Reducing Neonatal Mortality at Scale: Lessons for Targeting

Christine Valente (University of Bristol and IZA)*
ⓡ Hans H. Sievertsen (University of Bristol, VIVE, and IZA)
ⓡ Mahesh C. Puri (Center for Research on Environment, Health and Population Activities)

November 2022

Abstract

Neonatal sepsis kills over 400,000 children annually. Experimental estimates of the preventive use of chlorhexidine vary widely, leading to external validity concerns. We provide the first quasi-experimental estimates of the effect of chlorhexidine in "real life" conditions and apply machine-learning (ML) to analyze treatment effect heterogeneity in a nationally-representative, Nepalese observational dataset. We find that chlorhexidine decreases neonatal mortality by 43% and that a simple targeting policy leveraging heterogeneous treatment effects improves neonatal survival relative to WHO recommendations. Heterogeneous treatment effects extrapolated from our ML analysis are broadly in line with experimental findings across five countries despite significant implementation differences.

Keywords: neonatal mortality, chlorhexidine, Nepal
JEL Classification: I18, J13, O15

*Corresponding author: christine.valente@bristol.ac.uk. School of Economics, University of Bristol, Priory Road Complex, Priory Road, Bristol, BS8 1TU, UK. This paper supersedes previous versions circulated under the title "Saving Neonatal Lives for a Quarter". Funding from the University of Bristol's Global Challenges Research Fund is gratefully acknowledged. We thank Leela Khanal for sharing implementation details of the Chlorhexidine Navi(Cord) Care Program (CHX-NCP). For their useful comments and suggestions, we thank Douglas Almond, Clément de Chaisemartin, Thomas Dee, Lena Edlund, Mette Ejrnæs, Anna Folke Larsen, Sukjin Han, Xavier d'Haultfœuille, Stephanie von Hinke, Ulrik Hvidman, Grant Miller, Pauline Rossi, Helen Simpson, Pietro Spini, Stéphanie Vincent Lyk-Jensen, Miriam Wüst, and presentation participants at Royal Holloway, PAA Annual Meeting 2021, Essen Health Conference 2021, IAAE Conference 2021 and Workshop on Global Health, Environment and Labour (Royal Holloway). All errors are our own

1 Introduction

Contrary to a commonly held belief, prevention is not always desirable (Kowalski, 2021; Newhouse, 2021). In particular, prevention is unlikely to benefit everyone, so that recommendations regarding whether to- and who should use preventive measures evolve along with scientific evidence which may appear contradictory (World Health Organization, 2020; Kowalski, 2021). In this paper we study the prevention of a common cause of neonatal mortality in low-income settings for which scientific evidence and therefore policy recommendations have experienced much fluctuation in the past decade.

Each year, it is estimated that 400,000 children die of blood infection (or "neonatal sepsis") within the first 28 days of their lives, a condition which is especially frequent in very low-income settings where unsanitary delivery- and living conditions are common (Liu et al., 2016).[1] Starting fifteen years ago, three randomized controlled trials (RCT) in South Asia brought hopes that the simple preventive application of the disinfectant chlorhexidine (CHX) to the umbilical cord stump may eradicate this condition (see Mullany et al., 2006; El Arifeen et al., 2012; Soofi et al., 2012, who report a 20-38% decrease in neonatal mortality), leading the World Health Organization (WHO) to recommend in 2013 the application of CHX for home births in settings with neonatal mortality above 30 per 1000 (World Health Organization, 2015). Replication of the RCTs however failed in two further trials in Southeast Africa, leading experts to express doubts about the effectiveness of CHX at scale and the WHO to revise its recommendations in 2022 — to now recommend CHX cord care only in regions where the application of harmful substances such as mustard oil, turmeric or animal dung to the stump is common (Semrau et al., 2016; Sazawal et al., 2016; Osrin and Colbourn, 2016; Ponce Hardy, 2018; World Health Organization, 2022). While international estimates of the direct cost of implementing CHX are low, the WHO recommendation of restricting the use of CHX can be understood as balancing proven benefits against broader costs.[2] These include the risk of application of other, potentially dangerous substances, as well as the risk of diverting human, logistical, and financial resources away from other essential medicines and tasks in an area where the gap between recommended health care and practice is already large (Friberg et al., 2010; Requejo et al., 2015).

This paper answers three open questions: (i) What is the effect of CHX in "real-life" conditions?; (ii) What variables can, empirically, best account for the heterogeneity of the effect of

[1] Verbal autopsy estimates of causes of neonatal death carried out in various districts of Nepal outside experimental trials report between 38-47% of neonatal deaths due to perinatal infection or sepsis specifically, compared to 26-38% across selected areas of India, Malawi and Bangladesh (Fottrell et al., 2015; Khanal et al., 2011; Erchick et al., 2022). In the Sarlahi district of Nepal, Saya et al. (2022) further find that sepsis accounted for 19% of neonatal deaths in an observational study nested within a community-based trial providing comprehensive preventive and remedial health care, which is likely to have have decreased the incidence of sepsis.

[2] International estimates of the cost of implementing CHX range from US$0.23 for a single dose to US$2.9 when including all related fixed and variable costs (Hodgins et al., 2013; Federal Ministry of Health, 2016; Callaghan-Koru et al., 2019). This is roughly similar to the cost of introducing a single vaccine in low-income countries, which varies between $0.16 and $2.54 according to Vaughan et al. (2019).

CHX on neonatal survival?; and (iii) Could an alternative targeting policy to the WHO's past and current recommendations further reduce neonatal mortality?

Our first contribution is to provide the first estimates of the effect of CHX outside an experimental setting, which we do in a nationally representative sample for Nepal. Concerns about the scalability of experimental findings typically emphasize factors which lead to *smaller* treatment effects at scale — such as selected and non-representative samples, high compliance and adherence to protocol which cannot be replicated in "real-world" conditions (Al-Ubaydli et al., 2017). But in the case of trials of preventive health care measures such as CHX cord care, the treatment effect might be muted due to the lack of a pure control group. For instance, subjects involved in CHX trials are referred to the hospital where needed — e.g., if signs of cord infection (omphalitis) appear during the frequent research team visits. In addition, both treated and control groups typically receive a comprehensive package of measures preventing omphalitis which go well beyond the usual standard of care in low-income settings.[3,4] The additional remedial and preventive health care provided are likely to lead to smaller treatment effects than what would be expected outside trials (El Arifeen et al., 2012; Semrau et al., 2016). Indeed, in both trials finding no significant effect on mortality, the authors note that the neonatal mortality rate (NMR) was between 32 percent and 40 percent lower than in the most recent Demographic and Health Survey for the relevant area — even in the control group. Many factors may therefore lead to differences in CHX- and other preventive treatment effects in- and outside an experimental setting, in a direction that is unclear *a priori*.

Our second contribution is to apply machine-learning (ML) techniques recently developed by Athey et al. (2019) and Athey and Wager (2021) for use in observational data to understand how CHX treatment effects depend on community- and individual characteristics — which vary much across our nationally representative sample — and then identify an optimal targeting policy that takes this heterogeneity into account. Meta-analyses of existing randomized trials have concluded that CHX cord care was only effective for home deliveries, which is not surprising given that 90% or more of the births included in the South Asian trials took place at home vs. between 36% and 47% in the Southeast African trials (Imdad et al., 2013; Sankar et al., 2016; López-Medina et al., 2019). But these meta-analyses have important limitations due to the small number of included studies and the possibility that heterogeneous results by place of birth may be confounded by other differences across studies — such as the number of CHX applications or factors correlated with local economic development or cultural practices. The use of ML methods to study heterogeneous treatment effects has many advantages. In particular, it allows a high degree of flexibility in identifying sources of heterogeneity and has

[3]The typical package of services received by both control and treated subjects in CHX application trials are: a clean delivery kit, referral to clinic in the presence of danger signs, newborn health messages, antenatal clinic visits, and home visits starting soon after birth.

[4]Another potential channel through which CHX cord care may prevent neonatal death is by preventing neonatal tetanus (Bennett et al., 1997). In the Nepal CHX trial, this potential pathway to impact was also shut down by ensuring full maternal tetanus immunization at enrollment in the trial (Mullany et al., 2006).

embedded robustness checks in the form of cross-validation (Varian, 2014; Athey and Imbens, 2016).

Our third contribution is to take the findings obtained in our nationally-representative, Nepalese observational data, and use them to predict the effect of implementing the same program in the five regions in- and outside Nepal where CHX trials have been carried out. More specifically, we report predicted treatment effects of a program similar to that rolled out in Nepal if it were hypothetically extended to samples from the five RCT regions drawn from nationally representative surveys. To do so we use the extrapolation approach due to Dahabreh et al. (2020) as implemented in Tibshirani et al. (2022), which provides doubly robust treatment effect estimates that put more weight on estimates that are more similar to the out-of-sample observations. Given differences in the exact nature of the treatment between the Nepalese roll-out and the trials, as well as the absence of a pure control in the various trials, the effects we predict should not be expected to match experimental findings closely even if we had access to the experimental microdata and our heterogeneity analysis based on the national roll-out was fully externally valid.[5] This exercise however serves as a sanity check on our heterogeneity analysis as well as illustrates the informativeness of the treatment effect heterogeneity uncovered in our quasi-experimental setting for external samples.

The first country to introduce CHX cord cleansing nationwide is Nepal. We exploit plausibly exogenous variation in the timing of the CHX cord care program expansion across districts of Nepal using data from the nationally representative 2016 Nepal Demographic and Health Survey (DHS). After piloting the program in 4 out of 75 districts from late 2009, CHX cord application was quickly scaled-up across the rest of the country (see Figure 1). By 2015, 75 percent of the population was covered by the program (Department of Health Services, 2015). While the Chlorhexidine Navi(Cord) Care Program was integrated to the training, monitoring systems and operations of a recent national newborn health program ("Community-Based Newborn Care Program" or CB-NCP), the timing at which the CHX program was rolled out to a particular district largely depended on practical considerations such as presence of implementation partners on the ground and district government leadership.[6]

Two-way fixed effects estimates indicate that, overall, the CHX program decreased neonatal mortality by 1.8 percentage points or 43 percent compared to the control group mean. Our conclusions are robust to comprehensive robustness checks. In particular, we find that a placebo treatment "switching on" 6 months before the actual roll-out of CHX in the district has no ef-

[5]For studies where the same treatment is implemented in experimental and observational settings, a growing body of work develops methods to systematically combine observational and experimental data to address the shortcomings of one- and/or the other or reconcile them. Abstracting from the differences in treatment, our setting does not meet the data requirements of currently available methods, either due to lack of access to the experimental microdata (needed to implement the methods in e.g., Athey et al., 2020; Kowalski, forthcoming) or due to the absence of a suitable instrumental variable explaining the choice of study design (needed in Gechter and Meager, 2022). For an overview of this rapidly evolving literature, we refer the interested reader to the up-to-date literature review in Gechter and Meager (2022).

[6]Implementation partners were Care Nepal, Save the Children, Health Right International, UNICEF, ADRA and One Heart Worldwide (JSI Research & Training Institute, 2017).

Figure 1: CHX cord application roll-out across districts over time.

Notes: Districts where CHX-NCP has been rolled out are in green. Source: JSI administrative records.

fect, that the CHX treatment is not associated with a decrease in mortality between 2 and 12 months after birth, that estimates based on within-mother variation in treatment exposure are very similar to estimates exploiting within-district variation, and that the effect of the CHX program is observed independently of other neonatal health interventions. These checks narrow down the possibility of omitted variable bias to variables other than neonatal health interventions that would decrease neonatal mortality but not mortality after this short period of vulnerability to umbilical cord infection *and* which would affect neonatal mortality at the time of CHX roll-out but not shortly before that.

Turning to our heterogeneity analysis, two-way fixed effects estimates show that children only benefit from CHX application, on average, if they are predicted to be born at home, in line with WHO recommended use during 2013-2022. Place of delivery — which we correctly predict in 76 percent of cases for which we observe actual place of delivery — is however likely to proxy for risk factors such as hygiene conditions and healthcare at- and shortly after birth, health endowment at birth, and treatment compliance. To better describe the treatment effect heterogeneity we observe, we turn to machine learning. Our causal forest detects significant heterogeneity in treatment effects, and when comparing the lowest- with the highest treatment effect tertiles, we find large, statistically significant treatment effects in the two top

tertiles of treatment effect magnitude but no significant effect in the bottom tertile. Looking closer at the characteristics of the samples in the top- and bottom tertiles of the distribution of treatment effects, we find that children in the highest treatment effects tertiles are more often boys (unsurprisingly given that male newborns are more likely to die), and are often born to very young, less educated, rural mothers, are more often predicted to be born at home and are more likely to be born in districts with higher rates of neonatal mortality and higher prevalence of harmful substance application to the umbilical stump. Importantly, only a quarter of the variation in conditional treatment effects comes from differences in variables which the WHO has recommended considering to guide the use of CHX.

We then apply Athey and Wager (2021)'s approach to identify a targeting policy which would asymptotically result in the largest gains in neonatal survival which could be obtained from a targeting policy given the constraints imposed (or "policy class"), and compare this optimal policy — from the point of view of neonatal survival — to past and present WHO recommendations. We find that either of WHO recommended policies effectively target about a third of the population with returns to treatment as high as any other. However, both miss many children whose survival chances could significantly benefit from CHX treatment and therefore only achieve about half of the overall reduction in neonatal mortality achieved by a policy optimally targeting districts based on other easily-observed district-level indicators of, mostly, antenatal health care quantity and quality.

Finally, after applying the causal forest to our nationally representative Nepalese dataset, we take advantage of the international comparability of the DHS and predict doubly robust average treatment effects of implementing a program similar to the Nepalese CHX national roll out in five different DHS samples corresponding to the subnational regions and time periods where the five CHX trials took place. Unsurprisingly given the implementation differences previously discussed, the control neonatal mortality rates (NMR) and treatment effects found in the RCTs are both substantially smaller than the NMR and hypothetical treatment effects we obtain using nationally representative surveys for the same regions. But despite the differences in the interventions implemented in the Nepal scale-up and across RCTs and the fact that we have access to a different sample from broadly the same population rather than to the RCTs' microdata, we predict large, statistically significant decreases in neonatal mortality in the three regions where CHX trials led to significant decreases in mortality and we predict much smaller, statistically insignificant decreases in mortality in the two regions where CHX trials failed to decrease neonatal mortality.

Prior work studying the effect of health programs carried out at scale in developing countries has found no- or small decreases ($<$0.3 percentage points) in neonatal mortality (Lim et al., 2010; McKinnon et al., 2015; Powell-Jackson et al., 2015; Arulampalam et al., 2017; Van de Poel et al., 2016; Philibert et al., 2017; Fitzpatrick, 2018). Broad-based National Health Insurance systems introduced in the last few decades in a number of middle-income countries have been found to reduce infant mortality (see Conti and Ginja, forthcoming, and references

therein), but the few estimates on neonatal mortality are mixed (Bhalotra et al., 2019).[7,8] In this context, CHX appears to offer a valuable option to health policy makers looking for evidence-based, affordable, at-scale solutions to reduce neonatal mortality in low-income settings.

In the next section, we give an overview of early life mortality trends and CHX cord care in Nepal. Section 3 presents the data and identification strategy. The regression results for the mean treatment effect and robustness checks are reported in Section 4. Section 5 explores heterogeneity in the effect of CHX application using both the two-way fixed effects model and machine learning and derives lessons for policy targeting. Section 6 extrapolates our quasi-experimental heterogeneity analysis to samples drawn from the five RCT studies settings and compares these predictions with experimental estimates of the effect of the trialled interventions. Section 7 concludes.

2 Background

2.1 Early Life Mortality in Nepal

Nepal is a landlocked country situated between China and India which is home to 28.1 Million people. The country's Human Development Index ranks only 143 out of 191 (in 2021) and more than a third (36%) of children under age 5 are stunted. Notably, Nepal has seen sharp decreases in fertility over the past twenty years — from 4.7 children per woman in 1995 to 1.9 in 2020, and marked reductions in child mortality — from 106 deaths before the age of 5 per 1,000 births in 1995 to 28 in 2020.

However, progress in the NMR reduction in Nepal stalled in the early 2000s (at 33 per 1,000 both during 2002-2006 and 2007-2011) while under-5 mortality slowed down its downward trend, going from 61 to 54 per 1,000 during the same period.[9] This stagnation came to an end in 2012-2016 as NMR dropped to 21 per 1,000 — a 36% decline relative to the previous 10-year period.

2.2 Chlorhexidine Cord Care

The latest decrease in NMR observed since 2012 coincides with the acceleration of the roll-out of CHX cord application through the Chlorhexidine Navi(Cord) Care Program (CHX-NCP) (see Figure 2).

[7] PROGRESA, which paid cash transfers conditional to poor households conditional on, among others, regular prenatal checks, has been found to significantly decrease infant mortality, but not neonatal mortality (Barham, 2011).

[8] Historical evidence from today's developed countries has concentrated on infant mortality (see, e.g., Bauernschuster et al., 2017, and references therein).

[9] All the mortality and fertility figures in this sub-section are taken from Ministry of Health [Nepal] and New ERA and ICF (2017).

CHX-NCP was a $3.9 million program funded mainly by bilateral donors (US, Norway, Canada, UK) and the Bill & Melinda Gates Foundation. In partnership with the Nepalese Department of Health Services, international NGOs and a Nepalese pharmaceutical company which produced the CHX gel locally, the program was implemented by JSI Research & Training Institute, Inc. and was designed to support the Government of Nepal to scale up the use of CHX for cord care nationwide. This involved training as well as procurement, logistical, monitoring and technical support.

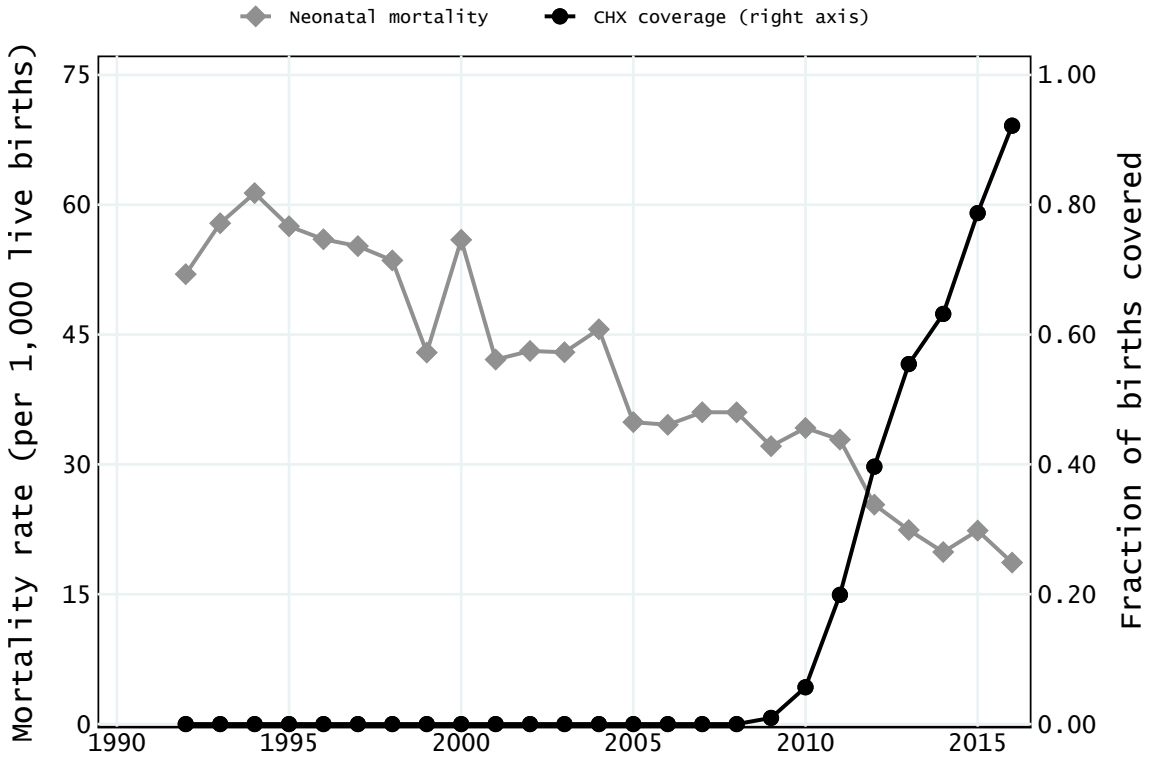


Figure 2: Neonatal mortality and CHX-NCP coverage

Notes: Authors calculations based on Nepal DHS 2016 microdata and JSI administrative records of the district roll-out of CHX-NCP.

The scaled-up intervention consisted of a single CHX gel application on the day of birth to all newborns irrespective of place of birth. For home births, CHX gel doses were distributed to pregnant women during antenatal contact — in general, during antenatal care visits by female community health workers in the last two months of pregnancy (Hodgins et al., 2019).[10,11] The CHX training of health workers lasted between three hours and one day and to reduce costs and increase program sustainability, an effort was made to integrate training and monitoring activities into broader maternal and newborn health programs, and more specifically

[10]Eighty four percent of women who gave birth in the five years leading to the 2016 DHS received antenatal care and 69 percent received four antenatal care visits or more (Ministry of Health [Nepal] and New ERA and ICF, 2017).

[11]Appendix Table A.1, report results showing that CHX-NCP was not accompanied by an increase (or decrease) in the number of antenatal care visits.

into the Community-Based Newborn Care Program (CB-NCP) (JSI Research & Training Institute, 2017; JSI, 2017; Hodgins et al., 2019). While this integration contributed to reducing costs, in Section 4.2.2 we estimate the effect of CHX in places where the CB-NCP and the other main neonatal health program rolled out in the later part of our study period (CB-IMNCI) were not in place and reach similar conclusions as in our main specification. We also test for complementarities between the three programs and find none.

Estimates of actual CHX application in program districts vary much and, for home deliveries, an important limitation is that there is no record of application and that maternal recall is unlikely to be reliable for non-salient events (Beckett et al., 2001).[12] Coverage estimates suggest that it may have peaked in 2014/2015, as estimates range from 75 percent of home deliveries and 96 percent of facility deliveries (HIMS (2014), as cited in Khanal (2015)) to 75 percent of all births according to Department of Health Services (2015) to only about 40 percent of home births and 90 percent of facility births in 2017 according to Hodgins et al. (2019) so that estimates presented in this study should be interpreted as intention-to-treat effects — arguably the parameter of interest from a policy point of view. The coverage is however consistently estimated to be higher among health facility deliveries, so that heterogeneity in treatment intensity cannot account for the larger decrease in NMR observed among predicted home births.

2.3 Other Neonatal and Child Health Programs

Nepal has a long history of active programmatic efforts to improve maternal and child health. To ensure that we capture the effect of chlorhexidine cord care independently of any other intervention, a thorough identification of programs that may have contributed to recent decreases in NMR was done by the Kathmandu-based Center for Research on Environment, Health and Population Activities (CREHPA) in two steps. First, all annual reports produced by the Department of Health since 2013 were analyzed in detail to identify candidate explanations for the recent decrease in NMR. Second, semi-structured interviews with 12 in-country neonatal and maternal health experts — from, among others, the Family Welfare Division of the Department of Health Services, the WHO, UNICEF, and Children and Maternity hospitals — were carried out in order to collect their specialist views on the most likely reason(s) for the NMR reduction. Eleven interventions were identified by key informants, including CHX-NCP. These are summarized in Table 1. Of these, three were being implemented in all districts prior to the roll-out of CHX cord care (CB-IMCI, Birth Preparedness Program and Safe Delivery Incentive Program), two were implemented in all districts of Nepal at the same time so that

[12] In the DHS, women who gave birth within five years of the interview are asked, among many other things, whether anything was placed on the stump after the umbilical cord was cut, and if so, what substance was applied. There is good reason to think that answers to these questions are not reliable: While CHX was neither available nor promoted in a district prior to the roll-out of CHX-NCP, as many as 30 percent report that CHX was applied to the stump of the newborn in *untreated* district-by-time cells. Meanwhile only 45 percent report that CHX was applied to the stump of the newborn in *treated* district-by-time cells, which is about half what is found in administrative records.

any effect they may have on neonatal mortality is captured by time fixed effects (Nyano Jhola and Aama and Newborn Care), and one (Rural Ultrasound Program) affects only 190 births in our sample, of whom 108 are also treated by the CHX cord care program (out of 3,255 treated observations). Three are nutrition programs targeting pregnant women and children up to two years old (Nepal Agriculture and Food Security Project, Sunaula and Suaahara) — therefore less likely to influence *neonatal* mortality specifically. In robustness checks, we control for all these programs except the ones whose implementation is subsumed in time fixed effects (see Appendix Figure B.1). The last one is a comprehensive program targeting neonatal health (CB-NCP), which was subsequently progressively integrated into CB-IMCI and rebranded "Community Based Integrated Management of *Newborn* and Childhood Illness" (CB-IMNCI). We control for the implementation of these two programs (CB-NCP and CB-IMNCI) throughout the main analysis, show that our results regarding CHX are robust to whether or not we control for these programs (or covariates more generally), robust to restricting the analysis to observations outside areas implementing either CB-NCP or CB-IMNCI, and show in Section 4.2.2 that there is no evidence of complementarities between CHX cord care and the presence of these programs in the district.

Table 1: Programs relevant to neonatal mortality

	Name	Overview	Implementation Detail
1	Community Based Integrated Management of Childhood Illness (CB-IMCI)	Management of multiple illnesses from birth to age 5	Rolled out to all 75 districts between 1997 and 2009
2	Birth Preparedness Program	Encourage institutional delivery, antenatal care and preparation for complications	Introduced in all districts in 2008/2009
3	Safe Delivery Incentive Program	Subsidy for institutional delivery	25 districts in 2006 then all districts from 2009
4	Aama and Newborn Program	Cash incentives for 4 ANC visits Free delivery care Free sick newborn care	Introduced in all districts from 2015/16
5	Nyano Jhola	Clothes to prevent hypothermia and infection	Introduced in all districts in 2015/16
6	Rural Ultrasound Program	Trained skilled birth assistants to use portable ultrasound machine	Rolled out from 2 to 14 districts between 2012 and 2017
7	Nepal Agriculture and Food Security Project	Combined agricultural and nutritional intervention	22 districts during 2013-2017
8	Community Action for Nutrition Project (Sunaula)	Improve nutrition and reduce exposure to smoking and indoor pollution during pregnancy	15 districts during 2012 to 2017
9	Suaahara I Project	Multisectorial intervention to improve nutrition from conception to 24 months	16 districts from 2011 then 41 districts from 2016
10a	Community Based Newborn Care Program (CB-NCP)	Prevent and manage newborn infections, hypothermia and low birth weight	Rolled out from 10 to 41 districts between 2009-14

Continued on the next page

Continued from the previous page

	Name	Overview	Implementation Details
		Manage asphyxia Referral of sick newborns	
10b	Community Based Integrated Management of Newborn and Childhood Illness (CB-IMNCI)	CB-NCP (10a) integrated into CB-IMCI (1)	Rolled out from 30 to 75 districts between 2014/15 and 2016/17
11	Chlorhexidine "Navi" (Cord) Care Program (CHX-NCP)	Introduction of CHX cord care for all births	Rolled out from 4 to 75 between 2009 and 2017

Source: Department of Health Services-Ministry of Health (2009/10-2016/17), USAID (2017), Bhattarai (2017).

3 Data and Identification Strategy

3.1 Data

The 2016 Demographic and Health Survey (DHS) of Nepal is a nationally representative survey that collected detailed pregnancy histories of all women age 15-49 found in sampled households, as well as comprehensive data on the demographic and socioeconomic characteristics of the household and its members (MoH, New ERA and ICF, 2017). The dataset includes, for each child ever born to the interviewed women, dates (month and year) of birth and death, if applicable. Detailed information on antenatal and postnatal care is also collected for births occurring within 5 years of the interview, including place of delivery. In the absence of comprehensive vital statistics systems, the DHS is the main source of information on child mortality in Nepal as in many other developing countries.

The survey collected data on a total of 26,028 births. We drop 366 multiple births and 118 births to mothers who are either less than 15 or 45 and above because the risk of neonatal mortality is much higher among these unusual births, and drop 116 births occurring within one month of the interview date and thus not fully exposed to the risk of neonatal death. While recall error is unlikely to be an issue for such a salient event in the life of a woman as the death of a newborn, we restrict our main analytical sample to births that occurred within 25 years prior to the date of interview, resulting in a sample of 23,465 births. Robustness checks varying this time window by 5 years on either side show that our findings are not sensitive to this sample selection criteria (see Section 4).

We merge the DHS microdata with administrative data on the implementation of all the main programs targeting maternal and newborn health in Nepal listed in Section 2.3. Dates of the district-level implementation of each program were collected from various Department of Health Annual Reports, controls included in the main analysis for the two health programs targeting newborns specifically (CB-NCP and CB-IMNCI) and in robustness checks for secondary programs whose coverage is not fully captured by time fixed-effects. For CHX-NCP,

which was administered by JSI, we obtained roll-out dates from the CHX-NCP program director.

The variable means for our sample, presented in Table 2, highlight that the sample at hand has very low levels of human development, with 57 percent of children having mothers with no formal education, 41 percent living in rural areas, and one in five children being born to a teenage mother. Forty-eight percent of children are female, which is close to what would be expected given the widely observed natural sex ratio at birth (51 percent male).

3.2 Identification Strategy

In our main specification, we estimate linear probability models of the form:

$$m_{idt} = \alpha + \beta CHX_{dt} + D_d'\Delta + T_t'\Gamma + X_{idt}'\Lambda + \varepsilon_{idt} \quad (1)$$

where m_{idt} is an indicator equal to 1 if child i dies by age one month (allowing for "heaping" at one month) and zero otherwise, CHX_{dt} is an indicator equal to 1 if CHX-NCP was rolled out in the child's district by the date the child was born, D_d is a vector of district fixed effects, T_t is a vector of time fixed effects, where time is defined at the month-by-year level (e.g., Ashwin 2066 in the Nepali calendar or October 2009, which we refer to as "month-year date of birth" fixed effects or just "month of birth" fixed effects interchangeably), X_{idt} is a vector of controls comprising all controls listed in panel A and B of Table 2, covering child, mother, household characteristics and district-time varying controls such as exposure to health programs other than CHX-NCP; $\alpha, \beta, \Delta, \Gamma$ and Λ are parameters to be estimated; and ε_{idt} is an error term allowing for arbitrary intra-district correlation.

Since we control for time- and district fixed effects, identification relies on the absence of time-varying omitted factors correlated with the timing of treatment. Regressing the treatment indicator on observable characteristics, we find that, other than the expected positive correlation between CHX-NCP and CB-NCP, the program on which CHX-NCP "piggy-backed" (Hodgins et al., 2019), the treatment is only weakly correlated with observable characteristics.[13] Treated newborns are significantly more likely to have a mother with an ethnicity from the residual "other" group, somewhat less likely to be found in rural areas and somewhat more likely to have a mother with primary education. However, these differences are small and there is no clear pattern of selection in terms of socio-economic status (See Figure 3 and Appendix Table A.2). In Section 4, we report on a number of robustness checks which indicate that our findings are unlikely to be biased by a correlation between district trends in early life health and the timing of CHX-NCP roll out.

Recent work has shown that, in the presence of heterogeneous treatment effects, two-way

[13] In Section 4.2.2, we test for interaction effects between the CHX program and CB-NCP. We find no evidence of complementarities and find that our results hold whether or not we include district-cells where CB-NCP is present.

Table 2: Variable means

	Mean
A. Demographics and SES	
Female	0.48
First born	0.34
Second born	0.28
Third born	0.18
Parity four or higher	0.21
Mother age 15-19y	0.20
Mother age 20-24y	0.41
Mother age 25-29y	0.26
Mother age 30-34y	0.10
Mother age 35-39y	0.03
Mother age 40-45y	0.01
Ethnicity: hill brahmin	0.09
Ethnicity: hill chhetri	0.23
Ethnicity: terai brahmin/chhetri	0.01
Ethnicity: other terai caste	0.14
Ethnicity: hill dalit	0.11
Ethnicity: terai dalit	0.04
Ethnicity: newar	0.02
Ethnicity: hill janajati	0.18
Ethnicity: terai janajati	0.10
Ethnicity: muslim	0.06
Ethnicity: other	0.00
Rural	0.41
Education: no education	0.57
Education: primary	0.18
Education: secondary	0.19
Education: higher	0.06
Wealth 0-20%	0.27
Wealth 20-40%	0.22
Wealth 40-60%	0.20
Wealth 60-80%	0.17
Wealth 80-100%	0.13
Altitude in 1st quintile	0.20
Altitude in 2nd quintile	0.20
Altitude in 3rd quintile	0.19
Altitude in 4th quintile	0.20
Altitude in 5th quintile	0.20
B. Health programs	
Program: CB-NCP	0.16
Program: CB-IMNCI	0.05
C. Child mortality	
Child died $\leq$1m	0.04
Child died $<$1m	0.03
Child died $\leq$12m	0.06
Child died $\leq$12m & $>$1m	0.01
Observations	23,465

Notes: Except for the variables measuring child gender and birth order, all variables in panel A are capturing mother characteristics. Panel B. shows means for whether the child was covered by the health programs CB-NCP and CB-IMNCI.

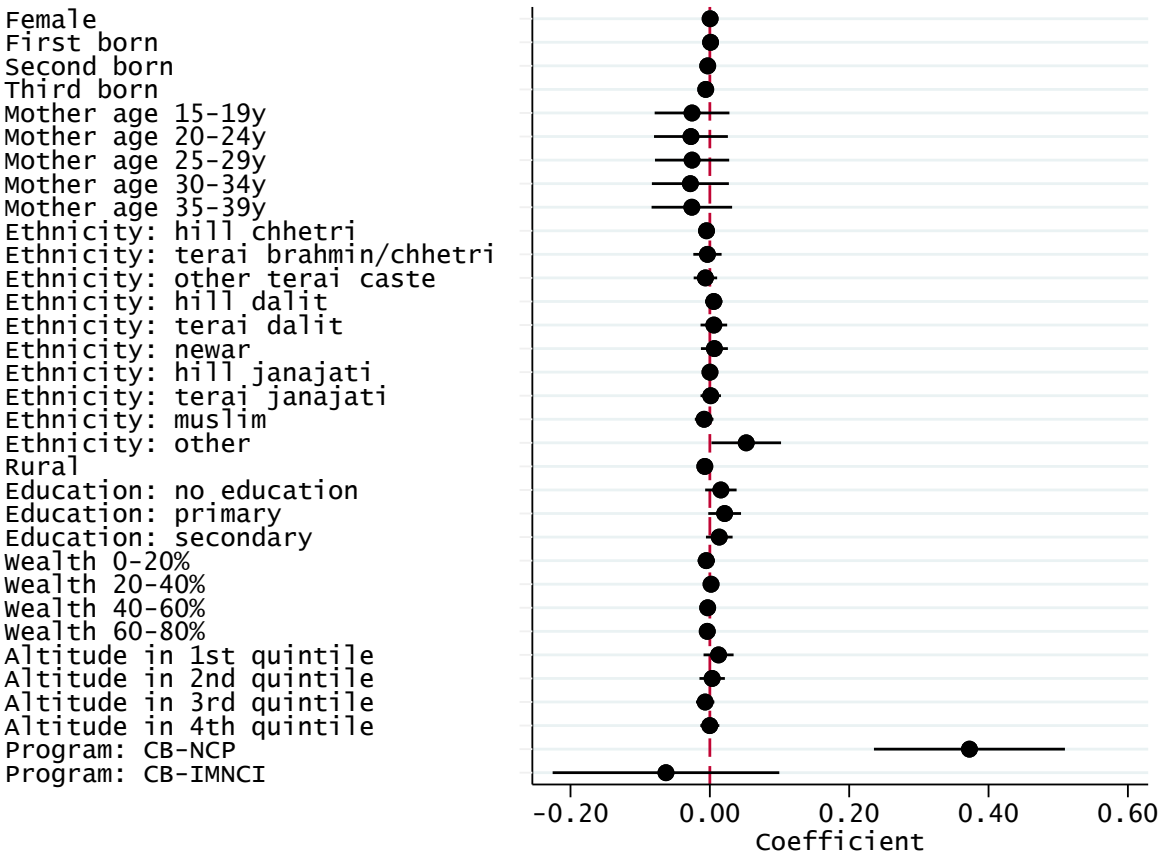


Figure 3: Covariate balance

Notes: Point estimates and 95% confidence intervals from estimating a regression with the treatment indicator as the dependent variable and all of the covariates listed in the figure as independent variables, as well as district and month-year date of birth fixed effects. The confidence intervals are calculated based on standard errors clustered at the district level. Appendix Table A.2 reports all coefficients.

fixed effects models such as the one we estimate can significantly depart from the average treatment effect (e.g., de Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021). We address this concern with several specification checks reported in Section 4.2.1.

4 Average Treatment Effect Results

4.1 Main Estimates

Table 3 reports our baseline estimates. In column (1) we show results from estimating a specification without any controls and find that CHX-NCP significantly reduces neonatal mortality by 1.4 percentage points. In column (2) we show the results when adding the full set of controls. Using this specification, we find that, conditional on controls, CHX-NCP decreases neonatal mortality by 1.8 percentage points or 43 percent of the control mean. This is larger than the 20-34% decreases observed in the three Southeast Asian RCTs, which suggests that

the additional preventive and remedial measures in place as part of the RCTs may have limited the benefits of CHX application. In column (3) we restrict the sample to children born before the treatment started in their district and include a placebo treatment which is equal to one if the child was born 6 months before CHX-NCP was rolled out in the district or later, and zero otherwise. As the table shows, the coefficient associated with this pre-treatment indicator is both small in magnitude and not significantly different from zero. Adding controls to the placebo specification does not alter the conclusions of no relationship between the lagged treatment indicator and neonatal mortality, as shown in column (4).

Table 3: Effect of CHX-NCP on neonatal mortality

	Dependent variable: mortality						
	month of death ∈ [0,1]					month ∈]1,12]	month ∈ [0,12]
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CHX	-0.014**	-0.018***			-0.020**	0.001	-0.019**
	(0.006)	(0.006)			(0.009)	(0.004)	(0.008)
CHX_{t-6}			0.002	-0.001			
			(0.011)	(0.011)			
Observations	23,465	23,465	20,321	20,321	21,209	22,571	22,571
Clusters	73	73	73	73	73	73	73
MDV	0.042	0.042	0.042	0.042	0.045	0.015	0.058
Sample	All	All	Pre	Pre	All	All	All
Controls	No	Yes	No	Yes	Yes	Yes	Yes
Month of birth FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	No	Yes	Yes
Mother FE	No	No	No	No	Yes	No	No

Notes: All specifications are estimated as linear probability models using OLS. Except for the mother FE specification, a "Yes" in the "Controls" row indicates that the regression also includes the full set of demographic, SES, and program controls. Demographic controls include birth order (three indicators), five year maternal age group indicators, and gender. SES controls include education (three indicators), wealth (four indicators), rural indicator, altitude quintile indicators, and ethnicity indicators. Program controls include controls for the CB-NCP and CB-IMNCI health programs. The mother FE specification excludes the SES controls as these are mother-invariant. All coefficients are reported in Appendix Table A.3. The "Pre" sample excludes those for whom *CHX* is equal to one. MDV is the mean of dependent variable among untreated individuals. Standard errors clustered at the district level in parentheses. Asterisks indicate significance at the following levels: * p<0.1, ** p<0.05, and *** p<0.01.

In Column (5) of Table 3, we include mother fixed-effects instead of district fixed effects and find similar results. This indicates that our district fixed-effects estimates are not biased by differential changes in the composition of mothers between treated and control districts (e.g., due to differential trends in maternal education or living standards between maternal cohorts).

We also carry out a falsification test based on the fact that cord infection (omphalitis) primarily affects neonates, but is uncommon among older infants (Painter and Feldman, 2019). CHX application, which narrowly targets omphalitis, should therefore decrease neonatal mortality but not mortality between 2 and 12 months of age — whereas unobserved time-varying improvements in maternal and child health should decrease both. In Column (6) of Table 3, we estimate Equation (1) using as dependent variable an indicator equal to 1 if the child died between 2 and 12 months of age and zero if they survived beyond infancy — the 12 first months

of life — and find that babies born under the CHX-NCP program are no more or less likely to die between 2 and 12 months (point estimate of 0.001). In the last column, we estimate the total effect of CHX-NCP on overall mortality in the first year of life and find a statistically significant decrease in infant mortality by 1.9 percentage points.[14]

4.2 Further Robustness Tests

4.2.1 Robustness of two-way fixed effects estimates

Recent work has shown that, in the presence of heterogeneous treatment effects, two-way fixed effects models such as the one we estimate can significantly depart from the average treatment effect (e.g., de Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021). Of particular concern is the fact that some of the treatment effects averaged over in the two-way fixed effects model can bear negative weights, which may lead to the sign of the treatment effect estimate obtained in a two-way fixed effects model to be the reverse of that of the true average treatment effect.

Table 4: Robustness of Two-Way Fixed Effects Estimates

	Average Treatment Effect							
	All controls		No controls					
	TWFE	BJS	TWFE	BJS	N_w	$N_{w<0}$	$\sum_{w<0} w$	σ_{FE}
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline	-0.018***	-0.015	-0.014**	-0.010***	1623	246	-0.043	0.020
	(0.006)		(0.006)	(0.004)				
Iteration 1	-0.018***	-0.018	-0.014**	-0.013***	1377	54	-0.005	0.026
	(0.006)		(0.006)	(0.004)				
Iteration 2	-0.018***	-0.019	-0.014**	-0.014***	1323	5	-0.000	0.028
	(0.006)		(0.006)	(0.004)				
Iteration 3	-0.018***	-0.019	-0.014**	-0.014***	1318	0	0.000	NA
	(0.006)		(0.006)	(0.004)				

Notes: TWFE is the Two-Way Fixed Estimator and BJS is the estimator by Borusyak et al. (2021) implemented in Stata with the *did_imputation* command. Columns (5) to (8) pertain to TWFE estimates. N_w indicates the number of cells, $N_{w<0}$ indicates the number of cells with negative weight, $\sum_{w<0} w$ indicates the sum of the negative weights, and σ_{FE} is the minimum standard deviation in the treatment effect across all district-month cells which would be required for the average ATT over all cells to in fact be zero. Except for the BJS estimator with controls, for which standard errors could not be obtained due to convergence issues, asterisks indicate significance at the following levels * p<0.1, ** p<0.05, and *** p<0.01.

We address this issue with four further analyses. First, we evaluate the stability of our estimates to restricting our sample to the district-month cells with non-negative weights —

[14]Given the small sample sizes we have in our data at the monthly level — the level at which treatment is defined, an event-study analysis leads to very imprecise estimates. For completeness, we report the estimates obtained from an event-study analysis at the quarterly level (Appendix Figure A.1), which show a noisy but largely flat and non-negative pattern prior to the introduction of the CHX program in the district, and then increasingly negative treatment effects after the program is rolled out.

where heterogeneity in treatment effects cannot lead to sign reversal. We compute the weights derived in de Chaisemartin and d'Haultfoeuille (2020) and find 246 cells with negative weights out of 1,623 (Columns (5) and (6) of Table 4). After dropping these cells, the estimated effect of CHX application is almost unchanged (from 0.0179 to -0.0185). In this new sample, the weights change and 54 out of the 1377 remaining cells now have negative weights. After three iterations of dropping cells with negative weights and re-estimating both our two-way fixed effects model and the remaining cells weights, we obtain a sample with no negative weights and the treatment effect on the remaining cells is -0.0184, compared to -0.0179 in the full sample, illustrating that our results are not driven by the negative weighting of some treatment effects.[15] Second, we compute the minimum standard deviation in the treatment effect across all district-month cells which would be required for the average ATT over all cells to in fact be zero, and assess the likelihood of the treatment effect exhibiting this much variation.[16] We find that the required amount of heterogeneity for the true effect to be zero is implausible. The minimum standard deviation required for a zero effect is between 0.0196 and 0.0283 (Table 4 Column (8)), whereas a random variable distributed between 0 and -0.0424 (the ATT needed for the complete eradication of NMR in our sample) can *at most* have a standard deviation of $\sqrt{\frac{1}{4}(0.0424)^2} = 0.0212$ (Popoviciu, 1935), which only occurs in the extreme case where half of the distribution is concentrated at 0 (no effect) and the other half at -0.042 (total eradication of NMR). If the absolute values of the ATTs for our 1,623 cells were drawn from a uniform distribution between 0 and -0.042, for instance, the standard deviation (SD) would only be 0.012.[17] Third, we report average treatment effects on the treated across all treated observations using the estimator proposed by Borusyak et al. (2021) (columns (2) and (4)) and confirm that our estimates are robust to allowing for dynamic heterogeneous treatment effects.[18] Finally, in Section 5.2 we provide estimates of the ATT using a causal forest approach instead of a linear two-way fixed effects regression and find very consistent estimates.

4.2.2 Other Programs

When a two-way fixed effects regression includes more than one variable with heterogeneous effects, the estimated average effect of each of these variables may be contaminated by that of the other(s) (de Chaisemartin and D'Haultfoeuille, 2022). A test of parallel pre-treatment trends however suggests that control variables other than time- and district- fixed

[15] Dropping cells with negative weights results in disproportionately dropping later time periods since these are more likely to carry negative weights. It is therefore related to what is proposed in Jakiela (2021).

[16] We compute both the weights and the minimum standard deviation using de Chaisemartin and d'Haultfoeuille (2020)'s *two-wayfeweights* command.

[17] If they were drawn instead from a normal distribution with mean 0.018 (our baseline two-way fixed-effects estimate) and SD 0.0196 — the minimum SD required for the average ATT over all cells to be zero in this specification, 29 percent of ATTs would have to be outside the [-0.0424,0] range, which is not plausible.

[18] We are constrained by which estimator we can use in our set-up where each district-month cell only has few observations. This rules out estimators which, contrary to Borusyak et al. (2021) only use the last pre-treatment period.

effects are not required for the parallel trends assumption to hold (column (3) of Table 3) and we find a similar CHX effect whether control variables are included or not (columns (1) and (2) of Table 3).

We also explore in detail the potential interaction between CHX and the broader healthcare programs concerned with neonatal mortality summarized in Table 1. First, we inspect visually the changes in neonatal mortality over time against the roll-out of both CHX care and the CB-NCP and CB-IMNCI (in which CB-NCP was later integrated). As shown in Appendix Figure A.2, there is no decrease in neonatal mortality between 2009 and 2011, even though the coverage of CB-NCP goes from 8% to 49% during this period. After that, while CB-NCP's coverage only increases by 14 percentage points between 2011 and 2016, neonatal mortality decreases steadily as CHX coverage goes from 20% to 92% coverage. CHX coverage increases linearly throughout its expansion, matched by a near-linear decrease in neonatal mortality. The very rapid scale-up of CB-IMNCI between 2013 and 2016 is not accompanied by an acceleration of the decrease in neonatal mortality. Second, we estimate the effect of CHX while allowing two-way interactions with CB-NCP and CB-IMNCI and a three-way interaction between all three. Results reported in Table A.6 show no sign of complementarities between CHX and the other programs, suggesting that our main specification is not capturing the combined effect of chlorhexidine cord care and wider neonatal care. For completeness, we also estimate the effect of CHX in the sample of births where neither CB-NCP nor CB-IMNCI are present (column (7)) and the effect of CB-NCP where neither CHX nor CB-IMNCI are present (column (8)) and confirm that CHX is effective in decreasing neonatal mortality independently of the presence of the other programs, while we find no evidence that CB-NCP has any independent effect — consistent with Paudel et al. (2017)'s findings that CB-NCP did not lead to significant improvements in newborn care practices.

In Appendix Section B, we report on further robustness tests including using alternative samples and covariates, not allowing for age of death heaping at one month old, using survey weights and using a logit- instead of a linear probability model. Our conclusions are robust across these alternative specifications.

In the next section, we go beyond average effects estimated across the whole sample to investigate the heterogeneity of the benefits of CHX and what it implies for optimal policy targeting.

5 Treatment Effect Heterogeneity and Lessons for Policy Targeting

Having established significant average beneficial effects of CHX-NCP at scale, we now turn to the question of treatment effect heterogeneity.

As previously discussed, the WHO recommendations are context-specific, in line with find-

ings from meta-analysis of the five existing trials (e.g. Imdad et al., 2013). Namely, between 2013 and 2022, WHO guidelines only recommended CHX cord care for home deliveries in areas with rates of neonatal mortality above 3 percent, and the 2022 revised guidelines only recommend CHX cord care in contexts where the application of harmful substances is common. Such application is more likely to occur for home births and is, in many contexts, not well documented (Coffey and Brown, 2017) and may therefore need to be approximated with more widely available data such as the prevalence of home births. Assessing treatment effect heterogeneity by place of delivery is therefore a useful first step to shed light on the desirability of these recommendations in "real-world" circumstances.

We then go beyond assessing the degree of treatment effect heterogeneity by place of delivery using a data-driven approach based on recent developments in the machine learning literature and compare our findings with factors associated with neonatal deaths caused by sepsis across six districts of Nepal studied in 2012/13 (Erchick et al., 2022).

5.1 Heterogeneity using the Two-Way Fixed Effects Approach

Place of delivery is only collected by the DHS for births in the five years leading to the survey. To use data covering a longer period of time and thus increase statistical power, we predict whether a child was delivered at home using a linear probability model regressing an indicator for being delivered at home on birth order, maternal age group, child gender, maternal ethnicity, altitude quintile, maternal education, rural location, wealth quintile, district fixed effects and date of birth — defined by Nepali month and year of birth — fixed effects (prediction model results reported in Appendix Table A.4).

In the sample for which we know the place of delivery, when predicting a home birth based on a probability of home delivery above 0.5, we predict place of birth correctly in 76 percent of cases (see Appendix Figure A.3). To account for the uncertainty in classifying births based on their predicted- rather than observed place of delivery, we obtain bootstrapped standard errors — clustered at the district level — by drawing 200 random samples from the original dataset, and, for each random sample, predicting whether the child is delivered at home or not and then re-estimating the relevant variant of Equation (1).

In Column (1) of Table 5 we repeat our baseline finding for ease of reference. In Column (2), we allow the effect of CHX to vary by predicted place of birth by including a control for predicted place of birth (1[P(home birth)>0.5]) and an interaction between predicted home birth and the CHX treatment variable. In this specification, the treatment effect is not significant for predicted facility deliveries but it is four times larger (2.8 percentage point) and statistically significant among predicted home deliveries. Finally, in Columns (3) and (4) we allow all the model coefficients to vary by predicted place of birth, which leads to a near-zero estimated effect of CHX among predicted facility deliveries (0.1 percentage point) while the estimated decrease in the probability of neonatal mortality among predicted home deliveries remains

equal to 2.8 percentage points — and we can reject the null of no difference in treatment effect between the two samples defined by predicted place of delivery (p-value: 0.031). In Appendix Table A.5, we repeat this heterogeneity analysis using a different prediction sample (see Appendix Table A.4 for prediction model estimates). Instead of using only the 2011-2016 recent births subsample of the 2016 DHS, for which we observe place of birth, in this sensitivity check we stack the recent births subsamples for which we observe place of birth for five DHS: 1996, 2001, 2006, 2011 and 2016 and use the parameters of this model to split our DHS 2016 main sample by predicted home or facility birth. We similarly conclude that CHX is, on average, only beneficial for predicted home births and that the average effect among those is large, but with a more marked contrast in treatment effects between the home- vs. facility birth samples and a marginally significant positive coefficient for the much smaller sample of predicted facility births.

Table 5: Heterogeneity by place of delivery

	Sample			
	All	All	P(home birth)	
			<0.5	>0.5
	(1)	(2)	(3)	(4)
CHX	-0.018**	-0.007	0.001	-0.028**
	(0.007)	(0.007)	(0.009)	(0.011)
1[P(home birth)>0.5]		-0.001		
		(0.005)		
CHX × 1[P(home birth)>0.5]		-0.021***		
		(0.008)		
CHX + CHX × 1[P(home birth)>0.5]		-0.028***		
		(0.008)		
Observations	23465	23465	10,860	12,605
Clusters	73	73	73	73
Control mean of dep. var	0.042	0.042	0.033	0.050
P-val (dif across sample)			0.031	

Notes: All specifications are estimated as linear probability models using OLS with the full set of demographic, SES, and program controls. Demographic controls include birth order (three indicators), five year maternal age group indicators, and gender. SES controls include education (three indicators), wealth (four indicators), rural indicator, altitude quintile indicators, and ethnicity indicators. Program controls include controls for the CB-NCP and CB-IMNCI health programs. All specifications are estimated with district and month-year date of birth fixed effects. We split the sample according to the predicted place of delivery, based on the linear probability model shown in Column (2) of Appendix Table A.4. Bootstrapped standard errors based on 200 iterations and clustered at the district level in parentheses. Asterisks indicate significance at the following levels: * p<0.1, ** p<0.05, and *** p<0.01.

Our results show clear evidence of beneficial effects of CHX-NCP on children predicted to being born at home, and no evidence of such benefits, on average, among other births. Heterogeneous effects between predicted place of delivery may come from different rates of compliance or different efficacy conditional on compliance, for which place of delivery acts as proxy. Here we estimate intention-to-treat effects, which are of interest to policy makers when

compliance cannot be enforced, as is the case for home births. As discussed in Section 2.2, compliance estimates vary but are close to 100% in the case of facility births, where treatment effect estimates are smallest, thus suggesting that different compliance rates are not a key driver of heterogeneity. In the next section, we investigate further the question of which newborns should be targeted by CHX cord care programs to achieve the largest neonatal survival gains.

5.2 Heterogeneity with Causal Forests

Place of delivery is likely to proxy for risk factors such as hygiene conditions including application of harmful substances to the stump, healthcare at- and shortly after birth and health endowment at birth (rather than differences in compliance, as just discussed). To better understand the treatment effect heterogeneity we observe and therefore potentially improve on current WHO recommendations for targeting, we turn to machine learning. We focus on characterizing heterogeneity along dimensions that can be realistically used by policy makers to target beneficiaries. District prevalence of harmful substance application is unlikely to be observed by the policy maker but since it is available in the Nepal DHS we include this variable in a stepwise fashion in the analysis so that we can assess its role as a potential driver of heterogeneous treatment effects.

We first use regression forests to "residualize" the treatment indicator and our outcome of interest (neonatal mortality) — i.e., to purge them of variation coming from, in our case, district- and month$\times$year of birth as captured by fixed effects and availability of the two main neonatal health programs. Using these residuals as outcomes, we then estimate a causal forest on a range of potential outcome predictors or "features". From a policy point of view, it may be more appealing to target whole communities than individuals, so in addition to the individual sample characteristics included as covariates in our two-way fixed effects regressions, we also consider district-level features measured in the five years prior to the survey. Namely, we use, at the individual level: birth order, gender, maternal education, maternal age, wealth, altitude, rural, predicted place of delivery, health programs, district, ethnicity; and, at the district level: the neonatal mortality rate, the average number of antenatal care (ANC) visits, the share receiving at least four ANC visits, the share receiving iron tablets during ANC visits, the share receiving at least one tetanus injection, the share of births protected against neonatal tetanus, the share receiving the first ANC visit during the first trimester, the share receiving prenatal care by a medical doctor or nurse, the share of newborns considered small at birth, the share initiating breastfeeding within one hour after birth, the share of births delivered at home, the share of births delivered at a public health facility, the share of births delivered at a private health facility, the share of births delivered at a non-private/non-public health facility, the share of deliveries assisted by a medical doctor or nurse, the share of deliveries unassisted by a health professional, the share receiving a postnatal check within two days after birth, the share receiving full immunization, and the share of births for whom the mother reports applying a

potentially harmful substance to the umbilical cord stump (e.g., mustard oil, turmeric, etc.). While the latter is not a salient event and therefore subject to recall error (Beckett et al., 2001), our measure of how "common" this application is is likely to be as good as any a policy maker can hope to have access to.

The building blocks of the causal forest are its trees. Each tree is created by partitioning a 50% draw of the sample into leaves defined by the value taken by a subset of features. The partitioning algorithm finds the combination of values taken by these features which maximizes treatment heterogeneity across leaves and penalizes treatment effect variance within leaves (Athey and Imbens, 2016). Following best practice, the fine-tuning of the algorithm is done optimally without researcher input based on cross-validation.[19]

Before reporting on the heterogeneity patterns uncovered by this exercise, we report results of diagnostic tests which indicate that the causal forest successfully captures both average and heterogeneous treatment effects. More specifically, in panel A of Table 6 we show results of Chernozhukov et al. (2020)'s omnibus test for heterogeneity modified to be applied in an observational setting following the procedure implemented in Tibshirani et al. (2021). Intuitively we are estimating a linear regression of the individual's treatment effect predicted by the forest on the average predicted treatment effect (Mean Forest Prediction) and the individual's predicted deviation from the average treatment effect (Differential Forest Prediction). If the forest captures the average treatment effect well and if there is treatment effect heterogeneity that is also captured by the causal forest, both coefficients should be 1. In our case both coefficients are close to 1 (and 1 is included in the confidence interval) and there is evidence of significant treatment effect heterogeneity since the coefficient associated with the differential forest prediction is statistically different from zero (at the 10% significance level).

Panel B of Table 6 shows the Augmented Inverse-Propensity Weighted (AIPW) Average Treatment Effects based on the causal forest. The AIPW is doubly robust, meaning that it is a consistent estimator of the ATE as long as at least one of (i) the propensity score *or* (ii) the outcome model, is correctly specified. Reassuringly, our AIPW estimates are similar to our two-way fixed fixed effects specification (full sample: -1.9 percentage points compared to -1.8 percentage points in Table 3, predicted home deliveries: -3 percentage points compared to -2.8 percentage points in Table 5), even though for predicted facility births, the AIPW is suggestive of CHX being somewhat effective (-0.7 percentage points, significant at 5% compared to an insignificant 0.1 percentage points in Table 5).

Figure 4 shows the distribution of individual Conditional Average Treatment Effects (CATEs) for, respectively: the full sample, the sample of predicted home births, and the sam-

[19]We estimate the causal forest using the *grf* package in R (Tibshirani et al., 2021) with 2000 trees and all other parameter settings selected based on cross-validation. We use 50% of the sample to grow each tree. The splitting structure of the trees is determined on a 50% sub-sample of the tree sub-sample, after which the tree is populated by the the other 50% to estimate the treatment effects. For the splits in the trees we consider 30 variables and we restrict the nodes to have at least 5 observations. Appendix C provides further details about the causal forest procedure.

Table 6: Causal forest fit & doubly robust average treatment effects

A. Omnibus diagnostic test for forest fit	
Mean Forest Prediction	1.123***
	(0.255)
Differential Forest Prediction	0.677*
	(0.479)
B. Doubly Robust Average Treatment Effects	
Full sample	-0.019***
	(0.003)
Predicted facility births	-0.007**
	(0.004)
Predicted home births	-0.030***
	(0.004)

Notes: Panel A shows the results for the omnibus test inspired by equation 3.1 in Chernozhukov et al. (2020) modified to the observational setting and implemented through the *test_calibration* function from the *grf* library in R. Panel B shows the Augmented Inverse-Propensity Weighted (AIPW) Average Treatment Effects. Standard errors in parentheses. Asterisks indicate significance at the following levels * p<0.1, ** p<0.05, and *** p<0.01. Note that, following Athey et al. (2019), the significance levels in panel A. are for the one-sided tests.

ple of predicted facility births. We observe that a large fraction of the sample is estimated to benefit from the treatment. However, the CHX programme is predicted to have a small or even harmful effect for a non-negligible part of the distribution. Indeed, as with most public health interventions, there are potential adverse consequences of CHX cord care. These include the risk of encouraging the application of other, potentially harmful substances by departing from the standard message of keeping the stump dry and clean, as well as the risk of diverting human, logistical, and financial resources away from other essential medicines and tasks in an area where the gap between recommended health care and practice is already large (Requejo et al., 2015). As expected, the distribution is shifted to the left for births that we predict to take place at home. Among those, very few are expected to have small or harmful treatment effects. However, we also note that a large share of children predicted to be born at a facility are estimated to benefit from the treatment. Figure 5 further illustrates the distribution of treatment effects across each subsample by reporting ATEs for each tertile of the overall- and birth place samples.

In sum, targeting treatment by place of delivery appears at first glance to be convenient and likely effective in avoiding adverse consequences. But the substantial overlap in the home- and facility births treatment effects distributions show that this targeting approach is a blunt policy tool which may be improved upon.

We now turn to a characterization of the treatment effect heterogeneity. In Table 7 we compare means for selected variables across the first and third tertiles. As expected from the distributions reported above, in the first tertile, where we observe large benefits of the treat-

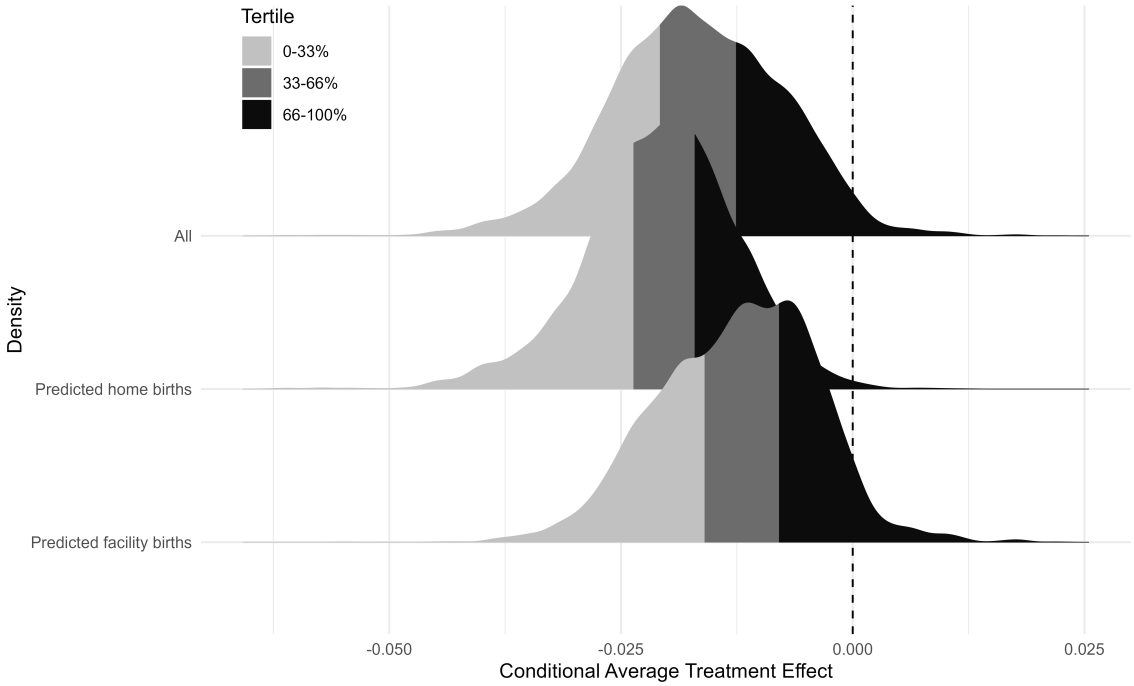


Figure 4: Distribution of CATEs by place of delivery

Notes: The distributions are estimated using bandwidth selected based on Silverman's 'rule of thumb' (Silverman, 1986) and a gaussian kernel.

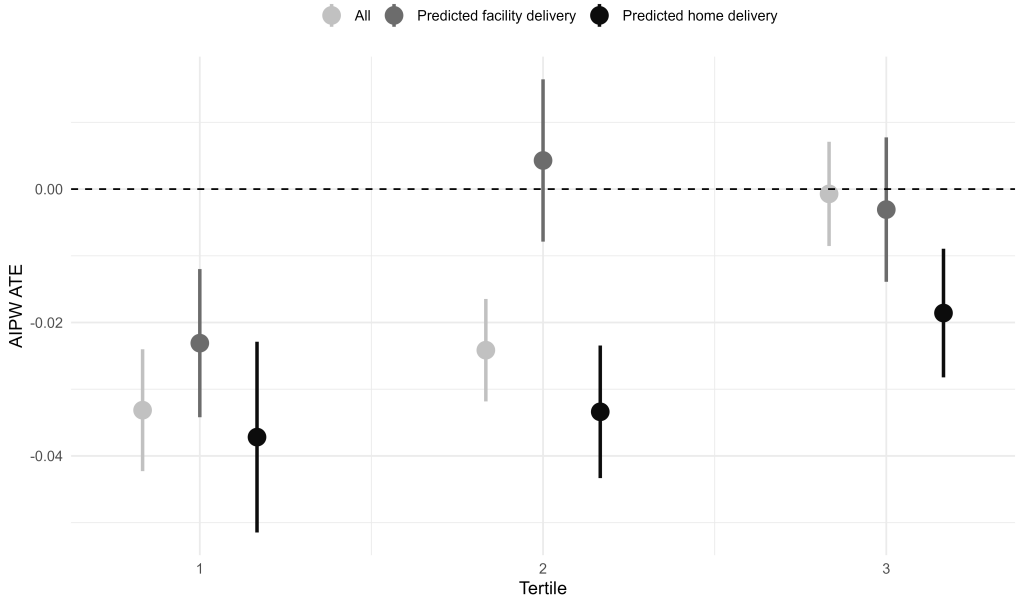


Figure 5: Doubly robust ATEs by tertiles of CATEs

Notes: The Augmented Inverse-Propensity Weighted (AIPW) Average Treatment Effects are estimated for tertiles of the conditional average treatment effects shown in Figure 4. The p-values for H_0 of equal treatment effects in the first and third tertiles are : $p<0.001$ for the full sample ($p<0.001$ for predicted home births and $p<0.001$ for predicted facility births).

ment, 76 percent of births are predicted to take place at home compared to only 25 percent in the third tertile. Moreover, children in the first tertile are more often boys (unsurprisingly given that male newborns are more likely to die), and are often born to very young, less educated, rural mothers and in districts with higher NMR. They are also more often found in districts where the application of harmful substances to the umbilical stump is more common, although the difference between tertiles (36% prevalence harmful substance application in the first tertile vs. 24% in the third tertile) is less pronounced than for other covariates — suggesting that targeting treatment based on this variable may not be optimal. While these characteristics are correlated, the treatment effect heterogeneity uncovered by the causal forest goes beyond well-known predictors of CHX effectiveness. For instance, being predicted to be born at home explains 19% of the variation in the predicted conditional average treatment effects (CATE), the district prevalence of harmful substance application explains 9% of the CATE variation, and district NMR explains 5% of the CATE variation whereas gender, maternal age, maternal education, birth order, wealth quintile, altitude quintile, ethnicity, rural location, being predicted to be born at home, district NMR and the district prevalence of harmful substance application together explain 61% of the variation in CATE (Figure A.4).

Table 7: Covariate means across tertiles of CATEs

	Tertile			
	First	Third	Difference	P-val
A. All births				
AIPW	-0.033	-0.004	0.029	<0.001
CATE	-0.033	-0.001	0.032	<0.001
Control Neonatal Mortality	0.057	0.024	-0.033	<0.001
Female	0.408	0.528	0.120	<0.001
Predicted home delivery	0.759	0.254	-0.506	<0.001
Age: 15-19y	0.278	0.130	-0.148	<0.001
Age: 20-24y	0.388	0.461	0.073	<0.001
Age: 25-29y	0.212	0.288	0.076	<0.001
Age: 30-34y	0.091	0.096	0.005	0.321
Age: 35-39y	0.026	0.022	-0.004	0.131
Age: 40-45y	0.005	0.003	-0.003	0.006
Education: No education	0.857	0.211	-0.647	<0.001
Education: Primary	0.101	0.202	0.101	<0.001
Education: Secondary	0.038	0.438	0.400	<0.001
Education: Higher	0.004	0.150	0.146	<0.001
Rural	0.536	0.277	-0.259	<0.001
District: harmful substance use	0.361	0.240	-0.121	<0.001
B. Predicted home births				
AIPW	-0.033	-0.004	0.022	<0.001

Continued on next page

Continued from previous page

	Tertile			
	First	Third	Difference	P-val
CATE	-0.037	-0.019	0.019	<0.001
Control Neonatal Mortality	0.063	0.039	-0.024	<0.001
Female	0.383	0.576	0.192	<0.001
Age: 15-19y	0.228	0.099	-0.130	<0.001
Age: 20-24y	0.390	0.369	-0.020	0.054
Age: 25-29y	0.244	0.346	0.102	<0.001
Age: 30-34y	0.105	0.133	0.027	<0.001
Age: 35-39y	0.029	0.046	0.018	<0.001
Age: 40-45y	0.005	0.008	0.003	0.068
Education: No education	0.928	0.525	-0.403	<0.001
Education: Primary	0.065	0.278	0.213	<0.001
Education: Secondary	0.006	0.169	0.162	<0.001
Education: Higher	0.000	0.028	0.028	<0.001
Rural	0.611	0.505	-0.106	<0.001
District: harmful substance use	0.388	0.282	-0.105	<0.001
C. Predicted facility births				
AIPW	-0.029	0.000	0.029	<0.001
CATE	-0.023	-0.003	0.020	<0.001
Control Neonatal Mortality	0.052	0.018	-0.034	<0.001
Female	0.405	0.517	0.112	<0.001
Age: 15-19y	0.430	0.112	-0.318	<0.001
Age: 20-24y	0.397	0.517	0.121	<0.001
Age: 25-29y	0.120	0.272	0.152	<0.001
Age: 30-34y	0.044	0.082	0.038	<0.001
Age: 35-39y	0.009	0.015	0.006	0.020
Age: 40-45y	0.001	0.001	0.001	0.479
Education: No education	0.637	0.068	-0.569	<0.001
Education: Primary	0.202	0.127	-0.075	<0.001
Education: Secondary	0.141	0.577	0.436	<0.001
Education: Higher	0.020	0.228	0.209	<0.001
Rural	0.282	0.186	-0.096	<0.001
District: harmful substance use	0.304	0.226	-0.078	<0.001

Notes: The table shows covariate means for the first and third tertile of the sample based on the estimated CATEs.

5.3 Lessons for policy targeting

We now ask what the optimal targeting policy is according to the heterogeneous doubly-robust treatment effects we estimate in the data and taking into account the uncertainty surrounding these estimates, using the approach proposed by Athey and Wager (2021). Concretely, we use the estimates of individual treatment effects from the causal forest (doubly-

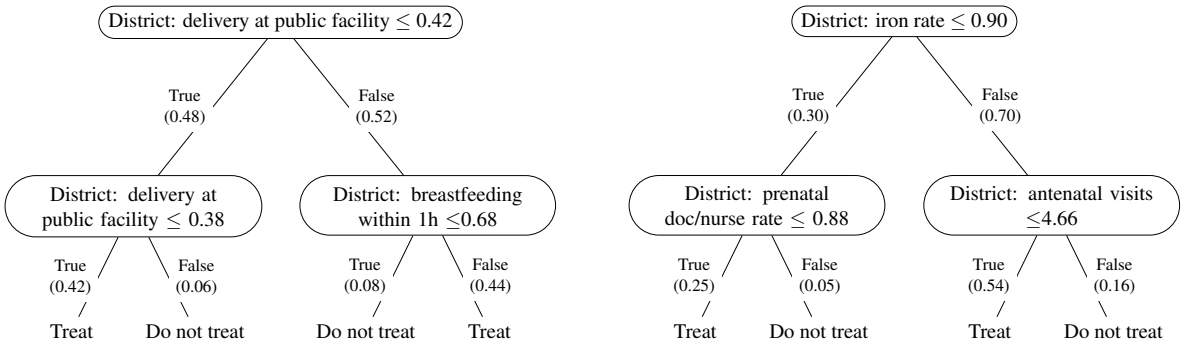
robust scores) to find the optimal policy, allowing this policy to use a wide range of antenatal-, delivery-, and postnatal care variables as well as, in some specifications, maternal characteristics and birth order. A policy consists of a treatment allocation rule based on covariates and the optimal policy is the allocation rule that leads to the difference in the expected utility from this policy and the maximum expected utility which could be achieved by a policy (in a given class Π and applied to the same population) being asymptotically "small".[20] We derive our optimal policies using the policy learning algorithm developed by Athey and Wager (2021) (and implemented in R with the policytree function due to Sverdrup et al., 2020). In particular, we split the sample into ten equally-sized folds and, for each fold, find the optimal policy using data from the other k-1 folds and then apply this optimal policy to the left-out k*th* fold.

We study policies obtained with three different sets of covariates. For the first policy, we allow the algorithm to select optimally who should be treated based on the full set of individual- and district-level variables. For the second policy, we only allow targeting based on district-level variables. For the third policy, we allow targeting based on all district variables except the district prevalence of harmful substance application. This variable is, indeed, not typically observed and so it may not be feasible to use it for targeting purposes. In addition, comparing the second- and third policies sheds light on the relevance of the prevalence of harmful substance application specifically, which is the focus of the latest WHO targeting recommendations.

To fix ideas, in Figure 6, we show the optimal policies obtained using all the district-level variables (but no individual characteristics) for the first and tenth folds. As illustrated in Figure 6, the resulting optimal assignment varies across folds and so in Table 8 we report the number of times each covariate is selected in an optimal policy. If only district-level variables are used for targeting, variables which capture the quantity and quality of antenatal care are the most commonly selected along with the share of newborns born in *public* facilities — interestingly, much more commonly than the share predicted to being born at home, the district prevalence of harmful substance application, or baseline neonatal mortality rates. The fact that the optimal targeting policies mostly select variables related to the quality and quantity of antenatal care is consistent with risk factors for neonatal death caused by *sepsis* identified using data from verbal autopsies carried out in 6 districts of Nepal in 2012/13. Erchick et al. (2022) indeed find that having fewer than four antenatal care visits is correlated with death by sepsis relative to birth asphyxia in multivariate regressions (while home delivery is not significantly more or less common for any specific cause of neonatal death).

In Table 9 we compare the optimal policies obtained with each set of covariates to the 2013-2022 WHO recommendation of treating only — here, predicted — home births in districts with NMR above 3 percentage points. We also compare to this policy the revised recommendation of treating only newborns in contexts where the application of harmful substances to the um-

[20] Where asymptotically "small" means "bounded on the order of $\sqrt{VC(\Pi)/n}$ with high probability", where $VC(\Pi)$ is the Vapnik-Chervonenkis dimension of class Π and n is the number of observations (Athey and Wager, 2021, , p.135).



(a) Fold 1 **(b)** Fold 10

Figure 6: Examples of optimal policies

Notes: Figure 6a shows the optimal policy obtained for the first fold of the data based on all district level variables: antenatal visits (timing and number), whether received/bought iron tablets during pregnancy, tetanus protection during pregnancy, place of delivery, postnatal visits, immunization rate, neonatal mortality rate, nurse/doctor delivery support, small baby, application of potentially harmful substance. Figure 6b shows the optimal policy obtained for the tenth fold of the data based on the same district level variables. Population shares are shown in parentheses.

bilical stump is "common". In the absence of a published threshold for this practice to be considered common, we use as threshold the centile of the distribution allowing us to compare the predicted effect on NMR of treating roughly the same proportion of newborns with the 2013-2022 and revised 2022 WHO policies. Without imposing any constraint on the share of treated newborns, the data-driven optimal policies treat between 81 and 83 percent of the sample compared to only 32 percent of the sample for the 2013-2022 WHO policy (and, by construction, also our operationalization of the 2022 revised policy). Since Figure 4 shows that there are significant benefits of treatment for at least 50 percent of the population, it is not surprising that the optimal policies with no further constraint than maximizing NMR reductions by applying up to two levels of targeting treat a larger share of the population and are able to reduce the neonatal mortality rate by more than the 2013-2022 WHO policy. This WHO policy would reduce neonatal mortality by 1.1 percentage point overall, whereas any of the three optimal policies we consider would reduce neonatal mortality by 1.9 percentage points, which is statistically significantly larger. Looking at the two first columns of Table 9, we can see that the average treatment effect on the treated tends to be larger with the WHO policies than the (so far unconstrained) optimal policies, so that the larger overall reduction in NMR is obtained with the optimal policies by exhausting most of the survival gains among those left untreated by the WHO policies. Interestingly, when imposing as we do that the optimal policy only use two levels of screening, the benefit of using all variables to design the optimal policy is negligible, compared to only using district-level variables which are more readily available to policy-makers.

Optimal policies vary across folds so we also study the effect of applying a single targeting

Table 8: Variables selected by optimal policies

	Unconstrained			Constrained			
	Individual & district variables	District variables only	District variables excl. harmful substance use	Individual & district variables	District variables only	District variables excl. harmful substance use	Total
Birth order	0	0	0	1	0	0	1
District: antenatal visits	8	4	4	0	0	0	16
District: antenatal visits in 1. trimester	0	2	2	3	10	10	27
District: antenatal visits$\geq$4	0	0	0	2	8	8	18
District: control NMR	0	2	2	1	2	2	9
District: delivery assisted by doc/nurse	0	1	1	0	1	1	4
District: delivery home	1	0	1	5	0	0	7
District: delivery in public facility	0	6	6	0	7	7	26
District: delivery private	1	1	1	0	0	0	3
District: full immunization	5	0	0	0	1	1	7
District: harmful substances	0	1	0	0	0	0	1
District: iron tablets during pregn.	3	1	1	2	0	0	7
District: postnatal check within 2days	0	3	3	0	0	0	6
District: prenatal care by doc/nurse	0	3	3	0	0	0	6
District: started breastfeeding $\leq$ 1h	1	4	4	0	0	0	9
District: tetanus protected	0	2	2	0	0	0	4
District: unassisted delivery	0	0	0	5	1	1	7
Maternal age	5	0	0	1	0	0	6
Maternal education	5	0	0	10	0	0	15
Predicted home delivery	1	0	0	0	0	0	1

Notes: this table shows how often each variable is included in the optimal policy across the ten folds for the six optimal policies indicated by the column headers.

rule using the two variables most commonly selected by the unconstrained optimal policies and their (rounded) cut-off values. Namely, we study the effect of applying a simple targeting rule based on the district average number of antenatal care visits being below 4.6 and the district share of deliveries occurring in public facilities being below 40%. A simple targeting rule treating newborns if either of these conditions holds successfully selects 74% of those treated

using the more complex set of optimal policies using district-level variables only. Applying this simple targeting rule results in treating 79% of newborns and would be predicted to achieve a similar reduction in NMR to that based on the more complex sets of optimal policies.[21]

Our — so far — unconstrained policies, which treat a much larger share of the population than the ones recommended by the WHO, are predicted to nearly double the mortality reduction but would however also be more expensive than the WHO recommended ones. We therefore also report results from a constrained optimization in Panel C of Table 9. The resulting constrained policies are based on an algorithm where we slowly increase the "price" of the treatment (set to zero in the default, unconstrained case) until the number of treated individuals is below or equal the number of treated individuals with the WHO policy. As the results show, the benefits are considerably lower than with the unconstrained policies and statistically indistinguishable from the 2013-2022 WHO policy. Focusing on the more feasible optimal policy targeting only based on district characteristics, the predicted reduction in neonatal mortality is 1.2 percentage points, obtained by treating just under 29% of births, which is statistically indistinguishable from the 1.1 percentage point reduction achieved with either of the two WHO policies treating each 32% of births.

Taken together, our results show that the WHO guidelines do an excellent job at targeting a third or so of newborns who would stand to benefit as much as any other from CHX, but also exclude many newborns whose chance of survival would be much improved by CHX cord care. Our optimal targeting exercise, which is based on a utilitarian criterion, indicate that it would be optimal under this criterion to treat many more newborns. This finding should however be understood with the caveat that a targeting that is optimal according to one welfare criterion may not be optimal according to a different one. Applying a welfare criterion which would put a much lower weight on the risk of not treating an individual who might benefit from treatment than on the risk of treating an individual who might be worse-off if treated may favor a policy rule closer to the WHO's — since, as is clear from Figure 4, part of the sample is predicted to have near-zero or even adverse treatment effects.[22]

In the next section, we extrapolate our heterogeneity analysis to other DHS samples in- and outside Nepal to assess the soundness of our findings so far and their informativeness beyond Nepal.

[21] An even simpler rule treating newborns if the district average number of antenatal care visits is below 4.6 would only result in a marginally smaller share treated and overall NMR reduction.

[22] See also Kowalski (2019), where finite-sample bounds are derived to infer quantities such as the number of individuals who would die if treated with a new drug based on data from a randomized trial.

Table 9: Reduced mortality and share treated under alternative targeting policies

	ATT	ATU	%treated	ΔNMR	ΔNMR-ΔNMR$_{WHO}$
A. Pre-defined policies					
WHO policy 2013-2022	-0.036***	-0.012***	32.3	-0.011***	
	(0.004)	(0.003)		(0.001)	
WHO policy 2022	-0.034***	-0.012***	31.8	-0.011***	0.000
	(0.003)	(0.004)		(0.001)	(0.001)
District average antenatal visits $\leq$ 4.6	-0.024***	-0.001***	79.2	-0.019***	-0.008***
or District share delivered in public fac. $<$0.4	(0.003)	(0.004)		(0.002)	(0.002)
B. Unconstrained optimal policies					
Individual & district variables	-0.023***	-0.003	81.3	-0.019***	-0.007***
	(0.003)	(0.004)		(0.002)	(0.002)
District variables only	-0.023***	-0.004	83.2	-0.019***	-0.007***
	(0.003)	(0.003)		(0.002)	(0.002)
District variables only (excl. harmful subst)	-0.023***	-0.003	83.4	-0.019***	-0.007***
	(0.003)	(0.003)		(0.002)	(0.002)
C. Constrained optimal policies					
Individual & district variables	-0.041***	-0.010***	29.6	-0.012***	-0.001
	(0.004)	(0.003)		(0.001)	(0.001)
District variables only	-0.039***	-0.011***	28.5	-0.011***	0.000
	(0.005)	(0.002)		(0.001)	(0.001)
District variables only (excl. harmful subst)	-0.041***	-0.011***	28.7	-0.012***	-0.000
	(0.005)	(0.002)		(0.003)	(0.001)

Notes: Rows labeled "Individual & district variables" show the reduction in NMR using the optimal policy based on all variables used in the causal forest (except gender and fixed effects) and district level variables: antenatal visits (timing and number), iron treatments, tetanus protection, place of delivery, postnatal visits, immunization rate, neonatal mortality rate, nurse/doc delivery support, small baby, share whose mothers self-reports applying a potentially harmful substance to the cord stump. Rows labeled "District variables only [(excl. harmful subst)]" show the reduction in NMR using the optimal policy based on all district level variables, except for the prevalence of harmful substance application if so indicated. "Constrained" optimal policies are obtained by adding a cost to the treatment until the proportion of treated individuals is below that treated under WHO 2013-2022 policy. The last column reports differences between the estimated change in NMR relative to WHO 2013-2022 policy. Standard errors in parenthesis. Asterisks indicate significance at the following levels * p$<$0.1, ** p$<$0.05, and *** p$<$0.01.

6 Extrapolating the Effect of Nepal's CHX Roll-Out Program across RCT Study Locations

The results of the causal forest suggest that there is substantial heterogeneity in the treatment effect of the Nepalese CHX national program (CHX-NCP), which echoes the fact that CHX trials were very successful in reducing NMR in three cases, but had no significant effect in two other ones. As discussed in the introduction, the treatment is not fully comparable between the RCTs and the Nepalese national roll-out because of differences such as the number of doses and who applied CHX, as well as, crucially, because RCT subjects in both control- and treatment groups received additional preventive and remedial health care, which also varied across RCT settings. This additional health care can explain the lower-than-expected mortality

rates observed in the trials' *control* groups and may have contributed to smaller treatment effects. The predicted effect of applying a CHX-NCP-like treatment to samples drawn from the regions where the RCTs took place should therefore not match the actual RCT treatment effects even if we could perfectly predict the effect of implementing CHX-NCP in these regions and the RCT samples and our DHS samples were equally representative of these regions. The aim of our exercise is therefore to see the extent to which, despite these limitations, the picture of heterogeneity we uncover in the observational Nepalese dataset matches the general pattern of experimental findings.

To do so, we follow the doubly-robust extrapolation approach due to Dahabreh et al. (2020) as implemented in Tibshirani et al. (2022). More specifically, we construct samples for each of the five subnational regions and time periods in which the RCTs were implemented based on the relevant national DHS surveys.[23] We then train a simplified causal forest in our nationally representative Nepalese dataset based on the restricted set of variables that we are able to observe in all five samples to predict the Conditional Average Treatment Effects (CATEs) and the corresponding doubly-robust treatment effects (AIPWs) for each RCT setting. Restricting the set of covariates affects the forest's overall performance in predicting treatment effect differences, but not its ability to predict the difference in average treatment effect between predicted home- and facility births, for instance (see column (2) relative to column (1) in Appendix Table C.2).[24]

Table 10 first shows the control group NMR rates reported in the RCT studies. Unsurprisingly given the additional health care provided as part of these RCTs, these NMR rates are between 0.5 and 3.2 percentage points smaller than those observed in the DHS samples (reported in Panel B), which holding all else equal should lead to smaller treatment effects. We then report the average estimated doubly-robust treatment effects (AIPWs) of implementing the same treatment as in the national Nepalese roll-out across the five samples. As expected, we predict larger decreases in neonatal mortality from extrapolating the effect of the national roll-out than those found in the trials. But we predict large, statistically significant average treatment effects in the three samples corresponding to areas where RCTs found that CHX significantly reduced neonatal mortality whereas, for the two samples corresponding to the regions where the RCTs show no significant effects of CHX cord care interventions, the predicted average treatment effects of a hypothetical national roll-out are smaller and statistically insignificant.

[23] Namely, MoH, New ERA and ICF (2017); NIPORT, Mitra and Associates and ICF (2013); NIPS and ICF (2013); MoH, MoHCDGEC, NBS, OCGS and ICF (2016); CSO, MoH and ICF (2014).

[24] We estimate a causal forest using the full roll-out in Nepal and the same orthogonalization as in the main results described above. However, to make the causal forest comparable across the five samples, the forest is estimated on a reduced set of variables consisting of birth order, gender, maternal age, rural location, maternal education, predicted place of delivery, and wealth quintile, as well as 14 district level variables observed in all samples. The fit and results for this forest are shown in column (2) of Appendix Table C.2. When computing AIPWs accounting for differences in the distribution of variables in our main Nepal sample and the five RCT settings subsamples, we further need to drop the district-level variables from the causal forest as these otherwise perfectly predict whether the observation is found in the main Nepal sample or not. The fit and results for this forest are shown in column (3) of Appendix Table C.2.

In the rest of Table 10, we report the average predicted CATEs which enter the computation of the doubly robust treatment effects, and characteristics of the different samples used in the forest to illustrate the variety of settings. In Figure 7, we report the distributions of predicted CATEs in each setting, which shows that there is substantial predicted heterogeneity both between-settings and within each setting too. In particular, sizeable shares of predicted zero- and even positive treatment effects are observed only in the Tanzanian and Zambian subsamples. There are also substantial differences across settings in the distribution of treatment effects *given* predicted place of delivery, as well as within each setting for a given predicted place of delivery — again suggesting that place of delivery is a useful but blunt proxy for the effectiveness of CHX cord cleansing at scale.

7 Conclusion

Neonatal mortality is an increasingly large contributor to early life mortality across the world, accounting for 45% of under-5 deaths in 2015 compared to 35% in 1980 (Wang et al., 2016), and most neonatal deaths are believed to be preventable at comparatively low cost (Bhutta et al., 2014). Hopes that CHX cord care would be a "game changer" (Hodgins et al., 2013) faded away as heterogeneous findings across randomized trials led experts to question its effectiveness at scale despite the fact that these trials, for obvious ethical reasons, cannot proceed with a pure control and may therefore underestimate the effectiveness of CHX in real-world circumstances.

In this paper, we estimate the effect of implementing a nationwide program training health personnel including community health workers to apply CHX to the umbilical stump and to distribute a single CHX dose to mothers who plan to deliver their baby at home. We find that the program led to a large reduction in neonatal mortality (43 percent), driven by reduced neonatal mortality among newborns predicted to have been born at home. This provides novel evidence of the effectiveness of CHX cord care outside an experimental setting, and one of the few instances of evidence of successful nationwide intervention targeting neonatal mortality in a low-income country.

Using recently developed machine learning techniques, we find evidence of substantial heterogeneity in treatment effects in our nationally representative Nepalese observational data. While place of delivery and average neonatal mortality are good proxies for large treatment effects, the optimal targeting we identify implies treating more than two-and-a-half times more births than the WHO recommendation based on these two variables, which prevailed during 2013-2022. In addition, we find no evidence that the recent 2022 revised WHO recommendation to treat only births in settings where the application of harmful substances to the umbilical stump is common is likely to improve targeting relative to the 2013-2022 recommendation. We indeed estimate very similar overall neonatal mortality improvements from either approach to targeting a similar share of newborns, and find that larger conditional treatment effects are less

Table 10: Extrapolating the effect of CHX-NCP across CHX trial locations

	Bangladesh 2007-2009	Nepal 2002-2005	Pakistan 2007	Tanzania 2011-2014	Zambia 2011-2013
RCT Data:					
Control Neo. Mortality	0.028	0.019	0.036	0.012	0.014
Treatment Effect	-0.006	-0.005	-0.013	-0.001	0.002
DHS Data:					
A. Doubly Robust Treatment Effects					
AIPW	-0.021**	-0.023***	-0.038***	-0.016	-0.007
	(0.009)	(0.005)	(0.005)	(0.017)	(0.005)
B.Variable means					
CATE	-0.017	-0.023	-0.024	-0.004	-0.004
Neonatal mortality	0.033	0.051	0.044	0.019	0.020
Predicted home delivery	0.942	0.642	0.909	0.239	0.061
Female	0.488	0.555	0.464	0.553	0.488
Mother age 15-19y	0.196	0.168	0.071	0.086	0.153
Mother age 20-24y	0.374	0.453	0.270	0.233	0.253
Mother age 25-29y	0.214	0.182	0.306	0.248	0.233
Mother age 30-34y	0.071	0.036	0.071	0.145	0.114
Mother age 35-39y	0.014	0.000	0.052	0.088	0.046
Birth order: 1	0.239	0.204	0.179	0.162	0.194
Birth order: 2	0.243	0.263	0.135	0.147	0.200
Birth order: 3	0.204	0.197	0.167	0.122	0.145
Birth order: $\geq$4	0.314	0.336	0.520	0.569	0.460
Education: none	0.325	0.854	0.873	0.395	0.061
Education: primary	0.353	0.073	0.079	0.258	0.535
Education: secondary	0.298	0.058	0.048	0.347	0.356
Education: higher	0.024	0.015	0.000	0.000	0.048
Wealth quintile: 1	0.294	0.058	0.631	0.019	0.131
Wealth quintile: 2	0.214	0.234	0.222	0.189	0.256
Wealth quintile: 3	0.195	0.409	0.091	0.277	0.288
Wealth quintile: 4	0.160	0.248	0.056	0.368	0.177
Wealth quintile: 5	0.138	0.051	0.000	0.147	0.148
Rural	1.000	0.409	1.000	0.889	0.704
Observations	637	137	252	476	854

Notes: Average CATE based on predictions using the causal forest estimated on the country-wide Nepal sample using a reduced set of variables as shown in Appendix Table C.2 Column (2). AIPW based on predictions using the causal forest estimated on the country-wide Nepal sample using a reduced set of variables which do not perfectly predict whether an observation is from the country-wide Nepalese sample or the RCT samples as shown in Appendix Table C.2 Column (3). The table also shows the neonatal mortality rate observed in the relevant DHS subsample and the averages of the demographic and SES variables used in the forests. The samples are taken from the same years and regions as those covered in the respective trials. Namely, we include 2007-2009 births in rural areas of Sylhet from DHS Bangladesh 2011, 2002-2005 births in Sarlahi district from DHS Nepal 2016, 2007 births in rural areas of Sindh from DHS Pakistan 2012-13, 2011-2014 births in Pemba Island from DHS Tanzania 2015-16, and 2011-2013 births in Southern Province from DHS Zambia 2013-14. Standard errors for the doubly-robust average treatment effect in parenthesis. Asterisks indicate significance at the following levels: * p<0.1, ** p<0.05, and *** p<0.01.

strongly associated with district prevalence of harmful substance application than with home delivery.

Our findings regarding optimal targeting come with two important caveats. First, the targeting of any policy (for which treating everyone is either not affordable or not desirable due to potential adverse consequences) should be regularly reviewed since, in many applications, the distribution of treatment effects may evolve over time. Targeting by place of birth may, for instance, become less appropriate if hospital quality deteriorates with increased demand relative to supply over time or if home births become less conducive to infection due to wider use of safe delivery kits. Second, our conclusions are based on a utilitarian criterion. But a targeting that is optimal according to one welfare criterion (e.g., utilitarian) may not be optimal according to a different one (e.g., one that puts unequal weights on the risks of not treating an individual who might benefit from treatment versus treating an individual who might be worse-off if treated).

Finally, we extrapolate the causal forest heterogeneity analysis carried out in our national Nepalese sample to five settings in as many countries. Despite substantial differences in the nature of the intervention and control group in- and outside trials as well as between trials, the doubly-robust predicted effects of implementing the same program as that rolled out in Nepal across these five settings matches the broad pattern of heterogeneous experimental results. This bolsters our confidence in the heterogeneity analysis based on the Nepalese roll-out and its relevance for settings outside Nepal and thus suggests that CHX may be much more widely beneficial in real-life circumstances than the current received wisdom would indicate.

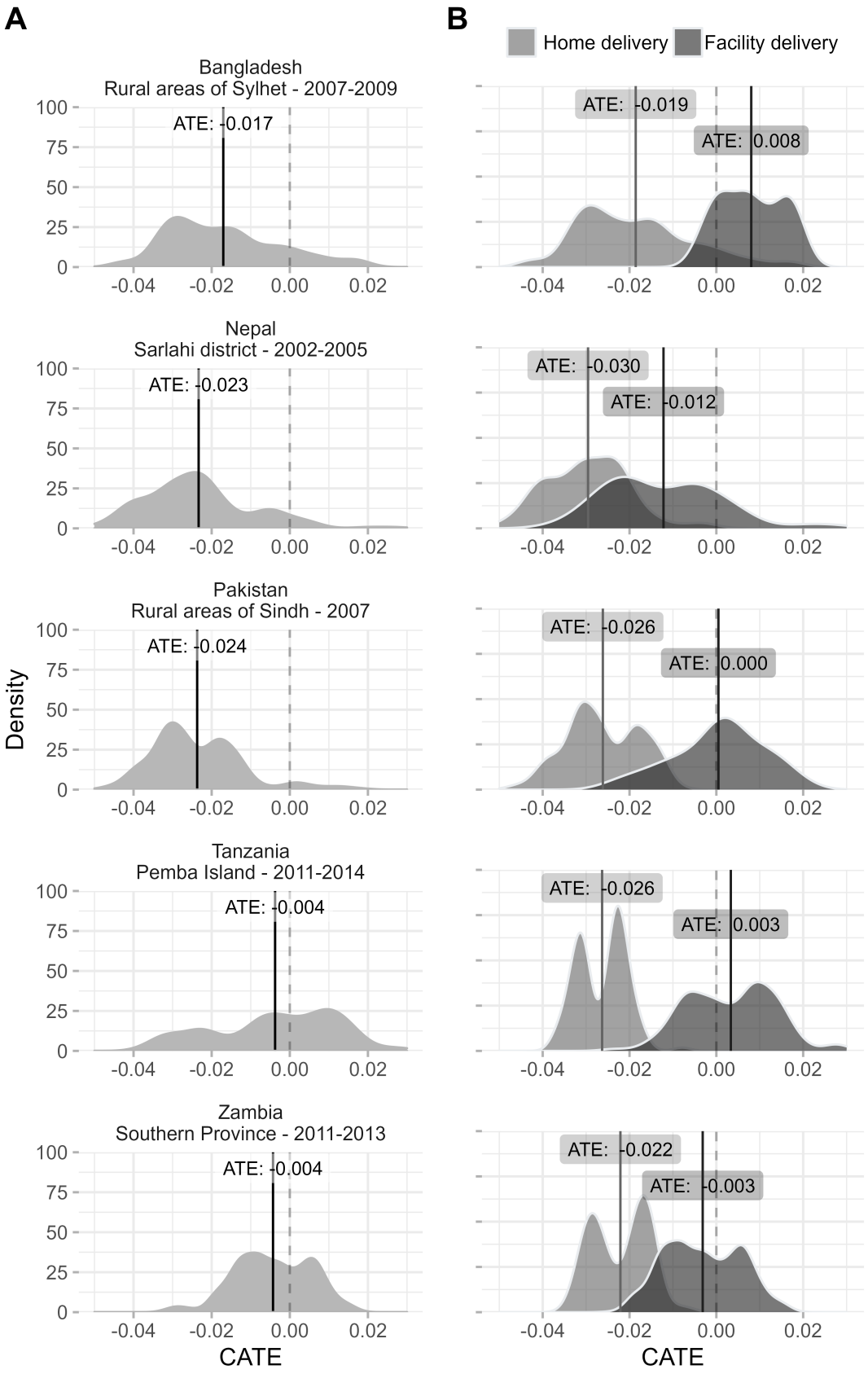


Figure 7: Distribution of predicted CATEs across DHS samples matching the RCT sites

Notes: The distributions are estimated using bandwidth selected based on Silverman's 'rule of thumb' (Silverman, 1986) and a gaussian kernel.

References

AL-UBAYDLI, O., J. A. LIST, AND D. L. SUSKIND (2017): "What can we learn from experiments? Understanding the threats to the scalability of experimental results," *American Economic Review*, 107, 282–86.

ARULAMPALAM, W., V. CORRADI, AND D. GUTKNECHT (2017): "Modeling Heaped Duration Data: An Application to Neonatal Mortality," *Journal of Econometrics*, 200, 363–377.

ATHEY, S., R. CHETTY, AND G. IMBENS (2020): "Combining experimental and observational data to estimate treatment effects on long term outcomes," *arXiv preprint arXiv:2006.09676*.

ATHEY, S. AND G. IMBENS (2016): "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360.

ATHEY, S., J. TIBSHIRANI, S. WAGER, ET AL. (2019): "Generalized random forests," *Annals of Statistics*, 47, 1148–1178.

ATHEY, S. AND S. WAGER (2021): "Policy learning with observational data," *Econometrica*, 89, 133–161.

BARHAM, T. (2011): "A healthier start: The effect of conditional cash transfers on neonatal and infant mortality in rural Mexico," *Journal of Development Economics*, 94, 74–85.

BAUERNSCHUSTER, S., A. DRIVA, AND E. HORNUNG (2017): "Bismarck's Health Insurance and the Mortality Decline," *Journal of the European Economic Association*.

BECKETT, M., J. DA VANZO, N. SASTRY, C. PANIS, AND C. PETERSON (2001): "The Quality of Retrospective Data: An Examination of Long-Term Recall in a Developing Country," *Journal of Human Resources*, 593–625.

BENNETT, J., J. MACIA, H. TRAVERSO, S. BANOAGHA, C. MALOOLY, AND J. BORING (1997): "Protective effects of topical antimicrobials against neonatal tetanus." *International journal of epidemiology*, 26, 897–903.

BHALOTRA, S. R., R. ROCHA, AND R. R. SOARES (2019): "Does Universalization of Healthwork? Evidence from Health Systems Restructuring and Expansion in Brazil," *IZA Discussion Paper #12111*.

BHATTARAI, M. (2017): "Nepal: Community Action for Nutrition Project (Sunaula Hazar Din) : P125359 - Implementation Status Results Report : Sequence 13 (English)," Tech. rep., Washington, D.C. : World Bank Group.

BHUTTA, Z. A., J. K. DAS, R. BAHL, J. E. LAWN, R. A. SALAM, V. K. PAUL, M. J. SANKAR, H. BLENCOWE, A. RIZVI, V. B. CHOU, ET AL. (2014): "Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost?" *The Lancet*, 384, 347–370.

BORUSYAK, K., X. JARAVEL, AND J. SPIESS (2021): "Revisiting Event Study Designs: Robust and efficient estimation," *arXiv preprint arXiv:2108.12419*.

CALLAGHAN-KORU, J. A., M. KHAN, M. ISLAM, A. SOWE, J. ISLAM, S. M. BILLAH, I. I. MANNAN, J. GEORGE, B. C. S. U. S. GROUP, ET AL. (2019): "Implementation Outcomes of the National Scale Up of Chlorhexidine Cord Cleansing in Bangladesh's Public Health System," *Journal of Global Health*, 9.

CHERNOZHUKOV, V., M. DEMIRER, E. DUFLO, AND I. FERNÁNDEZ-VAL (2020): "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," .

COFFEY, P. S. AND S. C. BROWN (2017): "Umbilical cord-care practices in low-and middle-income countries: a systematic review," *BMC pregnancy and childbirth*, 17, 1–21.

CONTI, G. AND R. GINJA (forthcoming): "Health Insurance and Child Health: Evidence from Mexico," *Journal of Human Resources*.

CSO, MOH AND ICF (2014): *Zambia Demographic and Health Survey 2013-14 [Dataset]. ZMBR61FL.DTA*, Central Statistical Office (CSO) [Zambia], `https://dhsprogram.com/data/dataset/Zambia_Standard-DHS_2013.cfm?flag=0`. Last accessed June 11, 2021.

DAHABREH, I. J., S. E. ROBERTSON, J. A. STEINGRIMSSON, E. A. STUART, AND M. A. HERNAN (2020): "Extending inferences from a randomized trial to a new target population," *Statistics in medicine*, 39, 1999–2014.

DE CHAISEMARTIN, C. AND X. D'HAULTFOEUILLE (2020): "Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110, 2964–96.

DE CHAISEMARTIN, C. AND X. D'HAULTFOEUILLE (2022): "Two-way Fixed Effects and Differences-in-Differences Estimators with Several Treatments," Tech. rep., National Bureau of Economic Research.

DEPARTMENT OF HEALTH SERVICES (2015): *Annual Report 2014/15*, Kathmandu.

DEPARTMENT OF HEALTH SERVICES-MINISTRY OF HEALTH, G. O. N. (2009/10-2016/17): "Annual Reports Department of Health Services, various editions from 2066/67 (2009/10) to 2072/73 (2015/2016)," .

EL ARIFEEN, S., L. C. MULLANY, R. SHAH, I. MANNAN, S. M. RAHMAN, M. R. R. TALUKDER, N. BEGUM, A. AL-KABIR, G. L. DARMSTADT, M. SANTOSHAM, ET AL. (2012): "The effect of cord cleansing with chlorhexidine on neonatal mortality in rural Bangladesh: a community-based, cluster-randomised trial," *The Lancet*, 379, 1022–1028.

ERCHICK, D. J., J. B. LACKNER, L. C. MULLANY, N. N. BHANDARI, P. R. SHEDAIN, S. KHANAL, J. R. DHAKWA, AND J. KATZ (2022): "Causes and age of neonatal death and associations with maternal and newborn care characteristics in Nepal: a verbal autopsy study," *Archives of Public Health*, 80, 1–10.

FEDERAL MINISTRY OF HEALTH (2016): *National Strategy For Scale-Up Of Chlorhexidine in Nigeria*, Abuja.

FITZPATRICK, A. (2018): "The Price of Labor: Evaluating the Impact of Eliminating User Fees on Maternal and Infant Health Outcomes," in *AEA Papers and Proceedings*, vol. 108, 412–15.

FOTTRELL, E., D. OSRIN, G. ALCOCK, K. AZAD, U. BAPAT, J. BEARD, A. BONDO, T. COLBOURN, S. DAS, C. KING, ET AL. (2015): "Cause-specific neonatal mortality: analysis of 3772 neonatal deaths in Nepal, Bangladesh, Malawi and India," *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 100, F439–F447.

FRIBERG, I. K., M. V. KINNEY, J. E. LAWN, K. J. KERBER, M. O. ODUBANJO, A.-M. BERGH, N. WALKER, E. WEISSMAN, M. CHOPRA, R. E. BLACK, ET AL. (2010): "Sub-Saharan Africa's Mothers, Newborns, and Children: How Many Lives Could Be Saved with Targeted Health Interventions?" *PLoS Med*, 7, e1000295.

GECHTER, M. AND R. MEAGER (2022): "Combining Experimental and Observational Studies in Meta-Analysis: A Debiasing Approach," .

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

HIMS (2014): *HIMS Database*, Health Information Management System, Nepal.

HODGINS, S., L. KHANAL, N. JOSHI, S. PENFOLD, S. TULADHAR, P. R. SHRESTHA, B. LAMICHHANE, P. DAWSON, T. GUENTHER, S. SINGH, ET AL. (2019): "Achieving and sustaining impact at scale for a newborn intervention in Nepal: a mixed-methods study," *Journal of Global Health Reports*, 3.

HODGINS, S., Y. PRADHAN, L. KHANAL, S. UPRETI, AND N. P. KC (2013): "Chlorhexidine for Umbilical Cord Care: Game-Changer for Newborn Survival?" *Global Health: Science and Practice*, 1, 5–10.

IMDAD, A., L. C. MULLANY, A. H. BAQUI, S. EL ARIFEEN, J. M. TIELSCH, S. K. KHATRY, R. SHAH, S. COUSENS, R. E. BLACK, AND Z. A. BHUTTA (2013): "The effect of umbilical cord cleansing with chlorhexidine on omphalitis and neonatal mortality in community settings in developing countries: a meta-analysis," *BMC public health*, 13, S15.

JAKIELA, P. (2021): "Simple Diagnostics for Two-way Fixed Effects," *arXiv preprint arXiv:2103.13229*.

JOHANNEMANN, J., V. HADAD, S. ATHEY, AND S. WAGER (2019): "Sufficient representations for categorical variables," *arXiv preprint arXiv:1908.09874*.

JSI (2017): *Monitoring and Evaluation of the Chlorhexidine "Navi" Care Program Technical Brief #1*.

JSI RESEARCH & TRAINING INSTITUTE (2017): *Use of Chlorhexidine for Cord Care, Social Change Communication, Experience from Nepal*, Kathmandu.

KHANAL, L. (2015): "Institutionalizing Chlorhexidine Program and Maintaining Coverage Chlorhexidine Cord Care Program in Nepal," https://www.healthynewbornnetwork.org/hnn-content/uploads/Institutionalizing-Chlorhexidine-Program-and-Maintaining-Coverage-in-Nepal.pdf.

KHANAL, S., P. DAWSON, R. HOUSTON, ET AL. (2011): "Verbal autopsy to ascertain causes of neonatal deaths in a community setting: a study from Morang, Nepal." *Journal of the Nepal Medical Association*, 51.

KOWALSKI, A. (2019): "Counting Defiers: Examples from Health Care," .

——— (forthcoming): "Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform," *Review of Economics and Statistics*.

KOWALSKI, A. E. (2021): "Mammograms and Mortality: How Has the Evidence Evolved?" *Journal of Economic Perspectives*, 35, 119–40.

LIM, S. S., L. DANDONA, J. A. HOISINGTON, S. L. JAMES, M. C. HOGAN, AND E. GAKIDOU (2010): "India's Janani Suraksha Yojana, a conditional cash transfer programme to increase births in health facilities: an impact evaluation," *The Lancet*, 375, 2009 – 2023.

LIU, L., S. OZA, D. HOGAN, Y. CHU, J. PERIN, J. ZHU, J. E. LAWN, S. COUSENS, C. MATHERS, AND R. E. BLACK (2016): "Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals," *The Lancet*, 388, 3027–3035.

LÓPEZ-MEDINA, M. D., M. LINARES-ABAD, A. B. LÓPEZ-ARAQUE, AND I. M. LÓPEZ-MEDINA (2019): "Dry care versus chlorhexidine cord care for prevention of omphalitis. Systematic review with meta-analysis," *Revista latino-americana de enfermagem*, 27.

MCKINNON, B., S. HARPER, J. S. KAUFMAN, AND Y. BERGEVIN (2015): "Removing user fees for facility-based delivery services: a difference-in-differences evaluation from ten sub-Saharan African countries," *Health policy and planning*, 30, 432–441.

MINISTRY OF HEALTH [NEPAL] AND NEW ERA AND ICF (2017): *Nepal Demographic and Health Survey 2016*, Ministry of Health [Nepal].

MOH, MOHCDGEC, NBS, OCGS AND ICF (2016): *Tanzania Demographic and Health Survey and Malaria Indicator Survey (TDHS-MIS) 2015-16 [Dataset]. TZBR7BFL.DTA*, Ministry of Health, Community Development, Gender, Elderly and Children (MoHCDGEC) [Tanzania Mainland], https://dhsprogram.com/data/dataset/Tanzania_Standard-DHS_2015.cfm?flag=0. Last accessed June 11, 2021.

MOH, NEW ERA AND ICF (2017): *Nepal Demographic and Health Survey 2016 [Dataset]. NPBR7HFL.DTA*, Ministry of Health [Nepal], https://dhsprogram.com/data/dataset/Nepal_Standard-DHS_2016.cfm?flag=0. Last accessed July 11, 2018.

MOH, NEW ERA AND ORC MACRO. (2002): *Nepal Demographic and Health Survey 2001 [Dataset]. NPKR41FL.DTA*, Family Health Division, Ministry of Health/Nepal, New ERA/Nepal, and ORC Macro. [NEPAL], https://dhsprogram.com/data/dataset/Nepal_Standard-DHS_2001.cfm?flag=0.Last accessed July 5, 2022.

MOHP, NEW ERA AND ICF (2011): *Nepal Demographic and Health Survey 2011 [Dataset]. NPKR61FL.DTA*, Ministry of Health and Population - MOHP/Nepal, New ERA/Nepal, and ICF International. [Nepal], https://dhsprogram.com/data/dataset/Nepal_Standard-DHS_2011.cfm?flag=0. Last accessed July 5, 2022.

MOHP, NEW ERA AND MACRO INTERNATIONAL (2007): *Nepal Demographic and Health Survey 2006 [Dataset]. NPKR51FL.DTA*, MOHP/Nepal, New ERA/Nepal, and Macro International. [Nepal], https://dhsprogram.com/data/dataset/Nepal_Standard-DHS_2006.cfm?flag=1. Last accessed July 5, 2022.

MULLANY, L. C., G. L. DARMSTADT, S. K. KHATRY, J. KATZ, S. C. LECLERQ, S. SHRESTHA, R. ADHIKARI, AND J. M. TIELSCH (2006): "Topical applications of chlorhexidine to the umbilical cord for prevention of omphalitis and neonatal mortality in southern Nepal: a community-based, cluster-randomised trial," *The Lancet*, 367, 910–918.

NEWHOUSE, J. P. (2021): "An Ounce of Prevention," *Journal of Economic Perspectives*, 35, 101–18.

NIPORT, MITRA AND ASSOCIATES AND ICF (2013): *Bangladesh Demographic and Health Survey [Dataset]. BDBR61FL.DTA*, National Institute of Population Research and Training [Bangladesh], https://dhsprogram.com/data/dataset/Bangladesh_Standard-DHS_2011.cfm?flag=0. Last accessed July 5, 2022.

NIPS AND ICF (2013): *Pakistan Demographic and Health Survey 2012-13 [Dataset]. PKBR61FL.DTA*, National Institute of Population Studies (NIPS) [Pakistan], https://dhsprogram.com/data/dataset/Pakistan_Standard-DHS_2012.cfm?flag=0. Last accessed June 11, 2021.

OSRIN, D. AND T. COLBOURN (2016): "No reason to change WHO guidelines on cleansing the umbilical cord," *The Lancet Global Health*, 4, e766–e768.

PAINTER, K. AND J. FELDMAN (2019): *Omphalitis*, https://www.ncbi.nlm.nih.gov/books/NBK513338/.

PAUDEL, D., I. B. SHRESTHA, M. SIEBECK, AND E. REHFUESS (2017): "Impact of the Community-Based Newborn Care Package in Nepal: a Quasi-Experimental Evaluation," *BMJ open*, 7, e015285.

PHILIBERT, A., M. RAVIT, V. RIDDE, I. DOSSA, E. BONNET, F. BEDECARRATS, AND A. DUMONT (2017): "Maternal and neonatal health impact of obstetrical risk insurance scheme in Mauritania: a quasi experimental before-and-after study," *Health policy and planning*, 32, 405–417.

PONCE HARDY, V. (2018): "Chlorhexidine for neonatal infection: the 'game-changer' we've all been waiting for?" *Tropical Medicine & International Health*, 23, 252–253.

POPOVICIU, T. (1935): "Sur les Équations Algébriques Ayant Toutes Leurs Racines Réelles," *Mathematica*, 9, 20.

POWELL-JACKSON, T., S. MAZUMDAR, AND A. MILLS (2015): "Financial incentives in health: New evidence from India's Janani Suraksha Yojana," *Journal of health economics*, 43, 154–169.

PRADHAN, AJIT, R. H. A. G. R. B. B. AND P. GOVINDASAMY (1997): *Nepal Family Health Survey 1995 [Dataset]. NPKR31FL.DTA*, Ministry of Health/Nepal, New ERA/Nepal, and Macro International.[NEPAL], https://dhsprogram.com/data/dataset/Nepal_Standard-DHS_1996.cfm?flag=1.Last accessed July 5, 2022.

REQUEJO, J. H., J. BRYCE, A. J. BARROS, P. BERMAN, Z. BHUTTA, M. CHOPRA, B. DAELMANS, A. DE FRANCISCO, J. LAWN, B. MALIQI, ET AL. (2015): "Countdown to 2015 and Beyond: Fulfilling the Health Agenda for Women and Children," *The Lancet*, 385, 466–476.

SANKAR, M., A. CHANDRASEKARAN, A. RAVINDRANATH, R. AGARWAL, AND V. PAUL (2016): "Umbilical cord cleansing with chlorhexidine in neonates: a systematic review," *Journal of Perinatology*, 36, S12–S20.

SAYA, A. R., J. KATZ, S. K. KHATRY, J. M. TIELSCH, S. C. LECLERQ, AND L. C. MULLANY (2022): "Causes of neonatal mortality using verbal autopsies in rural Southern Nepal, 2010–2017," *PLOS Global Public Health*, 2, e0001072.

SAZAWAL, S., U. DHINGRA, S. M. ALI, A. DUTTA, S. DEB, S. M. AME, M. H. MKASHA, A. YADAV, AND R. E. BLACK (2016): "Efficacy of chlorhexidine application to umbilical cord on neonatal mortality in Pemba, Tanzania: a community-based randomised controlled trial," *The Lancet Global Health*, 4, e837–e844.

SEMRAU, K. E., J. HERLIHY, C. GROGAN, K. MUSOKOTWANE, K. YEBOAH-ANTWI, R. MBEWE, B. BANDA, C. MPAMBA, F. HAMOMBA, P. PILINGANA, ET AL. (2016): "Effectiveness of 4% chlorhexidine umbilical cord care on neonatal mortality in Southern Province, Zambia (ZamCAT): a cluster-randomised controlled trial," *The Lancet Global Health*, 4, e827–e836.

SILVERMAN, B. (1986): "DensityEstimation Chapman and Hall," .

SOOFI, S., S. COUSENS, A. IMDAD, N. BHUTTO, N. ALI, AND Z. A. BHUTTA (2012): "Topical application of chlorhexidine to neonatal umbilical cords for prevention of omphalitis and neonatal mortality in a rural district of Pakistan: a community-based, cluster-randomised trial," *The Lancet*, 379, 1029–1036.

SVERDRUP, E., A. KANODIA, Z. ZHOU, S. ATHEY, AND S. WAGER (2020): "policytree: Policy learning via doubly robust empirical welfare maximization over trees," *Journal of Open Source Software*, 5, 2232.

TIBSHIRANI, J., S. ATHEY, R. FRIEDBERG, V. HADAD, D. HIRSHBERG, L. MINER, E. SVERDRUP, S. WAGER, AND M. WRIGHT (2021): *grf: Generalized Random Forests*, r package version 2.0.2.

TIBSHIRANI, J., S. ATHEY, E. SVERDRUP, AND S. WAGER (2022): *Estimating ATEs on a new target population*.

USAID (2017): "Chlorhexidine "Navi" (Cord) Program [Factsheet], Scaling-Up the Use of Chlorhexidine for Umbilical Cord Care: Nepal's Experience," Tech. rep., USAID.

VAN DE POEL, E., G. FLORES, P. IR, AND O. O'DONNELL (2016): "Impact of performance-based financing in a low-resource setting: a decade of experience in Cambodia," *Health Economics*, 25, 688–705.

VARIAN, H. R. (2014): "Big data: New tricks for econometrics," *Journal of Economic Perspectives*, 28, 3–28.

VAUGHAN, K., A. OZALTIN, M. MALLOW, F. MOI, C. WILKASON, J. STONE, AND L. BRENZEL (2019): "The Costs of Delivering Vaccines in Low-and Middle-Income Countries: Findings from a systematic review," *Vaccine: X*, 2, 100034.

WANG, H., Z. A. BHUTTA, M. M. COATES, M. COGGESHALL, L. DANDONA, K. DIALLO, E. B. FRANCA, M. FRASER, N. FULLMAN, P. W. GETHING, ET AL. (2016): "Global, regional, national, and selected subnational levels of stillbirths, neonatal, infant, and under-5 mortality, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015," *The Lancet*, 388, 1725–1774.

WORLD HEALTH ORGANIZATION (2015): "Postnatal care for mothers and newborns: Highlights from the World Health Organization 2013 Guidelines," *Available from: http://www.who.int/maternal_child_adolescent/publications/WHOMCA-PNC-2014-Briefer_A*, 4.

——— (2020): "Advice on the Use of Masks in the Context of COVID-19: Interim Guidance (5 June 2020)," Tech. rep.

——— (2022): *WHO recommendations on maternal and newborn care for a positive postnatal experience*, Geneva.

A Appendix (for Online Publication Only)

Table A.1: No effect of CHX-NCP on antenatal care and place of delivery

	Dependent variable:			
	Place of delivery			
	Actual	Predicted	ANC visits	ANC visits>p50
CHX	-0.020	-0.020	0.001	-0.012
	(0.027)	(0.018)	(0.141)	(0.030)
Observations	4,955	23,581	3,966	3,966
Clusters	73	73	73	73
MDV	0.411	0.554	4.362	0.415

Notes: All specifications are estimated as linear probability models using OLS with month of birth and district fixed effects. Standard errors clustered at the district level in parentheses. MDV is the mean of dependent variable among untreated individuals. Asterisks indicate significance at the following levels: * p<0.1, ** p<0.05, and *** p<0.01.

Table A.2: Balancing table. Dependent variable: CHX.

	(1)
Ethnicity: hill chhetri	-0.005
	(0.006)
Ethnicity: terai brahmin/chhetri	-0.003
	(0.010)
Ethnicity: other terai caste	-0.006
	(0.008)
Ethnicity: hill dalit	0.006
	(0.006)
Ethnicity: terai dalit	0.006
	(0.010)
Ethnicity: newar	0.007
	(0.010)
Ethnicity: hill janajati	0.000
	(0.006)
Ethnicity: terai janajati	0.001
	(0.007)
Ethnicity: muslim	-0.008
	(0.007)
Ethnicity: other	0.052**
	(0.025)
Rural	-0.007*
	(0.004)
Altitude in 1st quintile	0.013
	(0.011)
Altitude in 2nd quintile	0.003
	(0.009)
Altitude in 3rd quintile	-0.007
	(0.007)
Altitude in 4th quintile	-0.000
	(0.007)
Education: no education	0.016
	(0.011)
Education: primary	0.021*
	(0.012)
Education: secondary	0.014
	(0.010)
Wealth 0-20%	-0.005
	(0.006)
Wealth 20-40%	0.002
	(0.006)
Wealth 40-60%	-0.003
	(0.005)
Wealth 60-80%	-0.004
	(0.006)
Female	0.000
	(0.002)
First born	0.001
	(0.004)
Second born	-0.003
	(0.004)
Third born	-0.006
	(0.004)

Continued on next page

Continued from previous page

	(1)
Mother age 15-19y	-0.026
	(0.027)
Mother age 20-24y	-0.027
	(0.027)
Mother age 25-29y	-0.026
	(0.027)
Mother age 30-34y	-0.028
	(0.028)
Mother age 35-39y	-0.026
	(0.029)
Program: CB-NCP	0.373***
	(0.069)
Program: CB-IMNCI	-0.063
	(0.082)
Constant	0.095***
	(0.033)

Notes: The specifications are estimated with district and month of birth fixed effects. Standard errors clustered at the district level in parentheses. Asterisks indicate significance at the following levels * p<0.1, ** p<0.05, and *** p<0.01.

Table A.3: Effect of CHX-NCP on neonatal mortality - All coefficients

	Dependent variable: mortality						
	month ∈ [0,1]					month ∈]1,12]	month ∈ [0,12]
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CHX	-0.014**	-0.018***			-0.020**	0.001	-0.019**
	(0.006)	(0.006)			(0.009)	(0.004)	(0.008)
Female		-0.014***		-0.016***	-0.020***	-0.001	-0.016***
		(0.003)		(0.003)	(0.004)	(0.002)	(0.004)
First born		0.003		0.002	0.090***	-0.001	0.002
		(0.006)		(0.006)	(0.015)	(0.003)	(0.006)
Second born		-0.007		-0.009	0.060***	0.001	-0.006
		(0.005)		(0.005)	(0.011)	(0.003)	(0.005)
Third born		-0.008		-0.009	0.032***	0.002	-0.006
		(0.005)		(0.005)	(0.007)	(0.003)	(0.005)
Mother age 15-19y		0.019		0.020	-0.038	0.019***	0.038*
		(0.020)		(0.026)	(0.032)	(0.004)	(0.022)
Mother age 20-24y		-0.003		-0.004	-0.034	0.011***	0.007
		(0.019)		(0.026)	(0.031)	(0.003)	(0.021)
Mother age 25-29y		-0.011		-0.013	-0.019	0.012***	-0.002
		(0.019)		(0.025)	(0.030)	(0.003)	(0.021)
Mother age 30-34y		-0.008		-0.010	0.004	0.009***	0.000
		(0.020)		(0.025)	(0.028)	(0.002)	(0.021)
Mother age 35-39y		-0.006		-0.009	0.009	0.009**	-0.001
		(0.020)		(0.025)	(0.025)	(0.004)	(0.023)
Ethnicity: hill chhetri		-0.005		-0.005		0.003	-0.002
		(0.005)		(0.005)		(0.003)	(0.004)
Ethnicity: terai brahmin/chhetri		0.002		-0.013		0.001	0.001
		(0.015)		(0.011)		(0.005)	(0.018)
Ethnicity: other terai caste		0.002		0.003		0.006	0.009
		(0.008)		(0.009)		(0.004)	(0.007)
Ethnicity: hill dalit		-0.004		-0.003		0.001	-0.003
		(0.006)		(0.006)		(0.004)	(0.007)
Ethnicity: terai dalit		0.019		0.023*		0.011**	0.031**
		(0.011)		(0.013)		(0.005)	(0.013)
Ethnicity: newar		0.003		0.003		0.007	0.009
		(0.008)		(0.009)		(0.006)	(0.010)
Ethnicity: hill janajati		-0.006		-0.005		0.004	-0.002
		(0.006)		(0.006)		(0.002)	(0.006)
Ethnicity: terai janajati		0.007		0.008		0.009**	0.018**
		(0.007)		(0.008)		(0.004)	(0.009)
Ethnicity: muslim		-0.006		-0.008		0.011***	0.005
		(0.011)		(0.012)		(0.003)	(0.012)
Ethnicity: other		0.002		0.003		-0.003	-0.000

Continued on next page

Continued from previous page

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		(0.020)		(0.023)		(0.004)	(0.022)
Rural		0.003		0.003		0.003**	0.007*
		(0.003)		(0.004)		(0.001)	(0.004)
Altitude in 1st quintile		-0.014		-0.018		0.020***	0.005
		(0.012)		(0.014)		(0.007)	(0.016)
Altitude in 2nd quintile		-0.026**		-0.031**		0.018***	-0.009
		(0.011)		(0.013)		(0.006)	(0.014)
Altitude in 3rd quintile		-0.015*		-0.017**		0.013**	-0.001
		(0.008)		(0.008)		(0.005)	(0.009)
Altitude in 4th quintile		-0.007		-0.008		0.002	-0.004
		(0.007)		(0.007)		(0.005)	(0.009)
Education: no education		0.018***		0.021***		0.002	0.020***
		(0.005)		(0.006)		(0.003)	(0.006)
Education: primary		0.008*		0.011**		0.001	0.010*
		(0.005)		(0.005)		(0.003)	(0.006)
Education: secondary		0.007*		0.008		0.001	0.009*
		(0.004)		(0.005)		(0.003)	(0.005)
Wealth 0-20%		0.017**		0.016**		0.013***	0.030***
		(0.007)		(0.008)		(0.003)	(0.008)
Wealth 20-40%		0.019***		0.018***		0.004	0.023***
		(0.005)		(0.006)		(0.003)	(0.006)
Wealth 40-60%		0.013***		0.013**		0.002	0.014**
		(0.005)		(0.005)		(0.003)	(0.005)
Wealth 60-80%		0.004		0.003		-0.000	0.003
		(0.004)		(0.004)		(0.003)	(0.005)
Program: CB-NCP		0.006		0.008	0.018*	0.003	0.010
		(0.006)		(0.008)	(0.009)	(0.003)	(0.008)
Program: CB-IMNCI		-0.003		-0.001	0.016	0.013**	0.012
		(0.006)		(0.008)	(0.014)	(0.006)	(0.009)
CHX_{t-6}			0.002	-0.001			
			(0.011)	(0.011)			
Constant	0.042***	0.039*	0.042***	0.042	0.028	-0.021***	0.019
	(0.001)	(0.021)	(0.000)	(0.027)	(0.033)	(0.007)	(0.024)
Observations	23,465	23,465	20,321	20,321	21,209	22,571	22,571
Clusters	73	73	73	73	73	73	73
Control mean of dep. var	0.042	0.042	0.042	0.042	0.045	0.015	0.058
Sample	All	All	Pre	Pre	All	All	All
Controls	No	Yes	No	Yes	Yes	Yes	Yes
Month of birth FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	No	Yes	Yes
Mother FE	No	No	No	No	Yes	No	No

Notes: All specifications are estimated as linear probability models using OLS. Except for the mother FE specification, a "Yes" in the "Controls" row indicates that the regression also includes the full set of demographic, SES, and program controls. Demographic controls include birth order (three indicators), five year maternal age group indicators, and gender. SES controls include education (three indicators), wealth (four indicators), rural indicator, altitude quintile indicators, and ethnicity indicators. Program controls include controls for the CB-NCP and CB-IMNCI health programs. The mother FE specification excludes the SES controls as these are mother-invariant. MDV is the mean of dependent variable among untreated individuals. Standard errors clustered at the district level in parentheses. Asterisks indicate significance at the following levels: * $p<0.1$, ** $p<0.05$, and *** $p<0.01$.

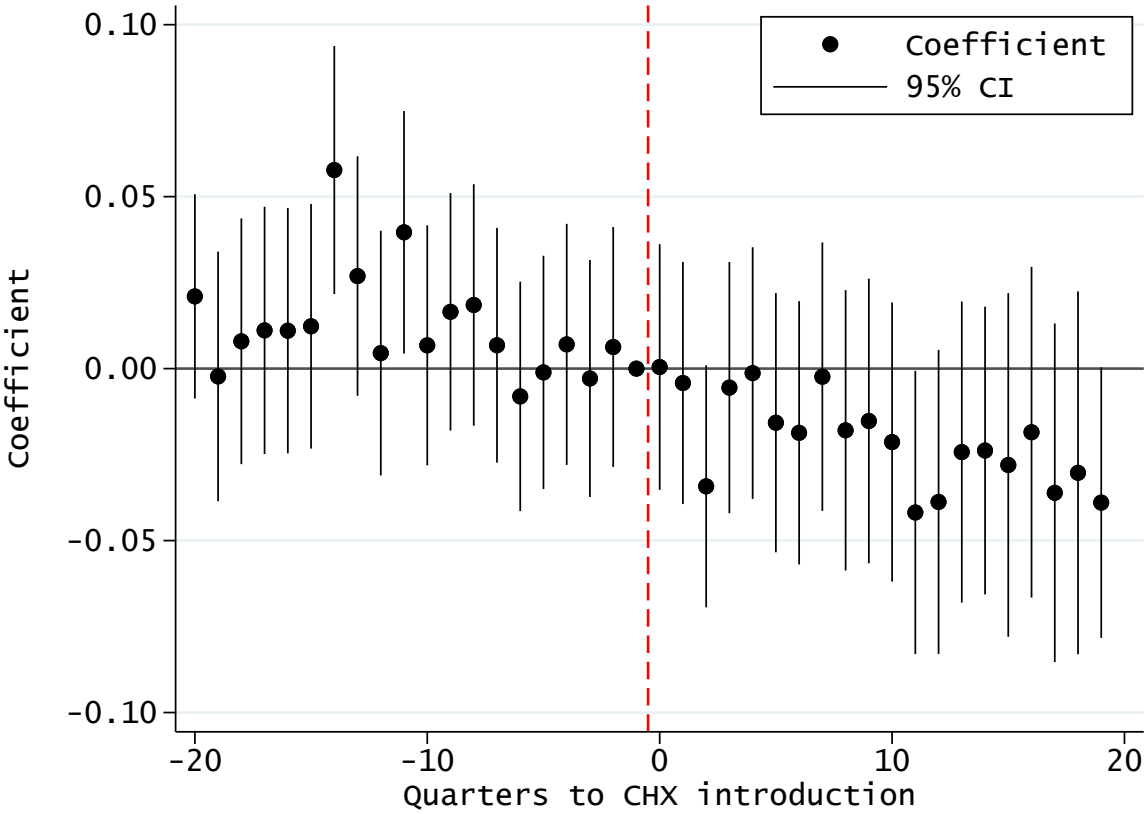


Figure A.1: Event study chart

Notes: The chart shows the estimates of a regression of an indicator variable for the child dying within one month of birth on the full set of month of birth and district fixed effects, the full set of control variables (see notes for Table 3) as well as quarter to the introduction of CHX-NCP program indicator variables. The specification is binned at 20 quarters to treatment start and 20 quarters after treatment start, such that the indicator for minus and plus 20 quarters is equal to one $\geq$ 20 quarters to treatment start and $\geq$ 20 quarters after treatment start, respectively. The chart shows the coefficients on the quarter to treatment indicators and associated 95% confidence intervals.

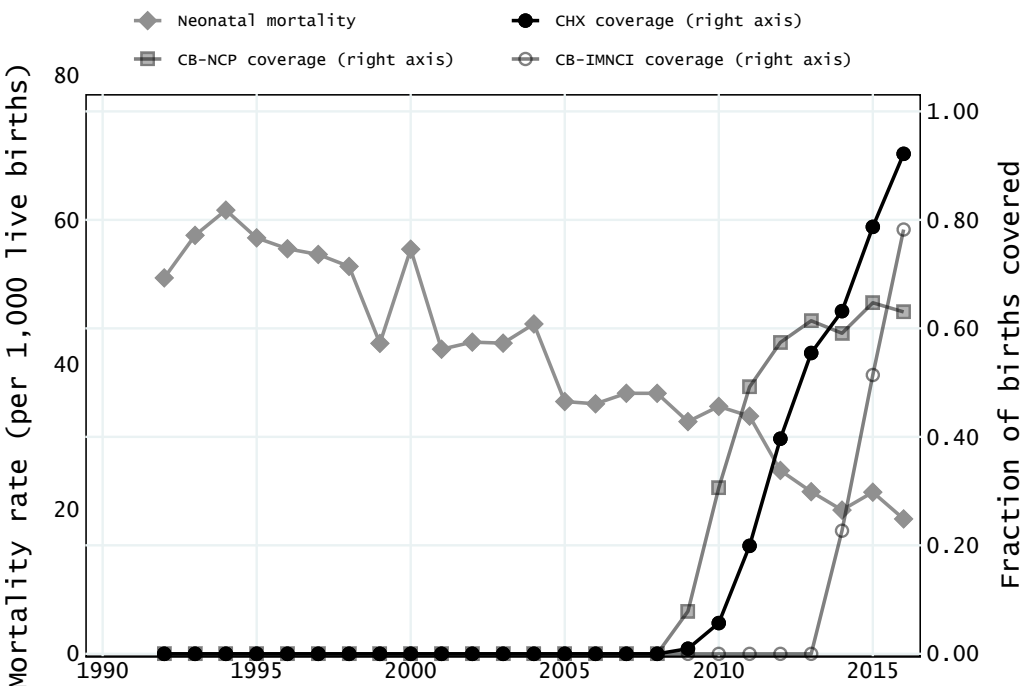
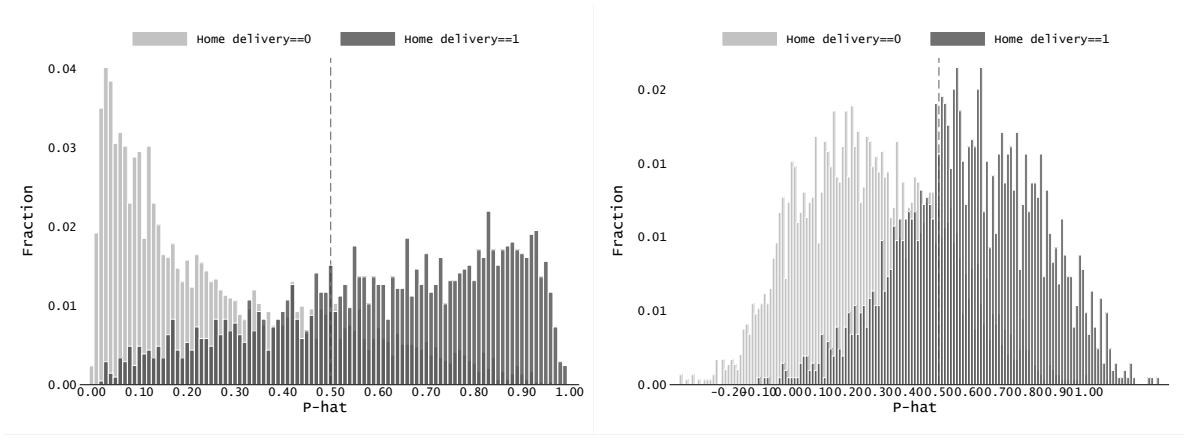


Figure A.2: Neonatal mortality, CB-NCP coverage, CB-IMNCI coverage, and CHX-NCP coverage

Table A.4: Predicting home deliveries

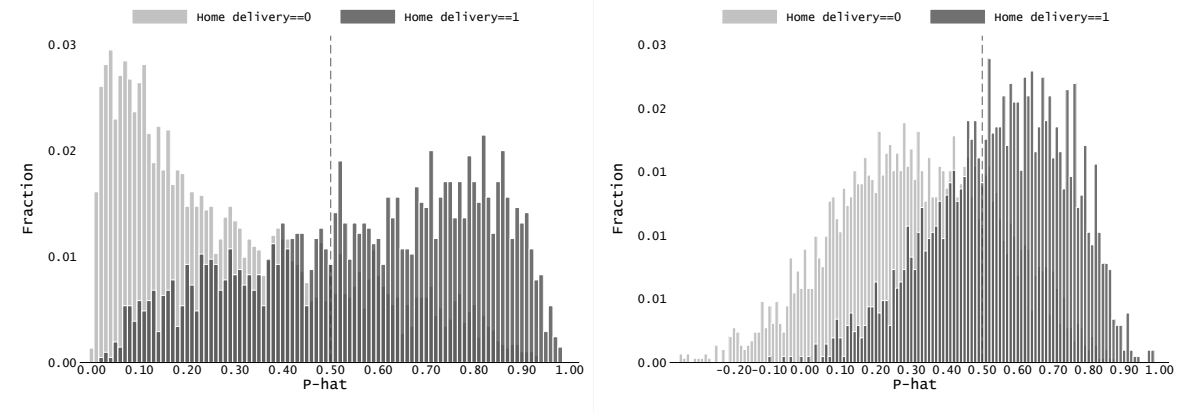
	DHS 2016		DHS 1996-2016	
	Logit	LPM	Logit	LPM
	(1)	(2)	(3)	(4)
Female	0.004	0.006	0.003	0.003
	(0.012)	(0.012)	(0.004)	(0.004)
First born	-0.248***	-0.267***	-0.184***	-0.203***
	(0.026)	(0.027)	(0.009)	(0.012)
Second born	-0.090***	-0.112***	-0.085***	-0.082***
	(0.019)	(0.022)	(0.007)	(0.008)
Third born	-0.022	-0.029	-0.044***	-0.036***
	(0.017)	(0.019)	(0.007)	(0.006)
Mother age 15-19y	0.069	0.067	0.130***	0.151***
	(0.062)	(0.065)	(0.023)	(0.019)
Mother age 20-24y	0.034	0.039	0.088***	0.094***
	(0.059)	(0.063)	(0.022)	(0.017)
Mother age 25-29y	0.015	0.018	0.045**	0.046***
	(0.057)	(0.062)	(0.021)	(0.015)
Mother age 30-34y	-0.048	-0.046	0.018	0.020
	(0.059)	(0.064)	(0.020)	(0.015)
Mother age 35-39y	-0.040	-0.048	0.004	0.005
	(0.060)	(0.064)	(0.024)	(0.017)
Rural	0.120***	0.130***	0.115***	0.171***
	(0.025)	(0.029)	(0.010)	(0.016)
Education: no education	0.146***	0.137***	0.255***	0.369***
	(0.030)	(0.027)	(0.013)	(0.017)
Education: primary	0.104***	0.082***	0.191***	0.302***
	(0.030)	(0.026)	(0.014)	(0.016)
Education: secondary	0.071***	0.035*	0.102***	0.142***
	(0.027)	(0.020)	(0.012)	(0.015)
Observations	4,911	4,956	26,975	26,986
Correct predictions (share)	0.759	0.761	0.726	0.719

Notes: Columns (1) and (2) are based on the recent births subsample of DHS 2016 (MoH, New ERA and ICF, 2017). In columns (3) and (4) the prediction is trained based on stacked recent births subsamples of DHS from 1996 to 2016 (Pradhan and Govindasamy, 1997; MoH, New ERA and ORC Macro., 2002; MoHP, New ERA and Macro International, 2007; MoHP, New ERA and ICF, 2011; MoH, New ERA and ICF, 2017). Columns (1) and (3) show average marginal effects from estimating a Logit specification. Columns (2) and (4) show point estimates from estimating a linear probability models. All regressions include district fixed effects and date of birth, defined by Nepali month and year of birth, fixed effects. Predictions are for the 2016 recent births subsample for comparability. Standard errors clustered at the district level in parentheses. Asterisks indicate significance at the following levels * $p<0.1$, ** $p<0.05$, and *** $p<0.01$.



(a) Logit - DHS 2016

(b) LPM - DHS 2016



(c) Logit - DHS 1996-2016

(d) LPM - DHS 1996-2016

Figure A.3: Estimated propensity score for the prediction of place of delivery.

Notes: See Table A.4 for estimation details.

Table A.5: Heterogeneity by place of birth as predicted based on stacked DHS samples

	Sample			
	All	All	P(home birth)	
			<0.5	>0.5
	(1)	(2)	(3)	(4)
CHX	-0.018***	-0.006	0.016*	-0.036***
	(0.006)	(0.006)	(0.008)	(0.010)
1[P(home birth)>0.5]		0.004		
		(0.006)		
CHX × 1[P(home birth)>0.5]		-0.024***		
		(0.007)		
CHX + CHX × 1[P(home birth)>0.5]		-0.030***		
		(0.007)		
Observations	23,465	23,465	5,978	17,479
Clusters	75	75	75	75
Control mean of dep. var	0.042	0.042	0.018	0.049
P-val (dif across sample)				0.000

Notes: All specifications are estimated as linear probability models using OLS with the full set of demographic, SES, and program controls. Demographic controls include birth order (three indicators), five year maternal age group indicators, and gender. SES controls include education (three indicators), wealth (four indicators), rural indicator, altitude quintile indicators, and ethnicity indicators. Program controls include controls for the CB-NCP and CB-IMNCI health programs. All specifications are estimated with district and month of birth fixed effects. We split the sample according to the predicted place of delivery, based on the linear probability model shown in Appendix Table A.4 Column (4). Bootstrapped standard errors based on 200 iterations and clustered at the district level in parentheses. Asterisks indicate significance at the following levels: * p<0.1, ** p<0.05, and *** p<0.01.

Table A.6: Testing for interactions across neonatal health programs - dependent variable: mortality month $\in$ [0,1]

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CHX	-0.016***	-0.016*	-0.024**	-0.018***	-0.017*	-0.024**	-0.029***	
	(0.006)	(0.008)	(0.010)	(0.006)	(0.009)	(0.010)	(0.010)	
CB-NCP	0.005	0.007	0.007	0.006	0.008	0.008		0.008
	(0.006)	(0.008)	(0.008)	(0.006)	(0.008)	(0.008)		(0.008)
CB-IMNCI	-0.004	-0.011	-0.011	-0.003	-0.012	-0.012		
	(0.006)	(0.010)	(0.012)	(0.006)	(0.010)	(0.011)		
CB-NCP $\times$ CHX		-0.004	0.005		-0.006	0.003		
		(0.010)	(0.011)		(0.010)	(0.011)		
CB-IMNCI $\times$ CHX		0.010	0.025		0.012	0.025		
		(0.010)	(0.017)		(0.010)	(0.017)		
CB-IMNCI $\times$ NCP			0.001			0.002		
			(0.012)			(0.013)		
CB-IMNCI $\times$ CB-NCP $\times$ CHX			-0.022			-0.020		
			(0.019)			(0.020)		
Observations	23,465	23,465	23,465	23,465	23,465	23,465	19,225	20,017
Clusters	73	73	73	73	73	73	73	73
MDV	0.042	0.042	0.042	0.042	0.042	0.042	0.044	0.043
Demogr. & SES Controls	No	No	No	Yes	Yes	Yes	Yes	Yes

Notes: Controls include birth order (three indicators), five year maternal age group indicators, gender, education (three indicators), wealth (four indicators), rural indicator, altitude quintile indicators, and ethnicity indicators. All specifications are estimated with district and month of birth fixed effects. Standard errors clustered at the district level in parentheses. MDV is the mean of the dependent variable among untreated individuals. Restricted sample to observations with CB-NCP=0 and CB-IMNCI=0 in Column (7) and to CHX=0 and CB-IMNCI=0 in Column (8). Asterisks indicate significance at the following levels: * $p<0.1$, ** $p<0.05$, and *** $p<0.01$.

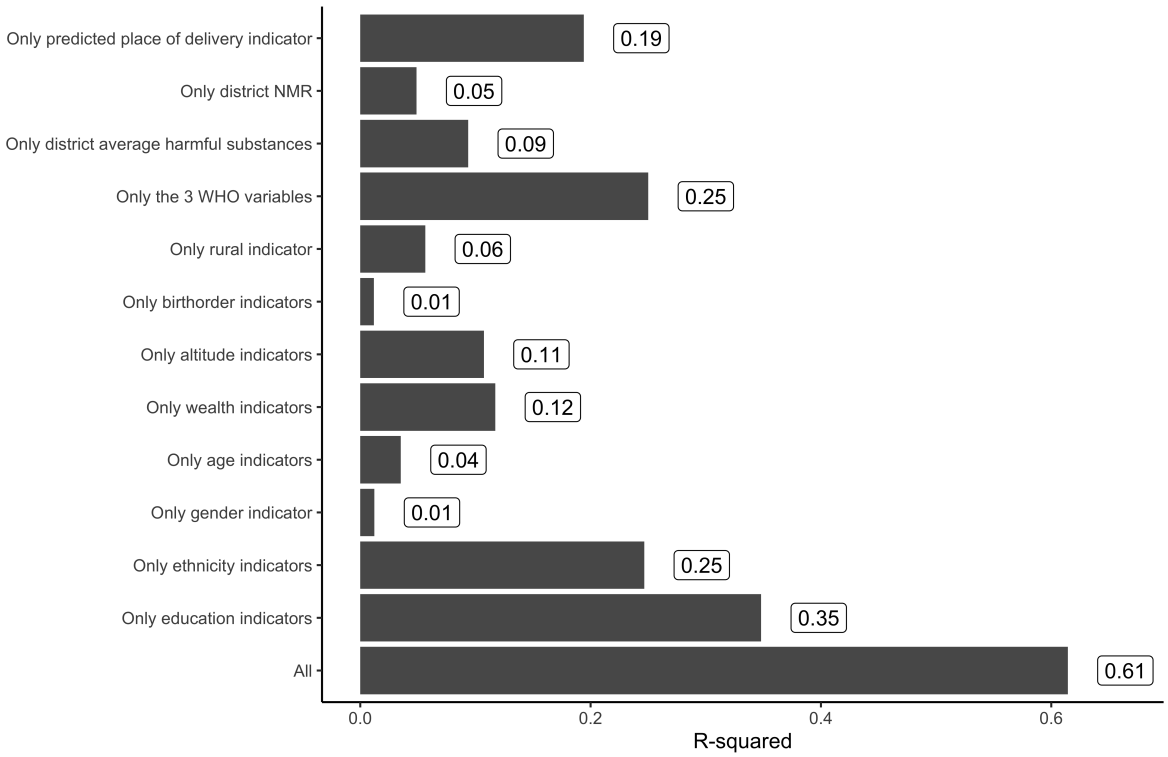


Figure A.4: Contribution to CATE variation

Notes: This Figure shows the R-squared from estimating an ordinary least squares regression of the CATE on the covariates listed on the vertical axis. The "3 WHO variables" are: predicted place of delivery indicator, district NMR and district average harmful substances application.

B Other Specification Checks

We also estimated a number of alternative specifications for Equation (1) and found no notable difference in estimates. In these alternative specifications, we removed all controls other than district and time effects, varied the subsets of controls included, added controls for additional health and nutritional programs, *in-utero* exposure to the severe earthquake which took place in 2015, controlled for an interaction term between baseline district neonatal mortality and a linear trend in month-year date of birth, and varied the sample in two ways: (i) changing the time period covered by the data — adding and removing five year cohorts to our baseline 25-year panel — and (ii) removing or not children for whom the district of birth cannot be established with certainty because their mothers were currently visiting the household surveyed or because the woman had moved to the district where she was interviewed after the CHX program was first introduced in the country. As depicted in Figure B.1, the estimated treatment effect is consistently between -0.010 and -0.019 across specifications and its associated p-value only goes slightly over 0.10 in some specifications using the smallest of the six samples we consider.

Our findings are also robust to adopting an alternative definition of neonatal mortality which is equal to zero for children reported to have died at exactly one month old and which are counted as having died within the neonatal period in the main analysis to allow for heaping (Appendix Table B.1). Weighted least squares estimates using the sampling weights provided by the DHS also lead to the same conclusions (Appendix Table B.2).

Finally, we fitted logit models to reflect the binary nature of our dependent variable of interest. The estimated treatment effects on neonatal mortality are, again, similar to our main specification and statistically significant (Appendix Table B.3), while the effect on infant mortality is almost identical in magnitude but slightly less precisely estimated.

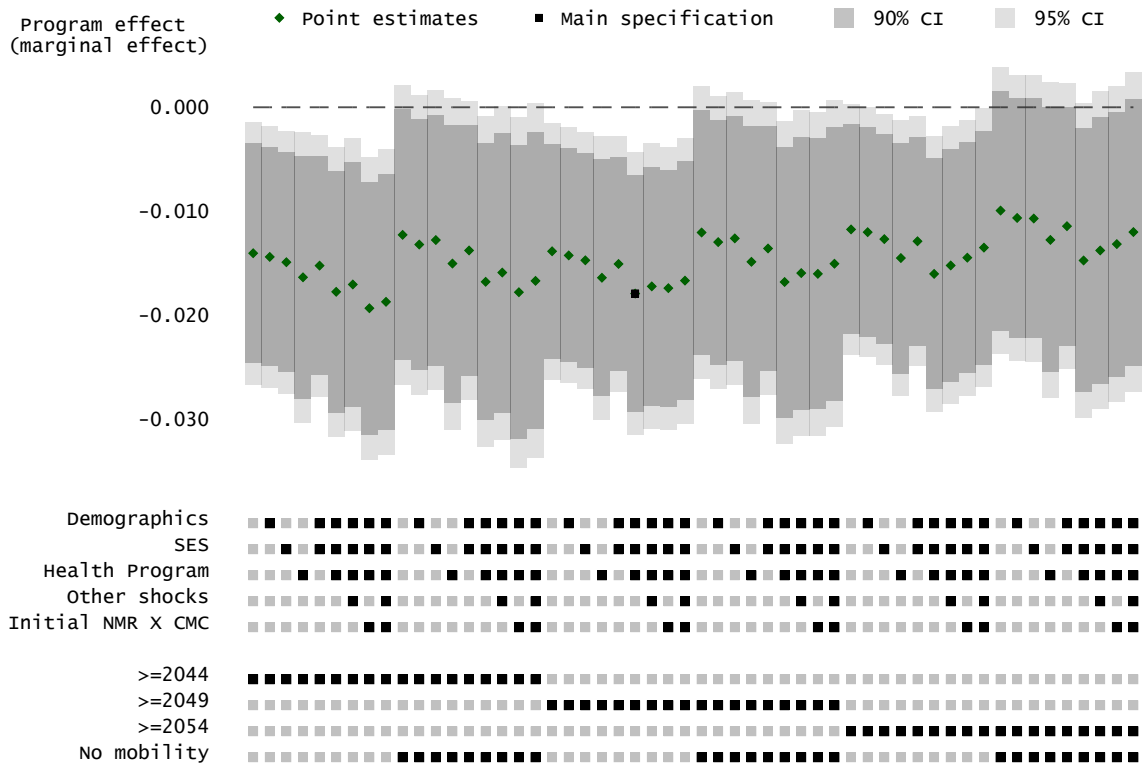


Figure B.1: Specification curve

Notes: This chart shows estimates from running 54 different specifications defined by the combination of markers bellow the chart. The red point indicates out main specification. Demographic controls include birth order (three indicators), five year maternal age group indicators, and gender. SES controls include education (three indicators), wealth (four indicators), rural indicator, altitude quintile indicators, and ethnicity indicators. Program controls include controls for the CB-NCP and CB-IMNCI health programs. Other shocks refer to the earthquake on 25 April 2015, the Community Action for Nutrition Project (Sunaula), an Integrated Nutrition Program (Suaahara), and the Safe Delivery Incentive Program. Initial NMR $\times$ CMC is the initial neonatal mortality times a quadratic time trend. The confidence intervals are based on standard errors clustered at the district level.

Table B.1: Effect of CHX-NCP on neonatal mortality excluding heaping at one month

	Dependent variable: mortality				
	month ∈ [0,1[				
	(1)	(2)	(3)	(4)	(5)
CHX	-0.011**	-0.014**			-0.014*
	(0.005)	(0.006)			(0.008)
CHX_{t-6}			0.003	0.001	
			(0.010)	(0.011)	
Observations	23,552	23,552	20,324	20,324	21,293
Clusters	73	73	73	73	73
Control mean of dep. var	0.037	0.037	0.037	0.037	0.039
Sample	All	All	Pre	Pre	All
Controls	No	Yes	No	Yes	Yes
Month of birth FE	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	No
Mother FE	No	No	No	No	Yes

Notes: All specifications are estimated as linear probability models using OLS. Except for the mother FE specification, a "Yes" in the "Controls" row indicates that the regression also includes the full set of demographic, SES, and program controls. Demographic controls include birth order (three indicators), five year maternal age group indicators, and gender. SES controls include education (three indicators), wealth (four indicators), rural indicator, altitude quintile indicators, and ethnicity indicators. Program controls include controls for the CB-NCP and CB-IMNCI health programs. The mother FE specification excludes the SES controls as these are mother-invariant. MDV is the mean of dependent variable among untreated individuals. Standard errors clustered at the district level in parentheses. Asterisks indicate significance at the following levels: * p<0.1, ** p<0.05, and *** p<0.01.

Table B.2: Effect of CHX-NCP on neonatal mortality using survey weights

	Dependent variable: mortality						
	month ∈ [0,1]					month ∈]1,12]	month ∈ [0,12]
	(1)	(2)	(3)	(4)	(5)		
CHX	-0.015***	-0.017***			-0.019**	0.001	-0.018**
	(0.005)	(0.006)			(0.009)	(0.005)	(0.008)
CHX_{t-6}			-0.003	-0.005			
			(0.012)	(0.012)			
Observations	23,465	23,465	20,321	20,321	21,209	22,571	22,571
Clusters	73	73	73	73	73	73	73
MDV	0.041	0.041	0.041	0.041	0.044	0.013	0.055
Sample	All	All	Pre	Pre	All	All	All
Controls	No	Yes	No	Yes	Yes	Yes	Yes
Month of birth FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	No	Yes	Yes
Mother FE	No	No	No	No	Yes	No	No

Notes: All specifications are estimated as linear probability models using weighted least squares. Except for the mother FE specification, a "Yes" in the "Controls" row indicates that the regression also includes the full set of demographic, SES, and program controls. Demographic controls include birth order (three indicators), five year maternal age group indicators, and gender. SES controls include education (three indicators), wealth (four indicators), rural indicator, altitude quintile indicators, and ethnicity indicators. Program controls include controls for the CB-NCP and CB-IMNCI health programs. The mother FE specification excludes the SES controls as these are mother-invariant. MDV is the mean of dependent variable among untreated individuals. Standard errors clustered at the district level in parentheses. Asterisks indicate significance at the following levels: * $p<0.1$, ** $p<0.05$, and *** $p<0.01$.

Table B.3: Effect of CHX-NCP on neonatal mortality - conditional logit and logit estimates

	Dependent variable: mortality						
	month ∈ [0,1]					month ∈]1,12]	month ∈ [0,12]
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Conditional Logit (beta coefficients)							
main							
CHX	-0.371*	-0.538**			-1.163***	0.304	-0.373
	(0.205)	(0.236)			(0.360)	(0.409)	(0.230)
CHX_{t-6}			0.139	0.027			
			(0.357)	(0.378)			
Observations	21,750	21,750	18,677	18,677	3,071	13,961	21,975
B. Logit (marginal effects)							
CHX	-0.015*	-0.021**				0.005	-0.019
	(0.008)	(0.009)				(0.006)	(0.012)
CHX_{t-6}			0.006	0.001			
			(0.015)	(0.016)			
Observations	21,750	21,750	18,677	18,677		13,961	21,975
Sample	All	All	Pre	Pre	All	All	All
Controls	No	Yes	No	Yes	Yes	Yes	Yes
Month of birth FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year of birth FE	No	No	No	No	No	Yes	No
District FE	Yes	Yes	Yes	Yes	No	Yes	Yes
Mother FE	No	No	No	No	Yes	No	No

Notes: Except for the mother FE specification, a "Yes" in the "Controls" row indicates that the regression also includes the full set of demographic, SES, and program controls. Demographic controls include birth order (three indicators), five year maternal age group indicators, and gender. SES controls include education (three indicators), wealth (four indicators), rural indicator, altitude quintile indicators, and ethnicity indicators. Program controls include controls for the CB-NCP and CB-IMNCI health programs. The mother FE specification excludes the SES controls as these are mother-invariant. In Column (6) we include year of birth fixed effects instead of month of birth effect because otherwise the variance-covariance matrix is highly singular in panel A, preventing the computation of clustered standard errors. However, point-estimates are very similar if we include month of birth fixed effects instead (0.282 in panel A and 0.006 in panel B). MDV is the mean of dependent variable among untreated individuals. Standard errors clustered at the district level in parentheses. Asterisks indicate significance at the following levels: * $p<0.1$, ** $p<0.05$, and *** $p<0.01$.

C Details of the Machine Learning Procedure

C. 1 Training the Causal Forest

To assess treatment effect heterogeneity we train a causal forest using the *grf* package in R (Athey et al., 2019; Tibshirani et al., 2021). Concretely, we proceed in the following two steps.

Step 1 We use regression forests to estimate the following two conditional mean functions

$$\mu_W = E[W|X = x] \tag{2}$$

$$\mu_Y = E[Y|X = x] \tag{3}$$

where W is equal to 1 if the child was born in a district and month where the CHX program was implemented and 0 otherwise, Y is 1 if the child died within the first month after birth and 0 otherwise, and X is a set of indicator variables capturing the district of birth, the month-year date of birth, whether the CB-IMNCI program is implemented, and whether CB-NCP is implemented in the district. Using the fitted conditional mean functions we construct the residuals, $W - \mu_W$ and $Y - \mu_Y$.

Step 2 We use the residuals from the first step to train a causal forest which we use to estimate the conditional average treatment effects (CATEs):

$$\tau(X) = E[Y(1) - Y(0)|X = x] \tag{4}$$

where $Y()$ are the potential outcomes and X contains birth order, maternal education, maternal age, wealth, district, altitude, rural, predicted place of delivery, health programs, district, ethnicity; and district-level averages for: antenatal care (ANC) visits (timing and number), whether iron tablets were received during ANC visits, tetanus protection, place of delivery, postnatal visits, immunization rate, neonatal mortality, nurse or doctor-assisted delivery, and whether the baby was considered small at birth.

For the categorical variables (ethnicity and district) we use the sufficient representation approach where we compute and include group means of the non-categorical variables based on the groups defined by the categorical variables.

In training the causal forest we tune all parameters by cross-validation. For non-tuned parameters we use the default settings, except that we set the forest to be clustered at the district level and we allow clusters to have different weights. The latter setting has very little practical implication in our setting. The chosen parameter settings are listed in Table C.1.

Having specified the parameter settings, we grow a tree as follows:

(i) We sample 50% of the original analysis sample and 30 of the variables.

(ii) The sample selected in (i) is split into two equally sized sub-samples. One sub-sample is used to find the splitting structure of the tree. The second sample is used for populating the trees.

(iii) The sample for splitting found in (ii) is split into two groups (nodes) using the variable among the 30 selected in (i) that creates the best split. The best split maximizes treatment effect heterogeneity across the two groups and minimizes the variance in treatment effect heterogeneity within the groups.

(iv) The tree is grown by repeating step (iii) on the created groups until there is no valid split (for example if the number of observations is smaller than 5) or if there is no split that improves the fit sufficiently. A group that is not split further is called a leaf.

(v) Using the splitting structure found in (iv) the tree is populated using the second sub-sample created in (ii) and the outcomes are predicted based on these observations. In other words the hold out sub-sample for populating the trees runs through the decision tree (the splitting) and these observations are then used to obtain an estimate of the leaves' treatment effects.

Steps (i) to (v) create a tree and these five steps are repeated 2000 times to create the forest. Having created the forest, an observation's predicted conditional treatment effect (CATE) is created based on the average predicted outcome for the leaves the observation ends up in across all trees where this observation was not used to split and populate the trees, i.e., based on the out-of-bag prediction.

Table C.1: Causal Forest Settings

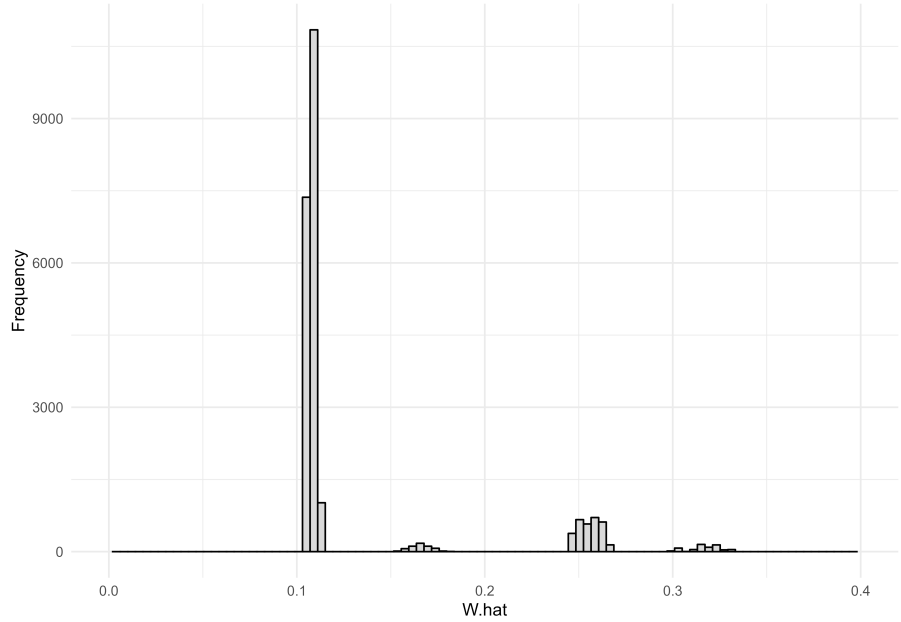
Setting	Value	Selection criteria
Number of trees	2000	Default
Clustering	District	Choice
Fraction of sample used to grow each tree	0.5	Cross-validation
Number of variables considered for each split	30	Cross-validation
Minimum size of a leaf node	5	Cross-validation
Fraction of sample used for splitting	0.5	Cross-validation
Prune empty leaves	True	Cross-validation
Maximum imbalance of a split (alpha)	0.05	Cross-validation
Penalization of imbalance splits	0	Cross-validation

Note: The table shows the parameter settings for the main causal forest. None of the parameters selected by cross-validation are different to the default setting.

C. 2 Distribution of propensity scores and covariates

Figure C.1 shows the distribution of propensity scores (i.e., the estimated values of the μ_W from Step 1 described above). These scores should be between 0 and 1 (not including 0 and 1), which is the case in our setting.

Figure C.1: Propensity scores for causal forest



Another important condition for the causal forest is that the features have common support across treatment status. Figures C.2 to C.5 suggest that this is the case in our setting.

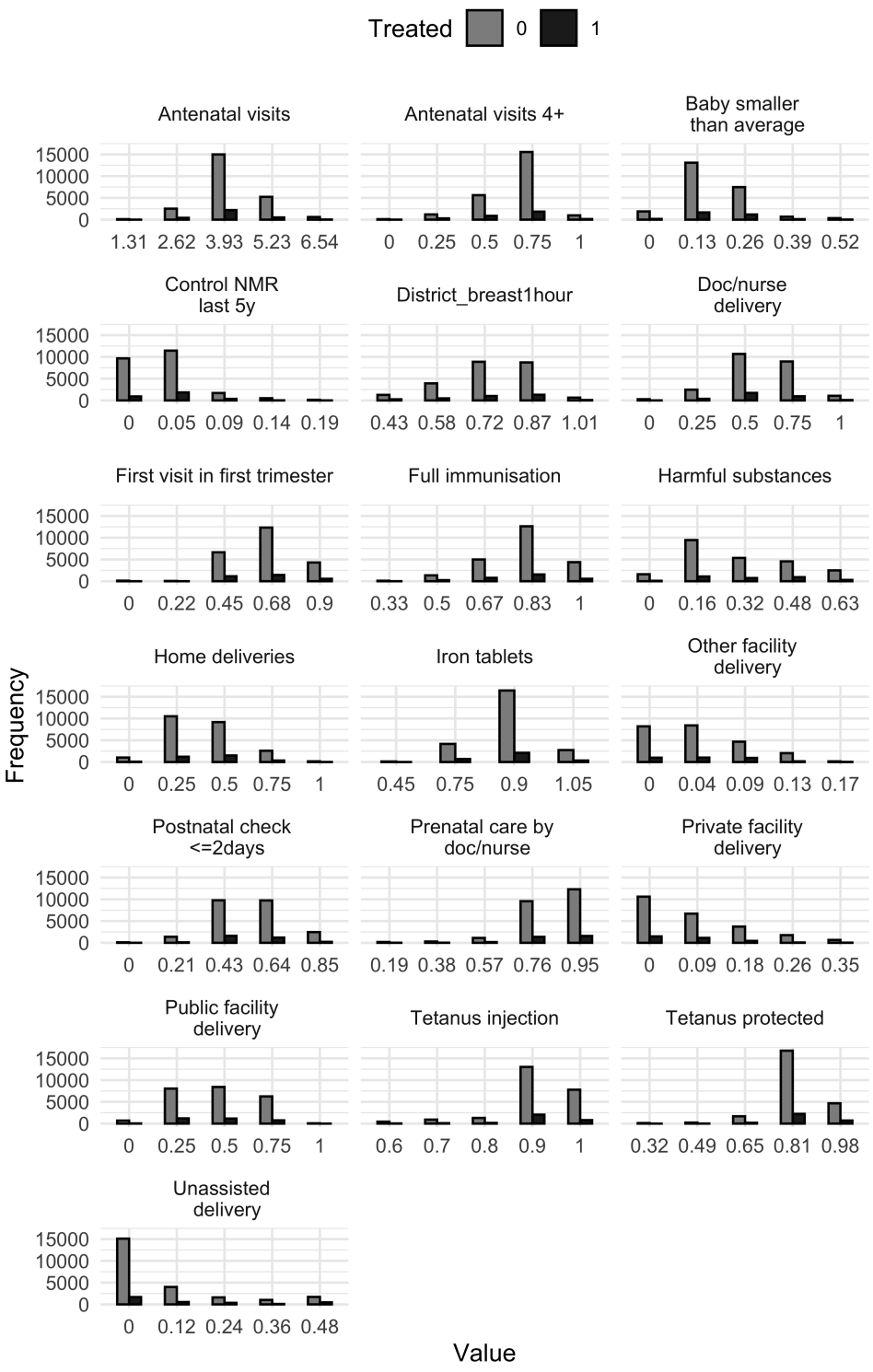


Figure C.2: Inverse-propensity score weighted distributions treated and control observations for individual covariates

Notes: This chart shows the distributions of all district level demeaned covariates by treatment status. Each observation is weighted by 1 divided by the estimated propensity for observing the actual treatment status.

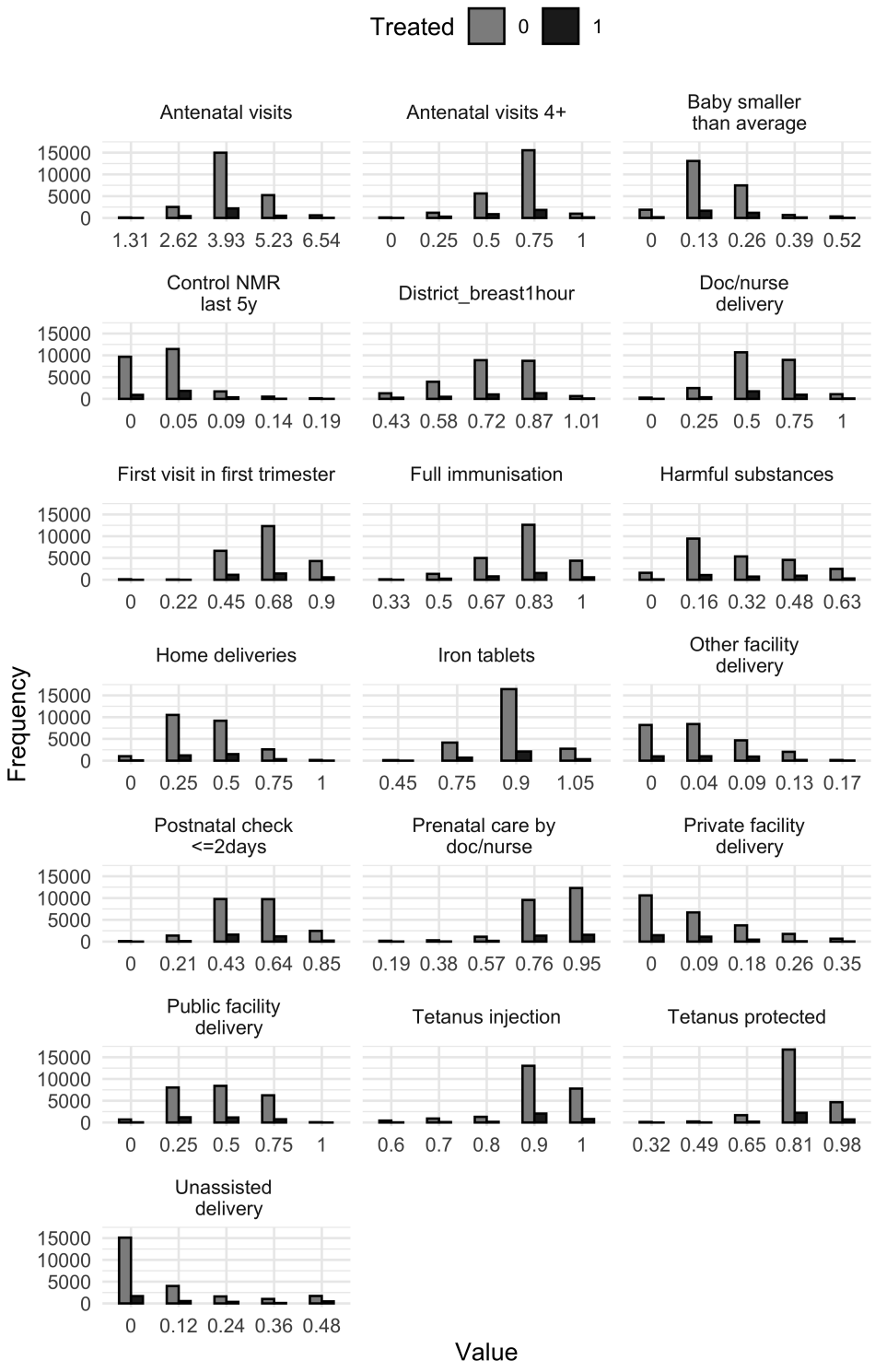


Figure C.3: Inverse-propensity score weighted distributions treated and control observations for district covariates

Notes: This chart shows the distributions of all district level demeaned covariates by treatment status. Each observation is weighted by 1 divided by the estimated propensity for observing the actual treatment status.

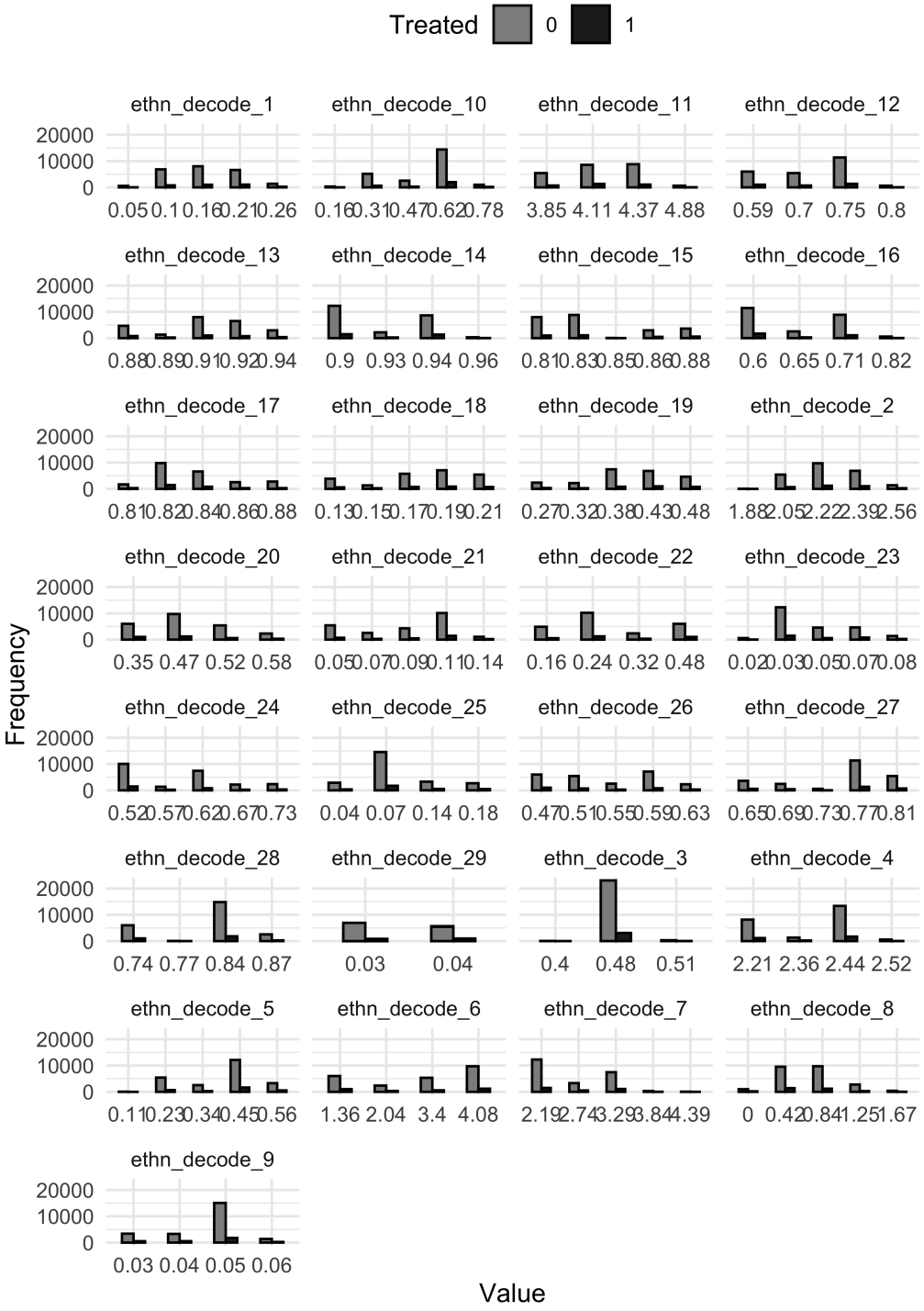


Figure C.4: Inverse-propensity score weighted distributions across treated and control observations for ethnicity demeaned covariates

Notes: This chart shows the distributions of all district level demeaned covariates by treatment status. Each observation is weighted by 1 divided by the estimated propensity for observing the actual treatment status.

C. 3 Alternative causal forest specifications

In Table C.2 we show results for three different specifications of the causal forest. Column (1) shows the main forest using the settings described above. Column (2) shows the results from training a forest using a smaller set of variables in step 2. This specification is used to obtain predictions of the CATEs for the five RCT locations. Column (3) shows the result of a specification based on the same variables used in specification (2), except for the district level variables, which perfectly predict whether the observation is in the main analysis sample (countrywide Nepal) or in the samples drawn from DHSs carried out in the RCT locations. This last specification is used to allow us to obtain the doubly-robust average treatment effects reported in Table 10.

Table C.2: Causal forest specifications for extrapolation exercise - diagnostic test and average treatment effects

	Main	Overlapping variables	Overlapping variables w/out dist. averages
	(1)	(2)	(3)
A. Omnibus diagnostic test for forest fit			
Mean Forest Prediction	1.123***	1.078***	1.051***
	(0.255)	(0.253)	(0.235)
Differential Forest Prediction	0.677*	0.312	0.205
	(0.479)	(0.385)	(0.217)
B. Doubly Robust Average Treatment Effects			
Full sample	-0.019***	-0.020***	-0.021***
	(0.003)	(0.003)	(0.003)
Predicted facility births	-0.007**	-0.008**	-0.007*
	(0.004)	(0.004)	(0.004)
Predicted home births	-0.030***	-0.031***	-0.033***
	(0.004)	(0.004)	(0.003)

Notes: The table shows results from estimating the causal forest using three different sets of variables. All specifications are based on the same orthogonalization based on district fixed effects, month-year of birth fixed effects as well as indicators for the CB-IMNCI and CB-NCB programs. Column (1) shows the results for the main specification where the forest is built based on indicators for CB-IMNCI and CB-NCB health programs, indicators for wealth and altitude quintiles, indicators for maternal education, indicators for birth order, indicators for maternal age group, a rural indicator, a child gender indicator, an indicator for predicted home delivery, and district level averages of: antenatal care (ANC) visits (timing and number), whether iron tablets were received during ANC visits, tetanus protection, place of delivery, postnatal visits, immunization rate, neonatal mortality rate, delivery support by a nurse or doctor, and share of newborns considered small at births. Moreover, following the means-encoding approach presented in Johannemann et al. (2019), in (1) the categorical variables district and ethnicity are included through demeaned versions of the other variables by, respectively, district and ethnicity. In Column (2) the forest is built on the same set of variables as in (1) except for the ethnicity and district demeaned variables, the indicators for the CB-NCB and CB-IMNCI programs, the district level measure of iron tablets received, tetanus protection, the measure of timing of antenatal visits, and postnatal visits, which are either inapplicable outside Nepal (in the case of the health programs) or not consistently available across all DHS location samples. Column (3) is showing results for a forest built only on birth order, gender of the child, maternal age, maternal age indicators, maternal education indicators, wealth quintile, and predicted place of delivery. Standard errors clustered at the district level in parentheses. Following Athey et al. (2019), the p-values for the omnibus diagnostic test are for the one-sided hypothesis test. Asterisks indicate significance at the following levels * $p<0.1$, ** $p<0.05$, and *** $p<0.01$.

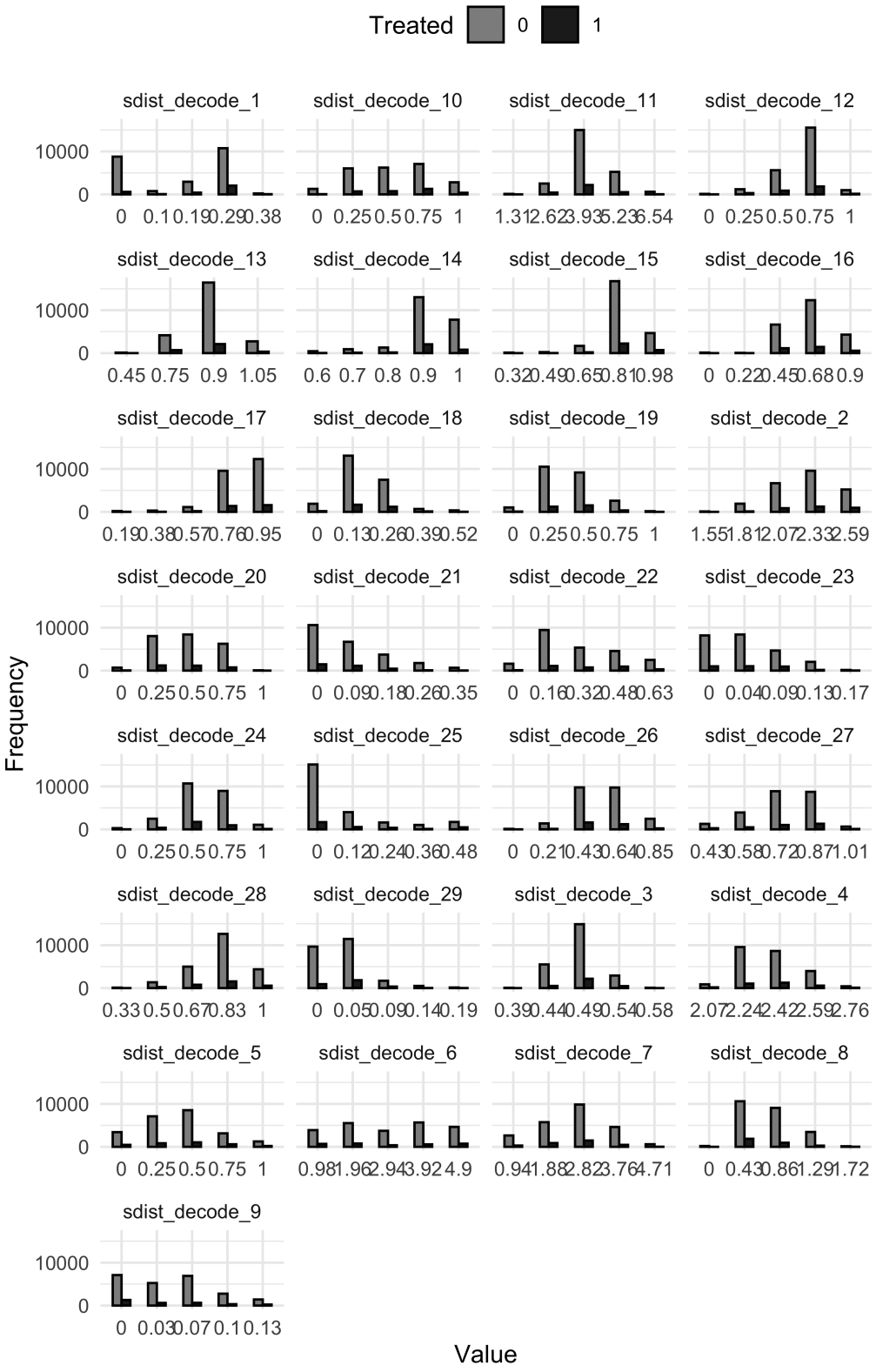


Figure C.5: Inverse-propensity score weighted distributions across treated and control observations for district demeaned covariates

Notes: This chart shows the distributions of all district level demeaned covariates by treatment status. Each observation is weighted by 1 divided by the estimated propensity for observing the actual treatment status.