ATT only compares children across treatment and control who are due in the same week, relaxing the unconditional parallel trends assumption. Since we only observe an individual's vaccination status if they are in the SEIR system, all our results should be interpreted as conditional effects. The child level ATT therefore gives the probability of getting further vaccines in the schedule, conditional on having at least one vaccine. We estimate the following specification:

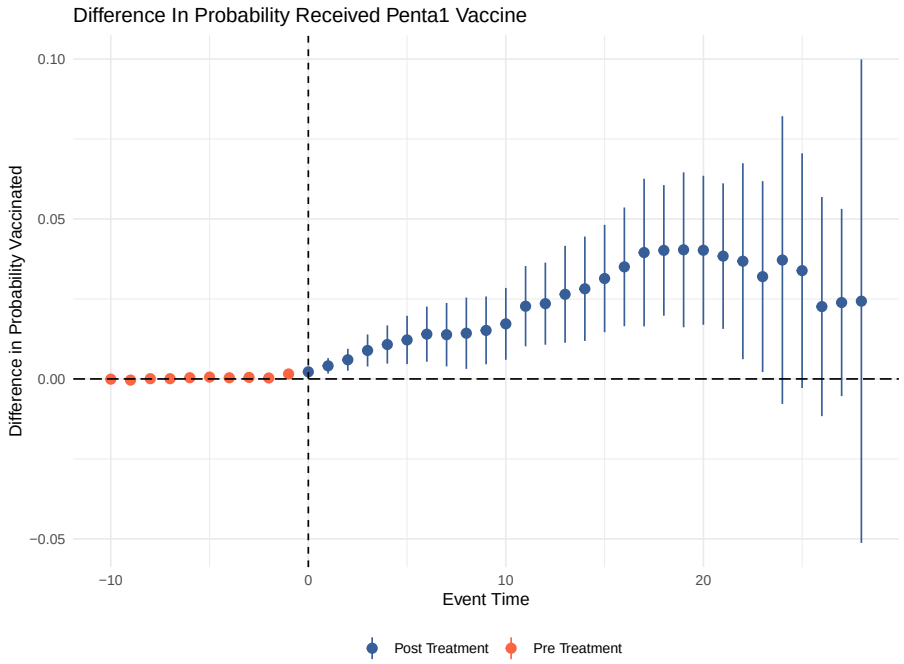
$$\widehat{ATT}^{ny}(g, t, b) = \frac{1}{N_g} \sum_{i=1}^{N_g} \Delta \mathbf{1}\{G_{gi} = 1, B = b\} - \frac{1}{N_{\text{nyt}(g)}} \sum_{i=1}^{N_{\text{nyt}(g)}} \Delta \mathbf{1}\{\text{nyt}(g)_i, B = b\}$$

Where $\Delta = V_{i,t} - V_{i,g-1}$ denotes an individual's $i's$ change in vaccination status at time t for those who first got treated at g. Since we condition on $B = b$ for both the treated and untreated difference our estimates make use of a conditional parallel trends assumption – we only require common trends to hold across districts for children with the same due-date. To aggregate back to a single ATT we calculate: $ATT(g, t) = \sum_{b \in \mathcal{T}} ATT_t(g, b) Pr(B_i = b)$. Since all our aggregations are just linear transformations of ATTs we can calculate standard errors by likewise transforming the empirical influence function discussed in Section 3.5.

We present results using individual-child level data for Pentavalent-1 and Measles-1 as outcomes in figures 8 and 9 and for other vaccines in the schedule in the online appendix section A10. In line with town-level estimates, our individual-level estimates don't show any pretends before the programme roll-out, and all the pre-treatment estimates are close to zero, satisfying the parallel trend assumption.

Conditional on the child's enrollment in the SEIR databases, there is a significant rise in the likelihood of receiving vaccination for Pentavalent-1 attributed to the implementation of the CCTs program. We find estimates within the range of 0 to 0.035, an equivalent of a 3.56% increase in the likelihood of receiving Pentavalent-1 vaccines attributable to the CCTs program. The findings suggest that individuals enrolled in the registry demonstrate a greater likelihood of receiving the Pentavalent-1 vaccine earlier in the schedule. Nevertheless, this effect shows a declining trend around 20 weeks and becomes statistically

insignificant by approximately 26 weeks, as illustrated in Figure 8. The individual level estimates for measles-1 indicate a significant increase due to the CCTs programme till week 5, but the effect diminishes after that (presented in figure 9). Our individual-level estimates are in line with aggregated town-level estimates (presented in section 4.1), though we find a smaller effect size. The results are also consistent for other vaccines including Pentavalent-2 and Pentavalent-3 (presented in figures A14 and A15).



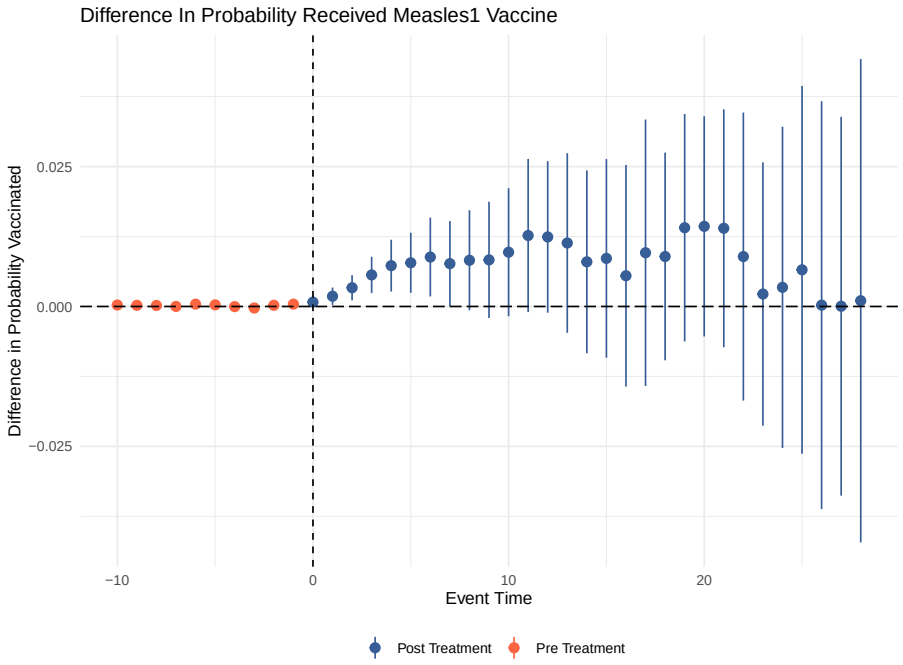
Notes: The figure displays the event-study plots constructed from difference in difference using (Callaway and Sant'Anna, 2021) method (in dots) using individual child level data. The outcome variable is the log of the number of Pentavalent-1 immunizations administered. We used a balanced panel for 29 towns treated for at least 52 weeks. 23 towns used as nevertreated units from additional, nonmCCT districts were used as a control gorup. We estimate 20 preperiods since estimating more preperiods dramatically increases the standard errors in the preperiod. Red dots indicate pre-treatment estimates, and blue indicates post-treatment estimates. The bars represent 95% percent confidence intervals. Standard errors are clustered at the district level, calculated using the wild cluster bootstrap.

Figure 8: Event Study estimates from Difference in Difference using Callaway and Sant' Anna: Individual-level estimates for Pentavalent-1

4.3.1 Heterogeneity in Effects: Enroled in the programme versus new enrolled

In this section we decompose our effects by those who are already enrolled in the ZM system before treatment begins, 'always-takers', and those who enrol after the program rolled in for their town, 'late-takers'.[33] We estimate the average treatment effect for those treated at g at time t who are born in b and

[33] These are a mix of 'always-takers' who would have taken the programme anyway, and 'compliers' who are induced into the programme by the incentive. Without additional assumptions we cannot disentangle the two in the latter group.



Notes: The figure displays the event-study plots constructed from difference in difference using (Callaway and Sant'Anna, 2021) method (in dots) using individual child level data. The outcome variable is the log of the number of Measles-1 immunizations administered. We used a balanced panel for 29 towns treated for at least 52 weeks. 23 towns used as nevertreated units from additional, nonmCCT districts were used as a control group. We estimate 20 preperiods since estimating more preperiods dramatically increases the standard errors in the preperiod. Red dots indicate pre-treatment estimates, and blue indicates post-treatment estimates. The bars represent 95% percent confidence intervals. Standard errors are clustered at the district level, calculated using the wild cluster bootstrap.

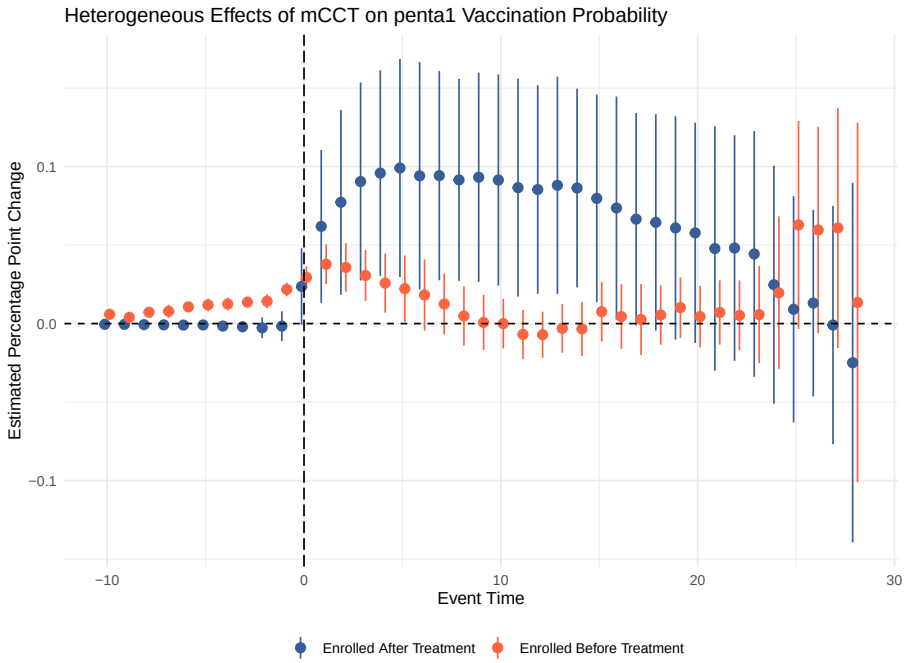
Figure 9: Event Study estimates from Difference in Difference using Callaway and Sant' Anna: Individual-level estimates for Measles-1

based on their enrolment using the following specification:

$$\widehat{ATT}^{ny}(g,t,b,\text{pre-enrol}=a) = \frac{1}{N_{ga}}\sum_{i=1}^{N_{ga}} \Delta\mathbf{1}\{G_{gi}=1, B=b, \text{pre-enrol}=a\} - \frac{1}{N_{\text{nyt}(g)}}\sum_{i=1}^{N_{\text{nyt}(g)}} \Delta\mathbf{1}\{\text{nyt}(g)_i, B=b\}, \quad a=\{0,1\}$$

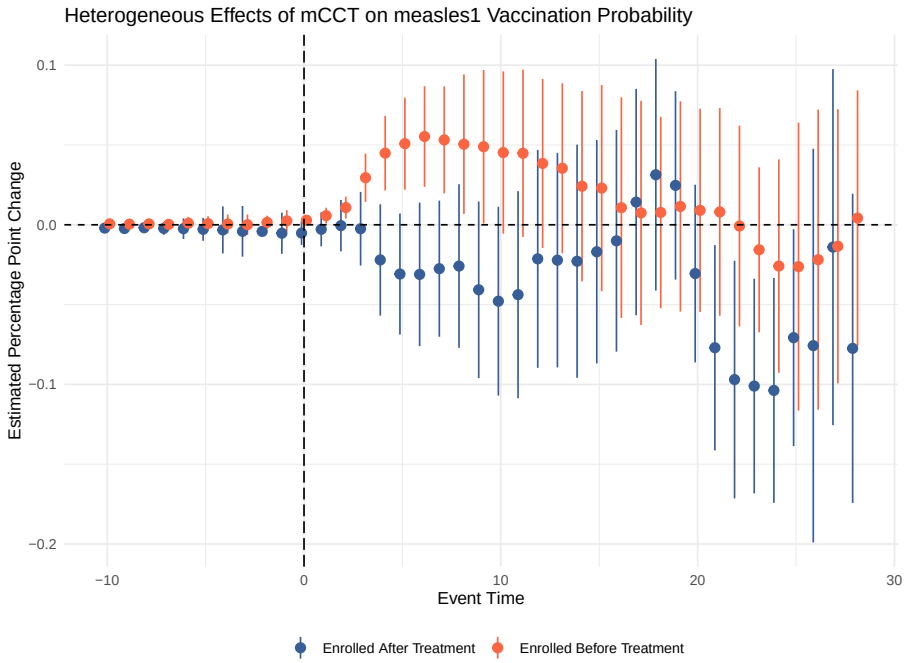
Separating out treatment effects by enrolment status lets us determine whether the program successfully incentivises pro-vaccine caregivers, who have already enrolled their children in ZM before they were eligible for the cash incentive, to continue later into the vaccine schedule. Post-enrolment estimates give a sense of how likely marginal caregivers are to persist in the ZM system once enrolled. We present heterogeneous individual child-level results for Pentavalent-1 and Measles-1 (presented in figures 10 and 11). Each graph presents the heterogeneous treatment for the two groups, i.e., those who were already enrolled in the system before the program started (represented by red dots) and those who were enrolled in the system after the program rolled out, i.e., they are newly enrolled (represented by blue dots). Our findings suggest that those who were enrolled in the system before the CCTs programme was rolled experienced an increase in Pentavalent-1 uptake and they are more likely to persist by 15%. As these children were already pro-vaccine and likely to receive timely vaccines, the program's impact diminished rapidly. Conversely, those who enrolled in the system after the program rollout exhibited a sharp increase in Pentavalent-1 vaccinations due to the program. We find that the programme is able to get these marginal children into the system but, they are less likely to persist after 25 weeks. Their initial motivation appeared to be driven by the incentive to get vaccinated, but sustaining their engagement in the system proved challenging.

In sum, our individual level analysis of heterogeneous treatment effect indicates that the scale up of the CCTs programme led to a 15% increase in enrollment in the vaccination registry. We observe notable heterogeneity among individuals, with marginal children induced to take first vaccine due to incentives but they are much less likely to persist. On the other hand, those who are inherently pro-vaccine, categorized as always-takers and enrolling pre-treatment, exhibit a much higher likelihood of completing the vaccination schedule.



Notes: The figure displays the event-study plots constructed from difference in difference using (Callaway and Sant'Anna, 2021) method (in dots) using individual child level data. The outcome variable is the log of the number of Pentavalent-1 immunizations administered. We used a balanced panel for 29 towns treated for at least 52 weeks. 23 towns used as nevertreated units from additional, nonmCCT districts were used as a control group. We estimate 20 preperiods since estimating more preperiods dramatically increases the standard errors in the preperiod. Red dots indicate pre-treatment estimates, and blue indicates post-treatment estimates. The bars represent 95% percent confidence intervals. Standard errors are clustered at the district level, calculated using the wild cluster bootstrap.

Figure 10: Event Study estimates from Difference in Difference using Callaway and Sant' Anna: Individual-level estimates for Pentavalent-1



Notes: The figure displays the event-study plots constructed from difference in difference using (Callaway and Sant'Anna, 2021) method (in dots) using individual child level data. The outcome variable is the log of the number of Measles-1 immunizations administered. We used a balanced panel for 29 towns treated for at least 52 weeks. 23 towns used as nevertreated units from additional, nonmCCT districts were used as a control group. We estimate 20 preperiods since estimating more preperiods dramatically increases the standard errors in the preperiod. Red dots indicate pre-treatment estimates, and blue indicates post-treatment estimates. The bars represent 95% percent confidence intervals. Standard errors are clustered at the district level, calculated using the wild cluster bootstrap.

Figure 11: Event Study estimates from Difference in Difference using Callaway and Sant' Anna: Individual-level estimates for Measles-1

4.4 Robustness

4.4.1 Crossovers and Spillovers

One concern with the validity of our estimates would be if caregivers realized that program districts were giving incentives and other districts were not and switched where they took their child to be immunized as a result. This would push up vaccinations in program districts and down in control districts and would give us a spurious positive impact of the program. We therefore tested to see whether the introduction of the mCCT program was associated with an increase in the number of caregivers switching between districts to benefit from the program. We find evidence of a slightly higher rate of switchers into program districts than into non-program districts suggesting that some people do change the clinic they bring their child to based on the payment. However the absolute numbers are very small and thus do not change the magnitude of the estimated effect size. In Figure 12, we plot the weekly proportion of vaccines that are 'switches' where a caregiver changes vaccination location between antigens for each type of switch. The dashed line shows the level of switching we'd expect if caregivers randomly chose their switching destination, using the pre-treatment rate of switching as a baseline rate.

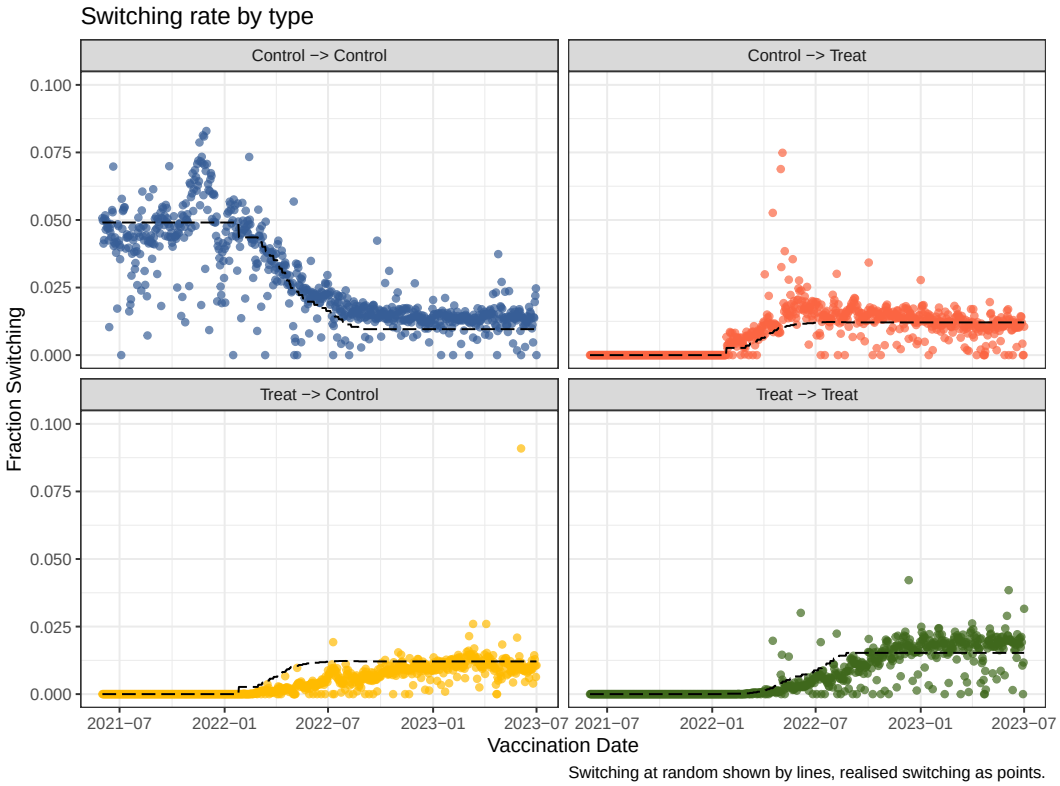


Figure 12: Fraction of Vaccines 'Switches' Between Districts

5 Discussion

In this section, we elaborate on the extent to which our findings have the potential to inform our understanding of the scale-up of pilot projects and conditional cash transfer programme on immunization today.

5.0.1 Sample and Prior Sensitivity

Another concern is that the districts least/extremely likely to be included in the RCT disproportionately drive results due to the inverse propensity score rising to high levels by the end of the period in the treatment/control terms respectively. To alleviate these concerns Appendix Figure A13 shows results are robust to excluding each district in turn, using a leave-one-district-out specification.

In Appendix Table A4 we find the prior specification has little impact on posterior treatment effects. Our main specification uses a conservative prior which places little density on regularising results more than 3 weeks apart. Appendix Table A4 shows results including a wider, less informative prior which spreads density out over a much greater range of shrinkage values. We find that Penta-1 and Measles-1 results are similar across both specifications: 15.17% (vs 13.86% for the conservative prior) and 13.59% (vs 13.47% for the conservative prior).

5.1 Validity Checks on Administrative Data

Another concern with the validity of our estimates is that they are based on administrative data which might systematically overrepresent vaccinations in program districts. In particular, if the introduction of the incentive program led vaccinators to create fake children or fake immunizations for real children in order to collect the incentive, this would bias upward the estimated impact of the program. The extent of fraud of this kind is limited by checks imposed by IRD which automatically flag, for further investigation by their in house monitoring team, when multiple children are linked to one phone number.[34] Nor can vaccinators create multiple accounts to collect incentives for fake children: anti-money laundering/anti-terrorist financing legislation makes it illegal to have more than five sim cards against one identity card.[35]

Nevertheless, to address concerns that the incentive leads to overreporting of number of children and immunizations in program districts we contacted a random sample of phone numbers (from program and non program districts) in the ZM database and a random sample of households with children under 3 years old in two of the program districts. The phone survey was designed to check whether the incentive

[34] In total, 89% of phone numbers are not repeated for different children in the ZM database; 9.4% are repeated for 2 children; 1.3% are repeated for 3 children and 0.5% are repeated for 4 or more children. This is consistent with the number of people who report sharing 1 phone number across multiple households in our household survey: 90% in Karachi East and 70% in Kambar report not sharing a phone number across households, while 2% of households in Karachi East and 11% of households in Kambar share phone numbers across 3 or more households.

[35] A computerized national identification card and bio-metric verification for each sim card makes this rule hard to get around.

had led to fake children being added to the register. The household survey tested whether additional immunizations had been added to records of real children.

The phone survey was conducted between November 2022 and February 2024 and sought to contact 2,131 randomly selected phone numbers from ZM (1,903 from program districts and 228 in non-program districts). The household survey was conducted in two program districts, one urban (Karachi East) and one rural (Kambar), chosen for their representativeness. As the household survey was not conducted in non program districts we cannot compare household survey findings from program and non program districts.[36] The household survey is completed with representative sample of 11,440 households.[37]

If a large number of fake accounts with phone numbers have been added to the ZM-EIR as a result of the incentive we would expect to either get lower than usual pick-up rates to our phone calls, or a large number of respondents reporting that they do not recognize the name of the child. These rates would also be systematically different in program and non program districts. We find evidence that households change phone number relatively frequently: the percent of successful calls declines the longer the time since the child's last vaccination and the percent of calls where the respondent does not recognize the child increases over time. We therefore focus mainly on calls made within 60 days of a child's most recent vaccination (the first two panels in Table 1), though we show the result of all calls for transparency. Surveys were successfully completed for 88% of randomly chosen phone numbers in program districts for children vaccinated within the previous 30 days (76% for those vaccinated in the last 31-60 days), a relatively high rate compared to other phone surveys we have conducted in Pakistan[38] The success rate is higher in program than non program districts for these time periods (p-value 0.027). We also find the rate of "child not recognized" is similar in program (5.1%) and non program districts (6.1%) for children vaccinated within 30 days and 7.5% in program vs 6.0% in non program for those vaccinated within 30-60 days. Overall, of 12 comparisons made between responses in program and non program districts in Table 1, three have p-values below 5: in two case program districts perform worse (they have a higher percentage of phones off in the 31-60 time period, and more child not recognized cases in 61-180 time period) and in one program districts perform better (there are more successful calls in the 31-60 time period).[39] These data suggest that the incentive has not led to a large number of fake children being added to the immunization registry as a result of the program which would undermine our estimate of the impact of the program.

[36] Kambar had an immunization rate in between that of the other two entirely rural districts and 95% of people speak the native (Sindhi) language. Karachi East is completely urban and offers considerable ethnic diversity (with 12% of the population speaking Sindhi language).

[37] The household survey and phone survey were conducted as part of an assessment of the program for the funder and was designed to corroborate the vaccination rate in these districts as well as check for signs of fake children or vaccines in the ZM data and incentive disbursement.

[38] In an ongoing study, Christina Brown and Maryiam Haroon contacted casual day labourers by phone after observing them on a full-day construction site for three consecutive days and found a pickup rate of 64.5% in Pakistan.

[39] The one case where the rates are very different in magnitude is child not recognized in the 61-180 time period. Here the non program district finding is out of line with the trend of rising child not recognized the longer the period since the child was vaccinated.

Time period	District	Phone off (%)	p-value	Successful calls (%)	p-value	Child not recognized (%)	p-value	N
0-30	nonmCCT	0	1.000	82.14	0.357	6.12	1.000	56
	mCCT	1.14		88.03		5.11		142
31-60	nonmCCT	4.25	0.054	64.89	0.027	5.97	0.829	94
	mCCT	10.90		76.10		7.53		431
61-180	nonmCCT	20.51	0.345	62.82	1.000	1.92	0.001	78
	mCCT	16.24		63.38		12.52		1330
180+	nonmCCT	24.53	0.457	47.83	1.000	18.67	0.226	115
	mCCT	20.98		48.22		24		591

Table 1: Phone survey results across mCCT and nonmCCT: Calls over time

For the household survey we used publicly available Union Council boundaries and satellite maps to divide Union Councils into equal sized segments and randomly selected one segment per Union Council. We conducted a census of all households within the selected segment. This listing was used to randomly select households with children under the age of three for a household survey. Following the survey, we matched surveyed households to the ZM (using QR codes on immunization cards and/or phone numbers). In addition to other tests for the existence of discrepancies in ZM data, we matched the immunization records collected from the household survey (from cards and caregiver recall) to the immunization record for the same child in the electronic record. The objective was to test whether there was evidence of systematic overreporting of immunizations of real children in ZM.

We are able to locate 68% of children from the household survey in the administrative data in Karachi East (57% because they had their vaccination cards and an additional 11% from phone numbers). The equivalent for Kambar is 54% (49% with vaccine cards). Among this subgroup we find a 91% match in the vaccination record between the two, in other words both report the same number of vaccinations in 91% of cases (see Appendix Section A6, Figure A7, and Figure A8). Where there is a mismatch, there are cases of a vaccine being recorded in the household data and not in ZM and vice versa. On average, BCG rates tend to be higher in household data than in ZM while other vaccines tend to be higher in ZM than in the household survey. This could be the result of overreporting in ZM for non-BCG vaccines or it could be that vaccinators do not record the vaccination on the card when they are busy and include them in the registry later. It is worth noting that the high match rate is calculated on a selected sample–i.e. those who we can find in the administrative data, often because they showed us their vaccination card. Nevertheless we take these results as suggesting there is not a widespread practice of adding fake vaccines to the records of real children. As the household survey was only conducted in

program districts, we do not know if this match rate is higher in non-program districts[40].

6 Conclusion

We present evidence from a scale up of a small incentive for immunization (known as mCCT) program designed to provide a nudge to encourage the take up of childhood immunization in Pakistan. The magnitude of the effect is large – a 14.6% increase in Penta-1 and 14.8% increase in Measles-1 vaccines. The incentive for vaccination helped those who had at least one immunization for their child to persist to receive subsequent vaccines and higher number of vaccines. It also brought a large number of new children into the system who would otherwise have not enrolled through anyother vaccination program.

Our findings suggest that a large scale government program produces impacts on immunizations consistent with the findings from evaluations of smaller scale, often researcher implemented, small conditional incentives (Holla and Kremer, 2009; Morris et al., 2004; Gibson et al., 2017; Banerjee et al., 2010). A review of the existing literature on small conditional transfers finds that payments of less than USD 3, complemented with reminders increased full immunization coverage in Kenya by 8 percentage points (4 percentage points versus SMS only) (Gibson et al., 2017), in-kind small transfers (in form of lentils and a set of plates) costing less than 1 USD per immunization increased full immunization coverage in India by 21 percentage points (Banerjee et al., 2010), small airtime CCTs of USD 0.5 per immunization increased coverage by 17 full immunization coverage over baseline estimates in another RCT from India (Banerjee et al., 2021) and food/medicine vouchers worth USD 2 doubled up-to-date DTP3 (Diphtheria, Tetanus, Pertussis) coverage at 18 months in Pakistan (Chandir et al., 2010).

Our estimates of 14.6% for Pentavalent-1 and 14.8% for Measles-1 resulting from an incentive of USD 1.25 per immunization are close to, though somewhat higher than, those found in the pilot study (Chandir et al. (2022), which finds results of 9.3% and 11.0% increases respectively). One reason a large scale, government program might have larger impacts than a small scale pilot is that information about the incentive could reach those caregivers who had not yet taken their child to receive any incentive. In contrast, caregivers in Chandir et al. (2022) were only enrolled in the RCT if they brought their child for an initial vaccine.

More broadly this study demonstrates the benefit of combining variation from a randomized control trial with that from units outside the RCT when the selection criteria for inclusion in the randomized study is known. Knowledge of the selection criteria allows researchers to exploit quasi random variation in whether a unit ends up above or below a cutoff and thus is or is not included in an RCT (and thus a program). This is particularly useful when evaluating the scale up of programs where an element of randomization can be included but implementation priorities take precedence over the need for the ideal

[40] In a separate study, we are enroling households into a new program and following the same process to match households from household survey with ZM and find that the match rates are quite similar. The enrolment in the new program is done by the research team

evaluation. In this study, a separate on the ground validation of the administrative date was carried out as part of the evaluation. However, similar evaluations could be carried out relying entirely on administrative data at low cost. We see potential to use the methodology in this paper to rigorously evaluate the roll out of large government programs with limited disruption and cost. The methodology can also be used to improve the precision and generalizability of estimates in smaller scale RCTs where administrative data is available for units outside the RCT.

References

ARKHANGELSKY, D., G. W. IMBENS, L. LEI, AND X. LUO (2023): "Design-Robust Two-Way-Fixed-Effects Regression For Panel Data," .

ATHEY, S. AND G. W. IMBENS (2022): "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," Journal of Econometrics, 226, 62–79.

BANERJEE, A., A. CHANDRASHEKHAR, E. DUFLO, S. DALPATH, J. FLORETTA, M. JACKSON, H. KANNAN, A. SCHRIMPF, AND M. SHRESTHA (2021): "Evaluating the impact of interventions to improve full immunisation rates in Haryana, India," .

BANERJEE, A. V., E. DUFLO, R. GLENNERSTER, AND D. KOTHARI (2010): "Improving immunisation coverage in rural India: clustered randomised controlled evaluation of immunisation campaigns with and without incentives," Bmj, 340.

BORUSYAK, K., X. JARAVEL, AND J. SPIESS (2024): "Revisiting Event Study Designs: Robust and Efficient Estimation," .

CALLAWAY, B. AND P. H. C. SANT'ANNA (2021): "Difference-in-Differences with Multiple Time Periods," Journal of Econometrics, 225, 200–230.

CALLAWAY, B. AND P. H. SANT'ANNA (2021): "Difference-in-differences with multiple time periods," Journal of Econometrics, 225, 200–230.

CENTRE FOR DISEASE CONTROL AND PREVENTION (2023): "Fast Facts on Global Immunization," Retrieved from: https://www.cdc.gov/globalhealth/immunization/data/fast-facts.html.

CHANDIR, S., A. KHAN, H. HUSSAIN, H. USMAN, S. KHOWAJA, N. HALSEY, AND S. OMER (2010): "Effect of food coupon incentives on timely completion of DTP immunization series in children from a low-income area in Karachi, Pakistan: a longitudinal intervention study," Vaccine, 28, 3473–3478.

CHANDIR, S., D. A. SIDDIQI, S. ABDULLAH, E. DUFLO, A. J. KHAN, AND R. GLENNERSTER (2022): "Small mobile conditional cash transfers (mCCTs) of different amounts, schedules and design to improve routine childhood immunization coverage and timeliness of children aged 0-23 months in Pakistan: An open label multi-arm randomized controlled trial," EClinicalMedicine, 50.

DE CHAISEMARTIN, C. AND X. D'HAULTFŒUILLE (2020): "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," American Economic Review, 110, 2964–2996.

FALETTO, G. (2023): "Fused Extended Two-Way Fixed Effects for Difference-in-Differences with Staggered Adoptions," .

GECHTER, M. AND R. MEAGER (2021): "Combining Experimental and Observational Studies in Meta-Analysis: A Mutual Debiasing Approach," Tech. rep., Mimeo.

GIBSON, D. G., B. OCHIENG, E. W. KAGUCIA, J. WERE, K. HAYFORD, L. H. MOULTON, O. S. LEVINE, F. ODHIAMBO, K. L. O'BRIEN, AND D. R. FEIKIN (2017): "Mobile phone-delivered reminders and incentives to improve childhood immunisation coverage and timeliness in Kenya (M-SIMU): a cluster randomised controlled trial," The Lancet Global Health, 5, e428–e438.

GLENNERSTER, R. (2022): "Effect of Mobile Conditional Cash Transfers (mCCTs) Roll out on Routine Childhood Immunization Coverage and Timeliness in Sindh, Pakistan," Pre-registered RCT in the AEA RCT Registry, accessed March 22, 2024. The study aims to evaluate the impact of mobile phone-based conditional cash transfers on routine childhood immunization coverage and timeliness in Sindh, Pakistan. It involves the use of the ZM Electronic Immunization Registry across Sindh for children aged 0-23 months against Vaccine Preventable Diseases. The program, with IRD as the monitoring partner, plans to introduce mCCTs in 7 high-risk districts.

GOODMAN-BACON, A. (2021): "Difference-in-Differences with Variation in Treatment Timing," Journal of Econometrics, 225, 254–277.

HOLLA, A. AND M. KREMER (2009): "Pricing and access: Lessons from randomized evaluations in education and health," Center for Global Development working paper.

MORRIS, S. S., R. FLORES, P. OLINTO, AND J. M. MEDINA (2004): "Monetary incentives in primary health care and effects on use and coverage of preventive health care interventions in rural Honduras: cluster randomised trial," The Lancet, 364, 2030–2037.

NATIONAL INSTITUTE OF POPULATION STUDIES (2019): "Pakistan demographic and health survey 2017-18," Retrieved from: https://dhsprogram.com/pubs/pdf/FR354/FR354.pdf.

ROSENBAUM, P. R. AND D. B. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika, 70, 41–55.

SHAO, J. AND D. TU (1995): "Bayesian Bootstrap and Random Weighting," in The Jackknife and Bootstrap, ed. by J. Shao and D. Tu, New York, NY: Springer, Springer Series in Statistics, 416–446.

SUN, L. AND S. ABRAHAM (2021): "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," Journal of Econometrics, 225, 175–199.

WORLD HEALTH ORGANIZATION (2023): "Immunization coverage," Retrieved from: https://www.who.int/news-room/fact-sheets/detail/immunization-coverage.

Appendix

A1 Treatment Status by District

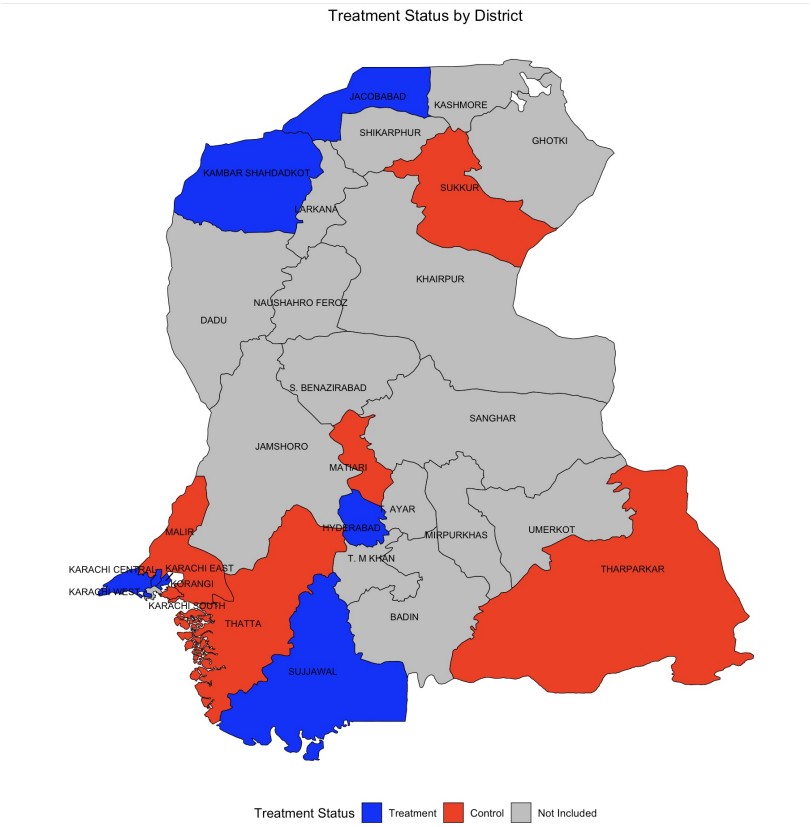


Figure A1: Treatment Status by District

A2 Timeline of the roll-out

Figure A2: Timeline for the rollout of the CCT programme

A3 Vaccination Card and Reminders

Figure A3: Vaccination card

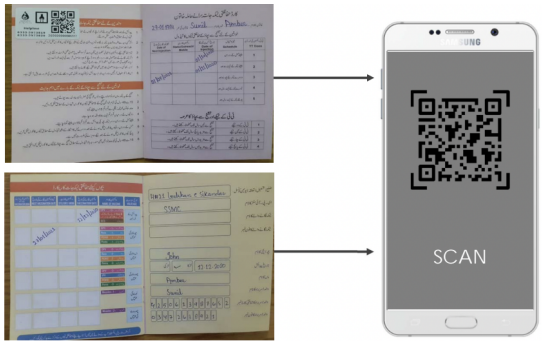


Figure A4: Vaccination card with QR code to update SEIR database

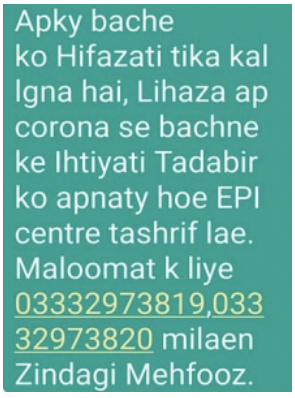


Figure A5: Reminder messages

A4 Vaccination Schedule

Pakistan's routine childhood immunization schedule includes the following vaccines to be administered in 6 visits:

(i). First visit: Bacille Calmette-Guérin (BCG), Oral Polio Vaccine-0 (OPV) and Hepatitis-B at birth;

(ii). Second visit: Pentavalent-1, Rotavirus-1, Oral Polio Vaccine-1 Pneumococcal Conjugate Vaccine (PCV)-1 at six-weeks;

(iii). Third visit: Pentavalent-2, Rotavirus-2, Oral Polio Vaccine-2 and Pneumococcal Conjugate Vaccine-2 at 10 weeks;

(iv). Fourth visit: Pentavalent-3, Oral Polio Vaccine-3, Pneumococcal Conjugate Vaccine-3, Inactivated Polio Vaccine (IPV)-1 at fourteen weeks;

(v). Fifth visit: Typhoid conjugate vaccine (TCV), Measles–Rubella-1, Inactivated Polio Vaccine-2 at nine months;

(vi). Sixth visit: Measles–Rubella-2 vaccine at 15 months.

Algorithm 1 $p_{i,g,t}$ Algorithm

Input: Set of dates $DATES$, set of vaccines $VACCINES$, coverage rates matrix $COVERAGE$ with rows representing districts and columns representing vaccines, and a list of towns within each district.
Output: Record of treatment for each town at each roll-in date.
for all dates d in $DATES$ **do**
  for all vaccine pairs v in $VACCINES$ **do**
    Permute through date d and vaccine v
    Calculate 20% threshold coverage rate $C_{\text{threshold}}$ for vaccine pair v and data dump generation date d
    Find districts $D_{<20\%}$ with coverage rates $< C_{\text{threshold}}$
    for all district D in $D_{<20\%}$ **do**
      Randomize order of district roll-in
      Randomize order of towns within district D
      for all roll-in dates $g \in \mathcal{G}$ **do**
        for all time periods $t \in \mathcal{T}$ **do**
          for all town i in districts $D_{<20\%}$ **do**
            return Record treatment status, H^s_{igt}, for permutation s, town i, roll-in date g, at time t.
          end for
        end for
      end for
    end for
  end for
end for
Calculate: $p_{i,g,t} = \frac{1}{S} \sum_{s=1}^{S} H^s_{igt}$

A5 Propensity Score Examples

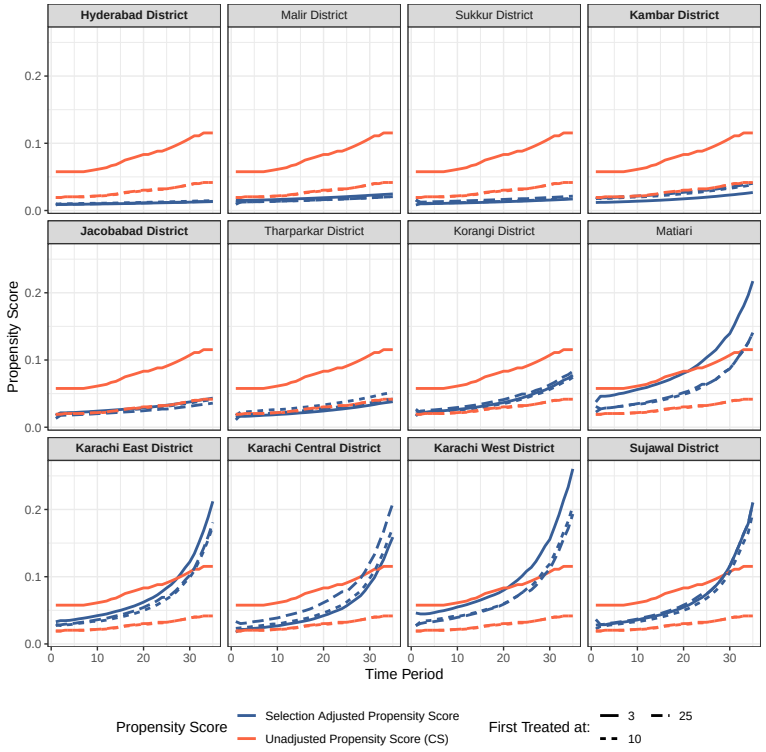


Figure A6: Comparison of Propensity Scores, Accounting for Selection vs Unadjusted

A6 Vaccination matches

We used the household survey data and matched it with SEIR data to link the vaccination records. We started by matching on QR codes when households had their vaccination card and then moved to match on phone numbers when households gave us phone numbers. There are several reasons why caregivers might be unable to provide their child's vaccination card: they could be misplaced, discarded, or the caregivers might be unwilling or unable to fetch the card. After linking the two datasets, we check to see if vaccination records are the same in the household survey as they are in the SEIR database. Reasons of lack of a perfect match include households forgetting about some vaccines (when match is on phone number) and vaccinators not putting vaccines into the system because the internet was down when the child was vaccinated, or the clinic was too busy and they forget afterwards. Using the subset of observations from the household survey matched through QR codes and (separately) those matched on phone numbers, we calculate the consistency between the household survey (relying on vaccination cards), and the SEIR records, at the individual vaccination level.

							Karachi East: QR					
	BCG	Penta 1	Penta 2	Penta 3	Measles 1	Measles 2	Polio (at Birth)	Polio 1	Polio 2	Polio 3	PCV 1	P
Vax Match+	97.1%	94.9%	92.3%	91.5%	92.0%	91.7%	66.5%	96.5%	93.4%	91.3%	94.9%	92
Only in HH++	81.3%	14.0%	19.8%	31.6%	26.7%	41.9%	99%	23.1%	25.7%	34%	12.3%	18
Vax. Date Match&	71.0%	72.9%	71.7%	73.0%	70.8%	70.4%	70.1%	71.4%	70.2%	72.2%	69.0%	74
Mean Diff. Days**	66.90	81.96	64.31	88.84	84.61	112.15	69.58	64.04	75.13	81.91	65.47	7
Overall N***	1,119											
Overall match rate ^	91.2%											

+ Percent of specified vaccine where status (vaccinated or unvaccinated) matches between SEIR and household survey data for subset matched on QR codes. Missing data are treated as having the status unvaccinated. Vaccines recorded in SEIR as being given after the survey date are excluded. Excludes observations where SEIR vaccination status or date is missing for the last vaccination recorded.
++ Percent of unmatched vaccines (one record indicates vaccine received, the other doesn't), where survey indicates vaccine was received, but value in SEIR is either missing or 0. A rate below 50% implies that the rate of specified vaccine is higher in household survey than SEIR.
& Of subset of vaccines that match, and are not missing dates, what % have the same date recorded for the vaccine.
** For the subset where a vaccine is recorded as received, but the date does not align between SEIR and household survey, how many days (magnitude) on average are between the dates recorded for the vaccine.
*** Consists of number of observations in survey.
^ Average of matching across all vaccines.

	Kambar : QR									
	BCG	Penta 1	Penta 2	Penta 3	Measles 1	Measles 2	Polio (at Birth)	Polio 1	Polio 2	P
Vax Match+	98.7%	93.0%	85.3%	82.4%	83.8%	87.0%	74.1%	94.3%	86.2%	82
Only in HH++	33.3%	8.7%	3.6%	3.5%	7.0%	10.8%	99.3%	21.5%	6.4%	4
Vax. Date Match&	69.6%	61.8%	63.3%	62.1%	61.4%	58.4%	67.4%	62.9%	61.4%	58
Mean Diff. Days**	101	104	108	110	112	162	104	85	102	
Overall N***	1,141									
Overall match rate ^	86.7%									

+ Percent of specified vaccine where status (vaccinated or unvaccinated) matches between SEIR and house
data for subset matched on QR codes. Missing data are treated as having the status unvaccinated. Vaccines
SEIR as being given after the survey date are excluded. Excludes observations where SEIR vaccination statu
missing for the last vaccination recorded.
++ Percent of unmatched vaccines (one record indicates vaccine received, the other doesn't), where survey
vaccine was received, but value in SEIR is either missing or 0. A rate below 50% implies that the rate of spe
vaccine is higher in household survey than SEIR.
& Of subset of vaccines that match, and are not missing dates, what % have the same date recorded for the
** For the subset where a vaccine is recorded as received, but the date does not align between SEIR and hou
survey, how many days (magnitude) on average are between the dates recorded for the vaccine.
*** Consists of number of observations in survey.
^ Average of matching across all vaccines.

Online Appendix

A7 Town-level Estimates: Callaway and Sant' Anna method for other vaccines with 5 additional districts as control

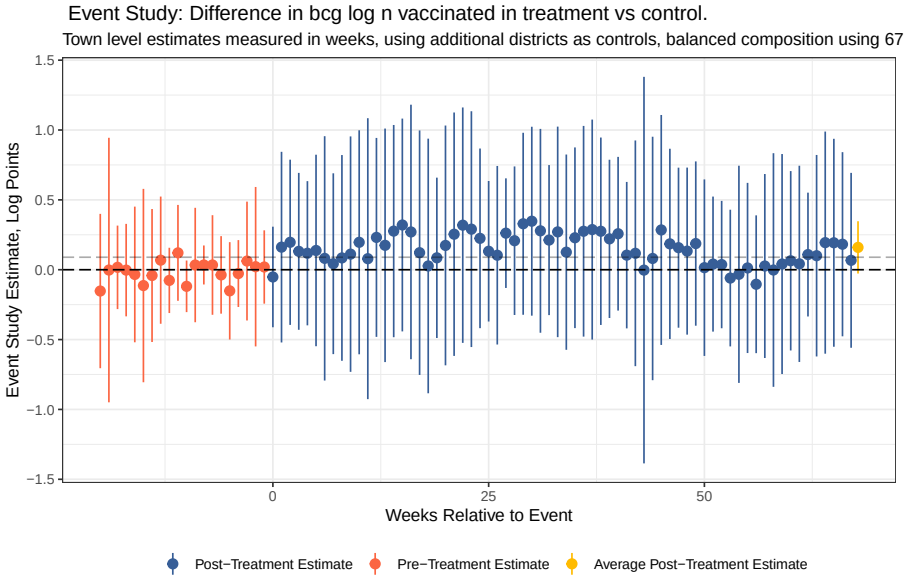
In this section, we present our estimations using the Callaway and Sant'Anna (2021) difference-in-difference method to estimate the impact of mCCTs with the log of the number of other vaccines, including Bacille Calmette-Guerin, Pentavalent-2, Pentavalent-3, and Measles-2 (presented in figures A9, A10, A11 and A12). Our estimates allow us to compare outcomes in a treated town after the roll-out for 1, 2, 3, etc. weeks, to outcomes in control towns where the programme has not yet been rolled out. For the estimations, we used seven treatment districts and five additional control districts to estimate the impact of the programme on immunization. Our estimations use a balanced panel of 29 treated towns observed for at least 67 weeks post-programme roll-out. We use 20 weeks before the treatment roll-out for the pre-programme period. In all estimations, we find that before the roll-out, our estimates do not exhibit pretrends and are consistent with parallel trend assumption (presented in figures A9 to A12). The CCT programme increased immunization for BCG by 15.94%, Pentavalent-2 by 9.82%, and Pentavalent-3 by 5.91%. The estimates show an initial increase in BCG with an overall increase of 13.14 percentage points. Our BCG estimates are also close to those of Chandir et al. (2022) from the pilot study (presented in figure A9). Additionally, Pentavalent-2 estimates show an overall increase of 6.68 percentage points in the number of vaccines due to the programme. The impact on Pentavalent-2 is also less than those found in the pilot study (Chandir et al., 2022). The overall increase in Pentavalent-3 is 3.4 percentage points due to the mCCT programme roll-out. However, on average, the estimates are close to zero over time, with a negative impact in a few weeks. Lastly, the overall impact of Measles-2 is negative by 0.3 percentage points, though the coefficient is small and close to zero. We summarise the overall effects in Table A1.

	ΔN Vax (%)	ΔVax Rate MICS (ppt)	ΔVax Rate EPI (ppt)	ΔN Vax N Vaccines, 52 wks
BCG	15.94 (10.8, 21.03)	13.14 (8.9, 17.33)	10.18 (6.9, 13.43)	63,395 (45,897, 78,739)
Penta-1	14.59 (10.76, 18.66)	10.95 (8.08, 14.02)	9.17 (6.77, 11.73)	57,895 (44,539, 71,178)
Penta-2	9.82 (6.21, 13.37)	6.68 (4.23, 9.11)	5.57 (3.52, 7.58)	30,183 (19,652, 40,820)
Penta-3	5.91 (1.4, 10.19)	3.4 (0.81, 5.87)	3.14 (0.74, 5.43)	10,628 (-2,746, 22,967)
Measles-1	14.75 (10.57, 19.06)	9.25 (6.63, 11.95)	7.76 (5.56, 10.02)	52,196 (39,546, 64,666)
Measles-2	-0.58 (-0.58, -0.58)	-0.3 (-0.3, -0.3)	-0.23 (-0.23, -0.23)	1,240 (-12,967, 14,970)

Table A1: CS Estimates - All Vaccines w/ Additional Districts

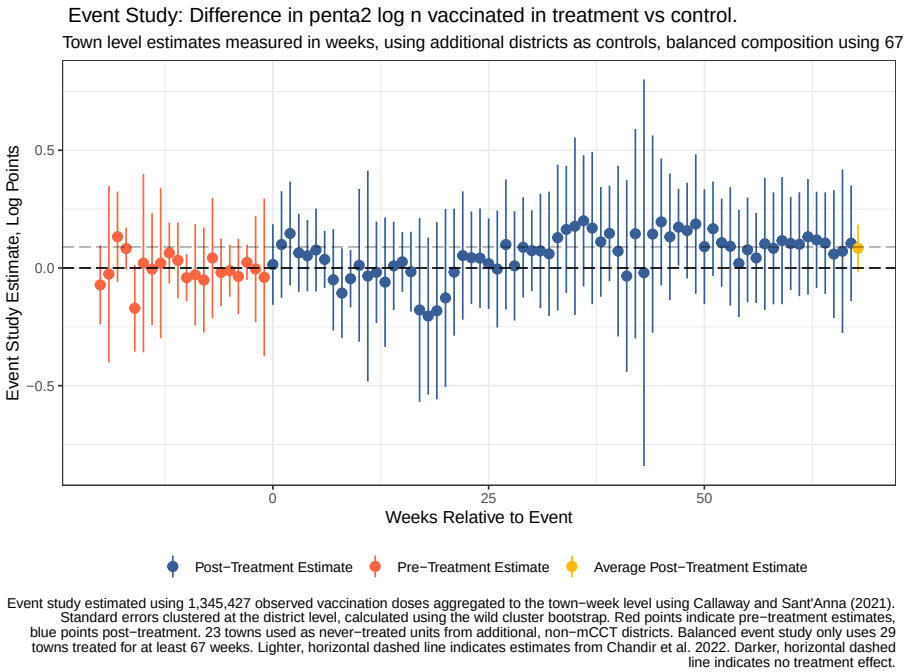
In sum, our findings for other vaccines in the schedule, including BCG, Pentavalent-2, Pentavalent-3, and

Measles-2, suggest that the mCCT programme had a positive and significant impact on other vaccines, which are earlier in the schedule, but as we move towards later vaccines, the estimate is close to zero and negative in the case of Measles-2.



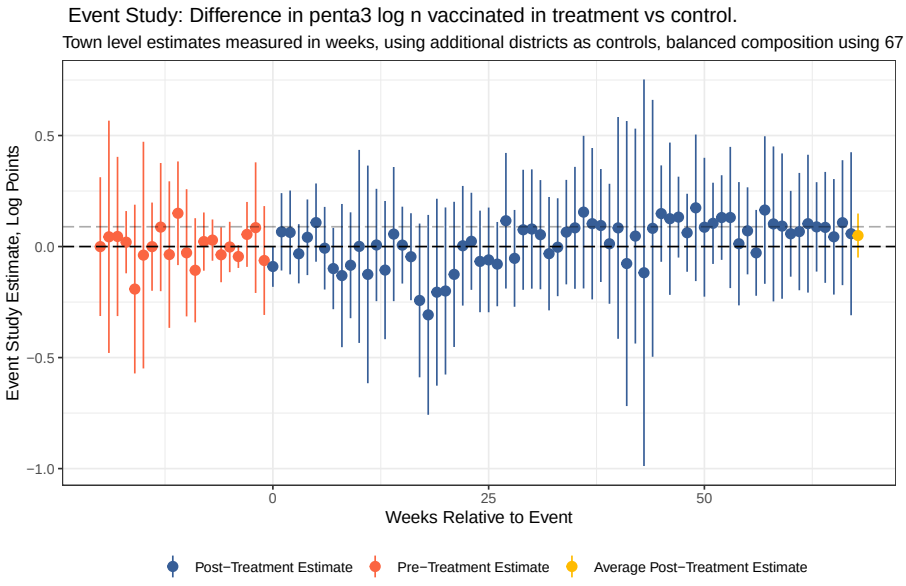
Notes: The figure displays the event-study plots constructed from difference in difference using (Callaway and Sant'Anna, 2021) method (in dots). The outcome variable is the log of the number of Bacille Calmette-Guerin immunizations administered. We used a balanced panel for 29 towns treated for at least 67 weeks. 23 towns used as nevertreated units from additional, nonmCCT districts were used as a control group. We estimate 20 preperiods since estimating more preperiods dramatically increases the standard errors in the preperiod. Red dots indicate pre-treatment estimates, and blue indicates post-treatment estimates. The orange dot indicates the average treatment effect averaging across time. The gray dashed line shows the estimated impact found in Chandir et al. (2022) for Pentavalent-3. The estimates are derived using 1,153,253 vaccinations observed at the individual level and aggregated to the town-week level. The bars represent 95% percent confidence intervals. Standard errors are clustered at the district level, calculated using the wild cluster bootstrap.

Figure A9: Event Study estimates from Difference in Difference using Callaway and Sant' Anna: Town-level estimates for Bacille Calmette-Guerin



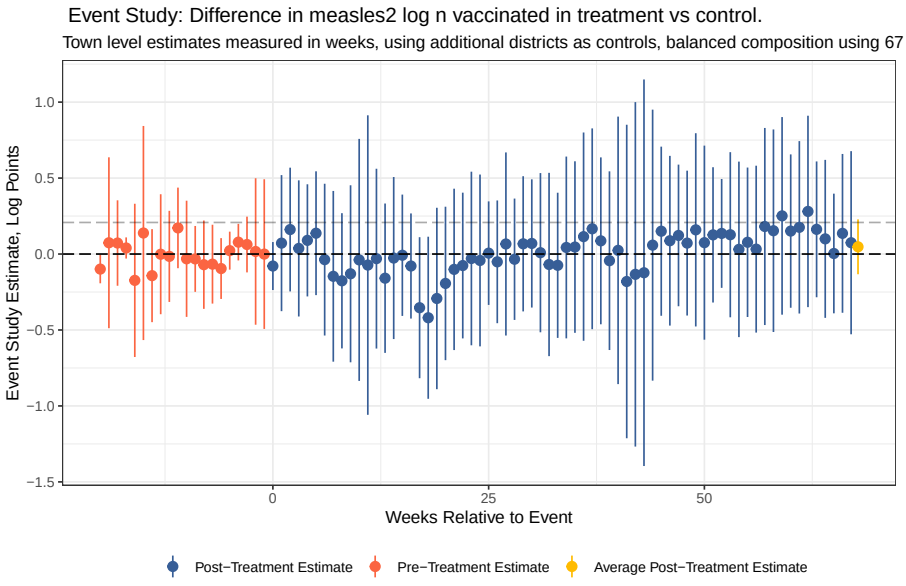
Notes: The figure displays the event-study plots constructed from difference in difference using (Callaway and Sant'Anna, 2021) method (in dots). The outcome variable is the log of the number of Pentavalent-2 immunizations administered. We used a balanced panel for 29 towns treated for at least 67 weeks. 23 towns used as nevertreated units from additional, nonmCCT districts were used as a control group. We estimate 20 preperiods since estimating more preperiods dramatically increases the standard errors in the preperiod. Red dots indicate pre-treatment estimates, and blue indicates post-treatment estimates. The orange dot indicates the average treatment effect averaging across time. The gray dashed line shows the estimated impact found in Chandir et al. (2022) for Pentavalent-3. The estimates are derived using 1,095,705 vaccinations observed at the individual level and aggregated to the town-week level. The bars represent 95% percent confidence intervals. Standard errors are clustered at the district level, calculated using the wild cluster bootstrap.

Figure A10: Event Study estimates from Difference in Difference using Callaway and Sant' Anna: Town-level estimates for Pentavalent-2



Notes: The figure displays the event-study plots constructed from difference in difference using (Callaway and Sant'Anna, 2021) method (in dots). The outcome variable is the log of the number of Pentavalent-3 immunizations administered. We used a balanced panel for 29 towns treated for at least 52 weeks. 23 towns used as nevertreated units from additional, nonmCCT districts were used as a control group. We estimate 20 preperiods since estimating more preperiods dramatically increases the standard errors in the preperiod. Red dots indicate pre-treatment estimates, and blue indicates post-treatment estimates. The orange dot indicates the average treatment effect averaging across time. The gray dashed line shows the estimated impact found in Chandir et al. (2022) for Pentavalent-3. The estimates are derived using 1,137,541 vaccinations observed at the individual level and aggregated to the town-week level. The bars represent 95% percent confidence intervals. Standard errors are clustered at the district level, calculated using the wild cluster bootstrap.

Figure A11: Event Study estimates from Difference in Difference using Callaway and Sant' Anna: Town-level estimates for Pentavalent-3



Notes: The figure displays the event-study plots constructed from difference in difference using (Callaway and Sant'Anna, 2021) method (in dots). The outcome variable is the log of the number of Measles-2 immunizations administered. We used a balanced panel for 29 towns treated for at least 52 weeks. 23 towns used as nevertreated units from additional, nonmCCT districts were used as a control group. We estimate 20 preperiods since estimating more preperiods dramatically increases the standard errors in the preperiod. Red dots indicate pre-treatment estimates, and blue indicates post-treatment estimates. The orange dot indicates the average treatment effect averaging across time. The gray dashed line shows the estimated impact found in Chandir et al. (2022) for Pentavalent-3. The estimates are derived using 864,321 vaccinations observed at the individual level and aggregated to the town-week level. The bars represent 95% percent confidence intervals. Standard errors are clustered at the district level, calculated using the wild cluster bootstrap.

Figure A12: Event Study estimates from Difference in Difference using Callaway and Sant' Anna: Town-level estimates for Measles-2

A8 Town-level Estimates: Callaway and Sant'Anna estimates without shrinkage

Specification	Vaccine	Change in # Vaccinated (%)
CS estimates - no shrinkage	BCG	0.16 (-0.01, 0.33)
CS estimates - no shrinkage	Penta-1	0.15 (0.03, 0.27)
CS estimates - no shrinkage	Penta-2	0.08 (-0.02, 0.18)
CS estimates - no shrinkage	Penta-3	0.05 (-0.06, 0.16)
CS estimates - no shrinkage	Measles-1	0.15 (0.03, 0.28)
CS estimates - no shrinkage	Measles-2	0.05 (-0.14, 0.24)

Table A2: CS Estimates - All Vaccines w/ Additional Districts: No Shrinkage Model

Notes:

A9 Town-level Estimates: Empirical Bayes Gaussian Process Prior for all vaccines

	ΔN Vax (%)	ΔVax Rate MICS (ppt)	ΔVax Rate EPI (ppt)	ΔN Vax N Vaccines, 52 wks
BCG	15.94 (10.8, 21.03)	13.14 (8.9, 17.33)	10.18 (6.9, 13.43)	63,395 (45,897, 78,739)
Penta-1	14.59 (10.76, 18.66)	10.95 (8.08, 14.02)	9.17 (6.77, 11.73)	57,895 (44,539, 71,178)
Penta-2	9.82 (6.21, 13.37)	6.68 (4.23, 9.11)	5.57 (3.52, 7.58)	30,183 (19,652, 40,820)
Penta-3	5.91 (1.4, 10.19)	3.4 (0.81, 5.87)	3.14 (0.74, 5.43)	10,628 (-2,746, 22,967)
Measles-1	14.75 (10.57, 19.06)	9.25 (6.63, 11.95)	7.76 (5.56, 10.02)	52,196 (39,546, 64,666)
Measles-2	-0.58 (-0.58, -0.58)	-0.3 (-0.3, -0.3)	-0.23 (-0.23, -0.23)	1,240 (-12,967, 14,970)

Table A3: CS Estimates - All Vaccines w/ Additional Districts: Shrinkage Model

Notes:

Prior	Estimation Specification	Balanced	Balanced Post Duedate Impact	Balanced Post-2 wk Impact	Balanced Post-4 wk Impact	Balanced Post-8 wk Impact	Unbalanced
BCG							
Conservative Prior (default)	Adding 5 control districts	16.86 (12.54, 21.12)	17.32 (12.80, 21.75)	17.19 (12.68, 21.52)	17.32 (12.80, 21.75)	17.39 (12.66, 22.04)	17.32 (13.39, 21.24)
Wide Prior	Adding 5 control districts	17.66 (13.22, 22.23)	17.70 (13.28, 22.41)	17.68 (13.24, 22.39)	17.70 (13.28, 22.41)	17.67 (13.10, 22.42)	17.60 (13.24, 22.04)
Penta-1							
Conservative Prior (default)	Adding 5 control districts	13.86 (10.66, 17.11)	14.32 (10.96, 17.78)	14.08 (10.80, 17.35)	14.17 (10.95, 17.58)	14.42 (10.99, 18.02)	15.70 (12.76, 18.83)
Wide Prior	Adding 5 control districts	15.17 (11.71, 18.75)	15.40 (11.93, 19.08)	15.26 (11.73, 18.85)	15.30 (11.91, 18.97)	15.50 (11.91, 19.17)	15.94 (12.97, 19.08)
Penta-2							
Conservative Prior (default)	Adding 5 control districts	6.37 (3.86, 8.84)	6.62 (4.12, 9.23)	6.48 (4.03, 9.03)	6.62 (4.12, 9.23)	6.97 (4.37, 9.74)	7.97 (5.77, 10.31)
Wide Prior	Adding 5 control districts	6.52 (3.95, 9.22)	7.01 (4.33, 9.71)	6.80 (4.22, 9.51)	7.01 (4.33, 9.71)	7.37 (4.66, 10.16)	8.17 (5.63, 10.64)
Penta-3							
Conservative Prior (default)	Adding 5 control districts	1.51 (-1.25, 4.25)	1.71 (-1.10, 4.47)	1.59 (-1.22, 4.35)	1.71 (-1.10, 4.47)	2.00 (-0.83, 4.84)	3.94 (1.51, 6.35)
Wide Prior	Adding 5 control districts	1.82 (-0.79, 4.52)	2.31 (-0.54, 5.10)	2.12 (-0.69, 4.87)	2.31 (-0.54, 5.10)	2.75 (-0.22, 5.62)	4.02 (1.66, 6.59)
Measles-1							
Conservative Prior (default)	Adding 5 control districts	13.47 (10.14, 17.02)	14.75 (10.57, 19.06)	13.67 (10.19, 17.31)	13.86 (10.37, 17.57)	14.19 (10.55, 18.03)	16.49 (13.33, 19.71)
Wide Prior	Adding 5 control districts	13.59 (9.92, 17.18)	14.66 (10.42, 18.77)	13.74 (9.98, 17.47)	13.88 (10.20, 17.61)	14.16 (10.52, 18.02)	16.49 (13.18, 19.86)
Measles-2							
Conservative Prior (default)	Adding 5 control districts	-0.66 (-4.29, 3.06)	2.67 (-1.99, 7.50)	-0.55 (-4.29, 3.19)	-0.45 (-4.34, 3.31)	-0.14 (-4.09, 3.83)	1.63 (-1.81, 5.05)
Wide Prior	Adding 5 control districts	-0.50 (-4.20, 3.13)	2.87 (-1.52, 7.30)	-0.15 (-3.93, 3.53)	0.11 (-3.72, 3.73)	0.60 (-3.30, 4.31)	1.70 (-1.80, 5.03)

Table A4: Robustness to Prior Choice

Notes: This table shows shrinkage results are robust to using a much wider prior for the length parameter of the Gaussian Process. Results are also robust across various specifications for calculating the post-impact effect, shown by different columns. Point estimates show posterior means whilst parentheses denote 95% credibility intervals.

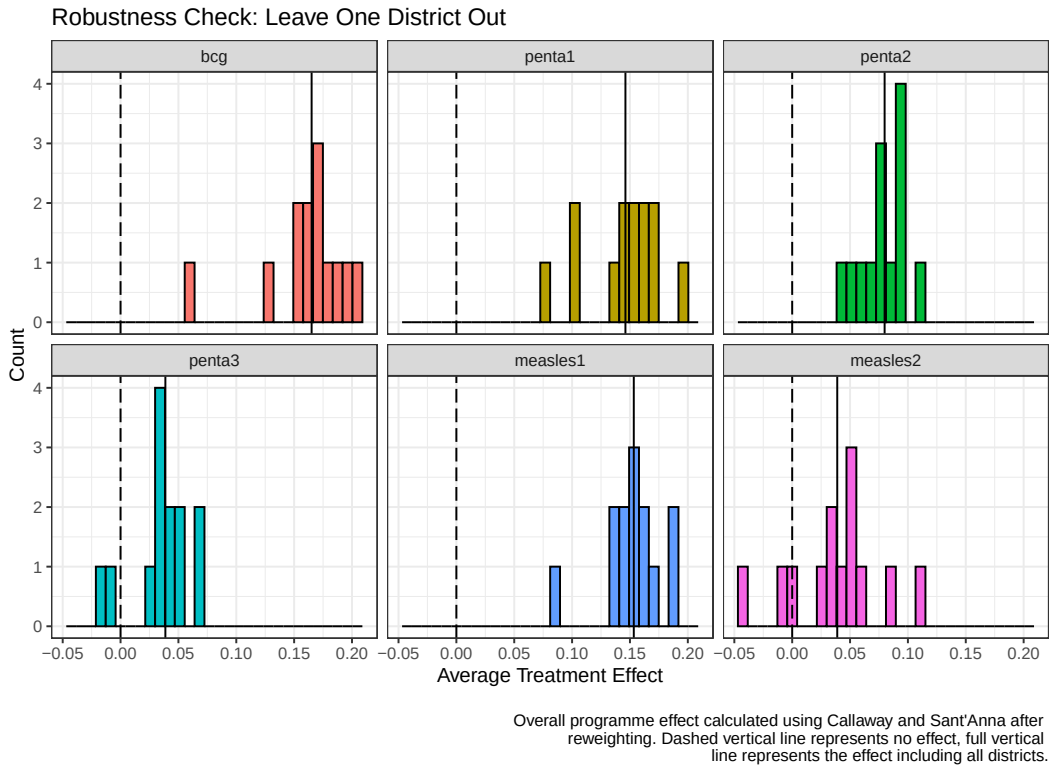
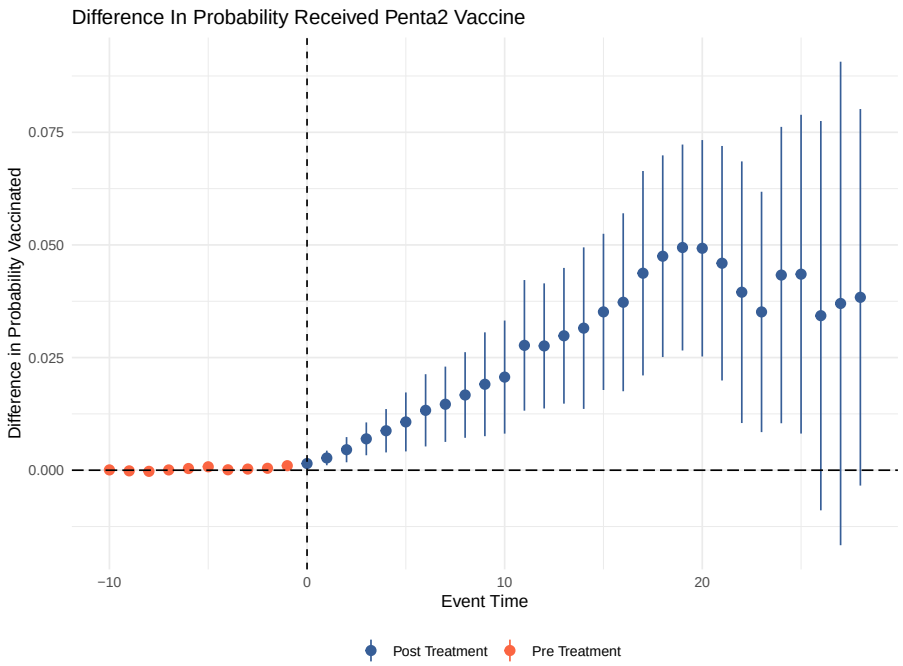


Figure A13: Robustness: Leave One District Out

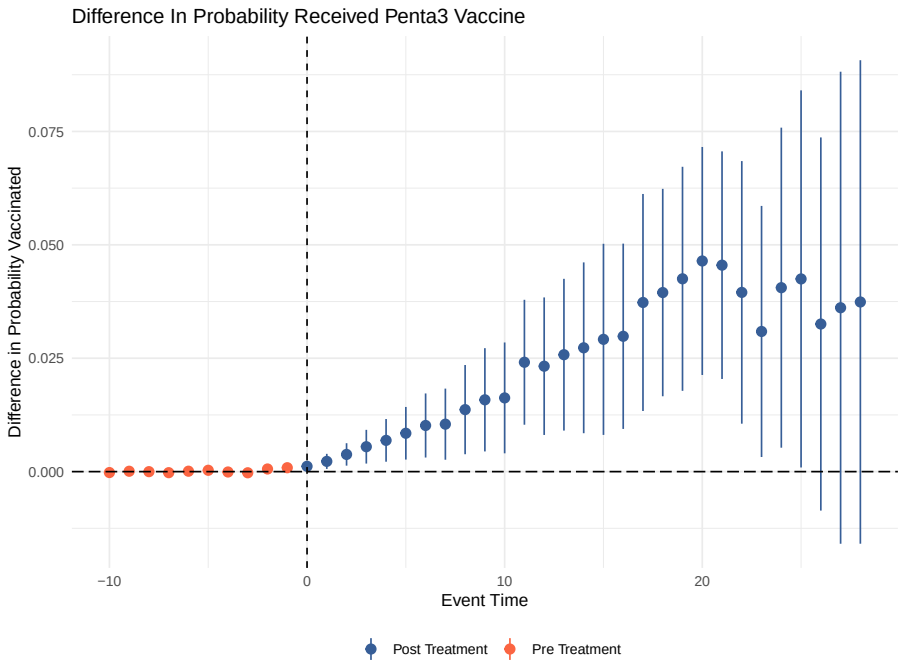
A10 Individual-level Estimates: Callaway and Sant' Anna method for other vaccines

We estimate our model using individual child-level data for other vaccines in the schedule to find the impact of CCT programme. We present these results for Pentavalent-2 and Pentavalent-3 in figures A14 and A15. For both Pentavalent-2 and Pentavalent-3, the preperiod estimates are close to zero and satisfy the parallel trend assumption. The individual level estimates show that the CCT programme has an impact on Pentavalent-2 vaccines in the range of 0 to 0.047 and a similar impact on Pentavalent-3 between 0 to 0.048. The effect is positive and statistically significant, with a 4.969% increase in Pentavalent-2 vaccines in 20^{th} week. Our estimates for Pentavalent-2 and Pentavalent-3 are consistent with those found for Pentavalent-1 (presented in section 4.3).



Notes: The figure displays the event-study plots constructed from difference in difference using (Callaway and Sant'Anna, 2021) method (in dots) using individual child level data. The outcome variable is the log of the number of Pentavalent-2 immunizations administered. We used a balanced panel for 29 towns treated for at least 52 weeks. 23 towns used as nevertreated units from additional, nonmCCT districts were used as a control group. We estimate 20 preperiods since estimating more preperiods dramatically increases the standard errors in the preperiod. Red dots indicate pre-treatment estimates, and blue indicates post-treatment estimates. The estimates are derived using xx vaccinations observed at the individual level. The bars represent 95% percent confidence intervals. Standard errors are clustered at the district level, calculated using the wild cluster bootstrap.

Figure A14: Event Study estimates from Difference in Difference using Callaway and Sant' Anna: Individual-level estimates for Pentavalent-2



Notes: The figure displays the event-study plots constructed from difference in difference using (Callaway and Sant'Anna, 2021) method (in dots) using individual child level data. The outcome variable is the log of the number of Pentavalent-3 immunizations administered. We used a balanced panel for 29 towns treated for at least 52 weeks. 23 towns used as nevertreated units from additional, nonmCCT districts were used as a control group. We estimate 20 preperiods since estimating more preperiods dramatically increases the standard errors in the preperiod. Red dots indicate pre-treatment estimates, and blue indicates post-treatment estimates. The estimates are derived using xx vaccinations observed at the individual level. The bars represent 95% percent confidence intervals. Standard errors are clustered at the district level, calculated using the wild cluster bootstrap.

Figure A15: Event Study estimates from Difference in Difference using Callaway and Sant' Anna: Individual-level estimates for Pentavalent-3