Article:
Kreif, Noemi, DiazOrdaz, Karla, Moreno Serra, Rodrigo et al. (2022) Estimating heterogeneous policy impacts using causal machine learning: a case study of health insurance reform in Indonesia. Health Services and Outcomes Research Methodology. pp. 192-227. ISSN: 1572-9400

https://doi.org/10.1007/s10742-021-00259-3

Health Services and Outcomes Research Methodology
https://doi.org/10.1007/s10742-021-00259-3

Estimating heterogeneous policy impacts using causal machine learning: a case study of health insurance reform in Indonesia

Noemi Kreif[1] · Karla DiazOrdaz[2] · Rodrigo Moreno-Serra[1] · Andrew Mirelman[1] · Taufik Hidayat[3] · Marc Suhrcke[1,4]

Received: 18 March 2021 / Revised: 23 September 2021 / Accepted: 28 September 2021

Abstract
Policymakers seeking to target health policies efficiently towards specific population groups need to know which individuals stand to benefit the most from each of these policies. While traditional approaches for subgroup analyses are constrained to only consider a small number of pre-defined subgroups, recently proposed causal machine learning (CML) approaches help explore treatment-effect heterogeneity in a more flexible yet principled way. Causal forests use a generalisation of the random forest algorithm to estimate heterogenous treatment effects both at the individual and the subgroup level. Our paper aims to explore this approach in the setting of health policy evaluation with strong observed confounding, applied specifically to the context of mothers' health insurance enrolment in Indonesia. Comparing two health insurance schemes (subsidised and contributory) against no insurance, we find beneficial average impacts of enrolment in contributory health insurance on maternal health care utilisation and infant mortality, but no impact of subsidised health insurance. The causal forest algorithm identified significant heterogeneity in the impacts of contributory insurance, not just along socioeconomic variables that we prespecified (indicating higher benefits for poorer, less educated, and rural women), but also according to some other characteristics not foreseen prior to the analysis, suggesting in particular important geographical impact heterogeneity. Our study demonstrates the power of CML approaches to uncover unexpected heterogeneity in policy impacts. The findings from our evaluation of past health insurance expansions can potentially guide the re-design of the eligibility criteria for subsidised health insurance in Indonesia.

Keywords Policy evaluation · Machine learning · Heterogenous treatment effects · Health insurance

✉ Noemi Kreif
Noemi.kreif@york.ac.uk

Extended author information available on the last page of the article

Published online: 09 November 2021

1 Introduction

Policymakers around the world are implementing health system policies to promote access to essential health care and to meet the health-related Sustainable Development Goals (Sachs 2012). Under limited budgets, maximising the impact of these policies on population health and health inequalities ideally requires evidence that explicitly acknowledges, and captures, how policy impacts might vary within a given population—i.e. the potential treatment effect heterogeneity. A key focus for health policies—and for the present paper—is heterogeneity in terms of observed *effect modifiers*, i.e. measured covariates that can modify the causal effect of a policy. In the typical setting where health policy impact evaluation is based on observational data, some evidence about effect modifiers can be obtained from subgroup analyses, by comparing the effects of interventions across different population groups, characterised, for instance, by their socio-economic status (Mackenbach 2003). However, impact evaluations of health policies tend not to present such comparisons, due to concerns that subgroup analysis, unless pre-specified, may produce spurious findings (Petticrew et al. 2012). Even when a treatment effect heterogeneity estimation *is* implemented, it typically involves including ad-hoc interaction terms in the models, thus necessitating parametric assumptions that are unlikely to hold (Hainmueller and Mummolo 2019).

Machine learning (ML) approaches are increasingly proposed as a way to pre-empt the criticism of arbitrariness, by estimating treatment effect heterogeneity based on a systematic exploration of the data (VanderWeele et al. 2019). These methods build on the growing body of methodological literature of 'causal machine learning' where supervised machine learning is used to help estimate causal parameters of interest (van der Laan and Rose 2011; Chernozhukov et al. 2018a). Some approaches aim to identify groups of individuals that have benefited most and least in terms of treatment effects (e.g. Imai and Strauss 2011, Chernozkhukov et al. 2018b), while others aim to flexibly capture how treatment effects vary according to observed covariates, by estimating a so-called 'conditional treatment effect function' (CATE) in a data-adaptive manner (e.g. Kunzel et al. 2019, Athey et al. 2019a, Fan et al. 2020). What is common to these approaches is that they combine the flexibility of ML methods with the rigour of semi-parametric statistical theory, resulting in valid inferences after data-adaptive estimation.

In this paper, we focus on a particular method, the so-called 'causal forests' (Wager and Athey 2018, Athey et al. 2019a, Nie and Wager 2021), to estimate individual treatment effects which can be aggregated to provide average causal treatment effects estimates for subgroups of interest. The first major benefit of this approach compared to traditional methods is in that researchers do not need to specify subgroups for stratified analysis or impose parametric assumptions about interactions. The second advantage is that it provides statistical tests to assess whether there is significant heterogeneity in the treatment effects that is explained by observed covariates, along with offering an indication of the variables that are most strongly associated with this heterogeneity.

Previous studies have demonstrated the good performance of the causal forest estimator and its modifications in simulation studies (e.g. Lechner 2018, Knaus et al. 2021, Nie and Wager 2021, Fan et al. 2020). A growing number of studies have been using causal forests for programme evaluation, typically in labour economics (Davis and Heller 2017, Knaus et al. 2020), but also for the evaluation of health interventions that have been randomised (e.g. Scarpa et al. 2019). To our knowledge, the approach has not been

applied in the setting of health policy evaluation where there is no randomisation and hence, statistical approaches are required to account for confounding.

In this paper, we demonstrate that the causal forests approach can provide beneficial information for health policymaking decisions aimed at improving overall health and reducing health inequalities. Specifically, we apply causal forests to explore effect heterogeneity in two types of public health insurance programmes in Indonesia: subsidised health insurance targeting the poor and the near-poor, and contributory health insurance for employees of the formal sector. We use the sequential implementation of health insurance that preceded the establishment of the unified National Health Insurance programme (Jamima Kesehatan National (JKN)) in 2014 as a natural experiment to investigate how changes in health insurance status have influenced health outcomes and health care utilisation. We first focus on infant mortality as the health outcome of interest, which is arguably quite sensitive to changes in access to health care services (Currie and Gruber 1996, Dow and Schmeer 2003). We would expect health insurance to reduce infant mortality through its effects on health care utilisation by insured mothers, as births attended by a skilled health professional have been found to be a predictor of reduced infant mortality in the neonatal stage (Lawn et al. 2005). We first estimate average causal treatment effects for these outcomes, using data from the Indonesian Family Life Survey (IFLS) (Strauss et al. 2004, 2009, 2016), a rich and high-quality longitudinal survey of Indonesian individuals and households that allows controlling for observed confounders of the causal relationship between health insurance as the 'treatment' and health care utilisation and health outcomes as the effects.

In light of the notable geographical, ethnic, and economic disparities within Indonesia, it is expected that average policy effects mask important within-country heterogeneity in the effect of health insurance programmes. For the optimal targeting of interventions, health policymakers need to know how the impact of health insurance varies across different subgroups, in particular for those groups most vulnerable in terms of disease burden and access to health care (Lagomarsino et al. 2012): mothers with low education, those in the bottom socioeconomic quintiles, and those living in remote, rural communities. We therefore also estimate individual treatment effects using the causal forests approach and aggregate these to estimate subgroup-average treatment effects. We do so both for ex-ante specified subgroups following traditional practice, and via a data-driven ML approach that characterises those variables most associated with heterogeneity.

This paper makes two main contributions. First, by evaluating the impact of health insurance on health care utilisation, we demonstrate the value of using novel causal ML methods for health policy evaluation. In particular, this is the first study that uses the causal forest approach in the context of a health policy evaluation to characterise the drivers of treatment heterogeneity. We highlight the challenges of conducting such evaluations, specifically the need to account for confounding due to observed covariates that affect individual participation in the health policy and outcomes, under the strong assumption of no unobserved confounding. Second, we offer novel and policy-relevant empirical evidence on the population-level effects of health insurance coverage, by characterising the heterogeneous impacts of public health insurance expansions on maternal health care utilisation and infant health, focusing on the specific context of Indonesia.

In the following sections, we first present the institutional setting in Indonesia (Sect. 2.1), briefly review the literature on the impact evaluations of health insurance (Sect. 2.2), and present the data used in the study (2.3). Then we describe the methods (Sect. 3), with a dual focus on the theory and practical implementation of causal forests. The results are presented in Sect. 4, followed by a discussion of the findings and future

avenues of research (Sect. 5). The online appendices include additional Tables (Appendix B) and Figures (Appendix C), as well as software code (Appendix D) to implement the causal forests analysis presented.

2 The evaluation of the national health insurance expansion in Indonesia

2.1 Institutional setting

With an estimated population of over 270 million in 2019,[1] Indonesia is the fourth most populous country in the world. Total health spending was 3.1% of gross domestic product (GDP) in 2016 (World health statistics 2019), with a relatively small share of total health expenditures being publicly funded (39%) (Mahendradhata et al. 2017). While on average, health indicators have improved significantly over the last decades—life expectancy rising from 63 to 71, and infant mortality falling from 41 to 26 deaths per 1000 live births, between 1990 and 2012 (Mahendradhata et al. 2017)—there remain considerable health inequalities (Agustina et al. 2019). To address unmet health care needs, and high out-of-pocket and catastrophic health spending, Indonesia launched an ambitious health system reform in 2014, the JKN, comprising a wide range of policies, including a unified benefit package and premium subsidies for the poor. The JKN reform was preceded by a series of health insurance expansions programmes—the focus of our study—starting from the 1990s. We briefly review these over our study period (2000–2014) (See Appendix B Table 1) describing the main contributory and subsidised health insurance schemes.

Historically, health insurance in Indonesia was available as contributory schemes, known as *Askes and Jamsostek*, for those employed in the formal sector and their family dependants (Achadi et al. 2014).[2] For poor households that were not eligible for these health insurance programmes, from 1994 a *Health Card* programme provided free basic health care at public health facilities (Johar 2009). The *Askeskin* scheme, established in 2005, was the first national, subsidised health insurance programme, basing eligibility on a combination of geographic and individual-level criteria (Sparrow et al. 2013). The insurance scheme covered a comprehensive package of health services (outpatient care and inpatient care, mobile health services, immunisation and medications), with the premium fully subsidised by central government (Sparrow et al. 2013). The scheme left a large group of households without health coverage, i.e. those not poor enough to be eligible but also not having access to contributory health insurance in the formal sector. In 2008 it was re-organised, and the resulting *Jamkesmas* scheme expanded the eligible population, targeting the poor and 'near poor', based on a combination of means testing (using 14 assets recorded in a National Poverty Census Survey indicators) and local government eligibility criteria (Harimurti, et al. 2013). However, not all households eligible for the programme possessed a membership card due to perceived stigmatisation from health care providers and concerns about long waiting times (Harimurti et al. 2013). Despite the means testing, a significant "leakage" occurred, resulting in households in higher income

[1] See, http://worldpopulationreview.com/countries/indonesia. Retrieved 2019–08-07.

[2] *Askes* was a mandatory health insurance programme for active and retired civil servants, and military personal, with a contribution of 2% from payroll salary or pension (Thabrany 2001). *Jamsostek* was an optional social security scheme for private employees, with a 3–6% salary contribution (Hidayat et al. 2004).

quantiles also receiving free health insurance (Harimurti, et al. 2013). While in principle, *Jamkesmas* provided a comprehensive package, in reality, the availability of services was limited, especially in rural areas, thereby contributing to large geographic inequalities in access (Harimurti, et al. 2013). To compensate for the large gaps in insurance status, district governments provided decentralised health care financing schemes offering subsidised health insurance, known as *Jamkesda* (Sparrow et al. 2017).[3] In 2014, all health insurance schemes were absorbed into a single national health insurance scheme, JKN, aimed at continuing to expand health insurance coverage to the total population, with the original stated objective to achieve universal health coverage by 2019.

2.2 Related literature on the impact of health insurance on health outcomes and utilisation

Evidence that health care utilisation increases as a result of providing health insurance is growing (e.g. Trujillo et al. 2010, Yilma et al. 2015). While country-level analyses have found that increasing health coverage through national-level health spending is beneficial for health, particularly within a system of risk-pooling (Moreno-Serra and Smith 2015), evaluations using less aggregated (subnational or individual level data) provide more mixed findings (Acharya et al. 2013, Erlangga et al. 2019a). Few studies found conclusive evidence of a health-improving impact (Wang et al. 2009, Mensah et al., 2010), with some finding either no evidence of a positive impact (e.g. Dow and Schmeer 2003, Chen and Jin 2012), or even adverse impacts (Fink et al. 2013). For Indonesia, quantitative impact evaluations of the different stages of health insurance expansions also reveal a mixed picture. Johar (2009) finds no evidence that the Health Card programme increased health care utilisation among the poor, and attributes this finding to inelastic demand amongst the recipients. Evaluations of the *Askeskin* programme found some increase in financial protection (Aji et al. 2013), but only a modest impact on health care utilisation among the beneficiaries (Sparrow et al. 2013). An evaluation of the early implementation of the JKN programme (between 2007 and 2014) found that while contributory health insurance increased both inpatient and outpatient utilisation, subsidised health insurance only increased inpatient utilisation, and to a smaller extent (Erlangga et. al. 2019b).

There are various reasons why impact evaluations of health insurance expansions may not always demonstrate measurable improvements in health outcomes. First, establishing the causal effect of health insurance programmes is challenging due to confounding: those receiving health insurance programmes are systematically different from those not receiving it. Correcting for such confounding requires either exploiting 'natural experiments' through quasi-experimental econometric techniques (Wagstaff 2010) or measuring enough variables to adjust for these differences. In addition, the availability of health insurance may affect specific sub-populations differently. Understanding this differential impact is crucial to inform health policymaking in Indonesia, where a large segment of the

[3] As of 2013, around 12% of the population was estimated to have been covered by the *Jamkesda* schemes (32 million covered in 2013 out of a population of 252 million in that year).http://gnhe.org/blog/wp-content/uploads/2015/05/GNHE-UHC-assessment_Indonesia-1.pdf, http://gnhe.org/blog/wp-content/uploads/2015/05/GNHE-UHC-assessment_Indonesia-1.pdf. A further subsidised scheme (*Jampersal*) aimed to cover uninsured pregnant women and newborns was launched in 2011 with the specific aim of filling the gap in delivery services for maternal and neonatal health (Achadi, Achadi et al. 2014), and this insurance status was universal and not means tested.

population remains uninsured. Erlangga et al. (2019b) looked at different impacts by subgroups, and found that the lowest income quintiles did not benefit from improved in-patient utilisation, with no effects in areas with low density of healthcare facilities. Anindya et al (2020) identified significant impacts of the JKN programme on maternal health care utilisation (skilled birth attendance, institutional deliver, antenatal care visits), and found that mothers from lower socioeconomic quintiles and more deprived regions benefitted more from health insurance.

2.3 Data

The IFLS household dataset includes respondents living in 13 out of the 27 Indonesian provinces, initially using the sampling frame of the 1993 national household socioeconomic survey (Survei Sosial Ekonomi Nasional—Susenas) from the Central Bureau of Statistics.[4] The first round was in 1993 (IFLS1), covering 7,224 households. Subsequent rounds were conducted with the same respondents and their new household members in 1997 (IFLS2), late 1998 (IFLS2 + with a 25% subsample), 2000 (IFLS3), 2007/2008 (IFLS4) and 2014/2015 (IFLS5). In order to exploit temporal variation in the availability of the health insurance schemes, we use the IFLS waves which were collected in the pre-*Askeskin* period (IFLS 3), in the pre-*Jamkesmas* period (IFLS4), and in the post-*Jamkesmas* period, covering the start of the *JKN* programme up to 2015 (IFLS5). We refer the reader to Appendix B Table 1 for the links between the various policy reforms and survey waves, and the construction of the analytical dataset, which we describe in some detail below. We constructed a birth-level dataset using the complete pregnancy histories available for women aged 15 to 49, including the date of birth of each child, whether the child is still alive, and if not, the age at death. Restricting the recall period to 6 years to minimise recall bias, we collated births between 2002 and 2007 from IFLS4, and between 2008 and 2014 from IFLS5. We use child birth as the unit of analysis throughout the paper.

We defined two treatment groups and a control group by assessing the mothers' insurance status and the type of health insurance in the year of each child birth. The first treated group, referred to as subsidised insurance, consists of births where in the year of the child birth, a mother reported enrolment in one of the following insurance schemes: *Health card (2002–2007)*, *Askeskin (2005–2007), Jamkesmas (2008–2014), Jamkesda (2008–2014)* or *JKN* (2014), where the years in parentheses indicate the years that a given type of insurance appears in our data sets. The second treatment group—contributory insurance—was defined as births where, in the year of the child birth, a mother reported enrolment in the *Askes* or *Jamsostek* or other employer provided insurance programmes, and these insurance types can be found throughout the study years.[5] Finally, 'uninsured' is defined as a birth where the mother has not reported any subsidised or contributory insurance in the year of the given child birth. Uninsured births, again, can be found throughout the study period. Those births for which a mother reports having both subsidised and contributory insurance

[4] The sample is stratified in provinces and rural–urban areas within provinces. There are some randomly selected enumeration areas (EA) within the strata and households within enumeration areas. The aim of the selection of the provinces was to be cost-effective given the size of the country without neglecting the representation of the population, the ability to illustrate the cultural and socioeconomic diversity of Indonesia. In addition, the survey was designed to have a panel structured at the household level.

[5] We have recoded mothers who have reported no health insurance if they were eligible based on the insurance status of their spouse or household head, and they reported being insured.

were excluded from the analysis,[6] as such double insurance, while it did occur in practice, was not formally allowed.[7]

We emphasise that we do not treat this dataset as longitudinal in the subsequent statistical analysis as individual births are the unit of analysis. We do allow both for women who only had one birth over the study period and for women who had repeated births, and allow these births to enter any of the three groups. We also use the time dimension of our data in several ways. First, due to the nature of the dataset, we are able to tell whether in the year of a given child birth, a woman is reported to be insured or not, and with which insurance type. Second, over time, there is an increasing number of insured women and births, due to the health insurance expansion. Third, we use year (birth cohort) dummies to allow for unobserved common time trends, and province fixed effects to allow for unobserved province level factors (see below).

We linked outcome information on births to insurance information of the mother, as well as her demographic information, her household and community. In line with conventions, the death of a child is classified as "infant death" if the death occurred before the first birthday. Our "skilled birth attendance" variable indicates whether the birth has been attended by either a midwife or a doctor, regardless of place of delivery (both in and out of hospital).

A common concern in the evaluation of health insurance programmes is that individuals self-select into health insurance, based on potential gains unobserved by the researcher (Currie and Gruber 1996, Wagstaff 2010). In the Indonesian setting, this problem might manifest in two ways. First, the eligibility assessment was complex, based on geographical and household level criteria that may not be fully captured by the information in our dataset. Second, insurance take-up was ultimately voluntary, leaving the possibility that those who are somewhat better off were less likely to opt for subsidised insurance, due to the perceived stigma and potentially lower quality of services, compared to those obtained in the private sector. To take this into account, we exploited the variation in the expansion of both health insurance schemes, across provinces and over time. Seeking to minimise bias due to selection on unobserved factors that could compromise comparisons of outcomes between insured and uninsured individuals, we adjust for a rich set of household-, individual- and community-level characteristics to approximate the institutional eligibility rules and process of selection into the health insurance schemes. We also control for province-specific effects that capture unobserved confounding factors that are common within provinces and time-invariant (see Sect. 3 for the formal assumptions).

When selecting variables to control for confounding, we focussed on the characteristics of the mothers, households and communities, which contribute to the eligibility and enrolment in the health insurance schemes, and which are considered to also be independently associated with health care utilisation or infant mortality.[8] Following previous studies (Dow and Schmeer 2003, Shrestha 2010), we included mother's education (categorised as primary, senior, secondary, and university), mother's literacy (ability to write a letter in

[6] Such double insurance constituted 5% of our overall sample.

[7] Because of the universal availability of the Jampersal programme, both insured and uninsured mothers may have reported "having" Jampersal. Hence, we did not include it in the definition of the health insurance variable.

[8] Variables that are only expected to affect enrolment in the subsidised or contributory health insurance schemes, but are unlikely to have a direct effect on infant mortality (or be affected by infant mortality themselves), were assessed as candidates for instrumental variables. However, none of them were found strong and valid at the same time.

Indonesian), age at birth, sex of the child, birth order of the child, and whether a household was urban or rural. To capture the means-testing eligibility criteria of the subsidised health insurance programmes (Johar 2009), we construct an asset index (O'Donnell et al. 2008), using principal component analysis (PCA) to classify households into wealth quintiles based on asset ownership and household characteristics (see Appendix A for specific variables used in the PCA). We also created a binary variable from the self-reported health of the mother (1 if good or excellent, 0 otherwise). Further indicators of socioeconomic deprivation are considered, in particular we capture participation in three major social assistance programs: a subsidised rice ("Raskin") programme, an unconditional cash transfer programme, and a "poor card" programme. We also added a variable capturing whether the household had been seriously affected by a natural disaster in the preceding five years. We also capture whether community members have access to a village midwife, a birth clinic, a hospital, a public health centre or private health care providers.[9] Indicators for province of residence for the mother, at the time of the survey are also included, in an attempt to control for unobserved confounding at the province level (e.g. in terms of worse access to health care in turn leading to worse outcomes). A year of birth variable seeks to control for time trends affecting changes in infant mortality (e.g. technological innovations in neonatal intensive care), that may have coincided with the gradual expansion of health insurance. For the pre-specified subgroup analysis, we selected three widely used socioeconomic proxies to be able to assess the impact of insurance for those most vulnerable in terms of disease burden and access to health care (Lagomarsino et al. 2012): mothers with low education, those in the bottom socioeconomic quintiles, and those living in remote, rural communities.

3 Methods

3.1 Notation and estimands

We are interested in estimating causal effects of a mother being enrolled in one of two health insurance types (subsidised or contributory) versus no health insurance, on one health outcome (infant mortality) and one health care utilisation outcome (skilled birth attendance) for a given birth, henceforth referred to as a unit. We conduct these analyses separately, and use a common notation Y for both outcomes, and W for both health insurance schemes. Denote the potential outcome for a given birth i by $Y_i(w)$, with $w \in 0,1$. The individual treatment effect is the difference between the two potential outcomes, $\tau_i = Y_i(1) - Y_i(0)$. Our main identifying assumption is that of unobserved confounding, requiring that$Y(1), Y(0) \perp W|X$, or that after adjusting for the sufficient variable set X the

[9] In order to strengthen our causal assumptions, we require that the observable variables included in our regressions are measured before a child is born, but also before a decision about enrolling in health insurance has been made. Hence, for births recorded in IFLS4 (2002–2007), we take measurements of individual and household level variables from IFLS3 (2000). Similarly, for births recorded in IFLS5 (2008–2014), we measure individual and household level variables from IFLS4 (2007/2008). For individuals who did not have a measurement in the previous wave, because they were not part of the IFLS sample yet (approximately 30% of the total sample), we take the current measurements as proxies. We follow a similar logic for missing household level covariates in the case of new people entering the IFLS sample (5% of the total sample missing). We construct indicator variables for these cases of missingness and include them in our analyses.

potential outcomes are independent of the observed insurance status$W$. Further assuming no interference, consistency,[10] and overlap,[11] we can identify the estimands of interest, the average treatment effect (ATE), the average treatment effect on the treated (ATT), and the average treatment effect on the controls (ATC).

These three estimands answer different policy evaluation questions. The ATE $=E[Y(1) - Y(0)]$ contrasts the potential outcomes in a world where everyone has a given insurance, and where no one has insurance, and takes the average of these causal contrasts over the pooled population of the uninsured and the insured. The ATT, defined as $E[Y(1) - Y(0)|W = 1]$ answers the question: how much did those who had a certain insurance type benefit from having that health insurance, compared to not having insurance? Finally, the ATC, defined as $E[Y(1) - Y(0)|W = 0]$ aims to answer the question: how much the uninsured would have benefitted from having a given insurance type? The ATC also allows us to contrast the impacts of the two insurance types, as the population for whom the benefits are calculated is held constant at the uninsured, representing a large portion of the population in Indonesia in the study period, including subgroups from all socioeconomic quintiles.

Beyond population average treatment effects, there is interest in the conditional average treatment effect (CATE) defined as

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

The CATE can be conceptualised as a function that takes a combination of observed covariates that are assumed to modify the effect of the treatment, at a selected covariate profile x, and outputs a treatment effect that corresponds to this covariate profile. In the context of health insurance, we expect that a range of the observed covariates can modify the treatment effect, beyond the socioeconomic factors listed above. The geographical availability of health services may be one such example.

3.2 Estimation of average treatment effects using a parametric double-robust approach

As a starting point we assume a linear predictor for each outcome of interest with identity link:

$$Y_i = \beta X_i + W_i\tau + \varepsilon_i \tag{1}$$

where Y_i indicates (a) the survival status of infant i born in year t at 12 months after birth (b) whether the birth was attended by health professional, and the vector $X_i = (Z_{mt}, Z_{ht}, Z_{ct}, \delta_p, \alpha_t)$ includes several components: Z_{mt} denotes the characteristics of the mother (e.g. education), Z_{ht} captures household characteristics (e.g. household asset quintile, social assistance), Z_{ct} captures community level variables (e.g. availability of hospital or birth clinic in the neighbourhood, or availability of a village midwife in the year of

[10] The no interference assumption requires that a unit's outcome is not affected by the treatment received by other units (Tchetgen Tchetgen and VanderWeele 2012). The consistency assumption requires that the observed outcome corresponds to the potential outcome under the observed treatment (VanderWeele 2009).

[11] The overlap assumption requires that there must be a positive probability to be enrolled in a given health insurance programme, but this probability must be strictly smaller than 1: no covariate combination should fully determine a mother's insurance status.

birth), δ_p are the effects of unobserved time-constant factors at the province level, and α_t is the birth cohort indicator capturing shocks over time. W_i is the treatment of interest, i.e. whether in the birth year of child$i$, the mother had a given health insurance ($W \in (0,1)$), τ is the treatment effect of interest. The residual term ε_i is assumed normally distributed, mean zero, and captures a composite of any unobserved province, community, household, mother and child level shocks**.** It follows from the previously stated assumptions that W_i is uncorrelated with ε_i implying that any unobserved health shock to the mother, or income shock to the household, beyond those captured by the year fixed effects is unrelated to whether a mother is enrolled in health insurance in a given year. The outcome regression, Eq. (1), assumes a homogenous additive treatment, hence τ cannot be directly interpreted as estimating either one of the ATE, ATT or ATC defined before. Moreover, the model assumes a linear relationship between the outcome and the covariates being correct (Ho et al. 2007), and the resulting regression model relies heavily on extrapolation.

To address these restrictions, we also estimate propensity scores (Rosenbaum and Rubin, 1983), defined as $p(X) = (W = 1|X)$, estimated via logistic regression including all the sufficient covariates as in [1],[12] for each health insurance status. Because using PS only for confounder adjustment would result in a different limitation—bias stemming from poor overlap—we use the inverse propensity scores to weight linear outcome regression models to construct the so-called Wooldridge double robust (DR) estimator (Wooldridge, 2007), where both the reliance on extrapolation and potential overlap problems are reduced. We implement this method using the teffects command in Stata, and obtain the estimated ATE, ATT and ATC. The unweighted (OLS) regression results are presented in Appendix B Table 4. These more traditional estimates are later contrasted to the causal forest estimates for average treatment effects (see next section). All models use the same set of covariates for confounder adjustment.

3.3 Estimation of heterogenous and average treatment effects using causal forests

The ATE, ATT and ATC estimands allow for the causal effects to be different for those insured and uninsured, but do not capture their variation over the observed X covariates. For this we focus on the CATE, $\tau(x)$. We begin by considering a partially linear model for the outcome of interest, as before, that is:

$$Y_i = f\left(X_i\right) + W\tau + \varepsilon_i \tag{2}$$

with $f(X)$ an unspecified function, and initially, that τ, the treatment effect, is constant in X. Following Robinson (1988), we can re-write this model in a "centred" or residualised form as follows

$$Y_i - m\left(X_i\right) = \left(W_i - p\left(X_i\right)\right)\tau + \in_i \tag{3}$$

where $p(X_i)$ is the propensity score as before, and $m\left(X_i\right) = E\left[Y_i|X_i\right]$ the conditional expectation of the outcome, marginalised over the treatment. The expressions $m(.)$ and $p(.)$ are often referred to as "nuisance functions", and they can be estimated with any prediction

[12] Instead of province dummies, we use region dummies to adjust for confounding due to geographic region, due to convergence issues experienced in the weighted parametric regression models used by the teffects package. The use of region vs. province dummies made no difference to the results.

algorithm, including ML methods. The causal effect τ can be estimated by solving Eq. (3), and plugging the predictions for $m\left(X_i\right)$ and $p(X_i)$ in the following formula:

$$\hat{\tau} = \frac{\sum_{i=1}^{n} \left\{(W_i - \hat{p}(X_i))(Y_i - \hat{m}(X_i))\right\}}{\sum \left\{(W_i - \hat{p}(X_i)\right\}^2} \tag{4}$$

This corresponds to running a regression of the Y-residual on the W-residual. Such "residualising" decreases the sensitivity of the resulting estimator to the errors in the estimates of the nuisance functions (Chernozhukov et al. 2018a). This can be extended to allow for heterogenous treatment effects, assuming a sufficiently small neighbourhood $N(x)$ such that $\tau(x)$ is constant, which allows us to rewrite Eq.(4) as

$$\hat{\tau}(x) = \frac{\sum_{\{i:X_i \in N(x)\}} \left\{(W_i - \hat{p}(X_i))(Y_i - \hat{m}(X_i))\right\}}{\sum_{\{i:X_i \in N(x)\}} \left\{W_i - \hat{p}(X_i)\right\}^2} \tag{5}$$

The main challenge for CATE estimation is how to choose $N(x)$. To solve this, Athey et al. (2019a) propose a generalised random forest approach, which conceptualises these neighbourhoods as a locally weighted set of neighbouring observations for a given value of x. The weights are estimated by performing a modification of the Random Forest algorithm (Breiman, 2001). In short, random forests calculate a predicted outcome for a unit by averaging the outcome of other units that are similar enough in covariates. The group of similar units are referred to as a leaf of a tree, and leaves are decided on by splitting the data based on cut-off values of the predictors, where the predictors to split on and cut-offs are decided so that the resulting splits minimise the prediction error in the sample. To reduce the noise stemming from using individual trees as predictors, this is done many times over bootstrapped samples of the data, and final predictions for each observation are obtained as the average of predictions over the bootstrap samples.

Generalised random forests build on this algorithm, but modify it in important aspects, to ultimately minimise the bias in the estimated CATE. First the outcomes and treatment are residualised as described before. Second, the splits of the data ("the causal trees") are formed by running the local linear regressions (Eq. 3) in each candidate split. Instead of choosing splits to minimise prediction error, they are chosen so that within a leaf, estimated treatment effects are similar (corresponding to homogenous treatment effects within a leaf), while between leaves, they differ (capturing treatment effect heterogeneity across units with differing X values). This procedure is performed on many bootstrap samples, thus forming causal forests. The causal forests are then used to calculate $\alpha_i(x)$ weights for each observation, based on how frequently an observation was used to estimate the treatment effect at x. The resulting weights are employed in an estimator of the CATE that modifies Eq. (4) as follows:

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x)\left\{(W_i - \hat{p}(X_i))(Y_i - \hat{m}(X_i))\right\}}{\sum \alpha_i(x)\left\{W_i - \hat{p}(X_i)\right\}^2} \tag{6}$$

Individual treatment effects $\hat{\tau}(X_i)$ can be estimated by evaluating $\hat{\tau}(\mathrm{x})$ at the covariate combination of each unit. Average treatment effects can also be obtained, by plugging in the estimated $\hat{\tau}(X_i)$ in a variant of the augmented inverse probability weighting estimator (Robins et al. 1994):

$$\hat{\tau} = \sum_{i=1}^{n^D} \hat{\tau}(X_i) + \frac{W_i - p(X_i)}{p(X_i)(1 - p(X_i)}((Y_i - m(X_i) - (W_i - p(X_i))\hat{\tau}(X_i) \quad (7)$$

where the summation is taken over n^D, that stands for the sample of the treated, the control or the treated plus control samples, depending on the whether the causal estimand is the ATT the ATC or the ATE, respectively. This formula also provides the subgroup average treatment effects, constraining the summation for units in the subgroups of interest (e.g. women with primary education only).

The causal forests approach, as implemented in the grf R package (Tibshirani et al. 2018), in its simplest form can be run using the causal_forest(X, Y, W) command, where X is the vector of confounders and potential effect modifiers, Y is the outcome, and W is the treatment. In order to improve the performance of the estimation, we follow the approach suggested by Athey and Wager (2019b). First, motivated by the double-machine learning literature (Chernozhukov et al. 2018a, b), this approach relies on an initial residualizing of the treatment and outcome variables (following Robinson 1988, as described in Eq. 3) in order to minimise confounding due to observed covariates. Second, the approach fits two sets of causal forests: first, a "pilot" causal forest, using all confounders as potential effect modifiers, then a final causal forest, using only those variables which were ranked highly in the variables importance analysis. This enables the final forest to make more splits on the most important features, even in situations where the heterogeneity in treatment effects is relatively weak (Athey and Wager (2019b).

The steps taken in this paper are described as follows:

1. **Estimate nuisance parameters**: Fit regression forests to estimate $m\left(X_i\right)$ and the $p\left(X_i\right)$, then calculate residualised outcomes using these quantities (see Eq. 3). We use 500 trees to select the tuning parameters, and 1000 trees to obtain the predictions.
2. **Train and fit causal forests:**
    a. Train an initial causal forest on 1000 bootstrap samples (with 500 trees to select tuning parameters), using the entire set of covariates for splits.
    b. Use the output from 2a, and rank variables in terms of variable importance in the initial causal forest (based on count of the proportion of splits on the given variable). Select those with higher than mean variable importance measure.
    c. Fit a second causal forest, using only those variables selected in Step 2b (with variable importance > mean variable importance). We use 500 trees for tuning, and 3000 trees for predicting ITEs.
3. **Estimate treatments effect**s: Estimate ITEs by evaluating the resulting $\hat{\tau}(x)$ function for each unit's own covariate values. Estimate ATEs, ATTs, ATCs, and subgroup ATCs for each pre-specified subgroup.
4. **Assess the heterogeneity** captured by the resulting causal forests:
    a. Plot the estimated individual level CATEs.
    b. Perform a test for the presence of overall heterogeneity captured by the $\hat{\tau}(x)$ estimate (Chernozhukov et al. 2018b). This test assesses whether $\hat{\tau}(x)$ captures any further information than simply using the ATE, $\hat{\tau}$ to "predict" the individual level treatment effects.

c. Assess the final ranking of the variable importance measure, and form further subgroups based on the top ranked variables, and contrast the differences in the average treatment effects across these subgroups.
d. Split individuals into two groups based on their estimated CATEs (below and above median), and describe these groups in a number of key characteristics.

We implement this approach for the skilled birth attendance outcome variable, and fit separate causal forests for the subsidised and contributory health insurance. The covariates used in Steps 1–2 include all variables used in the previous analyses. For the infant mortality outcome, we implement steps 1–3, and report average treatment effects.

4 Results

4.1 Descriptive statistics

While the majority of births recorded in our dataset were not covered by any insurance scheme (Appendix B Table 2), subsidised health insurance saw a steep increase from 2005, while infant mortality decreased and the proportion of births attended by a midwife or physician demonstrated a clear upwards trend (Appendix C Fig. 1). In Table 1, we contrast the observed characteristics of the three groups: births insured by subsidised insurance, births insured by contributory insurance, and births not covered by health insurance in the year of birth, comparing the means and standardised differences for each treatment group to the control group. Most variables display large differences (standardised differences $> 10\%$), with births under subsidised insurance being more likely to be from a rural household and from mothers who are older at birth, less likely to have studied at university and more likely to have only elementary school education, belong to lower wealth quintiles, and receive social assistance programmes, compared to those without subsidised insurance. By contrast, while those mothers with contributory insurance are also somewhat older at the time of birth than the uninsured, they are also more likely to have a university education, and are overrepresented among households within the highest asset index quintiles. A quarter of these mothers received subsidised rice, while only a small fraction received cash transfer (7%) or held a "Poor card" (4%). We interpret these large differences as indicative of a strong confounding of the relationship between health insurance and the outcomes of interest.

4.2 Average treatment effects

Table 1 and Appendix Fig. 2 describe the covariate balance achieved after inverse probability weighting using the estimated (logistic regression based) propensity scores for both treatment groups compared to the control group, and contrasts these to the unweighted balance. Using weights that aim to recreate the distribution for the treated (ATE, ATC and ATT weights), the balance improves for each covariate, and standardised differences stay above 10% for only a few covariates, and the ATE weights showing somewhat worse balance than the ATT and ATC weights. Appendix Fig. 3 displays the distributions of the estimated propensity scores. While there is a good overlap between the propensity score distributions for both insurance types, there is a large mass around zero for the uninsured,

Table 1 Descriptive statistics by insurance status, before and after propensity score reweighting

	Uninsured (n = 9,111)	Subsidised HI (n = 1511)			Contributory HI (n = 1454)			
	Mean	Mean	SMD (raw)	SMD (PS weighted, ATE)	SMD (PS weighted, ATT)	Mean	SMD (raw)	SMD (PS weighted, ATE)	SMD (PS weighted, ATT)
Age	27.14	28.01	14.0%	0.5%	− 0.3%	29.10	34.4%	− 9.4%	0.3%
Health (good)	0.87	0.83	− 8.8%	3.4%	− 4.5%	0.87	2.0%	0.5%	0.4%
Educ: primary	0.32	0.38	12.7%	5.9%	− 0.6%	0.09	− 59.0%	− 1.6%	0.3%
Educ: secondary	0.26	0.28	4.5%	− 2.0%	− 3.3%	0.14	− 30.6%	3.6%	− 0.1%
Educ: senior	0.33	0.29	− 8.6%	− 1.1%	3.5%	0.41	17.1%	− 2.2%	1.0%
Educ: higher	0.10	0.06	− 15.4%	− 5.3%	0.9%	0.36	66.6%	0.7%	− 1.2%
Writes (Indonesian)	0.96	0.95	− 2.8%	− 1.0%	0.4%	0.99	18.6%	5.0%	− 0.7%
Wealth quint 1	0.19	0.30	24.0%	2.6%	-0.9%	0.04	− 51.4%	− 0.3%	0.2%
Wealth quint 2	0.21	0.25	9.1%	4.2%	-3.4%	0.09	− 34.5%	2.7%	0.3%
Wealth quint 3	0.22	0.21	− 2.1%	0.3%	1.5%	0.17	− 14.0%	− 3.5%	0.2%
Wealth quint 4	0.20	0.16	− 11.1%	0.1%	1.9%	0.30	22.0%	0.9%	− 0.9%
Wealth quint 5	0.17	0.08	− 27.5%	− 8.7%	2.0%	0.41	55.3%	0.3%	0.4%
Raskin	0.50	0.72	45.6%	10.8%	− 4.4%	0.26	− 51.8%	− 2.0%	− 0.1%
Cash transfer	0.23	0.45	47.0%	7.7%	− 5.2%	0.07	− 46.1%	− 2.3%	0.2%
Poor card	0.09	0.20	31.6%	5.3%	− 2.6%	0.04	− 18.5%	1.7%	0.6%
Rural	0.48	0.47	− 2.3%	8.3%	− 1.3%	0.28	− 43.3%	− 0.7%	0.7%
Disaster	0.23	0.28	10.2%	7.8%	− 1.1%	0.24	1.5%	− 3.4%	0.4%
Birth clinic in comm	0.99	1.00	4.1%	4.3%	0.3%	0.99	0.4%	2.7%	− 0.4%
Health centre in comm	0.97	0.98	3.2%	5.0%	− 0.7%	0.97	− 1.4%	5.1%	− 1.6%
Private practice in comm	0.96	0.95	− 1.8%	1.2%	− 2.0%	0.94	− 7.9%	2.2%	0.4%
Hospital in comm	0.90	0.93	9.9%	− 3.2%	0.3%	0.89	− 4.6%	6.0%	− 2.0%
Midwife	0.82	0.83	3.0%	12.4%	− 0.7%	0.76	− 15.0%	1.8%	0.3%
1st child	0.68	0.52	− 33.7%	1.0%	0.9%	0.64	− 10.0%	6.3%	− 0.9%
2nd child	0.25	0.36	22.9%	− 0.5%	− 1.5%	0.30	9.8%	− 3.0%	0.7%

Table 1 (continued)

	Uninsured (n = 9,111)	Subsidised HI (n = 1511)			Contributory HI (n = 1454)			
	Mean	Mean	SMD (raw)	SMD (PS weighted, ATE)	SMD (PS weighted, ATT)	Mean	SMD (raw)	SMD (PS weighted, ATE)	SMD (PS weighted, ATT)
> = 3rd child	0.06	0.12	20.1%	− 1.0%	0.9%	0.07	1.5%	− 7.0%	0.4%
Female	0.49	0.49	1.5%	− 0.1%	1.0%	0.49	1.5%	0.5%	0.3%

educ education, *quint* quintile, *comm* community, *SMD* Standardised mean difference, *PS* Propensity score, *ATE* Average treatment effect, *ATT* Average treatment effect among the treated, *IPW* Inverse probability of treatment weighting

Table 2 Estimates of average treatment effects. Panel A: Wooldridge DR estimator Panel B: Causal Forests

Infant mortality	Panel A: Wooldridge DR estimator Estimate (SE)	Panel B: Causal Forests Estimate (SE)
Subsidised health insurance		
Unadjusted estimate	− 0.0025 (0.0047)	
ATE	− 0.0026 (0.0058)	− 0.005 (0.0051)
ATC	− 0.0058 (0.0055)	− 0.0048 (0.0055)
ATT	− 0.0026 (0.0052)	− 0.0061 (0.0049)
Contributory health insurance		
Unadjusted estimate	− 0.0126*** (0.0034)	
ATE	− 0.0147***(0.0033)	− 0.0120*** (0.0039)
ATC	− 0.0157***(0.0033)	− 0.0121*** (0.0041)
ATT	− 0.0101**(0.0041)	− 0.0100*** (0.0038)
Skilled birth attendance	Panel A: Wooldridge DR estimator Estimate (SE)	Panel B: Causal Forests Estimate (SE)
Subsidised health insurance		
Unadjusted estimate	0.0291*** (0.0104)	–
ATE	0.0206 (0.0136)	0.016 (0.0115)
ATC	0.0231 (0.0149)	0.016 (0.012)
ATT	0.0120 (0.0111)	0.011 (0.0093)
Contributory health insurance		
Unadjusted estimate	0.1279*** (0.0079)	–
ATE	0.0584***(0.0159)	0.055 (0.0109) ***
ATC	0.0639***(0.0176)	0.060 (0.012)***
ATT	0.0239***(0.0070)	0.024 (0.0058)***

SE Standard error, *ATE* Average treatment effect, *ATT* Average treatment effect among the treated, *ATC* Average treatment effect among the controls, *IPW* Inverse probability of treatment weighting, *** $p<0.01$, ** $p<0.05$, * $p<0.1$

implying that many of those who did not get the insurance were unlikely to get it based on their observed covariates.

Table 2 reports the average treatment estimates for both outcomes, both using the Wooldrige DR (Panel A) and the causal forests (Panel B) approaches, and contrast these to unadjusted estimates that simply compare the means for the treated and control groups. For subsidised health insurance, the unadjusted results for infant mortality are small and insignificant, and even after covariate adjustment using the Wooldrigde DR and the causal forests estimators, there is no evidence that they are different from the null, across all estimands. For the contributory health insurance, there is strong evidence of a large protective (i.e. infant mortality-reducing) insurance effect of around 1–2 percentage points, with the estimated ATE and ATC larger than the ATT, indicating that the uninsured would have benefitted more from the insurance than those who were actually insured. This pattern repeats with the skilled birth attendance outcome, for both insurance types: the benefit in terms of increased access is larger among the untreated than among the treated, while these estimated effects are significant ($p<0.01$) for the contributory health insurance, and not significant for the subsidised health insurance, and the causal forest estimates closely correspond to the DR IPW-regression estimates.

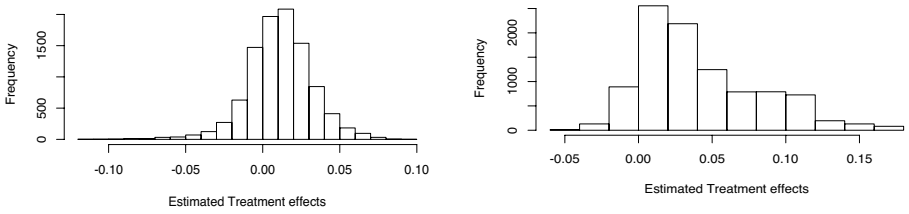


Fig. 1 Estimated (conditional) individual level treatment effects for skilled birth attendance. Left panel: subsidised health insurance; right panel: contributory health insurance

Table 3 Covariate importance in explaining treatment effect heterogeneity for skilled birth attendance

Ranking	Subsidised HI		Contributory HI	
	Variable importance measure for	Variable	Variable importance	Variable
1	0.126	Birth order > = 3	0.127	Province East Java
2	0.085	Birth year 2012	0.123	Higher education
3	0.084	Age > = 31	0.083	Wealth quantile 4
4	0.075	Past covariates imputed	0.069	Province South Kalimantan
5	0.066	Cash transfer	0.066	Rural community
6	0.065	Poor card	0.060	Wealth quantile 5
7	0.063	Birth year 2014	0.055	Province West Sumatra
8	0.062	Birth order = 2	0.049	Private practice in community
9	0.054	Province West Nusa Tenggara	0.048	Senior education
10	0.046	Natural disaster	0.045	Province Banten

10 highest importance covariates. The variable importance measure is based on the count of the proportion of splits on the given variable

4.3 Heterogenous treatment effects

Figure 1 presents the distribution of the estimated individual treatment effects, as histograms of the point estimates. The formal test for treatment effect heterogeneity indicates the presence of heterogeneity for the contributory health insurance ($p = 0.003$), but not for the subsidised health insurance ($p = 0.69$). Table 3 presents the ranking of covariates in terms of their importance in predicting treatment effect heterogeneity, in terms of utilisation of skilled attendance when giving birth. For the contributory health insurance scheme, these largely overlap with the pre-specified socioeconomic covariates: education, wealth quintiles, and the rurality of the household. The most important variable associated with the estimated heterogeneous effect was the indicator for East Java province: a relatively industrialised region of Indonesia. For the subsidised health insurance scheme, the most influential variables were mother's age and the birth order of the child, followed by being in receipt of cash transfers and possessing a poor card.

We then present the CATE among the controls for pre-specified subgroups (Fig. 2a) as well as subgroups constructed based on the variables suggested by the final causal forests variable importance Fig. 2b) (Also see Appendix Table 4). We detect large differences in subgroup ATCs for contributory health insurance corresponding to subgroups

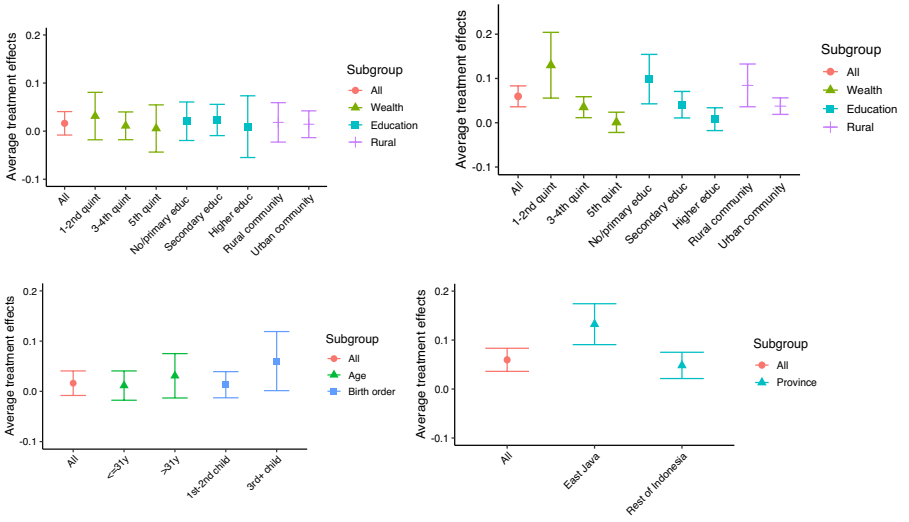


Fig. 2 **a** (*Top*) Estimated subgroup average treatment effects (ATC) for skilled birth attendance, for the overall population and the pre-specified subgroups. Left panel: subsidised health insurance; right panel: contributory health insurance. **b** (*Bottom*) Estimated subgroup average treatment effects (ATC) for skilled birth attendance, for the overall population and the subgroups suggested by the causal forest algorithm's variable importance measures. Left panel: subsidised health insurance; right panel: contributory health insurance

suggested by variable importance: there is a strong trend in terms of wealth quintiles in the estimated subgroup effects, and there are also considerable differences in ATC reported between those with different education levels, and between rural and urban communities. The differences in the subgroup effects, while showing a similar direction, are much less pronounced for the subsidised health insurance, and there is no evidence in support of a subgroup ATCs being different from zero. Among the subgroups suggested by the causal forest variable importance, for the subsidised scheme we found some evidence ($p < 0.05$) of treatment effect for the subgroup with the third or higher birth order. None of these results were found to be sensitive to the choice of tuning parameters for the causal forest algorithm, which were selected outside of the cross-validation algorithm (number of trees used for tuning, number of trees used for the final Causal Forests). We present the selected tuning parameters in Appendix Table 5. As a final, exploratory analysis, we compare the characteristics of mothers when they are grouped based on the estimated individual level CATEs, using the median value as the cut-off (Appendix Table 6), and using SMDs for the comparison. It appears that mothers who benefitted relatively more from the subsidised health insurance are older, more likely to be in lower wealth quintiles, and more likely to have received cash transfer or rice subsidy, than those in the lower half of the treatment effect distribution. Those benefitting most from contributory health insurance are also more likely to belong to the lower wealth quintiles, less likely to have had higher levels of education, and twice as likely to have received subsidies, compared to those in the lower half of the distribution. There is no difference in the availability of health services among the two groups. To investigate the surprising result of East Java being an important driver of heterogeneity of the impact of contributory health insurance, we followed the suggestion of Semenova et al. (2021) and explored the independent contribution of the East Java

variable, by running a linear regression on the double-robust score constructed from the estimated CATEs and the nuisance components, on a selected covariate vector: the pre-specified subgroups and the East Java variable. Here we find that for a child being born in East Java, the beneficial effect of contributory health insurance is relatively larger than for the overall control group, even after controlling for the heterogeneity that can be attributed to education, socioeconomic status and rurality.

5 Discussion

This paper is the first study to characterise the effect heterogeneity of a health policy intervention by employing causal forests, a causal machine learning approach that is quickly gaining popularity in economic and social science research. We highlighted the role of this approach for establishing heterogeneity of policy impacts, and further for suggesting the main observed covariates driving such heterogeneity. Our study also highlights a crucial challenge when using this approach to estimate treatment heterogeneity in an observational framework: the need to adjust for the key observable confounders in the institutional setting of interest. The causal forests algorithm allows to adjust for observed confounding in the first step of the analysis, by using the outcome and treatment residuals (from the corresponding models adjusting for the confounding variables), using flexible machine learning algorithms, in this case random forests. While in this study we use regression forests to estimate these residuals, other supervised learning algorithms, such as ensembling machine learning methods, may also be used for this step. Even when the nuisance parameters are estimated with non-forest-based algorithms, the causal forests method still utilises the extension of the random forest algorithm described earlier, for estimation of the CATEs. Ongoing work (Nie and Wager 2021) further generalises the causal forests approach and allows for any supervised learning algorithms that can solve a loss function designed to target the CATE estimand (the so-called R-learner loss function).

Moreover, our analysis demonstrated the ability of the causal forest algorithm to facilitate pre-specified subgroup analysis without having to re-fit propensity score and outcome regression models for the subgroups, but instead taking the estimated individual level CATEs, and plugging them in an augmented IPW estimator for average treatment effects. We then used formal statistical tests to assess the presence of treatment effect heterogeneity. It should be noted that the causal forest approach is not designed to identify and conduct inference on subgroups with the largest (or lowest) treatment effects. It can, however provide indicative evidence on which variables are most strongly associated with heterogeneity, using variable importance measures that characterise the resulting causal forest. We used these variable importance measures to select further variables to assess subgroup effects on. There is ongoing work on developing estimators specifically for group average treatment effects discovered by machine learning (Chernozhukov et al. 2018b).

This paper also contributes to the growing body of evidence on the impact of public health insurance on health outcomes and health care utilisation, by estimating the average and heterogenous treatment effects of two main types of health insurance in Indonesia on infant mortality, and on maternal health care utilisation at the time of delivery. We find that enrolment in contributory health insurance reduced infant mortality on average by 1.0 percentage points ($p<0.05$) among those who were insured, corresponding to a sizeable 30% reduction from the average infant mortality rate (i.e. infant deaths per 1000 live births) over the observation period. By contrast, we found no evidence of an effect of subsidised

health insurance. Our findings for the health care utilisation outcome may help explain these results: contributory insurance increased the expected probability of having a birth attended by a healthcare professional, while there was no such effect for the subsidised scheme. Our findings mirror the previous evidence that found small to negligible impacts of subsidised health insurance schemes on health services utilisation (Johar 2009, Sparrow et al. 2013, Erlangga et al. 2019a), but they are also consistent with the findings of Anindya et al. (2020) who found that the JKN programme improved the utilisation of skilled birth attendance, on a population that pooled subsidised and contributory recipients.

We delved deeper into this, by examining the heterogeneity in the effects for both insurance schemes. The estimated causal effects on health care utilisation among the uninsured appear to be higher than among the insured (that is, on average, we expect those uninsured in the study period would benefit from being insured more than the expected benefit estimated amongst those who are insured). Indeed, we found that the benefits, in terms of increased access to skilled birth attendance, are relatively higher among the more vulnerable subgroups, reflecting the findings of Anindya et al. (2020). While pre-specified socioeconomic variables ranked high in terms of being associated with treatment effect heterogeneity, we found further variables that according to the variable importance of the causal forest algorithm were more strongly associated with treatment heterogeneity: for example, women residing in certain provinces (e.g. East Java) would have benefited more than other subgroups, had they been insured (contributory vs remaining uninsured). Given East Java has development indicators above the county average (Unicef 2019), this result might hint towards the potential importance of non-health infrastructure in predicting the benefits of a demand side health policy such as health insurance. For subsidised health insurance, there was no strong evidence of a causal treatment effect for any of the pre-specified population subgroups. While we found no evidence of heterogeneity, the variable importance of the CF algorithm suggested the child's birth order may drive treatment heterogeneity, with the resulting subgroup of children who were third born or higher having the highest causal average treatment effect, with a 95% CI that excluded zero.

Our study has some limitations. Due to infant mortality being a rather rare event (approximately 300 events out of 12,000), we could not conduct a subgroup analyses for this outcome. Furthermore, because we use household survey data, covariate information was collected at discrete time points, which were assumed to provide valid baseline measurements for births that occurred closer to or further away from the survey dates. This measurement error can lead to a downward bias in the estimated coefficients.

Our analysis assumed no unobserved confounding, given the measured household and individual characteristics, year and province fixed effects. Previous impact evaluations of health insurance expansions in Indonesia also relied on adjustment for observed confounders: for example, Johar (2009) and Anindya et al. (2020) utilised a propensity score matching approach for the evaluation of the health card and JKN programmes, respectively. Sparrow et al. (2013) combined propensity score matching with differences-in-differences when evaluating the Askeskin programme, and Erlangga et al. (2019a) used the same approach for the evaluation of the JKN programme. By controlling for self-reported health, literacy and socioeconomic factors, we aimed to approximate the process of selection into subsidised health insurance and hence minimise any remaining bias due to selection on unobserved factors that could compromise comparisons of outcomes between individuals in subsidised insurance with those uninsured.

Nonetheless, we cannot fully rule out residual confounding biasing the results—this bias might appear both in the average treatment effect estimates, the estimated CATEs, and through them, in the resulting subgroup average treatment effects. For example, we

cannot rule out that for the contributory health insurance evaluation, some of the difference the ATC and the ATT, and between the CATC for the socioeconomic subgroups, may be due to remaining unmeasured confounding. One potential avenue to explore this would be to use methods that can explicitly handle unobserved confounding (instrumental variables (IVs) or regression discontinuity design). Under the strong assumption of an instrument being valid (predictive of health insurance, and meeting the exclusion criteria), we could re-estimate the average and subgroup average treatment effects, and comparing these to the causal forest estimates, potentially discover that in certain population subgroups, unobserved confounding played a relatively larger role. In order for unobserved confounding to explain the differences in the estimated CATEs in the current study, unmeasured confounding should be stronger in the more deprived population subgroups, something we don't have a priori reasons to believe. While for the current study we did not find IV that meet the exclusion criteria, this is an area of further investigation. A second approach, potentially promising to explore the sensitivity of heterogenous treatment effects to unobserved confounding is to employ sensitivity analysis to calculate bounds on the CATEs. This is an innovative, and currently developing area of methodological research (see e.g. Kallus et al. 2019), and is hence outside of the scope of our paper.

Despite these limitations, this paper provides a novel demonstration of the value of causal machine learning for public policy evaluation, in a setting where heterogenous treatments effects have the potential to critically inform policymakers. Based on our work, we can suggest at least two promising avenues for future related research. First, to address concerns of remaining unobserved confounding, Generalised Random Forests could be combined with instrumental variables, where valid instruments are available (Athey et al. 2019a). Second, the estimated individual treatment effects could be used to formulate so-called "optimal policy rules": treatment assignment mechanisms that maximise a pre-specified welfare function set by the decision-maker (Athey and Wager 2020). For the Indonesian context, such optimal policy rules could inform health policymaking in the country by guiding the re-design of the eligibility criteria for subsidised health insurance, which could help address the fiscal challenges brought by the move towards Universal Health Coverage. Beyond the specific case of Indonesia, causal machine learning may be used to help target policy efforts towards where the greatest potential benefits can be realised, thereby helping to pinpoint where adaptation of policy may be needed. In doing so, this could enable researchers to move policy impact evaluations beyond simple binary judgements on whether something 'works' or not, towards matters of for whom policies 'work' and how these can be improved.

Appendix A: Variables used in the principal component analysis to construct the wealth index

- Whether the household has electricity
- Access to piped water
- Types of stove
- Toilet inside the house
- Refrigerator
- Television
- House and land owned by household
- Ownership of other house

- Vehicles
- Household appliances
- Savings
- Receivables
- Jewellery

Appendix B

See (Tables 4, 5, 6, 7, 8, 9).

Table 4 Construction of the analytical dataset

	IFLS3	IFLS4	IFLS5
Survey data collection years	2000	2007/2008	2014/2015
Contribution to treatment group 1 (subsidised health insurance)	No	Yes Insurance programme: Health cards (2002–2007), Askeskin (2005–2007)	Yes Insurance programme: Jamkesda (2008–2014), Jamkesmas (2008–2014), JKN (2014)
Contribution to treatment group 2 (contributory health insurance)	No	Yes Insurance programme: Askes (2002–2007), Jamsostek (2002–2007), Employee provided insurance (2002–2007)	Yes Insurance programme: Askes (2008–2014), Jamsostek (2008–2014), Employee provided insurance (2008–2014)
Contribution to the control group	No	Yes (2002–2007)	Yes (2008–2014)
Contribution to covariate history	Yes (For births from the IFLS4 data)	Yes (For births from the IFLS 5data, And for IFLS4 data where IFLS3 history not available)	Yes (For births from the IFLS 5data, where IFLS4 history not available)

Table 5 Panel A: Insurance status of the mother a year of birth, Panel B: Absolute infant mortality, by year of birth

Panel A: insurance status by year of birth				
Year of birth	Uninsured	Subsidised insurance	Contributory insurance	Total
2002	615	8	79	702
2003	729	21	89	839
2004	687	40	104	831
2005	620	76	106	802
2006	603	79	94	776
2007	582	115	164	861
2008	875	59	116	1,050
2009	782	73	99	954
2010	877	93	107	1,077
2011	821	97	98	1,016
2012	773	151	139	1,063
2013	680	252	124	1,056
2014	467	447	135	1,049
Total	9,111	1,511	1,454	12,076
Panel B: Outcomes by insurance status				
Infant mortality (%)	2.57	2.31	1.12	2.37
Skilled birth attendance (%)	82.46	85.37	95.24	84.4

Table 6 Linear regression results for the effect of health insurance on Infant mortality and skilled birth attendance outcomes

Outcomes	Panel A: Subsidised HI			
	Unadjusted 1	Unadjusted 2	OLS1	OLS2
Infant mortality				
Estimate	− 0.0025	− 0.0020	− 0.0056	− 0.0055
(SE)	(0.0047)	(0.0052)	(0.0052)	(0.0053)
Observations	10,622	10,622	10,622	10,622
Skilled birth attendance				
Estimate	0.0291***	-0.0295***	0.0163	0.0183*
(SE)	(0.0104)	(0.0110)	(0.0108)	(0.0108)
Observations	9,834	9,834	9,834	9,834
	Panel B: Contributory HI			
	Unadjusted 1	Unadjusted 2	OLS1	OLS2
Infant mortality				
Estimate	− 0.0126***	− 0.0130***	− 0.0088**	− 0.0093**
(SE)	(0.0034)	(0.0035)	(0.0038)	(0.0039)
Observations	10,565	10,565	10,565	10,565
Skilled birth attendance				
Estimate	0.1279***	0.1225***	0.0237***	0.0294***
(SE)	(0.0079)	(0.0080)	(0.0080)	(0.0081)
Observations	9,732	9,732	9,732	9,732
Year dummies	N	Y	Y	Y
Covariates	N	N	Y	Y
Province dummies	N	N	N	Y

HI Health insurance, *OLSL* Ordinary least squares, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 7 Estimated conditional average treatment effects, for pre-specified subgroups and subgroups suggested by the causal forest algorithm's variable importance measures

	Subsidised HI		Contributory HI	
	CATC	SE	CATC	SE
All (uninsured)	0.0162	0.0125	0.0596***	0.012
Pre-specified subgroups				
1–2nd quint	0.0313	0.0252	0.1299***	0.0379
3–4th quint	0.011	0.0147	0.0351***	0.0121
5th quint	0.0055	0.025	0.0009	0.0117
No/primary educ	0.0206	0.0204	0.0985***	0.0285
Secondary educ	0.0232	0.0166	0.0408***	0.0153
Higher educ	0.0093	0.0328	0.008	0.0132
Rural community	0.0182	0.0209	0.0844***	0.0246
Urban community	0.0143	0.0142	0.0375***	0.0095
Subgroups suggested by CF				
$<=31$y	0.0115	0.0149		
>31y	0.0309	0.0225		
1st-2nd child	0.0132	0.0133		
3rd + child	0.0602**	0.030		
Java			0.1324***	0.0213
Rest of Indonesia			0.0482***	0.0137

HI Health insurance, *CATC* Conditional average treatment effect among the controls, *SE* Standard error, *quint* quintiles, *educ* education*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 8 Tuning parameters in the causal forest analysis

Tuning parameter	grf package argument in causal_forest() function	Values (subsidised HI analysis)	Values (contributory HI analysis)
Fraction of the data used to build each tree	Sample.fraction	0.472	0.500
Number of variables tried for each split	mtry	21	21
Minimum number of observations in each tree leaf	min.node.size	1	5
The fraction of data used for determining splits	Honesty.fraction	0.620	0.500
Prunes the estimation sample tree such that no leaves are empty	Honesty.prune.leaves	TRUE	TRUE
Maximum imbalance of a split	Alpha	0.091	0.05
Controls how harshly imbalanced splits are penalized	Imbalance.penalty	0.061	0

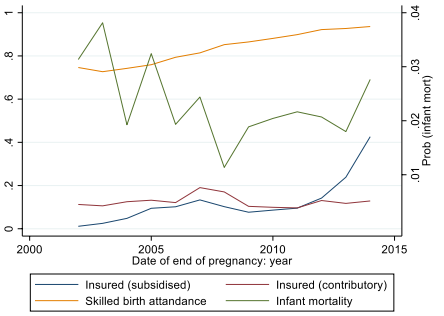
Table 9 Comparison of observed characteristics of mothers with low and high treatment effects

	Subsidised health insurance			Contributory health insurance		
	Below median CATE Mean (SE)	Above median CATE Mean (SE)	SMD[1]	Below median CATE Mean (SE)	Above median CATE Mean (SE)	SMD
Age at birth < = 23	0.34 (0.47)	0.26 (0.44)	0.167	0.22 (0.41)	0.35 (0.48)	0.299
Age at birth 23- < = 27	0.29 (0.45)	0.19 (0.39)	0.246	0.30 (0.46)	0.20 (0.40)	0.245
Age at birth 27- < = 31	0.23 (0.42)	0.20 (0.40)	0.067	0.23 (0.42)	0.21 (0.41)	0.063
Age at birth > 31	0.14 (0.35)	0.35 (0.48)	0.501	0.25 (0.43)	0.24 (0.43)	0.004
1st wealth quintile	0.20 (0.40)	0.22 (0.41)	0.049	0.08 (0.27)	0.26 (0.44)	0.507
2nd wealth quintile	0.20 (0.40)	0.23 (0.42)	0.07	0.11 (0.31)	0.28 (0.45)	0.451
3nd wealth quintile	0.23 (0.42)	0.22 (0.42)	0.009	0.17 (0.37)	0.27 (0.44)	0.252
4th wealth quintile	0.20 (0.40)	0.19 (0.39)	0.014	0.33 (0.47)	0.10 (0.31)	0.561
5th wealth quintile	0.17 (0.38)	0.13 (0.34)	0.11	0.32 (0.47)	0.08 (0.27)	0.632
No/primary education	0.28 (0.45)	0.38 (0.49)	0.216	0.09 (0.29)	0.48 (0.50)	0.964
Secondary education	0.27 (0.44)	0.25 (0.44)	0.026	0.11 (0.31)	0.37 (0.48)	0.645
Senior education	0.36 (0.48)	0.29 (0.45)	0.145	0.55 (0.50)	0.13 (0.34)	0.992
Higher education	0.10 (0.30)	0.08 (0.27)	0.076	0.25 (0.43)	0.01 (0.12)	0.739
Poor card	0.13 (0.33)	0.08 (0.28)	0.141	0.08 (0.27)	0.09 (0.28)	0.033
Received cash transfer	0.22 (0.42)	0.31 (0.46)	0.205	0.12 (0.33)	0.31 (0.46)	0.46
Received subsidised rice	0.49 (0.50)	0.57 (0.49)	0.157	0.34 (0.47)	0.60 (0.49)	0.535
Writes in Indonesian	0.97 (0.17)	0.94 (0.24)	0.153	0.99 (0.11)	0.93 (0.25)	0.282
Public health clinic in community	0.99 (0.09)	0.99 (0.09)	0.009	0.99 (0.08)	0.99 (0.10)	0.049
Hospital in community	0.91 (0.29)	0.90 (0.31)	0.046	0.89 (0.31)	0.90 (0.29)	0.053
Private practice in community	0.96 (0.21)	0.96 (0.20)	0.021	0.94 (0.23)	0.97 (0.17)	0.121

SE Standard error, *SMD* Standardised mean differences

Appendix C

See (Figs. 3, 4, 5).



Notes: Left axis: proportion of births covered by each insurance type, in a given year, proportion of birth attended by skilled health professional; right axis: probability of infant mortality

Fig. 3 Trends in the probability of infant mortality, health care utilisation, and in the proportion of births covered by subsidised and contributory insurance

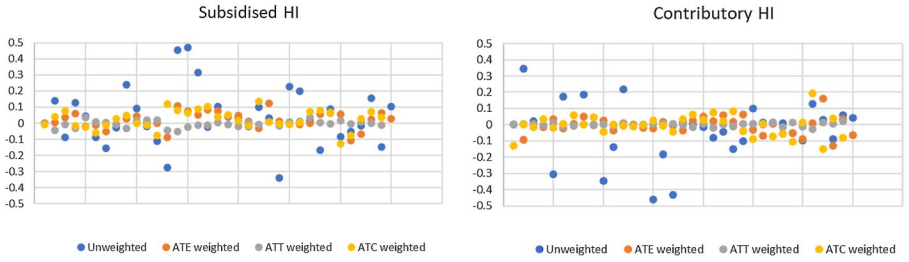
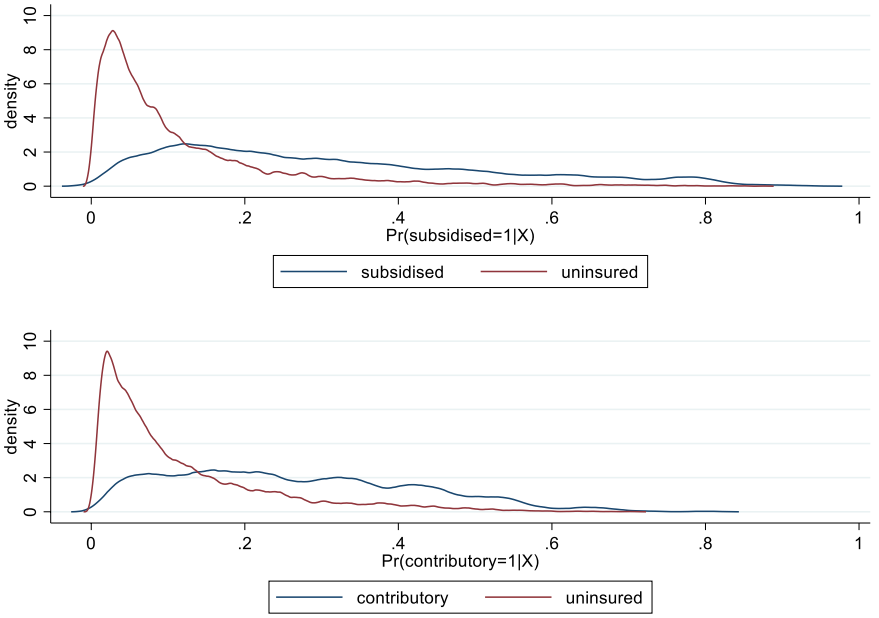


Fig. 4 Overview of balance after propensity score weighting: standardised mean differences of covariates involved in the propensity score analysis



Notes: The top panel shows the estimated probability of being insured among those who were in fact uninsured, while the bottom panel shows the corresponding probabilities for those who were insured. The right (top and bottom) panels show the corresponding estimates for the probability of contributory insurance.

Fig. 5 Estimated propensity score distributions

Appendix D: Software code

```
### FUNCTIONS TO ESTIMATE MODELS ###


#wrapper function for grf to predict the outcome on the confounders, used to
center, or residualised outcome for use in causal forest function

outcome.grf<-function(W, #W is the set of variables controlling for confounding
                      Y, #the outcome of interest
                      cc.weights = rep(1,dim(W)[1]), #, #the treatment
                      tune.num.trees =100,
                      tune.num.reps = 500,
                      num.trees = 500,
                      ci.group.size = 2 #The forest will grow ci.group.size
trees on each subsample.
)
{
  Y.forest = regression_forest(W,
                               Y,
                               sample.weights = cc.weights,
                               ci.group.size = ci.group.size,
                               tune.parameters = "all",
                               tune.num.trees = tune.num.trees,
                               tune.num.reps=tune.num.reps,
                               num.trees = num.trees)

  Yhat.W= predict(Y.forest, W, estimate.variance = TRUE)
  Yhat=as.data.frame(Yhat.W[,1])
  names(Yhat) <-cbind("Yhat")
  #variace
  Var.Yhat<-as.data.frame(Yhat.W[,2])
  names(Var.Yhat) <-cbind("Var.Yhat")

  ###results ####
  return(list=c(Yhat,Var.Yhat))
}


### wrap function for grf to obtain Propensity scores using generalised random
forest
exposure.model.grf<-function(W, #W is the set of variables controlling for
confounding
                     A, #the treatment
                     cc.weights = rep(1,dim(W)[1]), #the column of complete-case
analysis weights, used to address missing data,
                     tune.num.trees = 100,
                     tune.num.reps=500,
                     num.trees = 500,
                     ci.group.size = 2 #The forest will grow ci.group.size trees
on each subsample.
)
{


  W.forest = regression_forest(W,A,
                               sample.weights = cc.weights,
                               ci.group.size = ci.group.size,
                               tune.parameters = "all",
                               tune.num.trees = tune.num.trees,
                               tune.num.reps= tune.num.reps,
```

```
                              num.trees = num.trees)

g.RF = predict(W.forest)$predictions
g.tuned.param<-W.forest$tunable.params

gW=as.data.frame(g.RF)
names(gW) <-cbind("gW")


###results ####
return(list=c(gW,g.tuned.param))
}


#### Code to run subsidised health insurance analysis ###

## 1. Install packages / load libraries
install.packages("grf")
library(grf)

## 2. Run nuisance models ###
set.seed(1234)
cc.weights<-rep(1,dim(data)[1])        ### this can be made a parameter of the
whole function, at the end


## Predicted outcome (pooled)
set.seed(1234)
my.Yhat <- outcome.grf(W, #W is the set of variables controlling for confounding
                       Y, #the outcome of interest
                       cc.weights = rep(1,dim(W)[1]), #, #the treatment
                       tune.num.trees =100,
                       tune.num.reps = 500,
                       num.trees = 1000,
                       ci.group.size = 2 #The forest will grow ci.group.size
trees on each subsample.
)


set.seed(1234)
my.GW<-exposure.model.grf(W, #W is the set of variables controlling for
confounding
                          A, #the treatment
                          cc.weights = rep(1,dim(W)[1]), #the column of
complete-case analysis weights, used to address missing data,
                          tune.num.trees = 100,
                          tune.num.reps=500,
                          num.trees = 1000,
                          ci.group.size = 2 #The forest will grow ci.group.size
trees on each subsample.
)

#### run CF ######

### do these step by step, as I need to look into these results
set.seed(1234)
Rlearner.init<- causal_forest(W,
                              Y,
                              A,
                              Y.hat = my.Yhat$Yhat,
                              W.hat = my.GW$gW,
```

```
                                ci.group.size = 2,
                                tune.parameters = "all",
                                tune.num.trees=100,
                                tune.num.reps = 500,
                                num.trees = 1000)

varimp_forscreen = variable_importance(Rlearner.init)

selected.idx = which(varimp_forscreen > mean(varimp_forscreen))

names(mymatrix)[selected.idx]

### do the CF for the selected variables only ###
set.seed(1234)
Rlearner.model<- causal_forest(W[,selected.idx],
                               Y,
                               A,
                               Y.hat = my.Yhat$Yhat,
                               W.hat = my.GW$gW,
                               ci.group.size = 2,
                               tune.parameters = "all",
                               tune.num.trees=100,
                               tune.num.reps = 500,
                               num.trees = 3000)


Rlearner.predictions <- predict(object = Rlearner.model,
                                estimate.variance = TRUE)


tau.Rlearner <-Rlearner.predictions$predictions
Var.Rlearner.CATE<- Rlearner.predictions$variance.estimates

Rlearner.tuned.param<-Rlearner.model$tunable.params


Rlearner.ite<-as.data.frame(cbind(tau.Rlearner, Var.Rlearner.CATE))
names(Rlearner.ite)<-c("estimate", "Var")


### main ATE ####

Rlearner.ate <-average_treatment_effect(Rlearner.model,target.sample = "all")

ate<-as.data.frame(cbind(Rlearner.ate[1],Rlearner.ate[2]^2))
names(ate)<-c("estimate","Var")

### subgroup ATEs (example)

### wealth ##

s.wealth.1.atc <-average_treatment_effect(Rlearner.model,target.sample =
"control",subset=data$s.wealth==1)
s.wealth.1.atc.CIs<-as.data.frame(cbind(s.wealth.1.atc[1],s.wealth.1.atc[1]-
s.wealth.1.atc[2]*1.96,s.wealth.1.atc[1]+s.wealth.1.atc[2]*1.96))

s.wealth.2.atc <-average_treatment_effect(Rlearner.model,target.sample =
"control",subset=data$s.wealth==2)
s.wealth.2.atc.CIs<-as.data.frame(cbind(s.wealth.2.atc[1],s.wealth.2.atc[1]-
s.wealth.2.atc[2]*1.96,s.wealth.2.atc[1]+s.wealth.2.atc[2]*1.96))
```

```
s.wealth.3.atc <-average_treatment_effect(Rlearner.model,target.sample =
"control",subset=data$s.wealth==3)
s.wealth.3.atc.CIs<-as.data.frame(cbind(s.wealth.3.atc[1],s.wealth.3.atc[1]-
s.wealth.3.atc[2]*1.96,s.wealth.3.atc[1]+s.wealth.3.atc[2]*1.96))

## The p-value of the 'differential.forest.prediction'
#coefficient also acts as an omnibus test for the presence of heterogeneity ###

hettest <- test_calibration(Rlearner.model)
p.val.het<-hettest[2,4]
names(p.val.het)<-("p-val heterog")
### save the variable importance measures  #####


varimp <- as.data.frame(variable_importance(Rlearner.model)) %>%
  mutate(variable = colnames(
    Rlearner.model $X.orig)) %>%
  arrange(desc(V1))
names(varimp)<-c("importance","variable")
```

Acknowledgements We gratefully acknowledge the extremely valuable input from Ryota Nakamura, Mardiati Nadjib, Budi Hidayat, Darius Erlangga, and Peter Smith.

Author contributions All authors contributed to the study conception and design. The construction of the dataset and analysis was performed by Noemi Kreif. The first draft of the manuscript was written by Noemi Kreif and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This study was partly funded by the Bill and Melinda Gates Foundation, through the International Decision Support Initiative (iDSI), and by the National Institute for Health Research (NIHR) (16/137/90) using UK aid from the UK Government to support global health research. The views expressed in this publication are those of the author(s). KDO was supported by a Royal Society-Wellcome Trust Sir Henry Dale Fellowship (218554/Z/19/Z).

Availability of data and material The paper uses a publicly available data source (Indonesian Family Life survey). The code used to construct the analytical dataset can be shared upon request.

Code availability The code used to for the statistical analysis can be found in the Appendix.

Declarations

Conflict of interest The authors declare that they donot have any conflict of interest to declare.

References

Achadi, E.L., Achadi, A., Pambudi, E., Marzoek, P.: A study on the implementation of Jampersal policy in Indonesia.. *Open Knowledge Repository*. https://openknowledge.worldbank.org/handle/10986/20740(2014)

Acharya, A., Vellakkal, S., Taylor, F., Masset, E., Satija, A., Burke, M., Ebrahim, S.: The impact of health insurance schemes for the informal sector in low-and middle-income countries: a systematic review. World Bank Res. Obs. **28**(2), 236–266 (2013)

Agustina, R., Dartanto, T., Sitompul, R., et al.: Universal health coverage in Indonesia: concept, progress, and challenges. Lancet **393**(10166), 75–102 (2019). https://doi.org/10.1016/S0140-6736(18)31647-7

Aji, B., De Allegri, M., Souares, A., Sauerborn, R.: The impact of health insurance programs on out-of-pocket expenditures in Indonesia: An increase or a decrease? Int. J. Environ. Res. Public Health **10**(7), 2995–3013 (2013)

Anindya, K., Lee, J.T., McPake, B., Wilopo, S.A., Millett, C., Carvalho, N.: Impact of Indonesia's national health insurance scheme on inequality in access to maternal health services: a propensity score matched analysis. J. Global Health **10**(1), 010429 (2020). https://doi.org/10.7189/jogh.10.010429

Athey, S., Tibshirani, J., Wager, S.: Generalized random forests. Ann. Stat. **47**(2), 1148–1178 (2019)

Athey, S., Wager, S.: Estimating treatment effects with causal forests: an application. Obs. Stud. **5**(2), 37–51 (2019)

Athey, S., Wager, S.: *Policy Learning with Observational Data*https://www.econometricsociety.org/system/files/15732-4.pdf *(2020)*

Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)

Chen, Y., Jin, G.Z.: Does health insurance coverage lead to better health and educational outcomes? Evidence from rural China. J. Health Econ. **31**(1), 1–4 (2012)

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J.: Double/debiased machine learning for treatment and structural parameters. Economet. J. **21**(1), C1–C68 (2018a)

Chernozhukov, V., Demirer, M., Duflo, E., Fernandez-Val, I.: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India, NBER. https://doi.org/10.3386/w24678 (2018b)

Currie, J., Gruber, J.: Saving babies: The efficacy and cost of recent changes in the Medicaid eligibility of pregnant women. J. Polit. Econ. **104**(6), 1263–1296 (1996)

Davis, J., Heller, S.B.: Using causal forests to predict treatment heterogeneity: an application to summer jobs. Am. Econ. Rev. **107**(5), 546 (2017)

Dow, W.H., Schmeer, K.K.: Health insurance and child mortality in Costa Rica. Soc. Sci. Med. **57**(6), 975–986 (2003)

Erlangga, D., Ali, S., Bloor, K.: The impact of public health insurance on healthcare utilisation in Indonesia: evidence from panel data. Int. J. Public Health **64**(4), 603–613 (2019b)

Erlangga, D., Suhrcke, M., Ali, S., Bloor, K.: The impact of public health insurance on health care utilisation, financial protection and health status in low-and middle-income countries: a systematic review. PloS One **14**(8), e0219731 (2019a)

Fan, Q., Hsu, Y.C., Lieli, R.P., Zhang, Y.: Estimation of conditional average treatment effects with high-dimensional data. J. Bus. Econ. Stat. **11**, 1–5 (2020)

Fink, G., Robyn, P.J., Sié, A., Sauerborn, R.: Does health insurance improve health?: evidence from a randomized community-based insurance rollout in rural Burkina Faso. J. Health Econ. **32**(6), 1043–1056 (2013)

Hainmueller, J., Mummolo, J., Xu, Y.: How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. Polit. Anal. **27**(2), 163–192 (2019)

Harimurti, P., Pambudi, E., Pigazzini, A., Tandon, A.: The nuts & bolts of Jamkesmas, Indonesia's government-financed health coverage program for the poor and near-poor. UNICO Studies Series; No. 8. World Bank, https://openknowledge.worldbank.org/handle/10986/13305 License: CC BY 3.0 IGO (2013)

Hidayat, B., Thabrany, H., Dong, H., Sauerborn, R.: The effects of mandatory health insurance on equity in access to outpatient care in Indonesia. Health Policy Plan. **19**(5), 322–335 (2004)

Ho, D.E., Imai, K., King, G., Stuart, E.A.: Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit. Anal. **15**(3), 199–236 (2007)

Imai, K., Strauss, A.: Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. Polit. Anal. **19**(1), 1–9 (2011)

Johar, M.: The impact of the Indonesian health card program: a matching estimator approach. J. Health Econ. **28**(1), 35–53 (2009)

Kallus, N., Mao, X., Zhou, A.: Interval estimation of individual-level causal effects under unobserved confounding. In: The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, (2019)
Knaus, M.C., Lechner, M., Strittmatter, A.: Heterogeneous employment effects of job search programmes: A machine learning approach. J. Human Resour. (2020). https://doi.org/10.3368/jhr.57.2.0718-9615R1
Knaus, M.C., Lechner, M., Strittmatter, A.: Machine learning estimation of heterogeneous causal effects: empirical monte carlo evidence. Economet. J. **24**(1), 134–161 (2021)
Künzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B.: Metalearners for estimating heterogeneous treatment effects using machine learning. Proc. Natl. Acad. Sci. **116**(10), 4156–4165 (2019)
Lagomarsino, G., Garabrant, A., Adyas, A., Muga, R., Otoo, N.: Moving towards universal health coverage: health insurance reforms in nine developing countries in Africa and Asia. Lancet **380**(9845), 933–943 (2012)
Lawn, J.E., Cousens, S., Zupan, J.: Lancet Neonatal Survival Steering Team 4 million neonatal deaths: when? Where? Why? The Lancet. **365**(9462), 891–900 (2005)
Lechner, M.: Modified causal forests for estimating heterogeneous causal effects. arXiv preprint arXiv:1812.09487 (2018)
Mackenbach, J.P.: Tackling inequalities in health: the need for building a systematic evidence base. J. Epidemiol. Community Health **57**(3), 162 (2003)
Mahendradhata, Y., Trisnantoro, L., Listyadewi, S., Soewondo, P., Marthias, T., et al.: The Republic of Indonesia health system review. *Health Systems in Transition*, Vol-7 No.1. WHO Regional Office for South-East Asia https://apps.who.int/iris/handle/10665/254716 (2017)
Mensah, J., Oppong, J.R., Schmidt, C.M.: Ghana's national health insurance scheme in the context of the health MDGs: an empirical evaluation using propensity score matching. Health Econ. **19**(S1), 95–106 (2010)
Moreno-Serra, R., Smith, P.C.: Broader health coverage is good for the nation's health: evidence from country level panel data. J. Royal Stat. Soc. Series A (Stat. Soc.). 178(1):101 (2015)
Nie, X., Wager, S.: Quasi-oracle estimation of heterogeneous treatment effects. Biometrika **108**(2), 299–319 (2021)
O'donnell, O., Van Doorslaer, E., Wagstaff, A., Lindelow, M.: Analyzing health equity using household survey data: a guide to techniques and their implementation. World Bank, Washington, DC. https://openknowledge.worldbank.org/handle/10986/6896 License: CC BY 3.0 IGO (2008)
Petticrew, M., Tugwell, P., Kristjansson, E., Oliver, S., Ueffing, E., Welch, V.: Damned if you do, damned if you don't: subgroup analysis and equity. J. Epidemiol. Community Health **66**(1), 95–98 (2012)
Robins, J.M., Rotnitzky, A., Zhao, L.P.: Estimation of regression coefficients when some regressors are not always observed. J. Am. Stat. Assoc. **89**(427), 846–866 (1994)
Robinson, P.M.: Root-N-consistent semiparametric regression. Economet.: J. Economet. Soc. **56**: 931–54 (1988)
Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
Sachs, J.D.: From millennium development goals to sustainable development goals. Lancet **379**(9832), 2206–2211 (2012)
Scarpa, J., Bruzelius, E., Doupe, P., Le, M., Faghmous, J., Baum, A.: Assessment of risk of harm associated with intensive blood pressure management among patients with hypertension who smoke: a secondary analysis of the systolic blood pressure intervention trial. JAMA Netw. Open **2**(3), e190005 (2019)
Shrestha, R.: The village midwife program and infant mortality in Indonesia. Bull. Indones. Econ. Stud. **46**(2), 193–211 (2010)
Semenova, V., Chernozhukov, V.: Debiased machine learning of conditional average treatment effects and other causal functions. Economet. J. **24**(2), 264–289 (2021)
Sparrow, R., Budiyati, S., Yumna, A., Warda, N., Suryahadi, A., Bedi, A.S.: Sub-national health care financing reforms in Indonesia. Health Policy Plan. **32**(1), 91–101 (2017)
Sparrow, R., Suryahadi, A., Widyanti, W.: Social health insurance for the poor: targeting and impact of Indonesia's Askeskin programme. Soc. Sci. Med. **1**(96), 264–271 (2013)
Strauss, J., Witoelar, F., Sikoki, B.: The fifth wave of the Indonesia family life survey (IFLS5) overview and field report (2016)
Strauss, J., Witoelar, F., Sikoki, B., Wattie, A.: The fourth wave of the Indonesian family life survey (IFLS4): overview and field report (2009)
Strauss, J., Beegle, K., Sikoki, B., Dwiyanto, A., Herawati, Y., Witoelar, F.: The third wave of the Indonesia family life survey (IFLS): overview and field report." (WR-144/1-NIA/NICHD) (2004)
Tchetgen Tchetgen, E.J., VanderWeele, T.J.: On causal inference in the presence of interference. Stat. Methods Med. Res. **21**(1), 55–75 (2012)

Thabrany, H.: Private health sector in Indonesia: opportunities and progress. J. Indones. Med. Assoc. **5**, 1–3 (2001)
Tibshirani, J., Athey, S., Wager, S., Friedberg, R., Miner L., Wright, M.: grf: Generalized random forests (Beta). R package version 1.1.0.2018 (2018)
Trujillo, A.J., Vecino Ortiz, A.I., Ruiz Gómez, F., Steinhardt, L.C.: Health insurance doesn't seem to discourage prevention among diabetes patients in Colombia. Health Aff. **29**(12), 2180–2188 (2010)
Unicef: SDGs for children in Indonesia. Provincial snapshot: East Java. https://www.unicef.org/indonesia/sites/unicef.org.indonesia/files/2019-05/East_Java_ProvincialBrief.pdf
Van der Laan, M.J., Rose, S.: Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media https://link.springer.com/book/10.1007/978-1-4419-9782-1 (2011)
VanderWeele, T.J., Luedtke, A.R., van der Laan, M.J., Kessler, R.C.: Selecting optimal subgroups for treatment using many covariates. Epidemiology (Cambridge, Mass.) **30**(3):334 (2019)
VanderWeele, T.J.: Concerning the consistency assumption in causal inference. Epidemiology **20**(6), 880–883 (2009)
Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. J. Am. Stat. Assoc. **113**(523), 1228–1242 (2018)
Wagstaff, A.: Estimating health insurance impacts under unobserved heterogeneity: the case of Vietnam's health care fund for the poor. Health Econ. **19**(2), 189–208 (2010)
Wang, H., Yip, W., Zhang, L., Hsiao, W.C.: The impact of rural mutual health care on health status: evaluation of a social experiment in rural China. Health Econ. **18**(S2), S65-82 (2009)
World health statistics 2019: monitoring health for the SDGs, sustainable development goals. Geneva: World Health Organization; 2019. Licence: CC BY-NC-SA 3.0 IGO.
Wooldridge, J.M.: Inverse probability weighted estimation for general missing data problems. J. Economet. **141**(2), 1281–1301 (2007)
Yilma, Z., Mebratie, A., Sparrow, R., Dekker, M., Alemu, G., Bedi, A.S.: Impact of Ethiopia's community based health insurance on household economic welfare. World Bank Econo. Rev. **29**(supp_1), 164 (2015)

Authors and Affiliations

Noemi Kreif[1] · Karla DiazOrdaz[2] · Rodrigo Moreno-Serra[1] · Andrew Mirelman[1] · Taufik Hidayat[3] · Marc Suhrcke[1,4]

[1] Centre for Health Economics, University of York, Heslington, York YO10 5DD, UK

[2] Department of Medical Statistics, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK

[3] Center for Health Economics and Policy Studies (CHEPS), Faculty of Public Health, Universitas Indonesia, Depok, Indonesia

[4] Luxembourg Institute of Socio-Economic Research, 11 Porte des Sciences, 4366 Esch-sur-Alzette, Luxembourg