Investigating the causal effect of deforestation on infant health through soil characteristics: a comparison of traditional and machine learning mediation analysis using simulated and real data

ABSTRACT

Deforestation is a global threat that impacts the environment and biodiversity in several ways. Recently, some scholars' started investigating how deforestation affects human health. This paper aims to assess the causal effect of deforestation on infant health in Nigeria, evaluating the mediating role of soil characteristics. We analyse data relative to thousands of Nigerian children collected in 2008 and 2013 using two mediation analysis approaches: a traditional, regression-based approach, and a machine learning one. Given the novelty of the latter, the first part of the paper is devoted to illustrating the two approaches and comparing their performances in terms of estimates' bias and coverage rates through a simulation study. The second part of the work focuses on the analysis of deforestation data: in particular, we analysed the effect of deforestation on Nigerian children's probability of suffering from cough, dirrhea and malaria, through the pH, the organic carbon and cation levels of soil. The results show mixed evidence of the effect of deforestation on infant health, either because the results differ across the two approaches mainly in terms of significance, and because the significant ones show different signs. This leaves room for further analyses on this topic.

KEYWORDS

forest loss; indirect effects; malaria; Nigeria

1. Introduction

Deforestation is widely considered one of the major threats to the environment of our era. Although this phenomenon has characterised and fostered human development since ancient times, it has assumed alarming proportions in the last centuries, with approximately 10 million of forest hectares lost per year, on average, according to the Food and Agriculture Organisation.

The adverse effects of deforestation on the environment are well known (Chakravarty et al., 2012; Cook et al., 1990): first, it leads to soil degradation, with a decrease in soil organic matter, nitrogen, and other essential nutrients, resulting in lower soil fertility and productivity (Davari et al., 2020; Hajabbasi et al., 2004; Mainville et al., 2006); second, deforestation alters ecosystems, reducing fauna and microbial communities,

often favouring the spread of disease vectors (Morand and Lajaunie, 2020); third, a reduced number of trees leads both to an increase of local temperatures (Wolff et al., 2018) and a worsening of air quality (Butt et al., 2021). All these pose a threat for human health; nonetheless, the impact of deforestation on human illnesses is less studied. Some scholars have recently started analysing this link, focusing, for example, on the association between forest loss and malaria incidence, see Bauhoff and Busch (2020); MacDonald and Mordecai (2019); Pattanayak and Yasuoka (2012) and Tucker Lima et al. (2017) for a review. Reddington et al. (2015) and Chakrabarti (2021) studied the effect of deforestation on adult mortality in Brazil and infant mortality in Indonesia, respectively. Two recent studies (Alimi et al., 2021; World Economic Forum, 2020) have highlighted the association between deforestation and an increase of zoonotic and infectious diseases like Zika and Ebola.

Berazneva and Byker (2017a) investigated the causal effect of deforestation in Nigeria on three infant diseases: malaria, respiratory infections and diarrhea. In this paper, we deepen their analysis by including soil characteristics as possible mediating factors. Although other works have included mediators to study the relationship between forest loss and human health, they are always subject-specific variables like women's time devoted to children's care (Johnson et al., 2013) or net income Galway et al. (2018). To the best of our knowledge, this is the first paper assessing the mediating effects of soil characteristics. There exists a vast literature on approaches for estimating mediational effects, ranging from those that require the specification of parametric models (Doretti et al., 2022; Imai et al., 2010a), to the more flexible ones proposing semi/nonparametric inference (Bae et al., 2024; Roy et al., 2024; Tchetgen and Shpitser, 2012), and the most recent based on machine learning (Farbmacher et al., 2022; Wang and Huang, 2024). In this paper, we want to compare a traditional, regression-based approach (Imai et al., 2010b; Vanderweele, 2015) and a machine-learning-based approach (Farbmacher et al., 2022). Given the relative novelty of the latter, we run a simulation study to compare the performance of these approaches either in single- and multilevel scenarios. Also in this case, we are not aware of other papers carrying out a similar analysis. Our contribution is, then, twofold: on the one hand, we provide a methodological contribution, comparing a well-established approach to mediation

analysis based on regression models to a novel double machine learning (DML) approach; on the other hand, we analyse data that were already studied from a novel perspective.

The paper is structured as follows: we first provide an overview of the studies that addressed the relationship between deforestation and human health; then, in Section 3, we discuss causal mediation analysis and the two estimation approaches, which are compared in Section 4 through a simulation study. In the fifth section we analyse the data, and a discussion follows.

2. Literature review

Deforestation affects most countries in the world, in particular, tropical areas: according to the World Resources Institute (WRI), the tropics lost 3.7 million hectares in 2023 (World Resources Institute, 2024). For this reason, many studies investigating the relationship between forest loss and human health focused on diseases widely spread in tropical areas like malaria. Bauhoff and Busch (2020) and Estifanos et al. (2024) analysed the impact of deforestation on children's malaria rates in Africa, using data from different sources, among which, data from local surveys to estimate malaria prevalence among children under 5 years and satellite data to quantify forest loss. Both studies used multivariate statistical models to carry out the analyses, with fixed and random effects, respectively. They do not find any significant association between deforestation and malaria, but Estifanos et al. (2024) noted that a strong association can instead be found within the poorest strata of the population.

Rerolle et al. (2021) and Matsumoto-Takahashi et al. (2023) focused on malaria prevalence in the Lao People's Democratic Republic (Lao PDR). The former carried out a thorough spatiotemporal analysis using mixed-effect models and semiparametric models with splines for modelling possible nonlinear temporal trends. Their results show an increase of malaria cases in the short-term period after deforestation and a decrease in the long-term, a theory known as "frontier malaria". These results are consistent with those obtained by Matsumoto-Takahashi et al. (2023), who analysed data on deforestation and malaria from 2007 to 2015 through structural equation

modelling in relation to climate change.

Laporta et al. (2021) investigated how deforestation affected the abundance of anopheline infected with the main *Plasmodium* parasites responsible for malaria transmission in 12 municipalities in four Brazilian Amazon states between 2015 and 2017: using generalised linear models, they confirmed the frontier malaria theory. In contrast, Aguirre et al. (2024) used a spatial Durbin Watson model to assess direct and spillover effects of deforestation on malaria infections in the Peruvian Amazon region, and they found a positive effect of deforestaion on the number of new cases.

Although infectious diseases are one of the primary outcomes of interest related to forest loss, other health outcomes have been considered as well. For example, Fuentes Cordoba (2024) used multivariate regression analysis with clustered standard errors to examine the effect of prenatal exposure to deforestation in Cambodia on some children's health outcomes, i.e. birth weight, height-for-age z-score, and weight-for-age z-score; the findings show a negative impact on all the outcomes of interest. Santika et al. (2023) studied the impact of forest loss in Sumatra (Indonesia) on the prevalence of respiratory diseases, finding that their prevalence is higher in more deforested areas, whereas Bolton et al. (2022) performed an extensive analysis using data from 230 countries to assess the effect of deforestation on mental health, and their findings suggest a positive association, but only in low- and middle-income countries.

Other studies suggested that the impact of deforestation on human health may be indirect, conveyed, for example, by water quality (Ma et al., 2014), or increase of local temperatures (Palmer, 2022), but they do not present any formal mediation analysis. The studies mentioned so far are associational, but some scholars tried to address the relationship between deforestation and human illnesses from a causal perspective, leveraging either on data collected from experimental settings (Masuda et al., 2020), or using instrumental variables on observational data (MacDonald and Mordecai, 2019). In this paper, we will address mediation analysis within the counterfactual framework described in the next section.

3. Causal mediation analysis

In this section, we introduce the counterfactual framework and the two estimation approaches towards causal mediation analysis, which will be used throughout the paper.

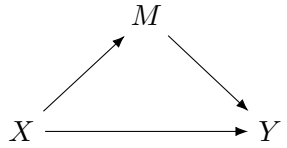
3.1. Counterfactual framework

There are different ways to define the causal effect of a treatment/exposure X on a response variable of interest Y, and one of the most widely employed frameworks is the counterfactual one (Morgan and Winship, 2014; Pearl, 2009b; Rubin, 1974). A counterfactual object (a sentence or a variable) involves a situation contrary to the factual one, i.e. contrary to the one that really occurred. The causal effect of X on Y can be defined by answering the question "Would Y have changed if X had changed?", or in other words, by comparing the values of Y under two different conditions of X, i.e. two counterfactuals. For example, if X is binary, to estimate the causal effect of X on Y, one should compare (generally via difference or ratio) the counterfactuals $Y(1)$ and $Y(0)$, i.e. the values assumed by the outcome if X were set to 1 and 0, respectively.

Determining the causal effect for a single subject is not straightforward; for this reason, many studies focus on the average causal effect (ATE), defined as $ATE = \mathbb{E}[Y_i(1) - Y_i(0)]$ for $i = 1, \dots, n$ subjects in a sample. It is clear that, for each subject, just one between $Y(1)$ and $Y(0)$ is actually observed, while the other is missing. This is the fundamental problem of causal inference and leads to the question about how to estimate the ATE. It can be proved that, if the relationship between X and Y is not affected by any unobserved confounders, formally written as $Y(x) \perp\!\!\!\perp X|Z$, where Z is a set of observed confouders, the average total *conditional* effect is identifiable from observed data and is equal to $ATCE = \mathbb{E}[Y_i \,|\, X = 1, Z] - \mathbb{E}[Y_i \,|\, X = 0, Z]$, an expression that does not involve any counterfactual quantity.

Assumption $Y(x) \perp\!\!\!\perp X|Z$ is called *conditional ignorability* or *exchangeability* and it holds by construction in experimental settings where X is randomised, while in observational settings this is not guaranteed, and it is generally necessary to adjust for a large number of confounders. See Morgan and Winship (2014); Pearl (2009b)

Figure 1. Basic mediation model.



and Hernán and Robins (2020) for a broader discussion on exchangeability and its meaning.

3.2. Causal mediation and natural mediational effects

Let us now move to a mediational setting. The basic idea is that the effect of X on Y is (at least partially) conveyed by an intermediate variable M called *mediator*, as shown in Figure 1

In mediational settings, in addition to the total effect, the interest lies in two other kinds of effects, the direct and indirect effects. The former is the part of the effect of X on Y entirely attributable to X or, in any case, not transmitted by M (although it can, in principle, be transmitted by other mediators), while the latter is the portion of effect conveyed by the mediator of interest M. In the literature, several definitions of mediational effects have been proposed, but in this paper, we focus on the oldest and more widely spread kinds of effects, introduced by Robins and Greenland (1992), which were called *natural* by Pearl (2001). Assuming a binary X, the natural direct effect (NDE) and the natural indirect effect (NIE) on the difference scale are given by

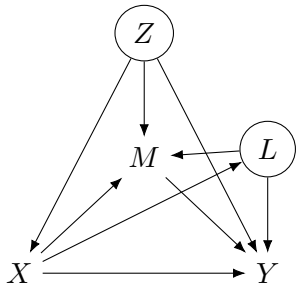
$$NDE = \mathbb{E}[Y(x, M(x)) - Y(x^*, M(x))] \tag{1}$$

$$NIE = \mathbb{E}[Y(x, M(x)) - Y(x, M(x^*))] \tag{2}$$

where, for $x, x^* \in \{0, 1\}$, $M(x)$ denotes the value of the mediator if X has been set to x, and $Y(x, M(x^*))$ denotes the value of the response if X has been set to x and the mediator to the *natural* value it would assume under x^*.

The NDE corresponds to the $X \to Y$ path in Figure 1 and expresses the average change in the outcome observed if the mediator were set to a fixed (counterfactual) value, and the value of X changed from x^* to x. Analogously, the NIE corresponds to

Figure 2. Mediation model with unobserved confounders.



path $X \to M \to Y$ and represents the average change in the outcome if the value of X were fixed and the mediator changed from the value it would assume in the scenario $X = x^*$ to the value it would have in the scenario $X = x$. The sum of the NDE and the NIE is equal to the ATE. Notice that it is implicitly assumed that the variables are temporally ordered, i.e. the exposure precedes the mediator, which in turn precedes the outcome. This is a key aspect of causality since it allows us to establish the direction of relationships among variables Pearl (2009a), even in the absence of longitudinal data, which are often more suitable to assess causal effects.

The natural direct and indirect effects can be identified under the following sufficient assumptions (Imai et al., 2010b; Vanderweele, 2015): (i) No confounders of the exposure-mediator and exposure-outcome relationships: $\{Y(x^*, m), M(x)\} \perp\!\!\!\perp X | Z = z$, (ii) No-confounders of the mediator-outcome relationship: $Y(x^*, m) \perp\!\!\!\perp M(x) | \{X = x, Z = z\}$ where Z are again additional observed covariates, and it is also assumed that $P(X = x|Z = z)$ and $P(M = m|X = x, Z = z)$ are both positive, $\forall\, x, m$ and z. Unobserved confounders of the exposure-mediator, exposure-outcome and mediator-outcome relationships, like Z in Figure 2 violate both assumptions. In addition, it is worth remarking that assumption (ii) is untestable and may be violated in different scenarios (Andrews and Didelez, 2020), but one of the most common is in the presence of observed *exposure-induced* or *post-treatment* mediator-outcome confounders like L in Figure 2, i.e. confounders of the M-Y relationship affected by the exposure. See VanderWeele et al. (2014) and Andrews and Didelez (2020) for an extensive discussion.

If assumptions (i)-(ii) are satisfied, it can be proved that the nested counterfactual quantity $\mathbb{E}[Y(x, M(x^*))]$, $x, x^* \in \{0, 1\}$, $x \neq x^*$, present in all effects (1)-(2), is

identifiable and is given by

$$\mathbb{E}[Y(x, M(x^*))] = \sum_m \mathbb{E}[Y \mid X = x,\, M = m] P(M = m \mid X = x^*), \tag{3}$$

which is known as *mediational formula* (Pearl, 2012) and allows us to easily obtain expressions for the NDE and the NIE.

3.3. Estimation procedures

Over the years, plenty of approaches to estimate natural mediational effects have been proposed in the literature. The most widely spread rely on the specification of parametric models for the mediator and the outcome and the estimation of effects using closed-form formulas, as done, for example, by Valeri and VanderWeele (2013), Doretti et al. (2022) and Samoilenko and Lefebvre (2021) in settings with binary mediators and/or outcomes, or by simulating counterfactuals to estimate the expectations required to obtain the NDE and the NIE (Imai et al., 2010a). Parametric approaches present several shortcomings since, as highlighted by Tchetgen Tchetgen (2013), they can be subject to model misspecification, leading to possibly biased estimates; in addition, even if the mediator and the outcome model are correctly specified, the use of link functions different from identity generates expressions for natural effects which are difficult to interpret. This induced many authors to move to more flexible estimation methods, i.e. semiparametric and nonparametric approaches. Tchetgen and Shpitser (2012) proposed semiparametric and doubly-robust estimators for natural effects based on inverse probability weighting, which showed good properties in simulations. The flexibility of semiparametric estimation was exploited also to estimate heterogeneous mediational effects (Kim et al., 2018), different types of model misspecification (Xia and Chan, 2021) and the presence of post-treatment confounders (Sun and Ye, 2023). Some of these approaches were extended to be fully nonparametric, especially in the context of Bayesian inference (Bae et al., 2024; Roy et al., 2024). In some applied contexts, the estimation of mediational effects is made more complex by the presence of high-dimensional mediators or confounders, and the possibly com-

plex causal structure among them, which is generally unknown. To address this issue, some scholars proposed to combine causal mediation analysis and machine learning: Wang and Huang (2024) combined deep neural network and penalized regression to handle nonlinear relationships between confounders and mediators and outcome, and high-dimensionality of mediators, respectively, while Nath et al. (2023) implemented an iterative approach based on the estimation of high-dimensional mediation models through machine learning, whose outputs are used in standard outcome regression models. Farbmacher et al. (2022) proposed the estimation of mediational effects using inverse probability weighting leveraging on machine learning in the context of high-dimensional confounders.

We will focus on the approach proposed by Imai et al. (2010a), implemented in the R `mediation` package and quite spread among practitioners, and on the recent approach discussed in Farbmacher et al. (2022).

Let us start by the approach in Imai et al. (2010b), which is based on sampling Monte Carlo draws of $Y(x, M(x^*))$ through the mediator and outcome models' predictions. Their algorithm can be summarised in the following steps:

1. Fit models for the observed outcome and mediator.
2. Simulate model parameters from their sampling distribution.
3. Repeat the following three steps: (a) simulate the potential values of the mediator, (b) simulate the potential outcomes given the simulated values of the mediator, (c) compute the causal mediation effects.
4. Compute summary statistics of interest (mean, confidence intervals) based on the simulated effects.

In other words, the algorithm simulates nested counterfactuals based on the fitted models for the mediator and the outcome. This approach can be extended also to semi- and non-parametric models, by using bootstrap.

Farbmacher et al. (2022) propose an alternative estimand for $\mathbb{E}[Y(x, M(x^*))]$:

$$
\begin{aligned}
&\mathbb{E}[Y(x, M(x^*))] = \mathbb{E}[\psi_x^*] \qquad \text{where} \\
&\psi_x^* = \frac{I\{X = x\}P(X = x|M, Z)}{P(X = x|M, Z)(1 - P(X = x|Z))} \cdot (Y - \mathbb{E}[Y|X = x, M, Z]) \\
&+ \frac{I\{X = x^*\}}{1 - P(X = x|Z)} \Big[\mathbb{E}[Y|X = x, M, Z] - \mathbb{E}\big[\mathbb{E}[Y|X = x, M, Z]\big|X = x^*, Z]\big]\Big] \\
&+ \mathbb{E}\big[\mathbb{E}[Y|X = x, M, Z]\big|X = x^*, Z\big]
\end{aligned}
\tag{4}
$$

Denoting by $\mathcal{W}$ the set $\{W_i, 1 \leq 1 \leq n\}$, where $W_i = (X_i, M_i, Y_i, Z_i)$ in an i.i.d sample of n observations, Farbmacher et al. (2022) propose the following algorithm to estimate $\mathbb{E}[Y(x, M(x^*))]$, based on K-fold cross-fitting:

1. Split $\mathcal{W}$ in K subsamples and, for each subsample k, denote by n_k, W_k and W_k^C, its size, the observations falling in the sample and the observations belonging to the complement of W_k, respectively.
2. For each k, use W_k^C to estimate the model parameters of $P(X = x|Z)$ and $P(X = x|M, Z)$. Split W_k^C into two nonoverlapping subsamples and use one to estimate the model parameters of $\mathbb{E}[Y|X = x, M, Z]$ and the other to estimate the nested conditional mean $\mathbb{E}\big[\mathbb{E}[Y|X = x, M, Z]\big|X = x^*, M\big]$. In $\mathcal{W}$, predict $P(X = x|Z), P(X = x|M, Z), \mathbb{E}[Y|X, M, Z]$ and $\mathbb{E}\big[\mathbb{E}[Y|X = x, M, Z]\big]$.
3. For each k and i in $\mathcal{W}_k$ estimate the score function $\hat{\psi}_{x,i}^{k*}$ by plugging in the predictions obtained in the previous steps in formula (4).
4. Average the estimated $\hat{\psi}_{x,i}^{k*}$ over all subjects across all subsamples: $\hat{\psi}_x^* = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\hat{\psi}_{x,i}^{k*}$.

The main difference between the two approaches is that the former requires the specification of mediator and outcome models to simulate the potential outcomes, while the latter relies on the estimation of propensity scores and conditional expectations. Although this paper focuses on the approach of Imai et al. (2010b) based on parametric models, they implemented algorithms also for semi- and nonparametric models; this implies that a correct model specification is crucial to obtain "good" simulated potential outcome and, as a consequence, unbiased estimates of natural effects. In contrast, in the approach proposed by Farbmacher et al. (2022), what matters is to correctly

predict the probabilities and expectations in Equation 4; to do so, in principle, it would be possible to use any estimation strategy, including parametric modelling. However, it is well known that machine learning approaches are particularly suited for providing accurate predictions, indeed, this is what Bodory and Huber (2018) implement in the R `causalweight` package using lasso as machine learner, although other choices are possible (Farbmacher et al., 2022). This suggests that the two approaches should be equivalent in the case of correctly specified mediator and outcome models, while DML-based mediation could be more accurate than traditional mediation if the mediator and/or outcome models are misspecified and in case of a high number of confounders.

4. Simulation study: empirical design and results

To compare the two estimation approaches, we evaluated their performance through a simulation study where we estimated the total, direct and indirect effects using either regression models and machine learning, and we assessed their bias and standard error and the coverage of confidence intervals at a 0.05 level.

4.1. Simulation design The simulation setting is similar to that in Farbmacher et al. (2022), but presents some differences.

The mediator and the outcome were simulated in two ways: in the first case, both from two Normal distributions with the following expectations

$$\mu_M = 0.5 \cdot X + \sum_{k=1}^{4} \beta_{zk}^T Z_k \tag{5}$$

$$\mu_Y = 0.5 \cdot X + 0.5 \cdot M + 0.5 \cdot XM + \sum_{k=1}^{4} \beta_{zk}^T Z_k, \tag{6}$$

respectively, where the covariates Z are drawn from a multivariate Normal with zero mean and identity covariance matrix, and the exposure is $I(\beta_{zk}^T Z_k + \varepsilon_x > 0)$ with ε_x from a standard Normal and $I(\cdot)$ being the indicator function. The β coefficients are obtained as $\beta_k = 0.3/k^2$ for $k = 1, \ldots, 4$. The specification of the mediator and outcome expectations in (5) and (6), of X and β's is equal to that adopted by Farb-

macher et al. (2022) in their simulation study, although we use a smaller number of covariates (4 vs 200). In the second case, the mediator is as before, while the outcome is from a Bernoulli with probability of success $\pi = \text{logit}^{-1}(\mu_Y)$, where μ_Y is as in (6). To make the study more complete, we also simulated the outcome from a Poisson distribution with expectation $\exp(\mu_Y)$, with μ_Y as in Equation 6. The results are reported in the Supplementary Material. For each setting, we generated $K = 1000$ data sets with sample sizes $n = 200$ and 1000; each scenario took approximately 55 minutes to run. It is worth remarking that, although our study is inspired by that of Farbmacher et al. (2022), it differs from it in some aspects. First, we considered the case of a Normal mediator, while the outcome was simulated first from a Normal and then from a Bernoulli, while they used a binary mediator and a Normal outcome. Second, we considered smaller sample sizes than Farbmacher et al. (2022), i.e. 200 and 1000, vs 100 and 4000, respectively; finally, our aim is different, since we want to compare the performance of DML mediation with that of the traditional regression-based approach, while Farbmacher et al. (2022) want to test their approach in high-dimensional settings.

Since the data of our interest are multilevel, we also considered this kind of design, simulating data with a clustered structure, varying the number of clusters $n_c = 20, 50, 100$ and the number of subjects $n_s = 10, 20$. Data follow a $2 \to 2 \to 1$ design, using the notation introduced by Krull and MacKinnon (1999, 2001), i.e. the exposure and the mediator are cluster-level (second-level) variables, while the outcome is assessed at the subject-level (first-level). Covariates are simulated at the first level as well, and this mimics the structure of the data, as we will see in the next section. In particular, the exposure is $I(U > 0)$, $U \sim N(0, 1)$ and the mediator and the outcome are generated from Normal distributions with the following expectations

$$\mu_{Mj} = 0.5 \cdot X_j \tag{7}$$

$$\mu_{Yij} = \alpha_j + 0.5 \cdot X_j + 0.5 \cdot M_j + 0.5 \cdot X_j M_j + \sum_{k=1}^{4} \beta_{z_k}^T Z_{kij}, \tag{8}$$

where $i = 1, \ldots, n_s, j = 1, \ldots, n_c$ denote the subject and the cluster, respectively,

and α_j is a random intercept generated from a Normal distribution with 0 mean and $\sigma = 0.2$.

The simulations were carried out using the R statistical software (version 4.3.0), through the packages `mediation` and `causalweight`, corresponding to the first and the second estimation approaches discussed, respectively. Specifically, to use `mediation` (Tingley et al., 2013), we first fitted the mediator and outcome models, which are used in the package to estimate regression coefficients and use them to simulate model parameters and the values of potential outcomes, as explained in Section 3. The `causalweight` package (Bodory and Huber, 2018) divides the data into a training and a test set and uses the functions `rlasso` and `rlassologit` to estimate the expectations in Equation 4.

Contrary to `mediation`, the `causalweight` package cannot accommodate clustered data structures; then, we considered what happens when clusters are totally ignored and when the cluster membership is included in the analysis as an additional covariate. To better interpret the results, we considered only the linear case.

4.2. Results

As already mentioned, we evaluated the performance of the two estimation approaches through the bias and the standard errors of the mediational effects, as well as the coverage of corresponding confidence intervals. These measures are quite standard in mediation analysis works comparing different approaches to estimate mediational effects, since the bias allows us to measure how much the estimate diverge from the true value of the parameter, the estimate standard errors provides a measure of variability, informing us on the estimator precision, and the coverage rate allows us to assess if the confidence intervals contain the true parameter a number of times consistent with the nominal level. See, for example, the study carried out by Smith et al. (2019) to assess the performance of khb estimators in ordinal mediation models, the simulations conducted by Burgos Ochoa et al. (2020) to compare traditional and causal mediation analysis in time-to-event settings, and the quite recent study by Falk et al. (2024), who compared the performance bootstrap and Bayesian approaches for estimating mediation effects in the context of multilevel data. Results of simulations for the first

Table 1. Results of the simulation study where both the mediator and the outcome are generated from a linear model, using regression-based and DML mediation to obtain mediational effects.

			Regression				DML			
n	Effect	True	Estimate	Bias	S.E.	Coverage	Estimate	Bias	S.E.	Coverage
200	ATE	1.000	**1.001**	**-0.001**	**0.201**	0.957	0.993	0.007	0.213	**0.962**
	NDE	0.500	**0.498**	**0.002**	**0.177**	**0.950**	0.495	0.004	0.200	0.941
	NIE	0.500	0.503	-0.003	**0.165**	**0.948**	0.497	0.003	0.190	0.934
1000	ATE	1.000	**1.003**	**-0.003**	0.085	**0.955**	1.010	-0.009	0.085	0.953
	NDE	0.500	**0.503**	**-0.003**	**0.075**	**0.958**	0.506	-0.006	0.076	0.946
	NIE	0.500	**0.500**	**0.000**	**0.069**	0.943	0.503	-0.003	0.073	**0.956**

Notes: Values in bold indicate a better performance.

Table 2. Results of the simulation study where the mediator is Normally distributed and the outcome is binary, using regression-based and DML mediation to obtain mediational effects.

			Regression				DML			
n	Effect	True	Estimate	Bias	S.E.	Coverage	Estimate	Bias	S.E.	Coverage
200	ATE	0.187	0.184	0.004	**0.068**	**0.958**	**0.187**	**0.001**	0.074	0.947
	NDE	0.097	**0.096**	**0.001**	**0.070**	**0.961**	0.094	0.003	0.082	0.951
	NIE	0.091	0.087	0.003	**0.034**	**0.947**	**0.093**	**-0.002**	0.047	0.914
1000	ATE	0.187	0.186	0.002	0.031	0.944	**0.187**	**0.000**	0.031	0.944
	NDE	0.097	0.096	0.001	**0.032**	0.939	**0.097**	**0.000**	0.033	**0.941**
	NIE	0.091	0.090	0.001	**0.015**	**0.952**	**0.091**	**0.000**	0.017	0.951

Notes: Values in bold indicate a better performance.

setting, where the data do not present a clustered structure, are shown in Tables 1 and 2.

It can be noted that both approaches perform well in all scenarios. When the mediator and the outcome are generated from Normal distributions (Table 1), the regression-based approach performs slightly better than DML, since the estimates are less biased and their standard errors smaller, moreover, the coverage rates are always in line with the nominal level 0.95. For $n = 1000$, the performance of the two estimation approaches is quite similar, although the regression-based one is still slightly superior. Moving to the nonlinear setting (2), the difference between regression-based and DML mediation is even less pronounced. Indeed, the effects' estimates are quite close to the true values in both cases; when $n = 200$, regression outperforms DML in terms of estimates' standard errors and confidence intervals' coverage rates, while for $n = 1000$ DML estimates show basically no bias. The standard errors and the coverage rates are comparable across the two approaches.

Results relative to the simulation study in the setting with clustered data are reported in Table 3. In all scenarios, the regression-based approach performs better than DML, especially in terms of bias and coverage rates, which for the former are always in line with the nominal level 0.95, while for the latter are always below, especially for

the NIE, whose coverage rates are between 30% and 60%. DML, either considering or not considering clusters, systematically underestimates the total and the indirect effects, and overestimates the direct effect, although both phenomena tend to attenuate as sample size increases. Within each approach, with the same cluster sizes, there are no remarkable differences among the results obtained with $n_s = 10$ and $n_s = 20$. The standard errors of DML estimates are smaller than those of regression-based estimates, as expected, since DML does not take into account the clustered structure of data, and, in general, we do not note any difference between the performance of DML when ignoring clusters and when taking them into account.

Table 3. Results of the simulation study for data with a clustered structure. The numbers of clusters and subjects are denoted by n_c and n_s respectively. Regression-based and DML mediation were compared, the latter either excluding and including cluster membership as an additional covariate.

				Regression				DML				DML with clusters			
n_c	n_s	Effect	True	Estimate	Bias	S.E.	Coverage	Estimate	Bias	S.E.	Coverage	Estimate	Bias	S.E.	Coverage
20	10	ATE	0.999	**1.024**	**-0.026**	0.478	**0.944**	0.921	0.077	**0.194**	0.764	0.915	0.083	0.203	0.764
		NDE	0.499	**0.480**	**0.019**	0.314	**0.978**	0.530	-0.031	**0.197**	0.864	0.542	-0.043	0.208	0.853
		NIE	0.500	**0.544**	**-0.044**	0.548	**0.923**	0.392	-0.031	**0.167**	0.609	0.373	0.127	0.169	0.598
	20	ATE	0.999	**1.010**	**-0.012**	0.451	**0.948**	0.923	0.075	**0.134**	0.638	0.913	0.086	0.139	0.642
		NDE	0.499	**0.497**	**0.001**	0.275	**0.987**	0.530	-0.031	**0.130**	0.800	0.545	-0.047	0.135	0.791
		NIE	0.500	**0.513**	**-0.013**	0.528	**0.928**	0.393	-0.031	**0.112**	0.505	0.368	0.132	**0.112**	0.511
50	10	ATE	1.001	**1.006**	**-0.005**	0.269	**0.954**	0.981	0.020	**0.117**	0.700	0.977	0.024	0.119	0.707
		NDE	0.501	**0.501**	**0.000**	0.174	**0.985**	0.509	-0.008	**0.110**	0.887	0.511	-0.010	0.112	0.898
		NIE	0.500	**0.505**	**-0.005**	0.301	**0.935**	0.472	-0.008	**0.099**	0.565	0.466	0.034	0.100	0.575
	20	ATE	1.001	**1.009**	**-0.008**	0.261	**0.958**	0.982	0.019	**0.082**	0.586	0.977	0.024	0.083	0.584
		NDE	0.501	**0.497**	**0.004**	0.159	**0.986**	0.507	-0.006	**0.076**	0.805	0.509	-0.008	0.077	0.808
		NIE	0.500	**0.512**	**-0.012**	0.301	**0.933**	0.475	-0.006	**0.069**	0.414	0.468	0.032	0.070	0.432
100	10	ATE	1.001	**1.001**	**0.001**	0.185	**0.959**	0.990	0.012	**0.082**	0.694	0.989	0.013	**0.082**	0.696
		NDE	0.502	**0.503**	**-0.001**	0.120	**0.981**	0.507	-0.005	**0.075**	0.874	0.507	-0.005	**0.075**	0.871
		NIE	0.500	**0.498**	**0.002**	0.207	**0.943**	0.483	-0.005	**0.069**	0.511	0.482	0.018	**0.069**	0.510
	20	ATE	1.002	**1.001**	**0.000**	0.180	**0.957**	0.992	0.010	**0.058**	0.521	0.990	0.011	**0.058**	0.521
		NDE	0.502	**0.503**	**-0.001**	0.109	**0.987**	0.507	-0.005	**0.052**	0.794	0.506	-0.005	0.053	0.789
		NIE	0.500	**0.498**	**0.002**	0.206	**0.935**	0.485	-0.005	**0.048**	0.352	0.484	0.016	**0.048**	0.366

Notes: Values in bold indicate a better performance.

Overall, the results, including the ones with a count outcome shown in the Supplementary Material, suggest that, although both approaches toward mediation perform quite well, the traditional one produces slightly better estimates than DML. This may be due to the specific functioning of the `causalweight` package, which relies on Lasso penalized regression. Indeed, it is well known that Lasso induces bias in the estimates, since it shrinks coefficients towards zero, causing underestimation of population coefficients, especially of those relative to interaction terms (Blackwell and Olson, 2022). Furthermore, in our simulation setting, we included a small number of covariates, and this may also have influenced the performance of the `rlasso` function. Indeed, DML is generally used in mediational settings with a large number of covariates: in fact, Farbmacher et al. (2022) conducted their simulations using 200 covariates. Another recent study by Wang and Huang (2024) compared the performance of their proposed deep-learning-based approach with the linear least squares mediation model (Guo et al., 2022) and the partially linear structural equation modeling approach Cai et al. (2022) in a setting with high-dimensional mediators, showing the superiority of machine-learning mediation. Our results suggest the need for more thorough simulations for evaluating the performance of machine learning in standard settings with a limited number of covariates.

5. Data analysis

As already mentioned, we analysed data concerning the health status of a sample of Nigerian children and the possible relationship with the environment in which they live, with a particular focus on the deforestation status of their region. We will first describe how data were collected, then we will move to the exploratory analysis and the mediation analysis based on the approaches discussed so far. Finally, we will present the results.

5.1. Data collection

Data were obtained by Berazneva and Byker (2017b) combining different sources: the core data set is made up of the 2008 and 2013 waves of Demographic and Health

Surveys (DHS) in Nigeria. The DHS primary sampling units in Nigeria are clusters within local government areas (LGAs), from which subjects belonging to different households (HH) were sampled. The study is not longitudinal, since subjects are not re-interviewed. The number of clusters in 2008 is 885, 888 in 2013, with approximately 30 thousand children assessed each year. At each wave, the survey provides information about under-five-years-old children's health status, as reported by their mothers (not assessed by a physician), i.e. whether they had fever (malaria), cough or diarrhea in the two weeks before the survey date, about their household (number of members, gender, education and marital status of the household head) and house characteristics (floor and heating systems of the house, distance from water sources, presence of a flush toilet and bednets).

Data about deforestation were collected from 2000 to 2014 by Earth observation satellites at a spatial resolution of 30 meters, see Hansen et al. (2013) for details. Forest loss is defined as the share of pixels with annual forest loss in a buffer zone around each DHS cluster with a radius of 5km. Information about the soil composition, pH, organic carbon and cation exchange capacity, which can be considered as proxies of soil fertility, refer to the two years of interest, and was obtained from the International Soil Reference and Information Centre.

5.2. Exploratory analysis

The data set contains more than eighty variables which are measured either at the cluster level, like soil characteristics, and at the first (respondent) level, like the outcome and the respondent's demographics. For ease of presentation, we will focus just on the variables of our interest. Forest loss ranges from 0 to 5% every year. Soil characteristics are positive variables: specifically, pH values range from 0, denoting maximum acidity, to 14, indicating basic compounds (in the data set, the scores are multiplied by 10). Organic carbon expresses how many organic components soil contains, and generally ranges from 0 to 14%, while cation exchange capacity indicates the soil's ability to retain and exchange positively charged ions (cations) such as calcium, magnesium, and potassium: the higher its value, the more fertile the soil is. Figure 3 displays the marginal densities of the mediating variables, which, as can be noticed, are positively

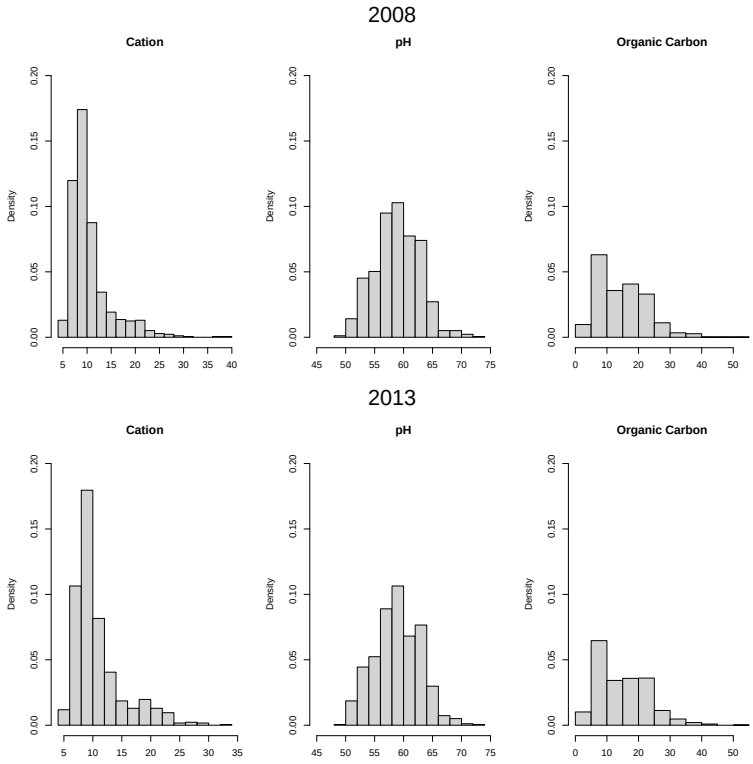


Figure 3. Marginal densities of the mediators, in 2008 and 2013.

skewed, except for pH, which is approximately symmetrical. Table 4 shows a summary of the average levels of exposure, mediators and outcomes computed by cluster.

To assess the relationships between deforestation, soil characteristics and infant diseases, we dichotomized the exposure by setting it to 1 if the share of pixels showing forest loss is positive, to 0 otherwise, since this is the variable that will be used also in the modelling phase. Figure 4 shows the distribution of soil characteristics conditionally on the deforestation status in the years of interest. It can be noticed that deforestation seems to affect organic carbon and pH more than cation exchange, specifically it is associated to increased levels of organic carbon and lower levels of pH, in both the periods considered. Figure 5 shows the relationship between forest loss and the proportion of children affected by the three diseases under investigation. Fever seems to be the least affected by deforestation, while the incidence of cough and diarrhea appears to be higher and smaller, respectively, in clusters that lost tree areas. Finally, Table 5 reports the correlations between soil characteristics levels and the prevalence of infant diseases, suggesting that some of these variables can be associated.

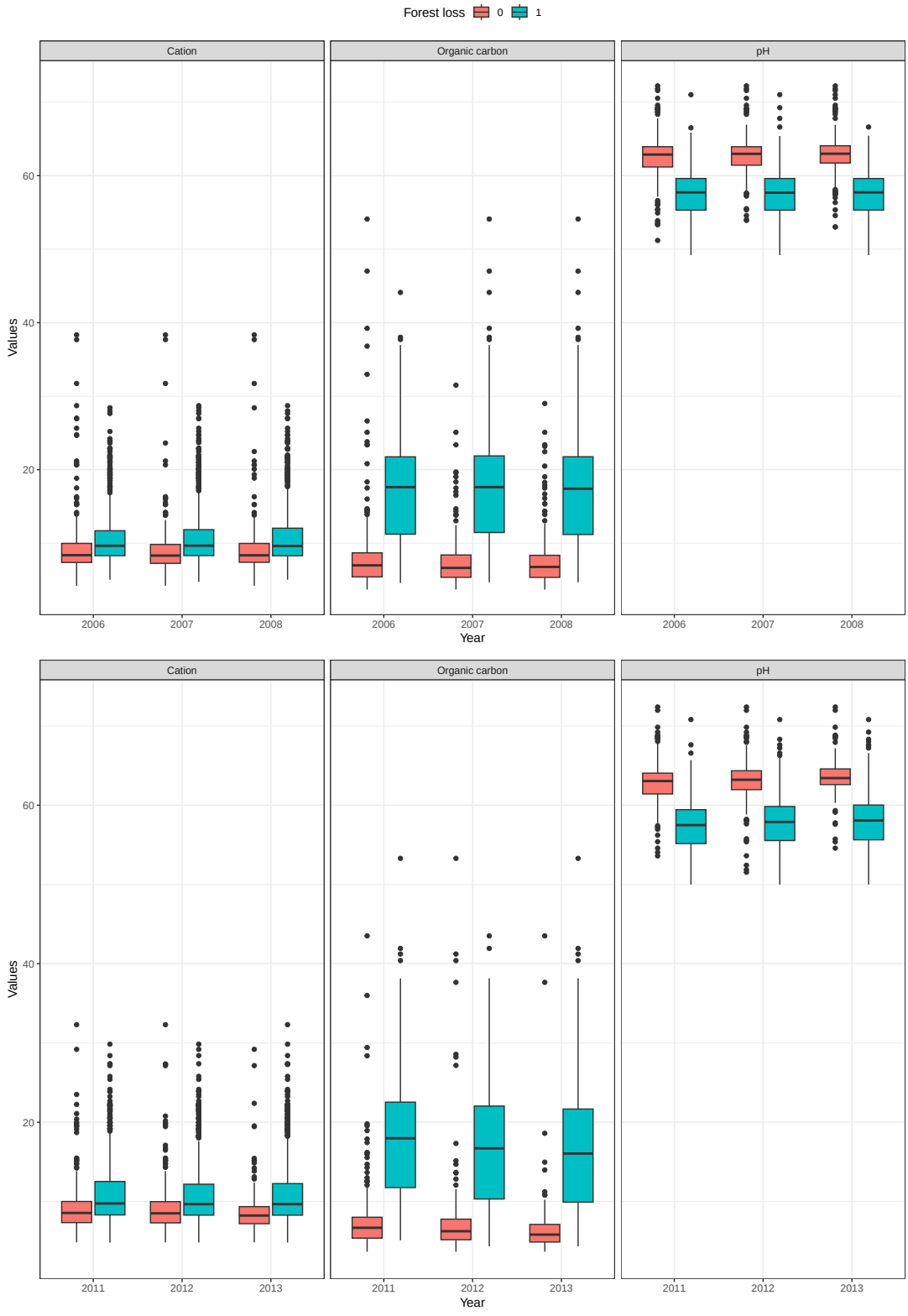


Figure 4. Distribution of soil cation exchange capacity, organic carbon concentration and pH, conditional on forest loss and years, from 2006 to 2008 on the top, from 2011 to 2013 on the bottom.

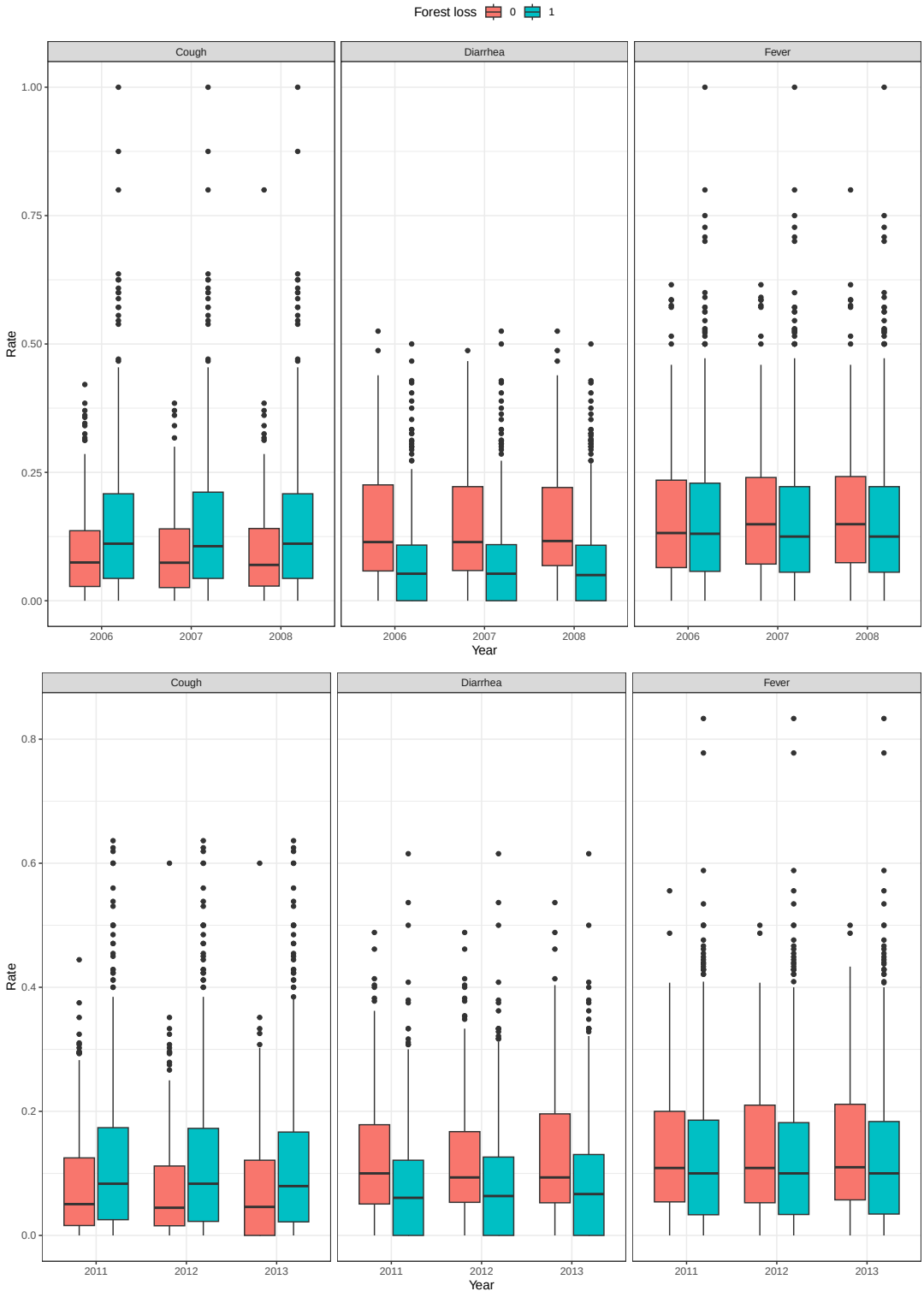


Figure 5. Distribution of infant diseases prevalence conditionally on deforestation and yesr, from 2006 to 2008 on the top, from 2011 to 2013 on the bottom.

Table 4. Summary statistics for the mean values of the key variables, computed at the cluster level, in 2008 and 2013 (first and second panel, respectively).

	FL 2006	FL 2007	FL 2008	Cation	pH	Org. carbon	Fever	Cough	Diarrhea
Min.	0.000	0.000	0.000	4.212	49.19	3.685	0.000	0.000	0.000
1st Qu.	0.000	0.000	0.000	7.936	56.49	7.954	0.059	0.034	0.000
Median	0.000	0.000	0.000	9.277	58.91	13.364	0.130	0.097	0.067
Mean	0.001	0.001	0.001	10.534	58.95	14.674	0.160	0.129	0.091
3rd Qu.	0.001	0.001	0.001	11.254	61.83	20.149	0.231	0.188	0.133
Max.	0.015	0.053	0.025	38.342	72.21	54.096	1.000	1.000	0.525
	FL 2011	**FL 2012**	**FL 2013**	**Cation**	**pH**	**Org. carbon**	**Fever**	**Cough**	**Diarrhea**
Min.	0.000	0.000	0.000	4.821	49.990	3.659	0.000	0.000	0.000
1st Qu.	0.000	0.000	0.000	8.063	56.370	7.526	0.040	0.018	0.011
Median	0.000	0.000	0.000	9.333	58.910	13.475	0.100	0.074	0.071
Mean	0.001	0.001	0.002	10.740	58.920	14.767	0.128	0.108	0.094
3rd Qu.	0.001	0.001	0.002	11.891	61.870	20.632	0.188	0.160	0.136
Max.	0.039	0.037	0.053	32.306	72.380	53.291	0.833	0.636	0.615

Table 5. Correlation between soil characteristics and infant diseases prevalence in 2008 and 2013.

	2008			**2013**		
	Fever	Cough	Diarrhea	Fever	Cough	Diarrhea
Cation	0.032	0.165	-0.064	-0.032	0.096	-0.097
pH	-0.069	-0.229	0.411	0.002	-0.126	0.34
Org. Carbon	0.089	0.267	-0.329	0.031	0.177	-0.271

5.3. Specification of traditional and DML mediation models

The mediational setting of interest involves then deforestation as the exposure, soil characteristics as mediators (one at a time) and the infant diseases as distinct outcomes. Since the `causalweight` package requires a binary exposure, we used a binary version of forest loss, obtained as already introduced in the exploratory analysis. As regards the `mediation` package, we fitted mixed-effect models for the outcomes, which included a random intercept dependent on clusters, and standard generalised linear models for the mediators: leveraging on the results from the descriptive phase, pH was modelled as a Gaussian, while cation and organic carbon were assumed to be from a Gamma distribution, and we used a log link function. The variables included in the mediator models are forest loss measured in the year of sampling (2008 and 2013), one year and two years before, as done by Berazneva and Byker (2017a), altitude and tree coverage in the DHS. They were used in the outcome models as well, in addition to subject-level covariates like the presence of a toilet and of kidnets in the house where the child lives, the distance from water sources and other characteristics of the house. We also adjusted for the month of survey and the region. In the `causalweight` package, the mediator's and outcome's distribution does not need to be specified, thus

we simply specify the role of each variable, including the cluster to which each subject belongs among confounders. To choose which variables to include in the models as confounders, we first excluded the ones which are not related to the variables of interest based on subject-matter knowledge, then we performed an AIC-based model selection, retaining the variables selected in the models with lowest AIC.

5.4. Results

Since the exposure and the mediators are cluster-level variables, while the outcomes are measured at the subject level, the setting is that of a $2 \to 2 \to 1$ mediation, as that addressed in the simulation study. In addition, the response variables are binary, so we will exploit also results from the simulation study in a single-level setting with a logistic outcome model. The results of the data analysis are shown in Tables 6-8.

Although the magnitude of the effects' estimates does not differ much between the two approaches, the associated p-values are quite diverse, with those of regression-based estimates being larger than those obtained with DML. This is not surprising, since the machine learning approach does not take into account the clustered structure, although we included the cluster membership as a covariate. This is in line with the results obtained in the simulation study.

What remains consistent between the two approaches are the total and direct effects of forest loss in 2007 on fever, for all mediators. Both effects are always negative. Also the indirect effect of forest loss in 2013 on the probability of cough and the total and indirect effect of forest loss in 2007 on diarrhea are significant.

In the light of the simulation results, we can reasonably assume that the effects estimated through DML are more biased, either because of the intrinsic bias characterising the estimation algorithm, based on penalisation, and the fact that it cannot account for the data clustered structure. Summarising the results, we can say that forest loss in 2007 directly affects the probability of having fever, independently of the mediators, since the indirect effects are not significant. Forest loss in 2012 and 2013 had a positive indirect effect on the probability of suffering from cough through pH in 2013, while in 2006, 2007 and 2008, it had a negative indirect effect on the probability of having diarrhea via pH. These results are consistent with the exploratory analysis,

Table 6. Results of the mediation analyses considering fever as outcome. FL stands for forest loss.

		DML			Regression		
Exposure		pH	Org. carbon	Cation	pH	Org. carbon	Cation
FL 2006	ATE	-0.009	-0.002	-0.002	-0.010	-0.013	-0.008
	NDE	-0.006	-0.002	0.000	-0.008	-0.007	-0.008
	NIE	-0.003	0.000	-0.002 ***	-0.002	-0.006	<0.001
FL 2007	ATE	-0.034 ***	-0.022 **	-0.027 **	-0.024 *	-0.029**	-0.021 *
	NDE	-0.034 ***	-0.027 ***	-0.027 **	-0.022 *	-0.022 *	-0.021 *
	NIE	-0.000	0.005 ***	0.001 *	-0.002	-0.006	< -0.001
FL 2008	ATE	-0.030 ***	-0.025 **	-0.027 **	-0.006	-0.010	-0.003
	NDE	-0.030 **	-0.026 **	-0.027 **	-0.003	-0.004	-0.003
	NIE	-0.000	0.001	0.000	-0.002	-0.006	<-0.001
FL 2011	ATE	0.017 **	0.024 **	0.025 **	-0.010	-0.016	-0.011
	NDE	0.023 **	0.024 **	0.024 **	-0.012	-0.013	-0.012
	NIE	-0.006	0.000	0.002	0.003	-0.003	0.001
FL 2012	ATE	0.009	0.017 *	0.019 *	-0.013	-0.019	-0.016
	NDE	0.013	0.017 *	0.019 *	-0.016	-0.017	-0.016
	NIE	-0.004	0.000	0.000	0.003	-0.002	0.000
FL 2013	ATE	-0.020 *	-0.007	-0.010	-0.012	-0.020	-0.017
	NDE	-0.013	-0.007	-0.006	-0.016	-0.017	-0.016
	NIE	-0.007 *	-0.001	-0.004 **	0.004	-0.003	-0.001

Notes: The number of * denotes the estimate significance, i.e. the corresponding p-value p, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7. Results of the mediation analyses considering cough as outcome. FL stands for forest loss.

		DML			Regression		
Exposure		pH	Org. carbon	Cation	pH	Org. carbon	Cation
FL 2006	ATE	0.007	0.011	0.009	-0.004	-0.012	-0.009
	NDE	0.009	0.011	0.011	-0.008	-0.008	- 0.008
	NIE	-0.002	0.000	-0.002 ***	0.004 *	-0.004	-0.001
FL 2007	ATE	0.014 *	0.016*	0.015 *	<0.001	-0.010	-0.006
	NDE	0.010	0.013 *	0.014 *	-0.005	-0.005	-0.005
	NIE	0.003	0.002 *	0.001 *	0.005	-0.005	< -0.001
FL 2008	ATE	0.009	0.010	0.010	0.009	0.000	0.003
	NDE	0.005	0.010	0.010	0.003	0.004	0.003
	NIE	0.004	0.000	0.001	0.006	-0.004	0.000
FL 2011	ATE	0.017 *	0.022 **	0.020 **	0.000	-0.004	-0.005
	NDE	0.031 **	0.020 **	0.024 **	-0.005	-0.006	-0.005
	NIE	-0.013 ***	0.001	-0.003 *	0.005	0.002	0.000
FL 2012	ATE	0.017 **	0.024 ***	0.026 ***	0.011	0.006	-0.006
	NDE	0.022 **	0.023 ***	0.023 ***	0.005	0.004	-0.005
	NIE	-0.004	0.001 *	0.002	0.006 *	0.002	0.000
FL 2013	ATE	-0.020 **	-0.017 *	-0.015 *	0.008	0.002	0.000
	NDE	-0.012	-0.016 *	-0.013	0.001	0.000	0.000
	NIE	-0.009 ***	-0.001 ***	-0.003 *	0.007 *	0.002	0.000

Notes: The number of * denotes the estimate significance, i.e. the corresponding p-value p, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8. Results of the mediation analyses considering diarrhea as outcome. FL stands for forest loss.

		DML			Regression		
Exposure		pH	Org. carbon	Cation	pH	Org. carbon	Cation
FL 2006	ATE	-0.031 ***	-0.026 ***	-0.026 ***	0.010	0.013	0.017
	NDE	-0.026 ***	-0.028 ***	-0.027 ***	0.017	0.016	0.016
	NIE	-0.004	0.002 ***	0.001 *	-0.006 **	-0.004	0.000
FL 2007	ATE	-0.021 **	-0.010	-0.012	-0.025 *	-0.022 *	- 0.019
	NDE	-0.012	-0.013	-0.013	-0.020 *	-0.019	-0.019
	NIE	-0.009 **	0.003 *	0.001 *	-0.005 *	-0.003	0.000
FL 2008	ATE	-0.022 **	-0.018 *	-0.018 *	-0.008	-0.004	0.000
	NDE	-0.019	-0.018 *	-0.018 *	-0.001	-0.001	0.000
	NIE	-0.003	0.000	0.000	-0.007 **	-0.003	0.000
FL 2011	ATE	0.013	0.021 **	0.024 **	-0.006	-0.007	-0.005
	NDE	0.023 **	0.019 *	0.023 **	-0.005	-0.006	-0.005
	NIE	-0.010 **	0.002 *	0.001	-0.001	-0.001	0.000
FL 2012	ATE	0.026 ***	0.038 ***	0.038 ***	0.003	0.004	0.005
	NDE	0.033 ***	0.037 ***	0.038 ***	0.004	0.005	0.005
	NIE	-0.007	0.001	0.000	-0.001	-0.001	0.000
FL 2013	ATE	-0.017 *	-0.014	-0.011	0.002	0.002	0.003
	NDE	-0.021 **	-0.014	-0.009	0.004	0.004	0.004
	NIE	0.004	0.000	-0.001	-0.002	-0.001	-0.001

Notes: The number of * denotes the estimate significance, i.e. the corresponding p-value p, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

which showed that deforestation was associated with lower rates of children suffering from diarrhea and higher rates of children affected by cough in both the years of interest. Although this may sound counterintuitive, it is in line with the results of previous studies (Guerra et al., 2006; Matsumoto-Takahashi et al., 2023; Ranjha and Sharma, 2021) that highlighted the protective role of deforestation against some human diseases, especially malaria. In our case, the negative indirect effect seems to be due to the fact that deforestation makes the soil more acidic, and diarrhea is associated to higher values of soil pH (more basic soil). The link is probably given by vegetables eaten by children, although we do not have such a piece of information. The negative effect of deforestation on the probability of developing diarrhea seems to be confirmed by the additional analysis that we performed by aggregating deforestation status over the five years before those of reference, and whore results are reported in the Supplementary Material. Indeed, the new exposure that we obtained expresses if an area was affected by deforestation in the last five years: we found that the effects on diarrhea are significant not only in 2008, but only in 2013, and this reminds of the frontier malaria theory mentioned in the Introduction, which may likely be applicable to other diseases: in the long run, deforestation may have a protective effect against

other diseases, not only malaria.

These results we obtained are influenced by many factors, first of all, the transformation of forest loss into a binary variable. Indeed, we are comparing the scenario with no forest loss with the scenario where forest loss occurred, although we did not quantify it. In the sample, the share of pixels "affected" by forest loss is small, so this could certainly affect the results. Indeed, if X is not binary, comparing a scenario with half of the pixels showing forest loss versus another one with total forest loss, the results (not shown here) are more neat and seem to indicate a larger impact of deforestation on infant diseases. In addition, the data we analysed are not longitudinal, preventing us from exploiting the temporal dependence among the variables of interest. Also, the outcome variables may be affected by measurement errors, since they are based on children's mothers' reports, not on clinical tests. Moreover, as in every causal study, we cannot know if identifiability assumptions are satisfied. We believe that it is plausible to assume that sequential ignorability holds in our analysis, since we included as many confounders as possible, both cluster- and subject-level. Clearly, we may have ignored possibly relevant variables like children's feeding habits or the abundance of mosquitoes in the areas where children live. However, these variables cannot play the role of confounders in the causal graph representing the relationships among the variables of our interest: at most, they can be in other mediating paths, being other mediators to investigate. Ignoring them does not mine the causal conclusions we drew.

6. Conclusions

In this paper, we investigated how deforestation causally affects the probability of infant diseases through soil characteristics. To do so, we used two approaches: regression-based and DML mediation. Since the latter is quite recent (Farbmacher et al., 2022), we compared its performance to one of the more established and widely spread regression-based approaches. When data are not clustered the two approaches behave quite similarly and both perform well in terms of bias and coverage rates, although the regression-based mediation shows in most cases better values of the selected metrics, probably

because of the intrinsic bias of the Lasso machine learner used in the `causalweight` package. In scenarios with clustered data, the regression-based approach outperforms DML since it can account for the clustered structure, while the `causalweight` package, used to implement DML, cannot accommodate clusters and even including subjects' cluster membership as a covariate did not produce significant improvements.

The data analysis led to different results in terms of the significance of the mediational effects, with those estimated through the regression-based approach being almost all insignificant compared to those obtained with DML. Only some indirect effects were significant for both approaches, like the one of forest loss on diarrhea mediated by soil pH. The effects are negative, indicating that deforestation reduces the probability of having diarrhea for children under five years through soil characteristics.

Our study is affected by many limitations and this does not allow us to draw final conclusions about the effect of deforestation on infant health. First of all, the study is not longitudinal, and this prevents us from analysing the long-term effects of deforestation. Second, children's diseases are reported by their mothers, not ascertained by a physician, and this can clearly influence the analysis. Moreover, we do not have important information about children's eating and clinical history. Despite these shortcomings, this study is, to the best of our knowledge, the first to explicitly investigate the role of soil characteristics in mediating the relationship between deforestation and human health. Further studies are needed to provide new evidence about this issue, extending the sample also to other countries.

Acknowledgements Not applicable.

Declarations

Disclosure statement The authors declare to have no conflict of interests.

Funding: No funding was received for conducting this study.

Data availability statement: Data used in this article are freely available in the ICPSR Repository, at the following link: https://www.openicpsr.org/openicpsr/project/113539/version/V1/view

Berazneva, Julia, and Byker, Tanya S. Replication data for: Does Forest Loss Increase Human Disease? Evidence from Nigeria. Nashville, TN: American Economic Association [publisher], 2017. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-10-12. https://doi.org/10.3886/E113539V1

References

Aguirre, J., Rojas, P., Yu, T. E., and Yenerall, J. (2024). Spatial Spillovers and the Relationship Between Deforestation and Malaria: Evidence from the Peruvian Amazon. *J Environ Plan Manag*, pages 1–17.

Alimi, Y., Bernstein, A., Epstein, J., Espinal, M., Kakkar, M., Kochevar, D., and Werneck, G. (2021). Report of the scientific task force on preventing pandemics. *Cambridge, MA: Harvard Chan C-CHANGE and Harvard Global Health Institute*.

Andrews, R. M. and Didelez, V. (2020). Insights into the Cross-world Independence Assumption of Causal Mediation Analysis. *Epidemiol*, 32(2):209–219.

Bae, W., Daniels, M. J., and Perri, M. G. (2024). A Bayesian Nonparametric Approach for Causal Mediation with a Post-Treatment Confounder. *Biom*, 80(3).

Bauhoff, S. and Busch, J. (2020). Does deforestation increase malaria prevalence? Evidence from satellite data and health surveys. *World Dev*, 127:104734.

Berazneva, J. and Byker, T. S. (2017a). Does Forest Loss Increase Human Disease? Evidence from Nigeria. *Am Econ Rev*, 107(5):516–521.

Berazneva, J. and Byker, T. S. (2017b). Replication data for: Does Forest Loss Increase Human Disease? Evidence from Nigeria. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-10-12.

Blackwell, M. and Olson, M. P. (2022). Reducing model misspecification and bias in the estimation of interactions. *Polit Anal*, 30(4):495–514.

Bodory, H. and Huber, M. (2018). The causalweight package for causal inference in R. Working Papers SES 493, Faculty of Economics and Social Sciences, University of Fribourg (Switzerland).

Bolton, A. V., Montag, D., and Gallo, V. (2022). Global forestry areas, deforestation and mental health: A worldwide ecological study. *J Clim Change Health*, 6:100109.

Burgos Ochoa, L., Rijnhart, J. J. M., Penninx, B. W., Wardenaar, K. J., Twisk, J. W. R., and Heymans, M. W. (2020). Performance of methods to conduct mediation analysis with time-to-event outcomes. *Stat Neerlandica*, 74:72–91.

Butt, E. W., Conibear, L., Knote, C., and Spracklen, D. V. (2021). Large air quality and public health impacts due to amazonian deforestation fires in 2019. *GeoHealth*, 5(7):e2021GH000429.

Cai, X., Zhu, Y., Huang, Y., and Ghosh, D. (2022). High-dimensional causal mediation analysis based on partial linear structural equation models. *Comput Stat Data Anal*, 174:107501.

Chakrabarti, A. (2021). Deforestation and infant mortality: Evidence from Indonesia. *Econ Hum Biol*, 40:100943.

Chakravarty, S., K., S., C., P., et al. (2012). Deforestation: Causes, Effects and Control Strategies. In Akais, O. C., editor, *Global Perspectives on Sustainable Forest Management*. InTech.

Cook, A. G., Janetos, A. C., and Hinds, W. T. (1990). Global effects of tropical deforestation: Towards an integrated perspective. *Env Conserv*, 17(3):201–212.

Davari, M., Gholami, L., Nabiollahi, K., Homaee, M., and Jafari, H. J. (2020). Deforestation and cultivation of sparse forest impacts on soil quality (case study: West Iran, Baneh). *Soil Tillage Res*, 198:104504.

Doretti, M., Raggi, M., and Stanghellini, E. (2022). Exact parametric causal mediation analysis for a binary outcome with a binary mediator. *Stat Methods Appl*, 31:87–108.

Estifanos, T. K., Fisher, B., Galford, G. L., and Ricketts, T. H. (2024). Impacts of deforestation on childhood malaria depend on wealth and vector biology. *GeoHealth*, 8:e2022GH000764.

Falk, C. F., Vogel, T. A., Hammami, S., et al. (2024). Multilevel mediation analysis in R: A comparison of bootstrap and Bayesian approaches. *Behav Res Methods*, 56:750–764.

Farbmacher, H., Huber, M., Lafférs, L., Langen, H., and Spindler, M. (2022). Causal mediation analysis with double machine learning. *Econom J*, 25(2):277–300.

Fuentes Cordoba, G. (2024). Deforestation and Child Health in Cambodia. *Econ Human Biol*, 52:101343.

Galway, L. P., Acharya, Y., and Jones, A. D. (2018). Deforestation and child diet diversity: A geospatial analysis of 15 Sub-Saharan African countries. *Health Pl*, 51:78–88.

Guerra, C., Snow, R., and Hay, S. (2006). A global assessment of closed forests, deforestation and malaria risk. *Ann Trop Med Parasitol*, 100(3):189.

Guo, X., Li, R., Liu, J., and Zeng, M. (2022). High-dimensional mediation analysis for selecting DNA methylation loci mediating childhood trauma and cortisol stress reactivity. *J Am Stat Assoc*, 117(539):1110–1121.

Hajabbasi, M., Jalalian, A., and Karimzadeh, H. (2004). Deforestation effects on soil physical and chemical properties, Lordegan, Iran. *Plant and Soil*, 190:301–308.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., and et al. (2013). High-resolution global maps of 21st-century forest cover change. *Sci*, 342:850–853.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton.

Imai, K., Keele, L., and Yamamoto, T. (2010a). Identification, inference and sensitivity analysis for causal mediation effects. *Stat Sci*, 25(1):51–71.

Imai, K., Tingley, D., and Keele, L. (2010b). A general approach to causal mediation analysis. *Psychol Methods*, 15(4):309–334.

Johnson, K. B., Jacob, A., and Brown, M. E. (2013). Forest cover associated with improved child health and nutrition: evidence from the Malawi Demographic and Health Survey and satellite data. *Glob Health Sci Pract*, 1(2):237–248.

Kim, C., Daniels, M., Li, Y., Milbury, K., and Cohen, L. (2018). A Bayesian Semiparametric Latent Variable Approach to Causal Mediation. *Stat Med*, 37:1149–1161.

Krull, J. L. and MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Eval Rev*, 23(4):418–444.

Krull, J. L. and MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivar Behav Res*, 36(2):249–277.

Laporta, G. Z., Ilacqua, R. C., Bergo, E. S., et al. (2021). Malaria Transmission in Landscapes with Varying Deforestation Levels and Timelines in the Amazon: A Longitudinal Spatiotemporal Study. *Sci Reports*, 11:6477.

Ma, L., Qin, X., Sun, N., and Yang, G. (2014). Human health risk of metals in drinking-water source areas from a forest zone after long-term excessive deforestation. *Human Ecol Risk Assess: Int J*, 20(5):1200–1212.

MacDonald, A. J. and Mordecai, E. A. (2019). Amazon deforestation drives malaria

transmission, and malaria burden reduces forest clearing: a retrospective study. *Lancet Planet Health*, 3(1):S13.

Mainville, N., Webb, J., Lucotte, M., Davidson, R., Betancourt, O., Cueva, E., and Mergler, D. (2006). Decrease of soil fertility and release of mercury following deforestation in the Andean Amazon, Napo River Valley, Ecuador. *Sci Tot Environ*, 368 1:88–98.

Masuda, Y. J., Garg, T., Anggraeni, I., Wolff, N. H., Ebi, K., Game, E. T., Krenz, J., and Spector, J. T. (2020). Heat exposure from tropical deforestation decreases cognitive performance of rural workers: An experimental study. *Environ Res Lett*, 15(12).

Matsumoto-Takahashi, E. L. A., Iwagami, M., Oyoshi, K., Sasaki, Y., Hongvanthong, B., and Kano, S. (2023). Deforestation inhibits malaria transmission in Lao PDR: a spatial epidemiology using Earth observation satellites. *Trop Med Health*, 51(1):60.

Morand, S. and Lajaunie, C. (2020). Outbreaks of Vector-Borne and Zoonotic Diseases Are Associated With Changes in Forest Cover and Oil Palm Expansion at Global Scale. *Front Vet Sci*, 8.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference*. Cambridge Univerity Press, New York, NY, 2 edition.

Nath, T., Caffo, B., Wager, T., and Lindquist, M. A. (2023). A machine learning based approach towards high-dimensional mediation analysis. *NeuroImage*, 268:119843.

Palmer, J. (2022). No escape from the heat: The cost of deforestation on human and animal health. *BioSci*, 72(11):1037–1043.

Pattanayak, S. K. and Yasuoka, J. (2012). Deforestation and malaria: Revisiting the human ecology perspective. In *Human Health and Forests*, pages 219–240. Routledge, London.

Pearl, J. (2001). Direct and indirect effects. In Kaufmann, M., editor, *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420.

Pearl, J. (2009a). Causal inference in statistics: An overview. *Stat Surv*, 3:96–146.

Pearl, J. (2009b). *Causality. Models, Reasoning, and Inference*. Cambridge University Press, New York, 2 edition.

Pearl, J. (2012). The Mediation Formula: A guide to the assessment of causal pathways in nonlinear models. In Berzuini, C., Dawid, P., and Bernardinelli, L., editors, *Causality: Statistical perspectives and applications*, pages 151–179. John Wiley and Sons, Chichester, England.

Ranjha, R. and Sharma, A. (2021). Forest malaria: the prevailing obstacle for malaria control and elimination in India. *BMJ Glob Health*, 6(5):e005391.

Reddington, C., Butt, E., Ridley, D., et al. (2015). Air quality and human health improvements from reductions in deforestation-related fire in Brazil. *Nat Geosci*, 8:768–771.

Rerolle, F., Dantzer, E., Lover, A. A., Marshall, J. M., Hongvanthong, B., Sturrock, H. J. W., and Bennett, A. (2021). Spatio-Temporal Associations Between Deforestation and Malaria Incidence in Lao PDR. *eLife*, 10:e56974.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiol*, 3(2):143–155.

Roy, S., Daniels, M. J., and Roy, J. (2024). A Bayesian Nonparametric Approach for Multiple Mediators with Applications in Mental Health Studies. *Biostat*, 25(3):919–932.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*, 66(5):688–701.

Samoilenko, M. and Lefebvre, G. (2021). Parametric-Regression–Based Causal Me-

diation Analysis of Binary Outcomes and Binary Mediators: Moving Beyond the Rareness or Commonness of the Outcome. *Am J Epidemiol*, 190(9):1846–1858.

Santika, T. et al. (2023). Deterioration of Respiratory Health Following Changes to Land Cover and Climate in Indonesia. *One Earth*, 6:290–302.

Smith, E. K., Lacy, M. G., and Mayer, A. (2019). Performance simulations for categorical mediation: Analyzing khb estimates of mediation in ordinal regression models. *Stata J*, 19(4):913–930.

Sun, B. and Ye, T. (2023). Semiparametric Causal Mediation Analysis with Unmeasured Mediator-Outcome Confounding. *Stat Sin*, 33:1–20.

Tchetgen, E. J. T. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Ann Stat*, 40(3):1816.

Tchetgen Tchetgen, E. J. (2013). Inverse odds ratio-weighted estimation for causal mediation analysis. *Stat Med*, 32:4567–4580.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2013). *mediation: R Package for Causal Mediation Analysis*. R package version 4.4.2.

Tucker Lima, J. M., Vittor, A., Rifai, S., and Valle, D. (2017). Does deforestation promote or inhibit malaria transmission in the Amazon? A systematic literature review and critical appraisal of current evidence. *Philos Trans R Soc London. Series B, Biol Sci*, 372(1722):20160125.

Valeri, L. and VanderWeele, T. J. (2013). Mediation Analysis Allowing for Exposure–Mediator Interactions and Causal Interpretation: Theoretical Assumptions and Implementation With SAS and SPSS Macros. *Psychological Methods*, 18(2):137–150.

Vanderweele, T. J. (2015). *Explanation in Causal Inference*. Oxford University Press, New York.

VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2014). Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder. *Epidemiol*, 25(2):300–306.

Wang, S. and Huang, Y. (2024). DP2LM: Leveraging deep learning approach for estimation and hypothesis testing on mediation effects with high-dimensional mediators and complex confounders. *Biostat*, 25(3):818–832.

Wolff, N., Masuda, Y., Meijaard, E., Wells, J. A., and Game, E. (2018). Impacts of tropical deforestation on local temperature and human well-being perceptions. *Glob Environ Change*.

World Economic Forum (2020). Outbreak Readiness and Business Impact: Protecting Lives and Livelihoods across the Global Economy. Technical report.

World Resources Institute (2024). Tropical Forest Loss Drops Steeply in Brazil and Colombia, but High Rates Persist Overall. `https://research.wri.org/gfr/latest-analysis-deforestation-trends`. Updated April 4, 2024.

Xia, F. and Chan, K. C. G. (2021). Identification, Semiparametric Efficiency, and Quadruply Robust Estimation in Mediation Analysis with Treatment-Induced Confounding. *J Am Stat Assoc*, 118(542):1272–1281.