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)

December 2024

Abstract

Neonatal mortality contributes an increasing share of under-5 mortality. Experimental estimates of a low-cost preventive measure (chlorhexidine cord care) vary widely across settings, leading to external validity concerns. We provide the first, quasi-experimental, estimates of the effect of a nationwide roll-out in Nepal and apply recently developed machine-learning techniques (ML) to analyze treatment effect heterogeneity. We find that the program decreases neonatal mortality by 36% 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.

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. Random author order generated using the AEA random author order tool (confirmation code: -nO5wvJwe_r8). 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 many presentation participants. All errors are our own.

1 Introduction

Mortality in the first 28 days of life accounts for over 2.5 Million deaths each year and contributes an increasing share of under-5 deaths globally (Wang et al., 2016). Most neonatal deaths are believed to be preventable at comparatively low cost (Bhutta et al., 2014), but simply reducing financial barriers to accessing general health care around birth in low-income countries often results in no- or small improvements (e.g., Powell-Jackson et al., 2015; Fitzpatrick, 2018).[1] Research is therefore needed to shed light on what specific interventions work at scale and for whom.

One leading cause of neonatal mortality is blood infection or "neonatal sepsis". It is estimated to kill 400,000 children annually, most of them in very low-income settings where unsanitary delivery- and living conditions are common (Liu et al., 2016).[2] Hopes of eradication were high after a simple preventive measure was found to massively reduce neonatal mortality in three randomized controlled trials (RCT) in South Asia (Mullany et al., 2006; El Arifeen et al., 2012; Soofi et al., 2012). But these hopes faded away after two further RCT in Southeast Africa found no effect of a similar intervention — namely, the preventive application of a local disinfectant called chlorhexidine (CHX) to the umbilical cord stump (Semrau et al., 2016; Sazawal et al., 2016; Osrin and Colbourn, 2016).

This paper answers three open questions: (i) Can a CHX cord care intervention reduce neonatal mortality outside experimental conditions?; (ii) What variables can, empirically, best account for the heterogeneity of the effect of CHX on neonatal survival?; and (iii) Could an alternative targeting policy to the World Health Organization's past and current guidelines further reduce neonatal mortality?

Our first contribution is to estimate the effect of a CHX cord care intervention 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*

[1] This is in contrast to historical evidence showing that health interventions around birth dramatically improved neonatal survival in high-income countries (Lazuka, 2018, 2021).

[2] 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).

treatment effects at scale (Al-Ubaydli et al., 2017). But in the case of trials of preventive health care measures, the treatment effect might be muted due to the lack of a pure control group (El Arifeen et al., 2012; Semrau et al., 2016). For instance, subjects involved in CHX trials are referred to the hospital if signs of cord infection appear during the frequent research team visits. Indeed, in both trials finding no significant effect on mortality, the authors note that the neonatal mortality rate (NMR) was much 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 recently developed machine-learning (ML) techniques (Athey et al., 2019; Athey and Wager, 2021) to understand how CHX treatment effects depend on observable characteristics of individuals and/or districts — 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 (Imdad et al., 2013; Sankar et al., 2016; López-Medina et al., 2019) have important limitations due to the small number of included studies and the possibility that heterogeneous results by place of birth — the main source of heterogeneity identified in existing meta-analyses — may be confounded by other differences across studies. The use of ML methods to study heterogeneous treatment effects has two main advantages: embedded robustness checks in the form of cross-validation and a high degree of flexibility in identifying sources of heterogeneity (Varian, 2014; Athey and Imbens, 2016). In our case, causal forest estimates show for instance that stark differences in average predicted treatment effects by home vs. facility delivery hide that many babies born in facilities may also benefit from CHX cord care, so that targeting implementation beyond babies born at home would further reduce neonatal mortality.

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 — including similar patterns of compliance conditional on covariates — if it were hy-

pothetically extended to samples from the five RCT regions. 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 based on data points that are more similar to the out-of-sample observations. Given differences in the exact nature of the intervention between the Nepalese roll-out and the trials, as well as, crucially, 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 Nepalese roll-out was fully externally valid.[3] Our exercise however serves as a sanity check on our heterogeneity analysis as well as illustrates the informativeness for external samples of the treatment effect heterogeneity uncovered in our quasi-experimental setting.

The first country to introduce CHX cord cleansing nationwide is Nepal. We exploit plausibly exogenous variation in the timing of the expansion of the Chlorhexidine Navi(Cord) Care Program 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, this CHX cord care program was quickly scaled-up across the rest of the country (see Appendix Figure A.1). By 2015, 75 percent of the population was covered by the program (Department of Health Services, 2015).

We estimate the effect of the Chlorhexidine Navi(Cord) Care Program (CHX-NCP) using the estimator developed by Borusyak et al. (2024) and find that, overall, the CHX program decreased neonatal mortality by 1.5 percentage points or 36 percent compared to the control group mean. Our conclusions are robust to comprehensive robustness checks. In particular, we find no significant differences in pre-treatment trends between treated and control cells, that the CHX program 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 ob-

[3] 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 between CHX trials and the Nepalese roll-out, lack of access to the relevant experimental microdata means that we are unable to implement the innovative methods proposed by Athey et al. (2020); Gechter and Meager (2022); Kowalski (2023).

served independently of other neonatal health interventions. We also reach similar conclusions based on "traditional" linear two-way fixed effects model estimates.

Turning to our heterogeneity analysis, we find that, when splitting the sample between predicted home- and institutional deliveries, 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 is however likely to proxy for risk factors such as hygiene conditions and healthcare at- and shortly after birth and health endowment at birth. 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. Importantly, only a quarter of the variation in conditional treatment effects comes from differences in variables which the WHO has ever 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 for a given level of policy complexity, and compare this optimal policy — from the point of view of neonatal survival — to past and present WHO recommendations. We find that WHO-recommended policies effectively target about a third of the population with returns to treatment as high as any other but that they miss many children whose survival chances could significantly benefit from CHX treatment.

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. 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.

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 average treatment effects and robustness checks are reported in Section 4. Section 5 explores heterogeneity in the effect of CHX application using machine learning techniques 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 trialed interventions. Section 7 concludes.

2 Background

2.1 Evolution of Neonatal 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).

The country saw a long period of reduction in NMR which ended in 2005, when it was followed by a period of stagnation until 2010 (Figure 1). This stagnation came to an end in 2011, as NMR dropped to 21 per 1,000 during 2012-2016 — a 36% decline relative to the previous 10-year period Ministry of Health [Nepal] and New ERA and ICF (2017).

The sharp decrease in NMR from 2011 to 2016 coincides with the acceleration of the roll-out of CHX cord application through the Chlorhexidine Navi(Cord) Care Program (CHX-NCP) (see Figure 1). Strikingly, since the completion of the program roll-out, there has been no further reduction of NMR according to the latest figures (Ministry of Health and Population, Nepal; New ERA; and ICF, 2022).

2.2 Details of the Chlorhexidine Cord Care Program

Program Objective and Components. CHX-NCP was a $3.9 million program funded mainly by bilateral donors (US, Norway, Canada, UK) and the Bill & Melinda Gates Foundation and was designed to support the Government of Nepal to scale up the use of CHX for cord care across districts nationwide. The aim of the program was for all newborns to receive a single CHX gel application on the day of birth irrespective of place of birth. The CHX roll-out program (i.e., our "treatment") consisted of all the technical support needed for: the delivery of

CHX doses, the training of staff to apply and counsel patients on CHX application, and promotion of CHX application — CHX cord care products were neither available nor promoted in a district prior to the program roll-out.

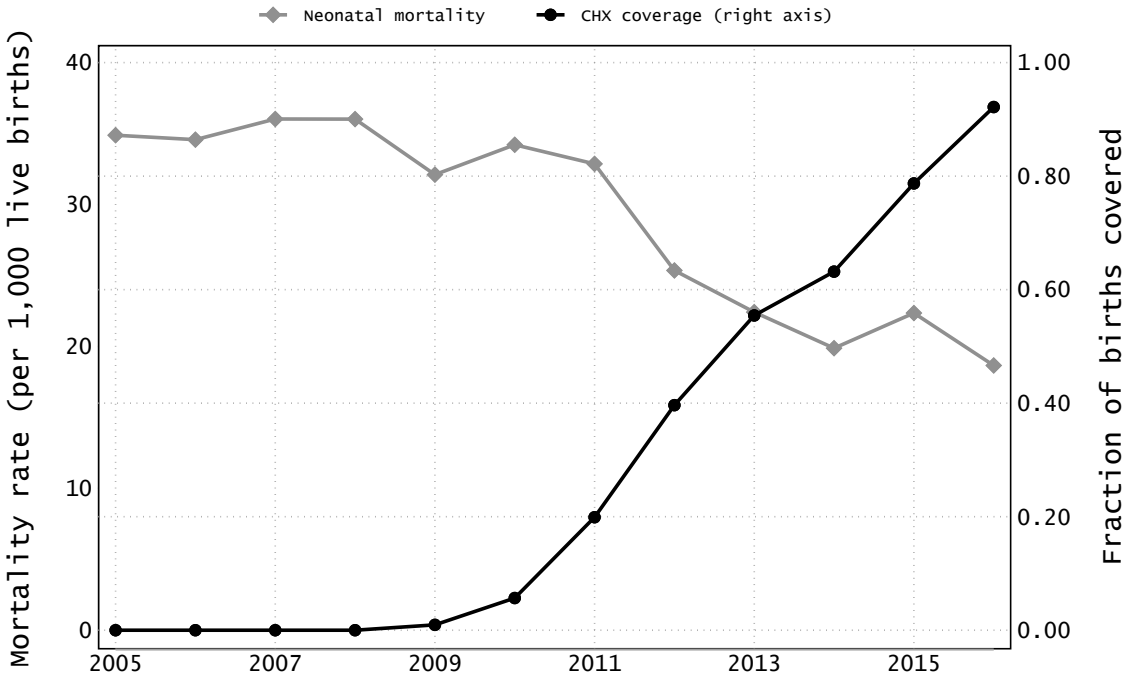


Figure 1: Neonatal mortality and CHX-NCP coverage

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

Implementation. The program was implemented by JSI Research & Training Institute, Inc in partnership with the Nepalese Department of Health Services, international NGOs and a Nepalese pharmaceutical company which produced the CHX gel locally. 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).[4,5] 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 into the Community-Based Newborn Care Program (CB-NCP) (JSI Research

[4]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).

[5]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.

& Training Institute, 2017; JSI, 2017; Hodgins et al., 2019). In Section 4.2.1 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.

Compliance with CHX Application. 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).[6] 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 our estimates of the effects of the program should be interpreted as intention-to-treat effects of actual CHX application — 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.

Why Did the Program Extend to Institutional Deliveries? The benefits of CHX cord care are likely to be larger for home births as more hygienic practices should generally be expected in birthing facilities than at home. There were, however, sufficient concerns about cord care to warrant implementation of the CHX cord care program irrespective of place of delivery. This could be due to the potential for infection to occur in hospital settings (e.g., due to insufficient hand hygiene, Khanal and Thapa, 2017) and/or the possibility of infection after the newborn has left the birthing facility.

[6] 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.

2.3 Other Neonatal and Child Health Programs

Identification of Relevant 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.

Relevant Programs and Implications for our Analysis. Eleven interventions were identified by key informants, including CHX-NCP. These are summarized in Appendix Table A.2. 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 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 A.2). 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 4.2.1 that there is no evidence of complementarities between CHX cord care and the presence of these programs in the district.

2.4 Evolution of World Health Organization Guidelines

WHO guidelines have evolved substantially in the last fifteen years. Prior to 2013, the only recommended approach to cord care was to keep the cord clean and dry. In 2013, the WHO started recommending the application of CHX in some cases — namely *home births in settings with neonatal mortality above 30 per 1000* (WHO, 2015). But since 2022, the WHO recommends 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* (WHO, 2022).

Why Not Treat All Births Everywhere? While international estimates of the direct cost of CHX application are low, the WHO's recommendation of restricting the use of CHX can be understood as balancing proven benefits against broader costs, in the same vein as its early stance on face masks during the COVID-19 pandemic (WHO, 2020).[7] Broader costs of CHX cord care include the behavioral risk of unintentionally encouraging the application of other, potentially dangerous substances, as well as the opportunity cost 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).

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

[7] 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).

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.[8] 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 4.2.2).

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. 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 Appendix Table A.3) 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 (49 percent female).

[8]The sample is 23,449 when using the Borusyak et al. (2024) estimator as 16 observations are dropped in the last two periods due to there being no never treated group.

3.2 Identification Strategy

We exploit the staggered roll-out of CHX-NCP across districts over time to estimate the average treatment effect on the treated (ATT) using the estimator developed by Borusyak et al. (2024) (henceforth: BJS). Under the standard parallel trends and no anticipation effect assumptions, the BJS estimator not only provides unbiased estimates in difference-in-differences designs such as ours even in the presence of heterogeneous causal effects, it is also efficient (Borusyak et al., 2024).[9] For completeness, in robustness checks we also report estimates from two-way fixed effects ordinary least squares, which lead to slightly larger estimates but otherwise similar conclusions.

The dependent variable 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. The treatment indicator, CHX_{dt} is an indicator equal to 1 if CHX-NCP was rolled out in the child's district by the month the child was born. We control for district fixed effects, D_d, and month$\times$year of birth fixed effects, T_t — e.g., one fixed effect for the equivalent of January 2012 in the Nepali calendar, another for February 2012, etc... We also control for a set of covariates, X_{idt}, comprising all controls listed in panels A and B of Table A.3, and which cover child-, mother-, household characteristics and district-time varying controls such as exposure to health programs other than CHX-NCP. We use the `did_imputation` implementation of the BJS estimator in Stata, where the ATT is estimated in the following three steps:

1. Using data on the not-yet-treated cells only, we estimate a model of potential outcomes using the common two-way fixed effects approach with controls. I.e. we estimate $m_{idt} = \alpha + \beta CHX_{dt} + D_d'\Delta + T_t'\Gamma + X_{idt}'\Lambda + \varepsilon_{idt}$.

2. Based on the estimated coefficients from 1 we extrapolate the outcomes for the treated units in the absence of treatment. We then compute the difference between the observed outcome and this predicted outcome: $\tau_{dt} = m_{dt} - \hat{m}_{dt}(0)$ for the treated units.

3. We calculate the average treatment effect by averaging τ_{dt} across the relevant cells.

[9] Borusyak et al. (2024)'s use of all pre-treatment cells is especially valuable in our set-up where each district-month cell has few observations. This rules out estimators which only use the last pre-treatment period to construct the counterfactual.

In the main specifications we present the ATT as the average τ_{dt} across all treated cells and report standard errors clustered at the district level in parentheses.

Identifying Assumptions. 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 the CHX cord care program (CHX-NCP) and the program onto which CHX-NCP added its operations whenever possible to save on costs (CB-NCP), the treatment is only weakly correlated with observable characteristics.[10] 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 Appendix Table B.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.

4 Average Treatment Effect Results

4.1 Main Estimates

Table 1 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 1.0 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 1.5 percentage points or 36 percent of the control mean. In the table we also report estimated coefficients for treatment effects in the six periods before CHX-NCP was introduced in the district. None of these estimates is significantly different from zero.

Having established significant average beneficial effects of CHX-NCP at scale, we assess treatment effect heterogeneity in columns (3) and (4) of Table 1. In line with findings from a

[10] In Section 4.2.1, we discuss the role of other programs. We find no evidence of complementarities and find that our results hold whether or not we include district-cells where CB-NCP is present.

meta-analysis of the five existing trials (e.g. Imdad et al., 2013), between 2013 and 2022 WHO guidelines only recommended CHX cord care for home deliveries in areas with rates of neonatal mortality above 3 percent. Assessing treatment effect heterogeneity by place of delivery is therefore a useful first step to shed light on the desirability of this recommendation outside experimental settings. Place of delivery is only collected by the DHS for recent births (i.e., within five years of 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. The covariates included and estimated coefficients are 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.4). 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.[11] In Columns (3) and (4) we split the sample between births predicted to take place in a facility (Column 3) and at home (Column 4). We find a near-zero estimated effect of CHX among predicted facility deliveries (0.2 percentage point) while the estimated decrease in the probability of neonatal mortality among predicted home deliveries remains statistically significant and increases to 2.7 percentage points.

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 (1) of Table 2, the dependent variable is 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 column (2), 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 2.0 percentage points. In column (3) we show results from a specification where we do

[11] Namely, we draw 500 random samples from the original dataset, and, for each random sample, predict whether the child is delivered at home or not and then re-estimate the ATT.

Table 1: The effect of CHX-NCP on neonatal mortality exploiting variation across districts and across time, by predicted place of birth

	(1)	(2)	(3)	(4)
$t \geq$ CHX introduction	-0.010***	-0.015***	0.002	-0.027**
	(0.004)	(0.006)	(0.010)	(0.013)
$t-1$	0.017	0.015	0.025	0.013
	(0.024)	(0.024)	(0.039)	(0.050)
$t-2$	0.001	-0.004	0.008	-0.012
	(0.018)	(0.018)	(0.023)	(0.033)
$t-3$	-0.007	-0.009	0.015	-0.044
	(0.018)	(0.018)	(0.030)	(0.037)
$t-4$	0.002	-0.001	-0.000	-0.008
	(0.032)	(0.031)	(0.037)	(0.051)
$t-5$	-0.010	-0.013	-0.024*	-0.006
	(0.027)	(0.027)	(0.015)	(0.059)
$t-6$	0.015	0.009	0.019	0.025
	(0.024)	(0.024)	(0.025)	(0.055)
P-value	0.978	0.983	0.473	0.929
Observations	23,449	23,449	10,855	12,481
MDV	0.042	0.042	0.033	0.050
Sample	All	All	$P_H < 0.5$	$P_H > 0.5$
Controls	No	Yes	Yes	Yes
Month of birth FE	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes

Notes: All specifications are estimated using the BJS (Borusyak et al., 2024) imputation estimator as described in the main text. P-value shows the p-value for the joint test that all lags are zero. MDV is the mean of the dependent variable among untreated individuals. P_H is the predicted probability of being born at home based on estimates reported in Column (2) of Table A.4. Standard errors clustered at the district level in parentheses. In columns (1) and (2) we report analytical standard errors and in columns (3) and (4) we report bootstrapped standard errors based on 500 iterations to account for the predicted place of delivery. Asterisks indicate significance at the following levels: * p<0.1, ** p<0.05, and *** p<0.01.

not allow for heaping at age at death at 1 month old and still find a comparable reduction in neonatal mortality of 1.2 percentage points.

In column (4) of Table 2 we show results from an OLS estimation of a linear two-way fixed effects model.[12] The coefficient on the treatment indicator is only slightly larger than the comparable results obtained with the BJS estimator in column 2 of Table 1. In Appendix B we show detailed linear two-way fixed effects results, including subgroup results by predicted

[12]Namely, we report results obtained when estimating linear probability models of the form $m_{idt} = \alpha + \beta CHX_{dt} + D_d'\Delta + T_t'\Gamma + X_{idt}'\Lambda + \varepsilon_{idt}$ including all observations irrespective of treatment status and availability or not of not-yet-treated cells.

Table 2: Alternative specifications and falsification tests

	Dependent variable: mortality				
	Falsification	Infant Mort.	No Heaping	OLS	
	month∈]1,12]	month∈[0,12]	month<1	month∈[0,1]	
	(1)	(2)	(3)	(4)	(5)
$t \geq$ CHX introduction	-0.001	-0.020***	-0.012**	-0.018***	-0.020***
	(0.003)	(0.007)	(0.005)	(0.006)	(0.004)
Observations	21,635	22,546	23,449	23,465	21,209
MDV	0.016	0.058	0.037	0.042	0.045
Controls	Yes	Yes	Yes	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
Estimator	BJS	BJS	BJS	OLS	OLS

Notes: BJS is the Borusyak et al. (2024) imputation estimator as described in the main text. OLS is the ordinary least squares estimator. MDV is the mean of the dependent variable among untreated individuals. Standard errors clustered at the district level in parentheses. The sample in column (1) excludes neonatal deaths. The samples in columns (1) and (2) exclude births that have taken place less than 12 months before the interview, since they are not fully exposed to the risk of infant mortality. Asterisks indicate significance at the following levels: * p<0.1, ** p<0.05, and *** p<0.01.

place of delivery. As a final robustness check, column (5) of Table 2 shows the results from estimating a specification where we exploit variation across siblings in the availability of CHX at birth by estimating a linear mother fixed effects model with OLS. This approach suggests a reduction in neonatal mortality by 2.0 percentage points.

4.2 Further Robustness Tests

4.2.1 Other Programs

We explore in detail the potential interaction between CHX and the broader healthcare programs concerned with neonatal mortality summarized in Table A.2. First, we inspect visually the changes in neonatal mortality over time against the roll-out of both CHX care and these broader programs (CB-NCP and CB-IMNCI, in which CB-NCP was later integrated). As shown in Appendix Figure A.3, 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. 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 the three health programs separately and report the results in Appendix Table A.7. For completeness, we also estimate the effect of CHX in the sample of births where neither CB-NCP nor CB-IMNCI are present and the effect of CB-NCP where neither CHX nor CB-IMNCI are present 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.

4.2.2 Further Specification Checks

In Appendix Figure A.2 we show that our conclusions are robust to the specification of control variables and the sample selection. In Appendix Table A.5 we show that conclusions are also robust to using survey weights, and in Appendix Table A.6 we show that conclusions also hold when using stacked DHS samples to predict the place of delivery instead of only using the 2016 DHS sample (See also Appendix Table A.4). Finally, Appendix Figure A.5 shows an event study chart depicting the ATTs by month. Although the estimates are somewhat noisy due to the small number of observations per month $\times$ year cells, we see a clear drop in neonatal mortality after the introduction of CHX.

In the next section, we investigate systematically the heterogeneity of the benefits of CHX and what it implies for optimal policy targeting.

5 Treatment Effect Heterogeneity and Lessons for Policy Targeting

In the previous section, we documented large average treatment effects on the treated, and when splitting the sample by place of birth, we found that the effects of CHX on neonatal

mortality were driven by the home birth sample. Place of birth is however likely to proxy for other factors.[13] We now investigate heterogeneous treatment effects more systematically 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 Investigating Heterogeneity with Causal Forests

Place of delivery is likely to proxy for risk factors such as hygiene conditions including the application of harmful substances to the stump, healthcare at- and shortly after birth and health endowment at birth. To better understand the treatment effect heterogeneity we observe and therefore potentially improve on current WHO recommendations for targeting, we also consider heterogeneity along other dimensions. Given the absence of a pre-analysis plan, we use a data-driven approach to study the heterogeneity by means of a causal forest (Athey and Imbens, 2016; Tibshirani et al., 2021). The causal forest gives us estimates of the individual Conditional Average Treatment Effect (CATE) — i.e., the ATE for observations with a given set of individual- or district characteristics, which allows us to identify and describe who benefits the most and the least from the treatment.

5.1.1 Conditional Average Treatment Effects Based on Causal Forests

Overview. 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×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 potential outcome predictors or "features". In addition to the individual sample characteristics included as covariates in the analysis in Section 4.1, we also consider a rich set of district-level features measured in the five years prior to the survey (see notes under Figure 3 for the full list).

[13] In principle, it could also proxy for compliance. But as discussed in Section 2.2, compliance estimates 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.

Building the Causal Forest. 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.[14]

Diagnostic Tests. 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 3 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).

Doubly-Robust Average Treatment Effects. Panel B of Table 3 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 baseline specification (full sample: -1.9 percentage points compared to -1.5 percentage points in Table 1, predicted home deliveries: -3 percentage points compared to -2.7 percentage points in Table 1), 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 0.002 percentage points in Table 1).

[14] 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 3: 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 one-sided tests where the null hypothesis is that the coefficient equals zero and the alternative is that the coefficient is strictly positive.

5.1.2 Conditional Average Treatment Effects Heterogeneity

Overview. We now describe the rich pattern of heterogeneity in Conditional Average Treatment Effects (CATEs). In doing so, we show which variables are most strongly associated with CHX impact — showing, in particular, that the three variables used in WHO guidelines are good predictors, but not the strongest predictors — and that there is much heterogeneity in treatment effects *within* groups defined by place of birth.

Heterogeneity Between- and Within Place of Birth. Figure 2 shows the distribution of individual CATEs for, respectively: the full sample, the sample of predicted home births, and the sample 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.[15]

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

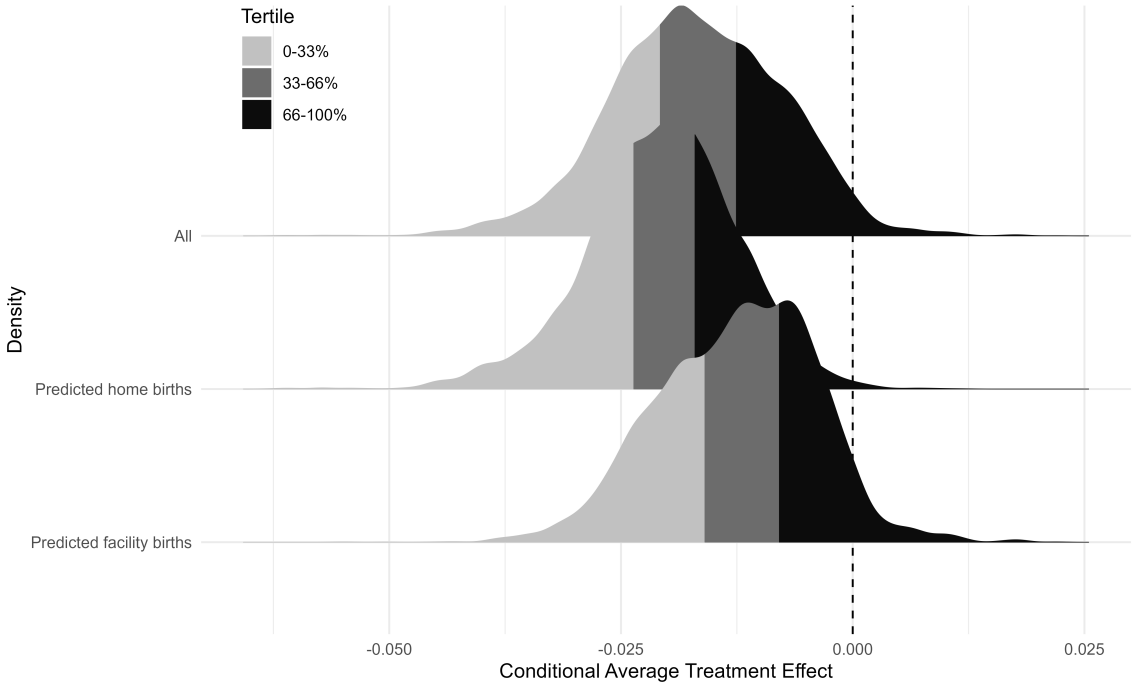


Figure 2: Distribution of CATEs by place of delivery

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

Individual- and District Predictors of Treatment Effects. We now turn to a broader characterization of the treatment effect heterogeneity. In Table 4 we compare means for selected variables across the first and third treatment effect tertiles. As expected from the distributions reported above, in the first tertile, where we observe large benefits of the treatment, 76 percent

[15] Appendix Figure A.7 shows the distribution of treatment effects across each sub-sample by reporting ATEs for each tertile of the overall- and birth place samples.

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, as per the latest WHO guidelines, may not be optimal.

Contribution of Variables Used in WHO Guidelines to Heterogeneity. The treatment effect heterogeneity uncovered by the causal forest goes beyond well-known predictors of CHX effectiveness, and in particular those used in WHO cord care recommendations (listed in Subsection 2.4). 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 maternal education, for instance, explains 35% of the variation in CATE (see Appendix Figure A.6).

5.2 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.

Optimal Policy Definition and Method. Following the approach proposed by Athey and Wager (2021), 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. 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

Table 4: Covariate means across tertiles of CATEs

	Tertile			
	First	Third	Difference	P-val
AIPW	-0.033	-0.001	0.032	<0.001
CATE	-0.027	-0.007	0.020	<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

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

being asymptotically "small" (for a given policy class Π and population).[16] 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.

Variables Used for Targeting. 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. Individual-level variables are likely to be of less practical use for policy targeting, but reporting estimates which also use these individual variables shows that targeting only by district-level variables does nearly as well. For the third policy, we allow targeting based on all district variables except the district prevalence of harm-

[16] 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).

ful substance application. 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 (issued in 2022).

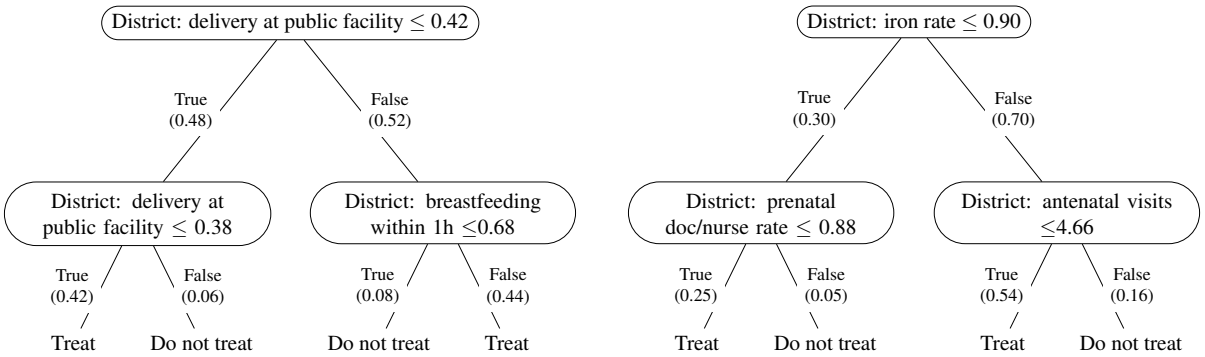
5.2.1 Comparing Optimal Policies with WHO Recommendations

Optimal Targeting Variables. To fix ideas, in Figure 3, 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 3, the resulting optimal assignment varies across folds.[17] 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).

Comparing NMR Reductions with WHO vs. Optimal Targeting. We now turn to predicting the effect on NMR of targeting different newborns in our sample. Each targeting policy selects different observations to be treated based on their characteristics. To obtain the Average Treatment Effects on the Treated (Untreated), we apply to each treated (untreated) observation the doubly-robust estimate or "AIPW" corresponding to their characteristics.[18] In Table 5 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 issued in 2022 of treating only newborns in contexts where the application of harmful substances to the

[17] In Appendix Table A.8 we report the number of times each covariate is selected in an optimal policy.

[18] Hence assuming that the rate of compliance is, across all targeting policies, the same as in the actual CHX program we evaluate in the preceding sections.



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

Figure 3: Examples of optimal policies

Notes: Figure 3a 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 3b 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.

umbilical 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 WHO 2013-2022 and WHO 2022 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 WHO 2013-2022 policy (and, by construction, also our operationalization of the WHO 2022 policy). As a result, the optimal policies would reduce the neonatal mortality rate by more than the WHO policies (namely, by 1.9 compared to 1.1 percentage points). Interestingly, the benefit of using both district- and individual-specific variables to design the optimal policy is negligible, compared to only using district-level variables which are more readily available to policy-makers.

5.2.2 Comparing Further Policies with WHO Recommendations

In this subsection, we consider two further targeting approaches. First, we consider a simple, practical targeting approach inspired by the optimal policies. Second, we restrict the share of births to be treated to be similar to the share treated under the WHO recommendations, and find the optimal targeting policy holding this share constant.

Simple Targeting Approach Inspired by the Optimal Policies. Optimal policies vary across folds so we also study the effect of applying a single targeting 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%. 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.[19]

Optimal Targeting of a Limited Share of Newborns. 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. When we constrain treatment to target no more than the number of treated individuals with the WHO policy, the benefits are considerably lower than with the unconstrained policies and statistically indistinguishable from the WHO 2013-2022 policy (Panel C of Table 5).

5.2.3 Conclusions for Policy Targeting

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 2, part of the sample is predicted to

[19] 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.

Table 5: Reduced mortality and share treated under alternative targeting policies

	ATT	ATU	%treated	ΔNMR	ΔNMR-ΔNMR$_{WHO}$
A. Pre-defined policies					
WHO 2013-2022	-0.036***	-0.012***	32.3	-0.011***	
	(0.004)	(0.003)		(0.001)	
WHO 2022	-0.034***	-0.012***	31.8	-0.011***	0.000
	(0.003)	(0.004)		(0.001)	(0.001)
District average antenatal visits ≤ 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, district fixed effects, and month times year 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. WHO 2013-2022 policy: treat children born at home in settings (here, districts) with NMR above 3 ppt. WHO 2022 policy: treat children in settings where the application of harmful substances is common. ATT (ATU) reports the AIPW for all individuals (not) treated with this policy. Standard errors in parenthesis. Asterisks indicate significance at the following levels * p<0.1, ** p<0.05, and *** p<0.01.

have near-zero or even adverse treatment effects.[20]

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

6 Extrapolating the Effect of Nepal's CHX Rolled-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

[20]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.

CHX trials were very successful in reducing NMR in three cases, but had no significant effect in two other ones.

Objective of the Extrapolation. 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, compliance with actual CHX application, 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.

Extrapolation Method. 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. 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.[21]

[21] We use 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. 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).

Comparison of Extrapolated Average Effects of CHX-NCP with RCT Findings. Table Appendix A.9 first shows the control group NMR rates reported in the RCT studies. Consistent with 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 Nepalese CHX rolled-out program 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 trials 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.

Heterogeneity Between- and Within RCT Locations. In the rest of Table A.9, 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 Appendix Figure A.8, 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, where RCTs found no significant effect. 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 a CHX program as implemented outside experimental circumstances.

In this paper, we estimate the effect of implementing a nationwide program rolling out CHX cord care. We find that the program led to a large reduction in neonatal mortality (36 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 a 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 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 and ours in particular, 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 beneficial in a much wider set of circumstances than the current received wisdom would indicate.

References

Al-Ubaydli, O., List, J. A., and Suskind, D. L. (2017). What can we learn from experiments? understanding the threats to the scalability of experimental results. *American Economic Review*, 107(5):282–86.

Athey, S., Chetty, R., and Imbens, G. (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *Annals of Statistics*, 47(2):1148–1178.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Beckett, M., Da Vanzo, J., Sastry, N., Panis, C., and Peterson, C. (2001). The quality of retrospective data: An examination of long-term recall in a developing country. *Journal of Human Resources*, pages 593–625.

Bhattarai, M. (2017). Nepal: Community action for nutrition project (sunaula hazar din) : P125359 - implementation status results report : Sequence 13 (english). Technical report, Washington, D.C. : World Bank Group.

Bhutta, Z. A., Das, J. K., Bahl, R., Lawn, J. E., Salam, R. A., Paul, V. K., Sankar, M. J., Blencowe, H., Rizvi, A., Chou, V. B., et al. (2014). Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost? *The Lancet*, 384(9940):347–370.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting Event-Study Designs: Robust and Efficient Estimation. *The Review of Economic Studies*, page rdae007.

Callaghan-Koru, J. A., Khan, M., Islam, M., Sowe, A., Islam, J., Billah, S. M., Mannan, I. I., George, J., Group, B. C. S. U. S., 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(2).

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2020). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *Mimeograph*.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernan, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39(14):1999–2014.

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., Mullany, L. C., Shah, R., Mannan, I., Rahman, S. M., Talukder, M. R. R., Begum, N., Al-Kabir, A., Darmstadt, G. L., Santosham, M., 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(9820):1022–1028.

Erchick, D. J., Lackner, J. B., Mullany, L. C., Bhandari, N. N., Shedain, P. R., Khanal, S., Dhakwa, J. R., and Katz, J. (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):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*, volume 108, pages 412–15.

Fottrell, E., Osrin, D., Alcock, G., Azad, K., Bapat, U., Beard, J., Bondo, A., Colbourn, T., Das, S., King, C., 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(5):F439–F447.

Friberg, I. K., Kinney, M. V., Lawn, J. E., Kerber, K. J., Odubanjo, M. O., Bergh, A.-M., Walker, N., Weissman, E., Chopra, M., Black, R. E., et al. (2010). Sub-saharan africa's mothers, newborns, and children: How many lives could be saved with targeted health interventions? *PLoS Med*, 7(6):e1000295.

Gechter, M. and Meager, R. (2022). Combining experimental and observational studies in meta-analysis: A debiasing approach. *Mimeograph*.

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

Hodgins, S., Khanal, L., Joshi, N., Penfold, S., Tuladhar, S., Shrestha, P. R., Lamichhane, B., Dawson, P., Guenther, T., Singh, S., 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., Pradhan, Y., Khanal, L., Upreti, S., and KC, N. P. (2013). Chlorhexidine for umbilical cord care: Game-changer for newborn survival? *Global Health: Science and Practice*, 1(1):5–10.

Imdad, A., Mullany, L. C., Baqui, A. H., El Arifeen, S., Tielsch, J. M., Khatry, S. K., Shah, R., Cousens, S., Black, R. E., and Bhutta, Z. A. (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(S3):S15.

Johannemann, J., Hadad, V., Athey, S., and Wager, S. (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, G. and Thapa, S. (2017). Awareness of hand hygiene among health care workers of chitwan, nepal. *SAGE Open*, 7(4):2158244017735141.

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., Dawson, P., Houston, R., 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(181).

Kowalski, A. (2019). Counting defiers: Examples from health care. *DOI: 10.48550/ARXIV.1912.06739*.

Kowalski, A. (2023). Reconciling seemingly contradictory results from the oregon health insurance experiment and the massachusetts health reform. *Review of Economics and Statistics*, 105(3):646–664.

Lazuka, V. (2018). The long-term health benefits of receiving treatment from qualified midwives at birth. *Journal of Development Economics*, 133:415–433.

Lazuka, V. (2021). It's a long walk: Lasting effects of maternity ward openings on labour market performance. *Review of Economics and Statistics*, pages 1–47.

Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., Lawn, J. E., Cousens, S., Mathers, C., and Black, R. E. (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(10063):3027–3035.

López-Medina, M. D., Linares-Abad, M., López-Araque, A. B., and López-Medina, I. M. (2019). Dry care versus chlorhexidine cord care for prevention of omphalitis. systematic review with meta-analysis. *Revista latino-americana de enfermagem*, 27.

Ministry of Health and Population, Nepal; New ERA; and ICF (2022). *Nepal Demographic and Health Survey 2022: Key Indicators Report.* Ministry of Health and Population, Nepal.

Ministry of Health [Nepal] and New ERA and ICF (2017). *Nepal Demographic and Health Survey 2016*. Ministry of Health [Nepal].

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., Darmstadt, G. L., Khatry, S. K., Katz, J., LeClerq, S. C., Shrestha, S., Adhikari, R., and Tielsch, J. M. (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(9514):910–918.

Osrin, D. and Colbourn, T. (2016). No reason to change who guidelines on cleansing the umbilical cord. *The Lancet Global Health*, 4(11):e766–e768.

Painter, K. and Feldman, J. (2019). *Omphalitis*. https://www.ncbi.nlm.nih.gov/books/NBK513338/.

Paudel, D., Shrestha, I. B., Siebeck, M., and Rehfuess, E. (2017). Impact of the community-based newborn care package in nepal: a quasi-experimental evaluation. *BMJ open*, 7(10):e015285.

Powell-Jackson, T., Mazumdar, S., and Mills, A. (2015). Financial incentives in health: New evidence from india's janani suraksha yojana. *Journal of Health Economics*, 43:154–169.

Pradhan, R. H. A., Regmi, G., Ban, B., and Govindasamy, P. (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., Bryce, J., Barros, A. J., Berman, P., Bhutta, Z., Chopra, M., Daelmans, B., De Francisco, A., Lawn, J., Maliqi, B., et al. (2015). Countdown to 2015 and beyond: Fulfilling the health agenda for women and children. *The Lancet*, 385(9966):466–476.

Sankar, M., Chandrasekaran, A., Ravindranath, A., Agarwal, R., and Paul, V. (2016). Umbilical cord cleansing with chlorhexidine in neonates: a systematic review. *Journal of Perinatology*, 36(1):S12–S20.

Sazawal, S., Dhingra, U., Ali, S. M., Dutta, A., Deb, S., Ame, S. M., Mkasha, M. H., Yadav, A., and Black, R. E. (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(11):e837–e844.

Semrau, K. E., Herlihy, J., Grogan, C., Musokotwane, K., Yeboah-Antwi, K., Mbewe, R., Banda, B., Mpamba, C., Hamomba, F., Pilingana, P., et al. (2016). Effectiveness of 4% chlorhexidine umbilical cord care on neonatal mortality in southern province, zambia (zam-cat): a cluster-randomised controlled trial. *The Lancet Global Health*, 4(11):e827–e836.

Silverman, B. (1986). *Density Estimation*. Chapman and Hall.

Soofi, S., Cousens, S., Imdad, A., Bhutto, N., Ali, N., and Bhutta, Z. A. (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(9820):1029–1036.

Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S., and Wager, S. (2020). policytree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software*, 5(50):2232.

Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Hirshberg, D., Miner, L., Sverdrup, E., Wager, S., and Wright, M. (2021). *grf: Generalized Random Forests*. R package version 2.0.2.

Tibshirani, J., Athey, S., Sverdrup, E., and Wager, S. (2022). *Estimating ATEs on a new target population*. https://grf-labs.github.io/grf/articles/ate_transport.html.

USAID (2017). Chlorhexidine "navi" (cord) program [factsheet], scaling-up the use of chlorhexidine for umbilical cord care: Nepal's experience. Technical report, USAID.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Vaughan, K., Ozaltin, A., Mallow, M., Moi, F., Wilkason, C., Stone, J., and Brenzel, L. (2019). The costs of delivering vaccines in low-and middle-income countries: Findings from a systematic review. *Vaccine: X*, 2:100034.

Wang, H., Bhutta, Z. A., Coates, M. M., Coggeshall, M., Dandona, L., Diallo, K., Franca, E. B., Fraser, M., Fullman, N., Gething, P. W., 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(10053):1725–1774.

WHO (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.

WHO (2020). Advice on the use of masks in the context of covid-19: Interim guidance (5 june 2020). Technical report.

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

Appendix - For Online Publication

A Additional results

Figure A.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.

Table A.1: CHX introduction does not predict place of delivery or antenatal care

	Dependent variable:	
	Home Delivery	ANC visits
$t \geq$ CHX introduction	-0.032	0.222
	(0.029)	(0.139)
MDV	0.411	4.361
Observations	3,614	2,938

Notes: Both specifications are estimated as BJS (Borusyak et al., 2024) without controls. 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: 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 Manage asphyxia Referral of sick newborns	Rolled out from 10 to 41 districts between 2009-14
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 (1617), USAID (2017), Bhattarai (2017).

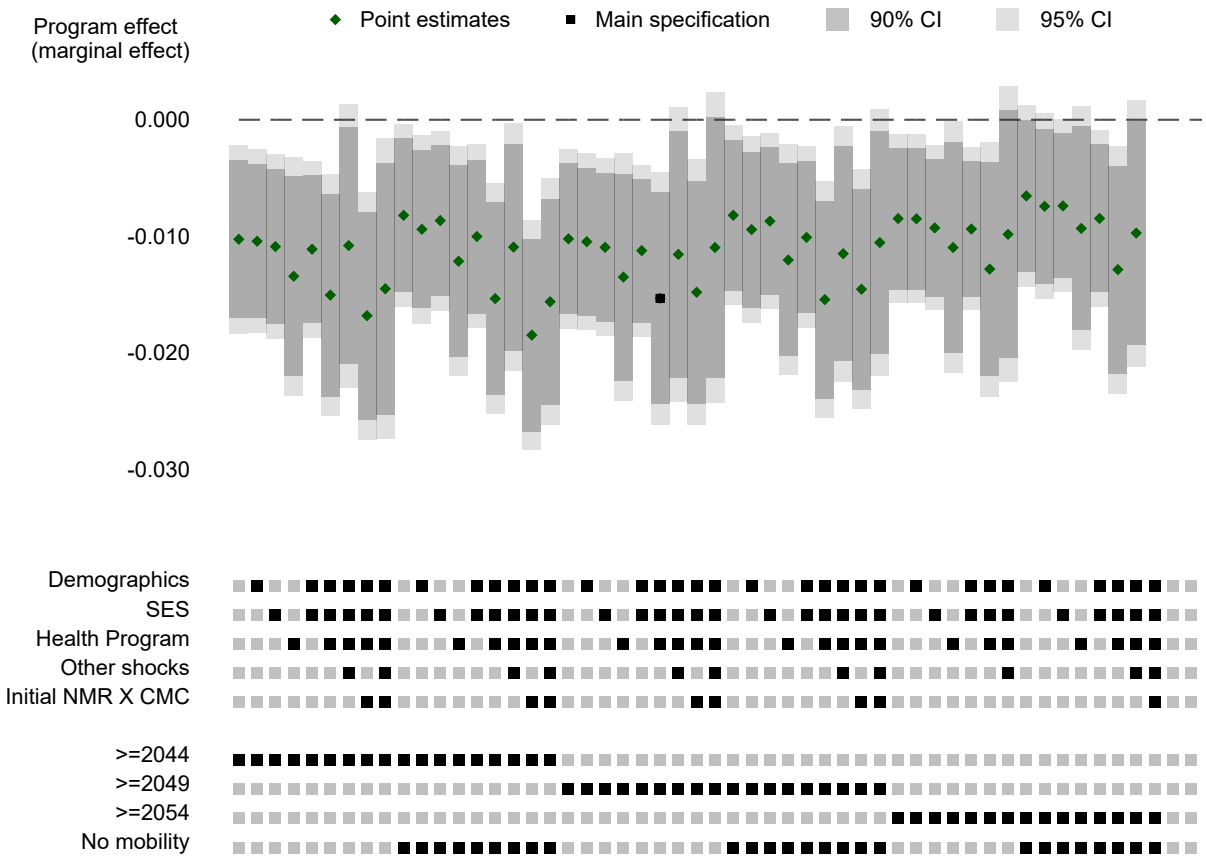


Figure A.2: Specification curve

Notes: All specifications are estimated using the BJS (Borusyak et al., 2024) imputation method. This chart shows estimates from running 54 different specifications defined by the combination of markers bellow the chart. The black square marker 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 rows ">=2044", ">=2049", and ">=2054" indicate the birth cohorts included, in Nepali calendar years. "No mobility" indicates that we restrict the sample to individuals who have not moved in the last eight years. The confidence intervals are based on standard errors clustered at the district level.

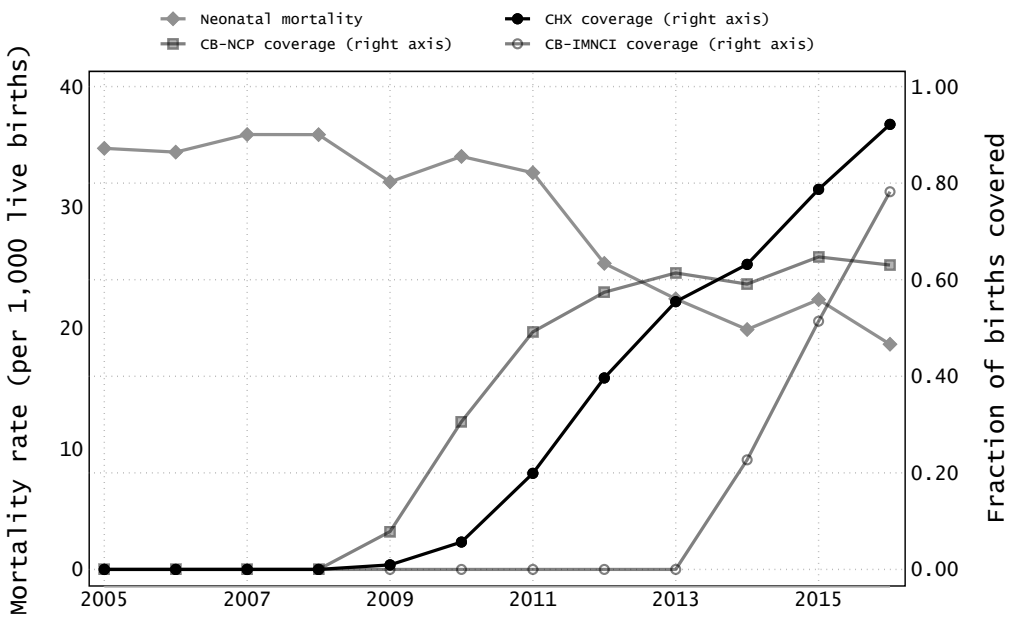


Figure A.3: Neonatal mortality, CB-NCP coverage, CB-IMNCI coverage, and CHX-NCP coverage

Table A.3: Variable means

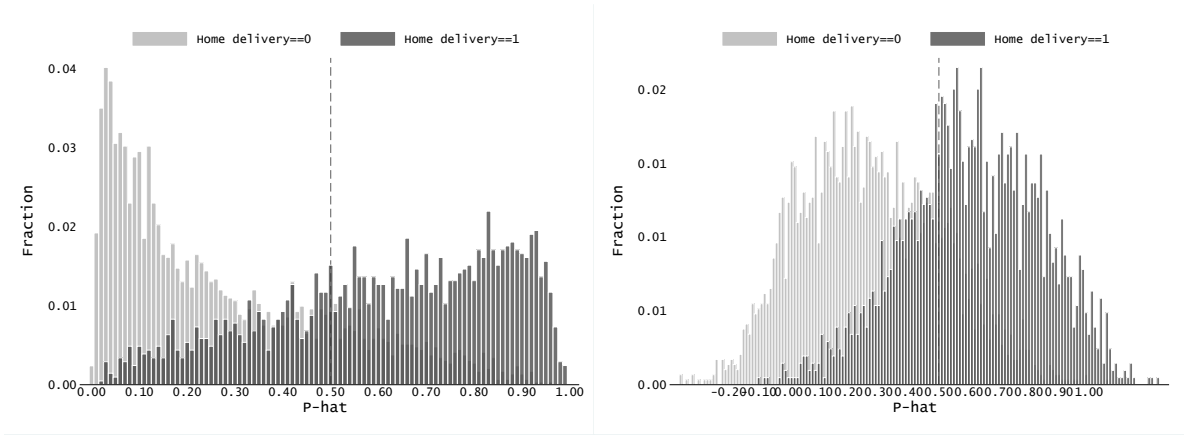
	Mean
A. Demographics and SES	
Female	0.48
First born	0.34
Second born	0.28
Third born	0.18
Parity four or higher	0.21
Mother age 15-19y	0.20
Mother age 20-24y	0.41
Mother age 25-29y	0.26
Mother age 30-34y	0.10
Mother age 35-39y	0.03
Mother age 40-45y	0.01
Ethnicity: hill brahmin	0.09
Ethnicity: hill 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.

Table A.4: Predicting home deliveries

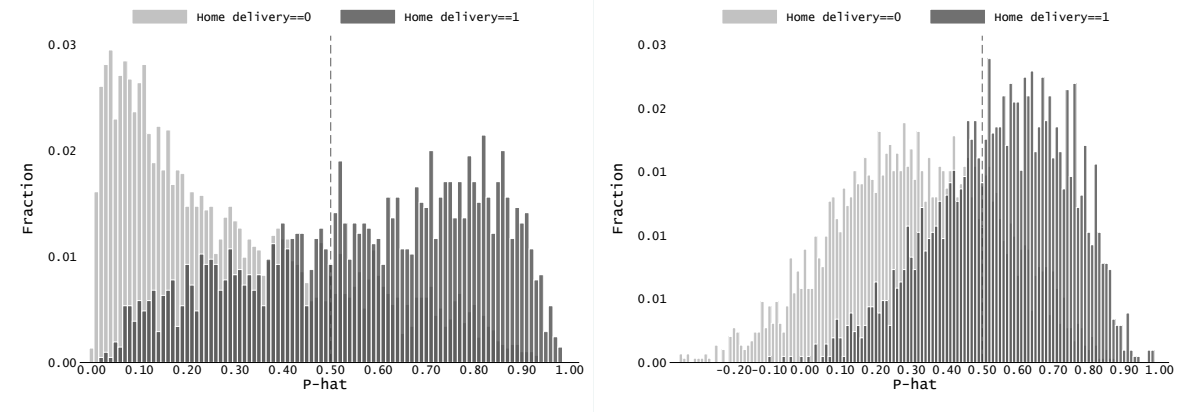
	DHS 2016		DHS 1996-2016	
	Logit	LPM	Logit	LPM
	(1)	(2)	(3)	(4)
Female	0.004	0.006	0.003	0.003
	(0.012)	(0.012)	(0.004)	(0.004)
First born	-0.248***	-0.267***	-0.184***	-0.203***
	(0.026)	(0.027)	(0.009)	(0.012)
Second born	-0.090***	-0.112***	-0.085***	-0.082***
	(0.019)	(0.022)	(0.007)	(0.008)
Third 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 et al., 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.4: Estimated propensity score for the prediction of place of delivery.

Notes: See Table A.4 for estimation details.

Table A.5: Effect of CHX-NCP on neonatal mortality using survey weights

	(1)	(2)	(3)	(4)
$t \geq$ CHX introduction	-0.012**	-0.014*	0.002	-0.020
	(0.005)	(0.007)	(0.010)	(0.014)
$t-1$	0.014	0.014	0.017	0.039
	(0.025)	(0.027)	(0.035)	(0.050)
$t-2$	-0.004	-0.008	0.010	-0.019
	(0.022)	(0.020)	(0.025)	(0.042)
$t-3$	-0.010	-0.010	0.005	-0.059***
	(0.016)	(0.016)	(0.011)	(0.022)
$t-4$	0.005	0.004	0.000	0.013
	(0.036)	(0.036)	(0.038)	(0.050)
$t-5$	-0.017	-0.021	-0.028	-0.022
	(0.025)	(0.026)	(0.019)	(0.052)
$t-6$	-0.002	-0.005	-0.000	0.023
	(0.022)	(0.022)	(0.013)	(0.060)
P-val	0.966	0.965	0.800	0.169
Observations	23,449	23,449	10,855	12,387
MDV	0.042	0.042	0.032	0.051
Sample	All	All	$P_H < 0.5$	$P_H > 0.5$
Controls	No	Yes	Yes	Yes
Month of birth FE	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes

Notes: P-value shows the p-value for the joint test that all lags are zero. MDV is the mean of dependent variable among untreated individuals. BJS is the Borusyak et al. (2024) imputation estimator as described in the main text. Bootstrapped standard errors clustered based on 500 iterations 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: Effect of CHX-NCP on neonatal mortality based on stacked predictions

	(1)	(2)	(3)	(4)
$t \geq$ CHX introduction	-0.010*	-0.015**	0.022*	-0.038***
	(0.005)	(0.007)	(0.011)	(0.013)
$t-1$	0.017	0.015	-0.010	0.044
	(0.026)	(0.026)	(0.031)	(0.052)
$t-2$	0.001	-0.004	0.015	-0.013
	(0.020)	(0.019)	(0.029)	(0.028)
$t-3$	-0.007	-0.009	0.013	-0.031
	(0.019)	(0.019)	(0.025)	(0.033)
$t-4$	0.002	-0.001	-0.013	0.051
	(0.033)	(0.032)	(0.026)	(0.076)
$t-5$	-0.010	-0.013	-0.020	-0.012
	(0.029)	(0.028)	(0.012)	(0.062)
$t-6$	0.015	0.009	0.012	0.003
	(0.025)	(0.025)	(0.029)	(0.047)
P-val	0.987	0.988	0.668	0.682
Observations	23,449	23,449	5,949	17,425
MDV	0.042	0.042	0.019	0.043
Sample	All	All	$P_H < 0.5$	$P_H > 0.5$
Controls	No	Yes	Yes	Yes
Month of birth FE	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes

Notes: P-value shows the p-value for the joint test that all lags are zero. MDV is the mean of dependent variable among untreated individuals. BJS is the Borusyak et al. (2024) imputation estimator as described in the main text. Bootstrapped standard errors clustered based on 500 iterations 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.7: CHX-NCP effects are independent of other neonatal health programs

	Dependent Variable: Neonatal Mortality				
	(1)	(2)	(3)	(4)	(5)
A. The effect of CHX					
CHX	-0.011***	-0.008*	-0.014***	-0.025***	
	(0.004)	(0.005)	(0.004)	(0.004)	
Observations	23,449	19,781	22,295	19,223	
B. The effect of CB-NCP					
CB-NCP					0.008
					(0.006)
Observations	23,465	23,465	23,465	23,465	20,321
B. The effect of CB-IMNCI					
CB-IMNCI					0.005
					(0.006)
Observations	23,465	23,465	23,465	23,465	20,303
Observations|CHX==1	3144	639	2290	302	
Sample	All	CB-NCP==0	CB-IMNCI==0	CB-NCP==0 & CB-IMNCI==0	CHX==0

Notes: All estimates are obtained with the BJS Borusyak et al. (2024) imputation estimator using the full set of demographic and SES 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. 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.

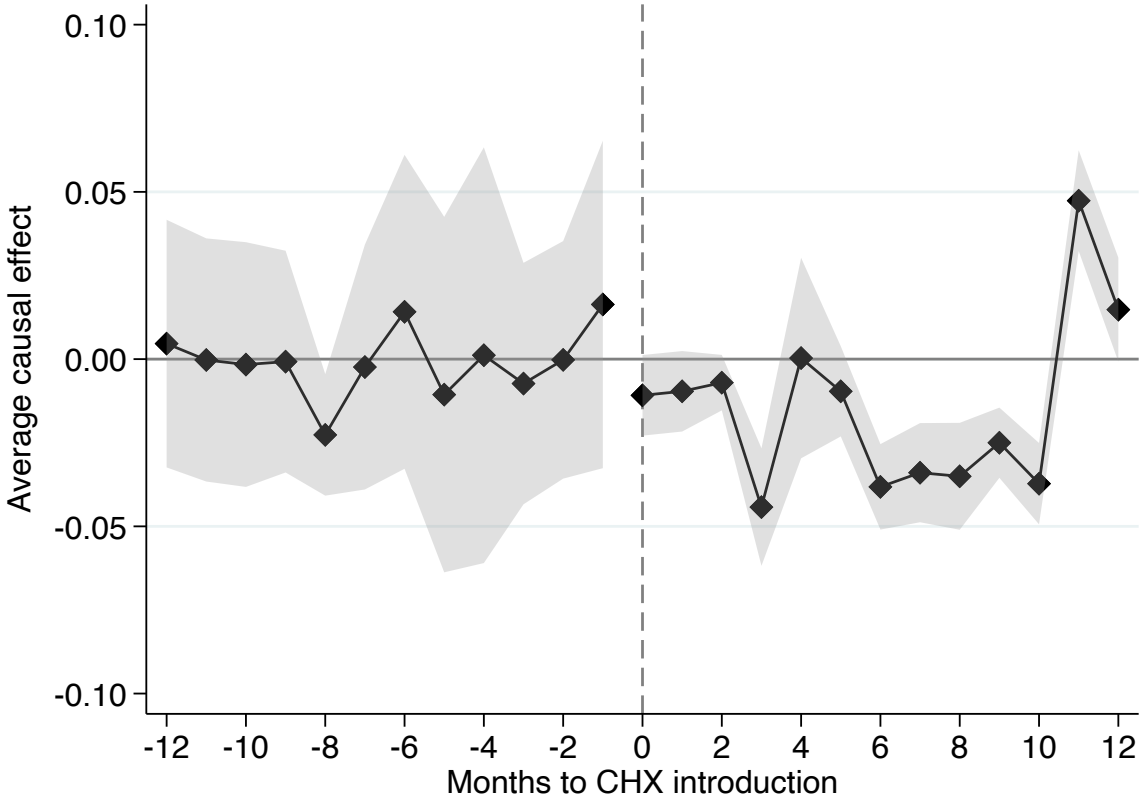


Figure A.5: Event study chart

Notes: The chart shows the time to event effects based on the BJS (Borusyak et al., 2024) imputation estimator. The shaded areas show the 95% confidence intervals based on standard errors clustered at the district level.

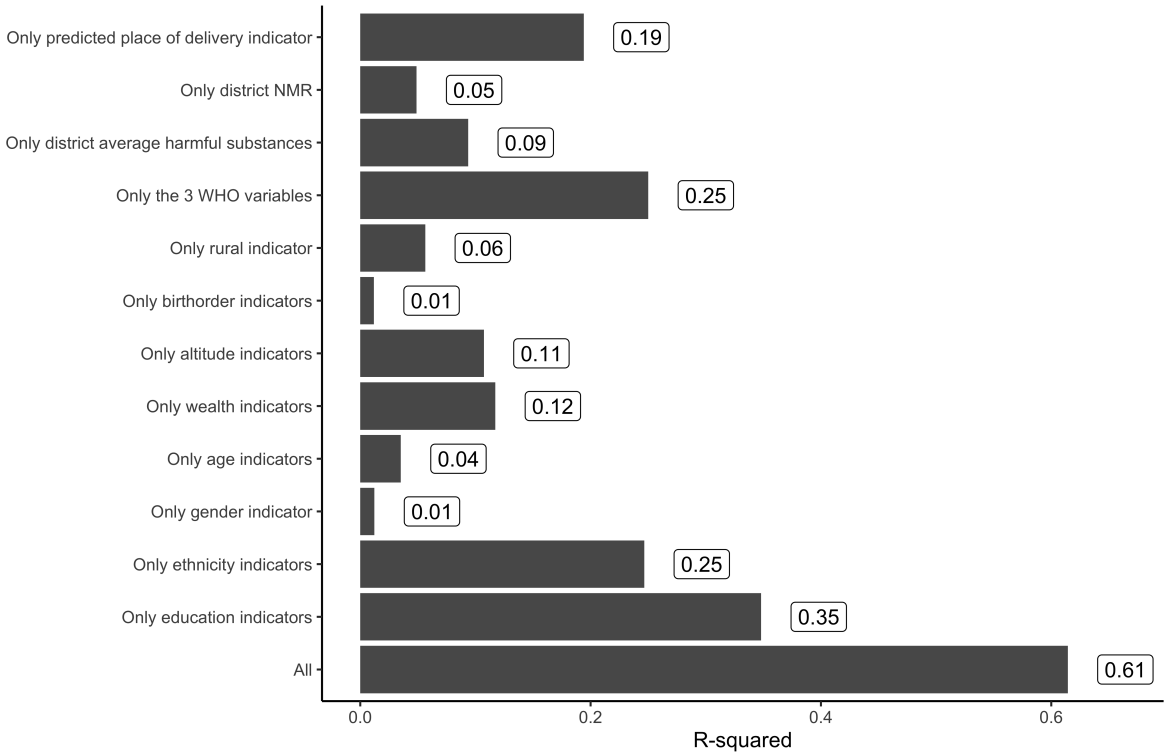


Figure A.6: Contribution to CATE variation

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.

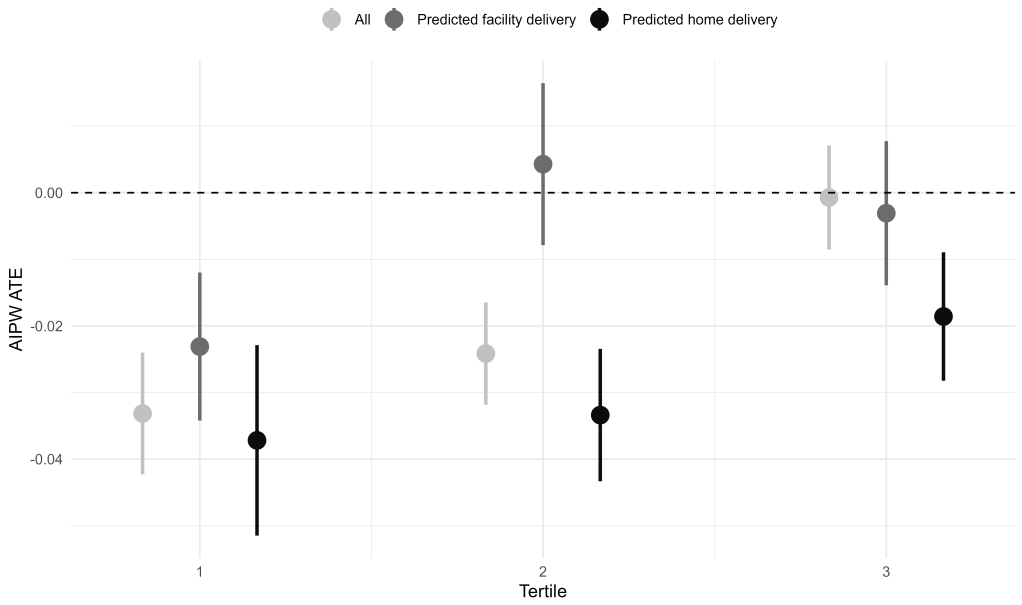


Figure A.7: Doubly robust ATEs by tertiles of CATEs

Notes: The Augmented Inverse-Propensity Weighted (AIPW) Average Treatment Effects are estimated for tertiles of the conditional average treatment effects shown in Figure 2. 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).

Table A.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.

Table A.9: 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 *** $p<0.01$.

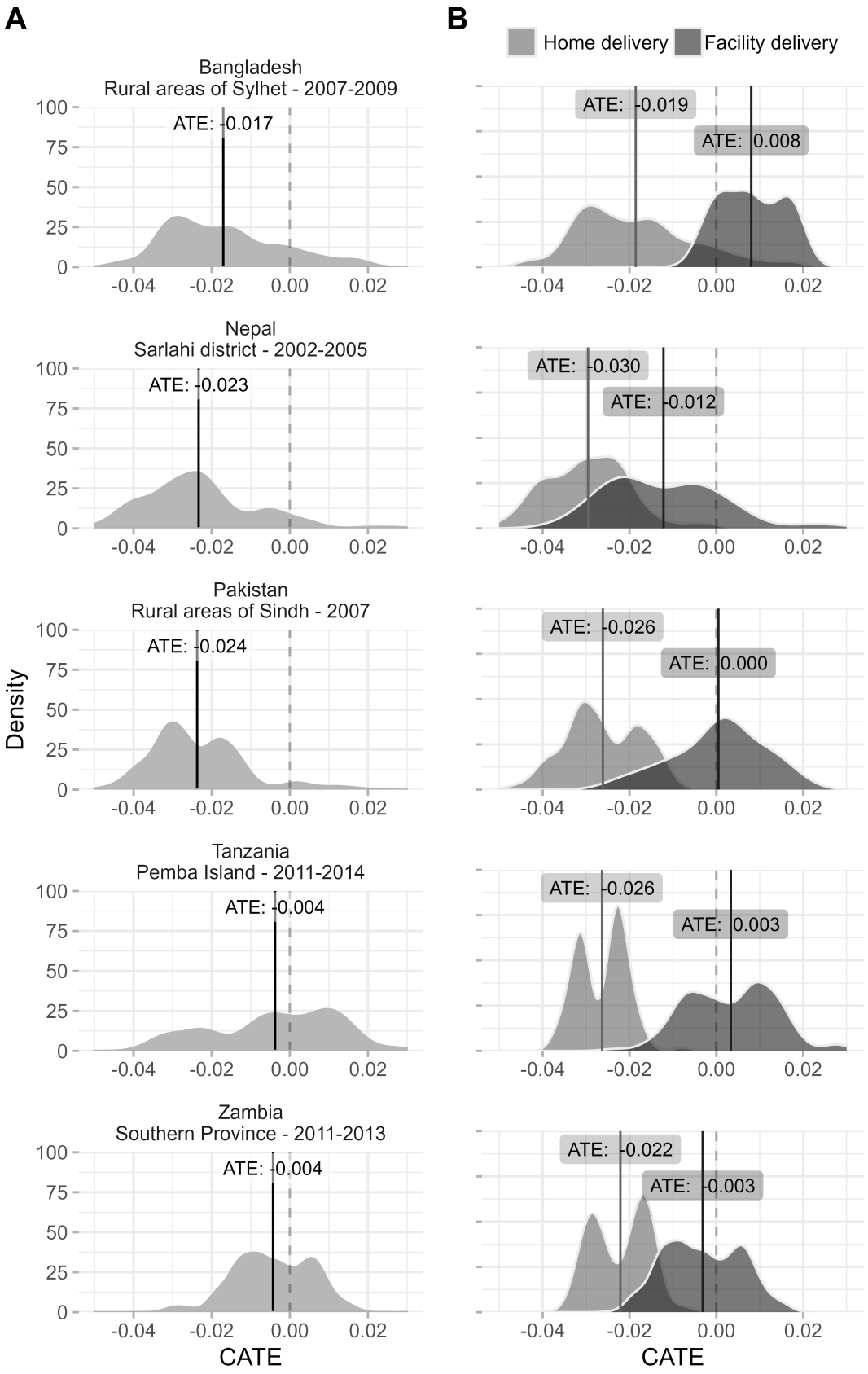


Figure A.8: Distribution of predicted CATEs across DHS samples matching the RCT sites

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

B Two-Way Fixed Effects (TWFE) Estimates

Table B.1: TWFE regression results - the effect CHX on neonatal mortality

	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	23,465	23,465	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 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 B.2: TWFE Balancing table. Dependent variable: CHX.

	(1)
Female	0.000
	(0.002)
First born	0.001
	(0.004)
Second born	-0.003
	(0.004)
Third born	-0.006
	(0.004)
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)
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)
Education: no education	0.016
	(0.011)
Education: primary	0.021*
	(0.012)
Education: secondary	0.014
	(0.010)
Wealth 0-20	(0.006)
Wealth 20-40	(0.006)
Wealth 40-60	(0.005)
Wealth 60-80	(0.006)
Altitude in 1st quintile	0.013
	(0.011)

Continued on next page

Continued from previous page

	(1)
Altitude in 2nd quintile	0.003
	(0.009)
Altitude in 3rd quintile	-0.007
	(0.007)
Altitude in 4th quintile	-0.000
	(0.007)
Program: CB-NCP	0.373***
	(0.069)
Program: CB-IMNCI	-0.063
	(0.082)
Constant	0.095***
	(0.033)
Observations	

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.

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

$$\mu_W = E[W|X = x] \tag{1}$$
$$\mu_Y = E[Y|X = x] \tag{2}$$

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):

$$\tau(X) = E[Y(1) - Y(0)|X = x] \tag{3}$$

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

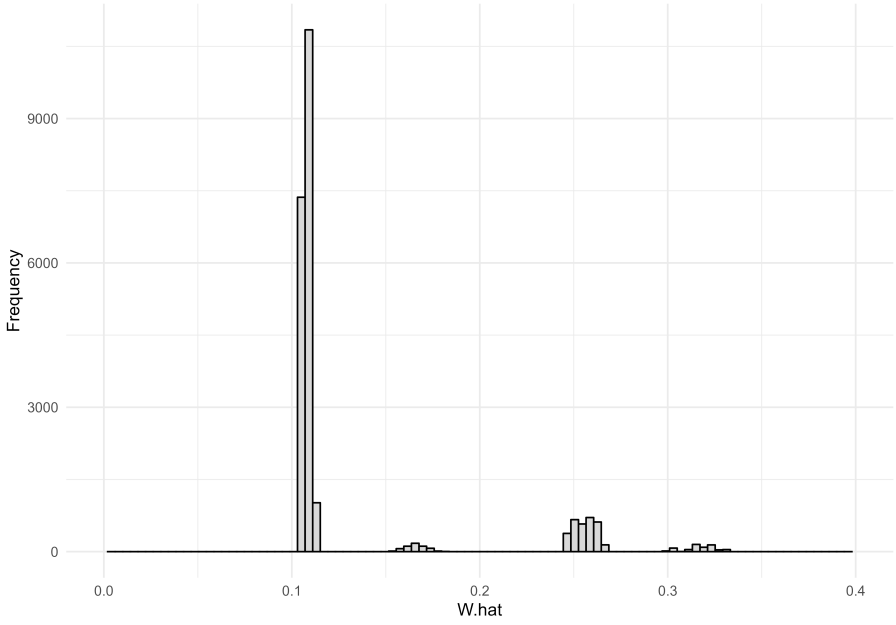
Setting	Value	Selection criteria
Number of trees	2000	Default
Clustering	District	Choice
Fraction of sample used to grow each tree	0.5	Cross-validation
Number of variables considered for each split	30	Cross-validation
Minimum size of a leaf node	5	Cross-validation
Fraction of sample used for splitting	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.
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.1: Propensity scores for causal forest



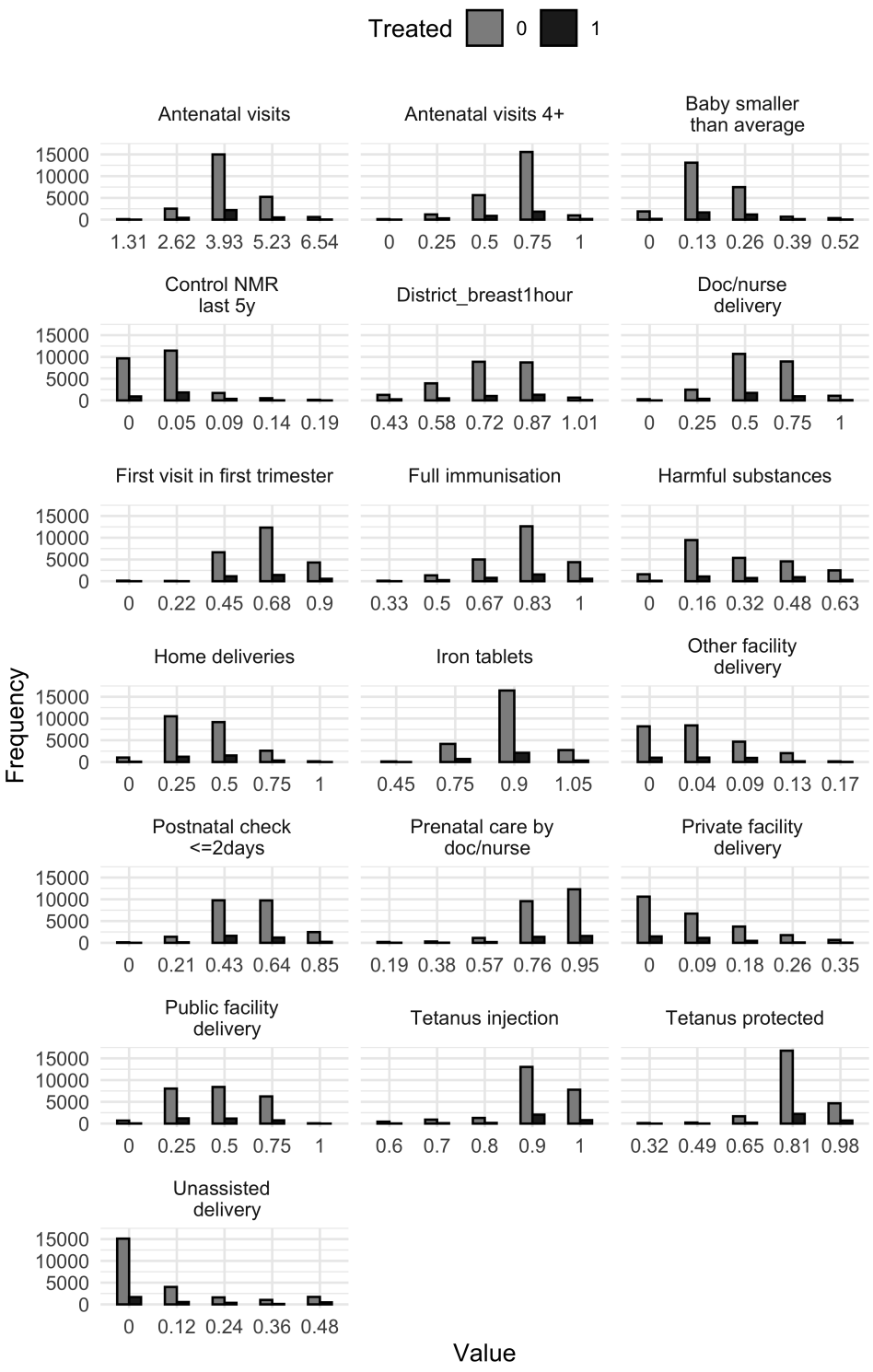


Figure C.2: Inverse-propensity score weighted distributions treated and control observations for individual covariates

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.

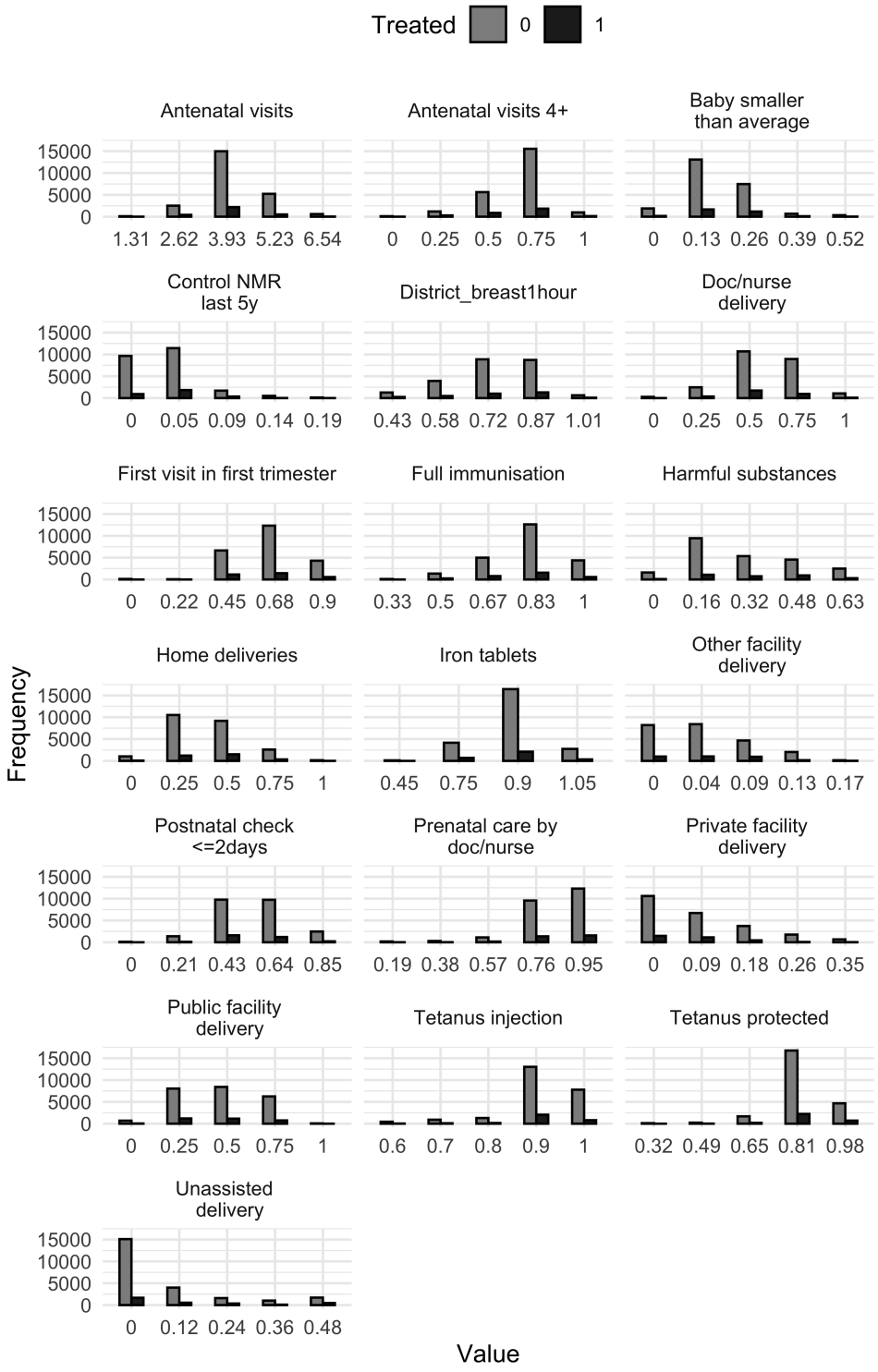


Figure C.3: Inverse-propensity score weighted distributions treated and control observations for district covariates

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.

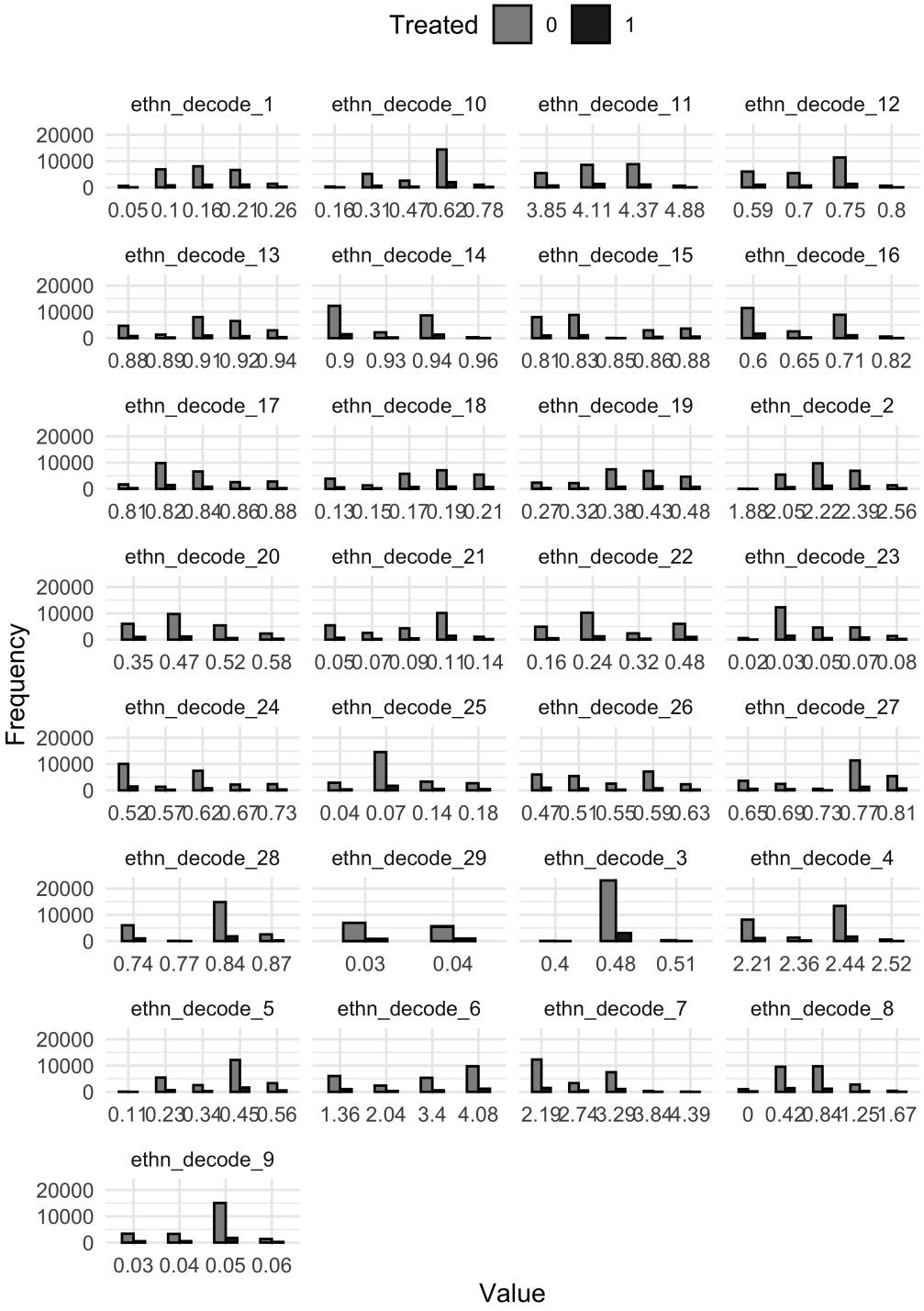


Figure C.4: Inverse-propensity score weighted distributions across treated and control observations for ethnicity demeaned covariates

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 A.9.

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 *** $p<0.01$.

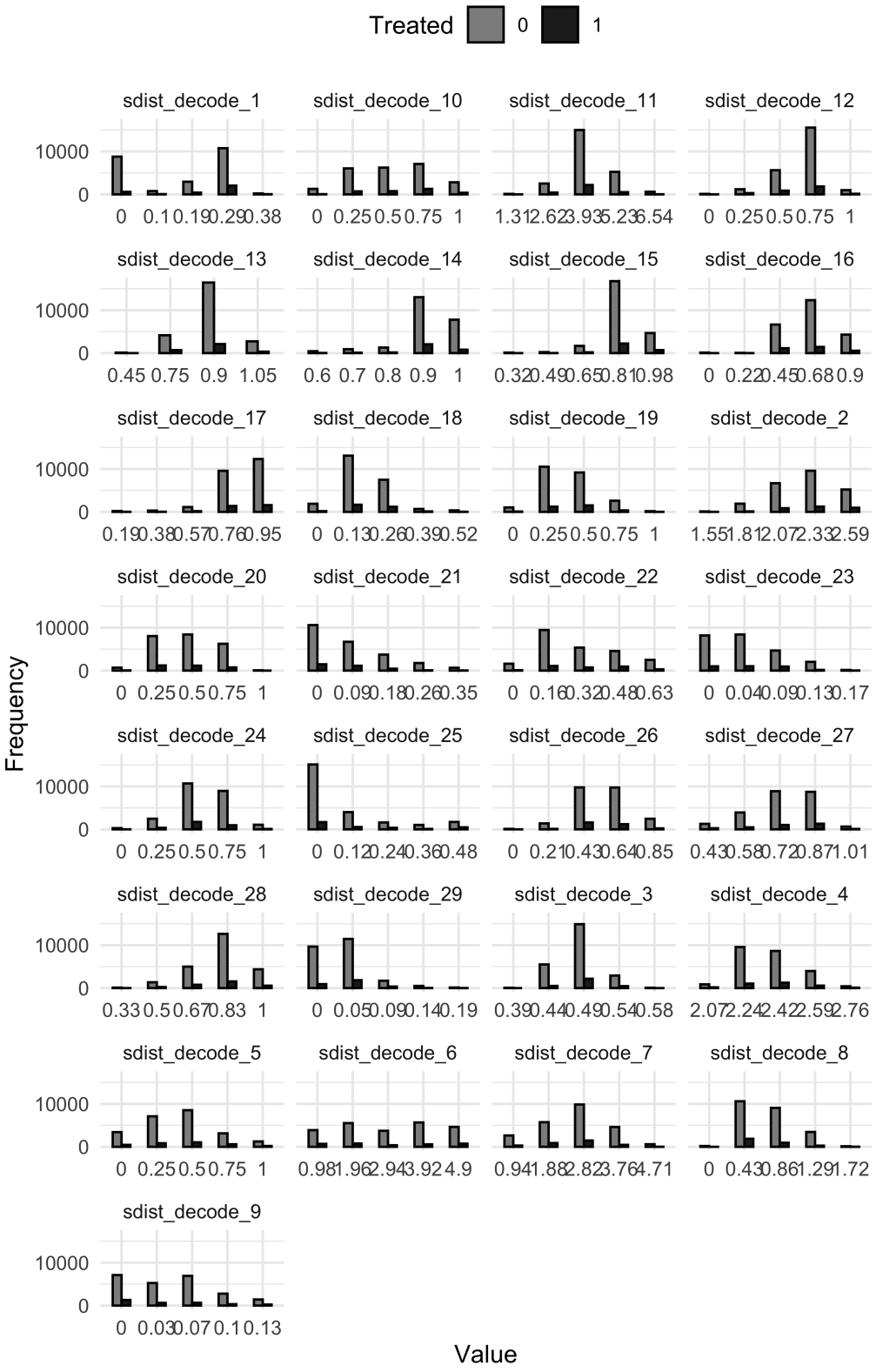


Figure C.5: Inverse-propensity score weighted distributions across treated and control observations for district demeaned covariates

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.