

Uncovering Sources of Heterogeneity in the Effects of Maternal Smoking on Infants' Health at Birth*

Johanna Kutz[†]

Abstract

Smoking during pregnancy poses significant risks to infants' health at birth, but individual drivers of effect sizes are largely unknown. This study is the first to uncover the heterogeneity in its impact on birth weight and other health measures, using a novel machine learning based effect decomposition. In a large US sample, older maternal age increases smoking's negative effects, while high pre-pregnancy BMI and excessive weight gain can reduce its impact on birth weight but have mixed effects on other health outcomes. These insights can improve targeted smoking cessation for pregnant women, potentially reducing healthcare costs and enhancing newborn health.

JEL classification: C21, I12, J13

Keywords: Smoking, infant health, machine learning, heterogeneous treatment effects, distributional treatment effects, causal forest, effect decomposition

*I am very grateful to Tom Zimmermann and Daniel Wiesen for their invaluable mentorship and advice. I want to expand special thanks to Michael Darden for helpful comments and support. I thank conference and seminar participants at University of Cologne, AIES XXVI National Conference, 15th RGS Doctoral Conference, Tilburg University, Royal Holloway University, EUHEA Conference Oslo, EEA-ESEM Milan, 23rd European Workshop on Econometrics and Health Economics, 2022 Annual Health Econometrics Workshop, CINCH Monday Seminar, University of St. Gallen, and Nuremberg Research Seminar in Economics.

[†]University of St. Gallen, johanna.kutz@unisg.ch

1 Introduction

Understanding the drivers of heterogeneity in the effects of maternal smoking during pregnancy is crucial for developing effective policies aimed at improving neonatal health outcomes. Mothers who smoke are at high risk of preterm delivery, and stillbirth, and have infants with low birth weight, which are primary causes of death, disability, and disease among newborns (e.g. Almond et al., 2005; Baba et al., 2013). In the US, neonatal costs attributable to smoking were estimated to be almost \$750,000,000 (in year 2025 dollars) (Adams et al., 2002). While there is a well-established link between maternal smoking during pregnancy and adverse birth outcomes, the heterogeneity in these effects is not well understood. However, identifying and understanding sources of heterogeneity is crucial for personalized care and targeted support in smoking cessation for those mothers most at risk. This would not only contribute to improving infants' health but also lead to savings in healthcare costs.

In this paper, I propose a novel approach to identify key factors of heterogeneity in the effect of smoking on infants' health at birth, by combining machine learning and econometrics. The key idea is to decompose the distribution of conditional average treatment effects (CATE), to isolate driving factors of heterogeneity. This decomposition uses counterfactual distributions (Chernozhukov et al., 2013). For estimation of the effect of smoking on birth weight conditional on the mother's characteristics, I use the causal forest (Athey et al., 2019). I decompose differences in the distribution of CATE for groups into structural and compositional effects. The structural effect reveals the effect differences solely associated with the variable of interest while keeping group characteristics comparable. The compositional effect captures differences due to different characteristics between groups. This way I can learn about the observable factors strongly associated with the heterogeneity and offer a structured identification of the drivers of heterogeneity in treatment effects.

In a comprehensive data set of infants born in the US, I find strong modifying effects

for mother’s age on the effect of smoking on birth weight, low birth weight¹, Apgar score² and gestation length. Increased mother’s age amplifies the effect of smoking on health at birth. It can explain up to 75 grams³ of difference in treatment effects. For increased prepregnancy BMI and weight gain, I find no clear effect on health at birth. Increased BMI and weight gain mitigate the effect of smoking on birth weight, however, no such effects can be found regarding Apgar Score or gestation length. This suggests that birth weight increased through the mother’s excessive calorie intake or unhealthy BMI does not result in better health outcomes for the infant.

To be able to learn about drivers of heterogeneity, I propose a novel way to decompose estimated treatment effects. While machine learning techniques have proven very useful in predicting heterogeneous treatment effects with respect to observables (e.g., Athey and Imbens, 2019, 2016; Belloni et al., 2013; Chernozhukov et al., 2018c; Wager and Athey, 2018), the estimated treatment effect function depending on individual characteristics may be very complex and hard to describe. The proposed decomposition allows me to isolate the effect of a change in a single variable while keeping other characteristics comparable. To do so, I make use of counterfactual distributions. The counterfactual distribution does not arise from any observable population, it is rather constructed by integrating the conditional distribution of CATE for one group with respect to the distribution of characteristics of another population Chernozhukov et al. (2013). The distribution thus shows the treatment effect for one group, that would have prevailed in case they faced the other group’s characteristics. This allows me to keep characteristics comparable in the decomposition. I decompose the distribution of CATEs into structural and compositional effects to find observables, for which

¹A birth weight below 2500 grams is considered low birth weight.

²The Apgar score is a measure to quickly judge the health condition of the newborn right after birth.

³When comparing mothers aged 23 or younger and mothers older than 34

large structural differences in treatment effects exist⁴. The structural effect is the main effect of interest, as it captures the effect solely attributable to a single variable while keeping all characteristics comparable.

The decomposition allows identification of sources of differences in treatment effects across groups, such as age groups. The structural effect captures how much of the treatment effect difference can be attributed to differences in responsiveness, showing what would happen if all age groups had the same characteristics but differed only in age. The composition effect, on the other hand, reflects differences in covariates across age groups, showing how much of the observed treatment effect is because, for instance, younger and older individuals differ in education or other baseline characteristics. Overall, the decomposition helps to describe whether a single characteristic matters because it changes how people respond, or because age groups differ in who they are.

My analysis demonstrates that optimizing the allocation of smoking cessation programs by targeting key factors of heterogeneity can significantly reduce healthcare expenditures. To highlight how identified drivers of heterogeneity can be leveraged for optimized policies, I utilize them in an example of targeting smoking cessation programs for mothers most in need. In a scenario where 10,000 smoking pregnant women are offered intensified assistance, prioritizing those with higher risk factors (such as age and BMI) over random assignment can increase savings by nearly 80%. This approach effectively identifies mothers whose babies are most at risk of low birth weights, leading to substantial cost savings and improved infant health outcomes if these mothers quit smoking. Knowledge of these drivers of heterogeneity appears especially useful in settings, where no large data sources on health outcomes and smoking are available, and easy targeting rules are needed.

⁴Ideas for this type of decomposition go back to Oaxaca (1973) and Blinder (1973) who decomposed differences in wage distribution into a discrimination effect, which arises when comparing men and women with the same characteristics and a compositional effect which arises due to differences in the characteristics of men and women.

The identified drivers of heterogeneity are robust to several alternative specifications, ruling out issues such as dose dependency, underreporting of smoking, limited overlap in propensity scores, and alternative health outcomes. To capture additional aspects of health at birth, I analyze heterogeneity in the effects of smoking on additional outcomes, such as gestation length, Apgar score, non-standardized birth weight, and low birth weight birth (LBW). Rerunning the analysis for heavy smokers shows that results on heterogeneity are not driven by smoking intensity. Additionally, using prepregnancy smoking as a non-stigmatized treatment, I find that concerns regarding underreporting are negligible since resulting treatment effect estimates do not differ a lot from the main analysis, and patterns of heterogeneity remain the same.

This paper contributes to the fast-emerging literature of causal machine learning in economics (Athey et al., 2019; Belloni et al., 2013; Chernozhukov et al., 2018a; Nie and Wager, 2020; Wager and Athey, 2018), with a special focus on effect heterogeneity. Other than traditional methods, the causal forest and other machine learning based approaches provide more flexibility and a structured way of identifying heterogeneity. Prediction of heterogeneous effects are more flexible using those advanced procedures, as they can search over high-dimensional functions of covariates and do not necessarily impose parametric assumptions about covariate interactions or functional form. Given these advantages over classical methods, there is also a growing number of applications to health-related questions (Deryugina et al., 2019; Elek and Bíró, 2021; Kreif et al., 2022). This paper also adds to the strand of literature focusing on methodological advances to highlight heterogeneity in treatment effects (Bearth and Lechner, 2024; Chernozhukov et al., 2018b; Farbmacher et al., 2022). I add to this literature by describing the CATE function in a new way using a decomposition based on the counterfactual distribution and therefore propose a new approach to overcome the interpretability problem of machine learning methods for CATE estimation. Machine learning methods prove very useful in the estimation of CATE, albeit the interpretation of driving factors of heterogeneity is not always clear. Therefore, I propose using a decompo-

sition technique to uncover the driving factors of heterogeneity.

This paper also contributes to the large economic and medical literature on smoking and health outcomes. Here, the effect of maternal smoking on health outcomes is studied mostly focusing on average effects and birth weight as outcomes of interest (e.g., England et al., 2001; Almond et al., 2005; Lien and Evans, 2005; Cattaneo, 2010; Bharadwaj et al., 2014)⁵. I contribute to this literature by analyzing heterogeneity in the effect of smoking during pregnancy and providing new evidence of possible drivers of heterogeneity in this effect. I can provide causal interpretation based on a much more comprehensive data set than previous studies. Additionally, I go beyond the effect of smoking on birth weight and widen the range of possible proxies for health at birth to capture additional details of the effect of smoking on newborn health. Additionally, I introduce new, cutting-edge machine learning techniques to the economics of smoking and the economics of early human capital formation.

2 Background

Smoking during pregnancy is associated with adverse birth outcomes, such as decreased birth weight, LBW, premature birth, neonatal mortality, and sudden infant death, leading to higher medical care costs (Cnattingius, 2004; England et al., 2001; Almond et al., 2005). The Surgeon General’s 1990 (US Surgeon General, 1990) report states that maternal smoking reduces birth weight by an average of 200 grams and doubles the likelihood of LBW, contributing to about 20% of LBW cases in the US. A large body of economic literature confirms these findings, with Almond et al. (2005) and Bharadwaj et al. (2014) reporting significant reductions in birth weight due to smoking, although no consistent effects on Apgar scores or infant mortality were found. Currie et al. (2009) showed increased LBW risk only in preg-

⁵Figure B.1 in the supplemental material provides an overview of treatment effect estimates for the effect of maternal smoking during pregnancy on birth weight in the literature.

nancies where the mother smoked. Studies using cigarette tax hikes as an instrument (e.g., Lien and Evans (2005); Simon (2016)) corroborate these findings. When treating smoking as exogenous, estimates in Lien and Evans (2005) do not change much compared to the IV approach, suggesting that the omitted variable bias is not severe in models where smoking is treated as exogenous. While the link between smoking, birth weight, and prematurity is clear, the impact on other outcomes like Apgar scores is inconsistent (Almond et al., 2005; Gilman et al., 2008).

Only few studies in the medical literature have examined the possibility that mothers' characteristics influence the effects of maternal smoking, with mixed and often contradictory results due to small or imbalanced samples and lack of causal identification. Maternal age significantly impacts the effect of smoking on birth weight, with older smokers experiencing greater fetal growth reduction and higher SGA risk (Cnattingius, 1989; Wu Wen et al., 1990). Parity also matters; multiparas are at higher risk for LBW and preterm delivery, while age effects are more pronounced in nulliparas (Cnattingius et al., 1993; Misra et al., 2005). While increased weight gain during pregnancy raises birth weight in non-smokers, smoking reduces appetite and maternal weight gain but does not alter the smoking-fetal growth relationship (D'Souza et al., 1981; Haworth et al., 1980). The role of pre-pregnancy BMI is debated, with mixed findings on whether maternal overweight mitigates smoking's adverse effects (Haworth et al., 1980; La Merrill et al., 2011; Zhang and Yang, 2019). The "fragile male" hypothesis suggests male fetuses are more affected by smoking-induced growth retardation, though evidence is mixed (Spinillo et al., 1994; Zarén et al., 2000).

3 Data

For this study I use the annual natality micro data by the National Center for Health Statistics (2011-2018). Natality data from the National Vital Statistics System of the NCHS provides data on births occurring each year in the United States, based on information

abstracted from birth certificates. The data contains detailed information on socioeconomic and demographic background of the mother and detailed health information on the infant. Importantly, the data contains information on the mother’s smoking behavior. Every woman who reported to have smoked at least one cigarette per week in any of the three trimesters of pregnancy is classified as a smoker.

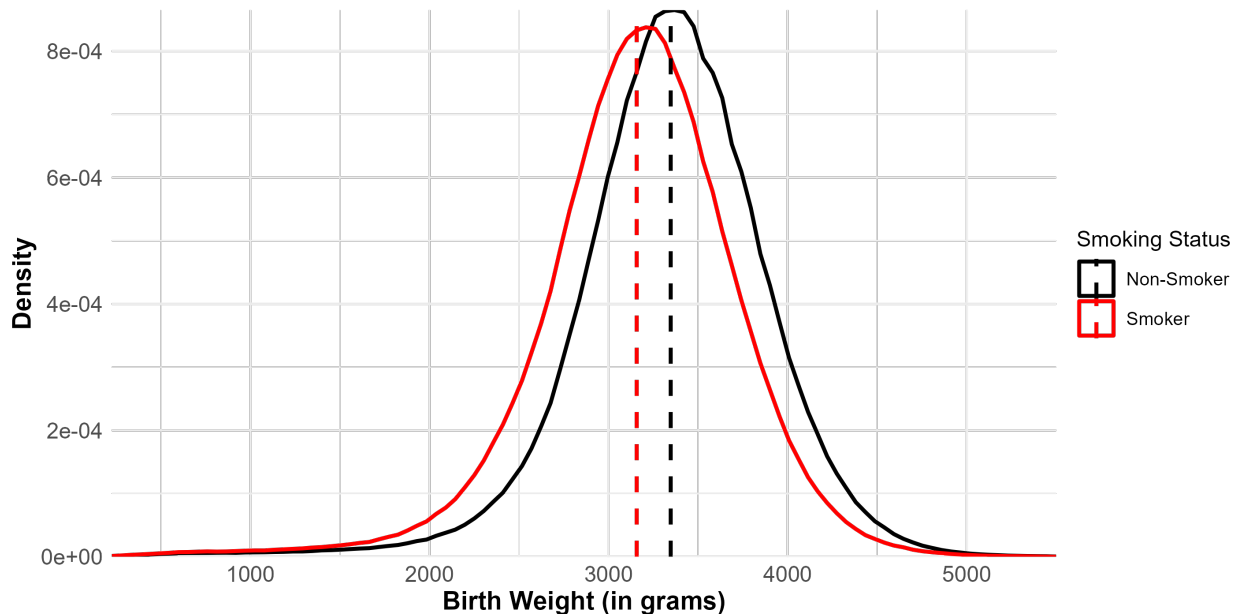


Figure 1: Density of Birth Weight for Smokers and Non-Smokers for 2011-2018

Note: Solid lines show density estimates of birth weight for smokers and non-smokers for births between 2011 and 2018. The dashed lines indicate the mean birth weight of each group.

The main outcome of interest in this study is health at birth, which cannot be measured directly. Hence, I use several proxy measures for health at birth: (Standardized) birth weight, an indicator for low birth weight birth, 5-minute Apgar score and gestation length. Apgar Score is a key measure of infant health right after birth, which is able to capture different dimensions of health at birth than birth weight, since it focuses more on activity and respiration of the newborn (American College of Obstetricians and Gynecologists., 2015). Following Cnattingius (1989), I standardize birth weight for gestational age and sex of newborn. None of the used proxy measures is a perfect proxy for health at birth, but evaluating a wide range of outcomes enables to capture several aspects of health at birth.

My sample contains 17,398,750 pregnancies which occurred in the US between 2011-2018.

I restricted the sample to singletons born in the US. Additionally, I discard all observations with missing information in the variables of interest. The data is rich and includes general demographic information of the mother and father, as well as mothers pregnancy history. Further, I consider information on prenatal care, medical and other risk factors, infections and characteristics of the newborn ⁶.

The density plot in Figure 1 for data from 2011 to 2018 clearly shows the shift of the entire distribution for smokers to lower birth weights. Mean birth weight for smokers is 3157.57 grams, which is 189.72 grams lower than the mean birth weight for babies born to non-smokers.

Table 1 represents summary statistics of relevant variables, grouped by mother’s demographic information, pregnancy related risk factors and infant characteristics for smoker and non-smoker (see online Appendix Table A.1 for an extension). Smoking mothers giving birth are on average 2.55 years younger, are less likely to be married and report to have less educational attainment than non-smoking mothers. Medicaid is the most common payment source for smoking mothers, whereas private insurance is the most common payment source for non-smokers. For smokers, the average number of daily cigarettes declines with advancing pregnancy from 10.3 in the first trimester to 6.8 in the third. Birth weight of infants born to smoking mothers is lower than that of infants born to non-smoking mothers. Also, the low-birth-weight rate increases sharply for smokers. LBW rate is 5% for non-smokers and increases to 10% for smokers.

⁶For a detailed overview on the data, refer to Appendix A in the supplemental material.

Table 1: Summary Statistics for Smoker and Non-Smoker for Births between 2011-2018

	Non-Smoker		Smoker		Norm.
Variables	Mean	SD	Mean	SD	Diff.
Infant Characteristics					
Birth Weight in grams	3347.29	537.65	3157.57	555.51	-0.25
Low Birth Weight	0.05	0.22	0.10	0.30	0.13
5-min. APGAR Score	8.82	0.74	8.78	0.85	-0.04
Gestation Length	38.87	2.15	38.70	2.51	-0.05
Fraction Male	0.51	0.50	0.51	0.50	0.00
Risk Factors					
Smoking during Pregnancy	0.00	0.00	1.00	0.00	-
Cigarettes Before Pregnancy	0.26	2.20	14.03	9.88	1.36
Cigarettes 1st Trimester	0.00	0.00	10.30	8.17	1.26
Cigarettes 2nd Trimester	0.00	0.00	7.65	7.47	1.02
Cigarettes 3rd Trimester	0.00	0.00	6.79	7.31	0.93
Weight Gain in pounds	30.31	14.34	30.60	17.05	0.01
Prenatal Care Visits	11.62	3.76	10.94	4.21	-0.12
Month Prenatal Care Began	2.84	1.40	3.18	1.61	0.16
Total Delivery Order	2.28	1.43	2.66	1.65	0.18
BMI (prepregnancy)	26.39	6.37	26.88	7.07	0.05
Previous Preterm Birth	0.02	0.15	0.05	0.21	0.09
Mothers Demographic Information					
Mother's Age	28.79	5.64	26.24	5.31	-0.33
Married	0.72	0.45	0.38	0.48	-0.51
Race - White	0.79	0.41	0.89	0.31	0.19
Race - Black	0.12	0.32	0.08	0.27	-0.09
Race - American Indian / Eskimos	0.01	0.09	0.02	0.14	0.06
Race - Asian / Pacific Islander	0.08	0.27	0.01	0.10	-0.24
Hispanic Origin - None	0.79	0.41	0.95	0.21	0.36
Hispanic Origin - Mexico	0.14	0.34	0.02	0.15	-0.30
Hispanic Origin - Puerto Rico	0.01	0.11	0.01	0.10	-0.02
Hispanic Origin - Cuba	0.01	0.07	0.00	0.04	-0.05
Hispanic Orig. - C./S. America	0.03	0.16	0.00	0.04	-0.14
Education - up to 12th grade	0.03	0.17	0.02	0.13	-0.06
Education - Highschool	0.08	0.27	0.20	0.40	0.27
Education - College without degree	0.21	0.41	0.41	0.49	0.31
Education - Associate degree	0.20	0.40	0.26	0.44	0.09
Education - Bachelor's degree	0.09	0.29	0.06	0.24	-0.07
Education - Master's degree and PhD	0.24	0.43	0.03	0.18	-0.45
Payment - Medicaid	0.34	0.47	0.69	0.46	0.53
Payment - Private Insurance	0.57	0.49	0.25	0.43	-0.49
Payment - Self-Pay	0.04	0.20	0.02	0.14	-0.09
Number of Observations	16,298,727		1,100,023		

Source: National Center for Health Statistics (2011-2018)

4 Empirical Strategy

4.1 Setup and Identification

I define causal effects via the potential outcomes model by Imbens and Rubin (2015). For each observation $i, i = 1, \dots, N$, the outcome under potential treatment is $Y_i^{(T_i)}$, $T_i \in \{0, 1\}$, where $Y_i^{(0)}$ denotes the potential outcome if individual i did not receive the treatment and $Y_i^{(1)}$ denotes the potential outcome if i did receive the treatment. For a single observation, only one potential outcome can be observed, which is given by the observed outcome $Y_i^{obs} = T_i Y_i^{(1)} + (1 - T_i) Y_i^{(0)}$. So for each individual I observe (Y_i^{obs}, T_i, X_i) , with X_i being the vector of individual characteristics.

The average treatment effect is defined as $\tau_{ATE} = E[Y_i^{(1)} - Y_i^{(0)}]$. The main effect of interest in this analysis are conditional average treatment effects (CATE) which differ across individuals by their characteristics defined as

$$\tau(x) = E[Y_i^{(1)} - Y_i^{(0)} | X_i = x]. \quad (1)$$

To ensure causal interpretation I assume exogeneity of confounders, SUTVA, common support (a discussion of the common support assumption can be found in section 5)

$$0 < P(T_i = 1 | X_i) < 1 \quad (2)$$

and unconfoundedness (random treatment status conditional on X):

$$\{Y_i^{(0)}, Y_i^{(1)}\} \perp T_i | X_i, \quad (3)$$

where $\perp$ denotes independence of two random variables.

Unconfoundedness is a very strong assumption. The dataset provides a rich set of covariates, that the literature has found to be associated with both the smoking decision and birth

outcomes (Smedberg et al., 2014). But only relying on rich data does not ensure unconfoundedness. While I cannot test for unconfoundedness directly, I can provide several justifications that it is likely to hold. The first one is a sensitivity analysis of Oster (2019), which relies on observables to assess the importance of omitted variables in regression analysis. Table ?? in the supplemental material shows that the bounded estimates are very close to the causal forest estimates, suggesting that results are not driven by unobserved confounding. Second, I estimate the effect of smoking on several placebo outcomes, namely infectious diseases, which might proxy for risky behavior and should be unrelated to smoking. While coefficients appear significant in the placebo analysis, this can be attributed to very large sample size. The actual effect size is indeed negligible, supporting the validity of the assumption. Third, I match my estimates with external evidence on causal effects from randomized controlled trials Lumley and Watson (2009), controlling for unobserved maternal factors by comparing the same mothers in different pregnancies (Currie et al., 2009), or quasi-experimental evidence using instruments (Lien and Evans, 2005; Bharadwaj et al., 2014). Figure B.1 in the supplemental material shows that the estimated average treatment effect is well in line with the literature. Taken all of these approaches together unconfoundedness is likely to hold. For a detailed discussion of the result refer to the supplemental material.

4.2 Causal Forest

To estimate the conditional average treatment effect as in (1), I use the causal forest by Athey et al. (2019). The causal forest is a tree based machine learning algorithm building on the idea of classical random forests (Breiman, 2001). The starting point for the idea behind the causal forest is a partially linear model as in Robinson (1988), in which the treatment effect is considered to be constant and causal effect τ can be estimated via

$$\hat{\tau} = \frac{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{m}(X_i)) (T_i - \hat{p}(X_i))}{\frac{1}{n} \sum_{i=1}^n (T_i - \hat{p}(X_i))^2}, \quad (4)$$

where $\hat{p}$ and $\hat{m}$ are estimates of the so-called nuisance parameters $p(x) = \mathbb{E}[T_i|X_i = x]$, the propensity score, and $m(x) = \mathbb{E}[Y_i|X_i = x]$, the conditional mean of Y . These can be estimated by two separate regression forests. Building on this idea, Nie and Wager (2020) derived an “R-learner” objective function for treatment effects that differ with respect to covariates

$$\hat{\tau}(\cdot) = \arg \min_{\tau} \left(\sum_{i=1}^n ((Y_i - \hat{m}^{(-i)}(X_i)) - \tau(X_i) (T_i - \hat{p}^{(-i)}(X_i)))^2 + \Lambda_n(\tau(\cdot)) \right), \quad (5)$$

where $\Lambda_n(\tau(\cdot))$ is a regularization term controlling for the complexity of $\hat{\tau}(\cdot)$.

Thinking of the classic random forest algorithm as an adaptive kernel method, where weights $\alpha_i(x)$ essentially measure how often the i -th training observation falls in the same leaf as the test point x , Athey et al. (2019) suggest a combination of (4) and (5) to estimate the conditional average treatment effects. The grf algorithm, first builds a forest to derive weights $\alpha_i(x)$ which are then used to estimate the CATE function as follows:

$$\hat{\tau}(x) = \frac{\sum_{i=1}^n \alpha_i(x) (Y_i - \hat{m}^{(-i)}(X_i)) (T_i - \hat{p}^{(-i)}(X_i))}{\sum_{i=1}^n \alpha_i(x) (T_i - \hat{p}^{(-i)}(X_i))^2}. \quad (6)$$

To arrive at the weights $\alpha_i(x)$, the grf implementation of the causal forest fits two separate regression forests in order to estimate propensity scores $\hat{p}(x)$ and marginal outcomes $\hat{m}(x)$ and makes out-of-bag predictions (denoted by $^{(-i)}$). These are then used to residualize outcomes and treatments, on which causal trees are grown. The splits of the trees are chosen, so that treatment effect within each leaf is similar or homogeneous, while there is heterogeneity in treatment effects between leaves. For details on choosing a splitting rule see Athey et al. (2019).

To judge whether there is heterogeneity and whether the estimated $\tau(x)$ on average captures the ATE well, a calibration test is used. For this test the following regression is used, inspired by the BLP in Chernozhukov et al. (2018c), but adapted for estimation in

observational settings:

$$Y - \hat{m}(x) = \beta_1(T_i - \hat{p}(X_i))\bar{\hat{\tau}}(x) + \beta_2(T_i - \hat{p}(X_i)) * (\hat{\tau}(x) - \bar{\hat{\tau}}(x)), \quad (7)$$

where $(T_i - \hat{p}(X_i))\bar{\hat{\tau}}(x)$ refers to the mean forest prediction and $(T_i - \hat{p}(X_i)) * (\hat{\tau}(x) - \bar{\hat{\tau}}(x))$ to the differential forest prediction. A coefficient β_1 close to 1 suggests that the mean forest prediction is correct. If β_2 is significantly different from 0, we can reject the null of no heterogeneity.

4.3 Effect Decomposition

I want to identify individual's characteristics for which there are large differences in treatment effects, while keeping all other characteristics fixed. To do so, I split the sample into non-overlapping groups based on a partitioning characteristic (i.e., sex: male and female population), and decompose the treatment effect for individuals in these populations⁷. I decompose the treatment effect difference into the structural effect, which arises due to differences in treatment effect for individuals with the same characteristics leaving out the partitioning characteristic, and the compositional effect arising from differences in the characteristics between the groups. Chernozhukov et al. (2013) proposed this decomposition that makes use of the counterfactual distribution. I apply the decomposition with focus on changes in the distribution of the treatment effect $\tau(x)$, depending on the outcome Y and treatment indicator T . Using this decomposition on the estimated heterogeneous treatment effect allows for detecting the direct effect of a variable on the treatment effect (structural effect) and helps to identify sources of heterogeneity. The proposed decomposition can in principle be applied in different scenarios and is not tied to a only selection on observables

⁷In principal, one could also use multiple variables for building the groups, as long as they describe non-overlapping groups. These variables should be chosen by the researcher to balance characteristics between groups of interest.

setting. It can be applied in scenarios with estimated CATE based on a rich set of covariates and includes settings like RCT and IV.

Considering two non-overlapping populations $j, k \in \mathcal{K}$, the observed differences in the treatment effects can be decomposed as

$$\underbrace{F_{\tau\langle k|k \rangle}(y) - F_{\tau\langle j|j \rangle}(y)}_{\text{Total Effect}} = \underbrace{[F_{\tau\langle k|k \rangle}(y) - F_{\tau\langle j|k \rangle}(y)]}_{\text{Structural Effect}} + \underbrace{[F_{\tau\langle j|k \rangle}(y) - F_{\tau\langle j|j \rangle}(y)]}_{\text{Compositional Effect}}. \quad (8)$$

$F_{\tau\langle k|k \rangle}(y)$ is the estimated distribution function of $\tau(x)$ for population k . $F_{\tau\langle j|k \rangle}(y)$ represents the counterfactual distribution function of τ . This can be interpreted as the distribution function of τ that would have prevailed for population j , if they faced population k 's covariate distribution. The derived decomposition shows the variation along the whole distribution of treatment effects. It is also possible to decompose quantile effects. Quantile effect decomposition is easier to interpret and directly shows the magnitude of effect differences. The resulting decomposition looks as follows:

$$\underbrace{Q_{\tau\langle k|k \rangle}(p) - Q_{\tau\langle j|j \rangle}(p)}_{\text{Total Effect}} = \underbrace{[Q_{\tau\langle k|k \rangle}(p) - Q_{\tau\langle j|k \rangle}(p)]}_{\text{Structural Effect}} + \underbrace{[Q_{\tau\langle j|k \rangle}(p) - Q_{\tau\langle j|j \rangle}(p)]}_{\text{Compositional Effect}}, \quad (9)$$

$$\forall k, j \in \mathcal{K}, p \in (0, 1),$$

where $Q_{\tau\langle j|k \rangle}(p) := F_{\tau\langle j|k \rangle}^{-1}(p)$, $p \in (0, 1)$ is the left-inverse function of $F_{\tau\langle j|k \rangle}$.

The counterfactual distribution, the key feature to the decomposition, can be thought of as either the result of a change in the distribution of a set of covariates X , or a change in the relationship between the covariates with the outcome of interest (Chernozhukov et al., 2013). It does not arise from any observable population. It is derived by integrating the conditional distribution of the outcome Y or in our setting $\tau(x)$ with respect to the

distribution of characteristics of another population.

$$F_{\tau\langle j|k\rangle}(y) := \int_{\mathcal{X}_k} F_{\tau_j|X_j}(y|x) dF_{X_k}(x), \quad (10)$$

$$Q_{\tau\langle j|k\rangle}(y) := F_{\tau\langle j|k\rangle}^{-1}(p), \quad p \in (0, 1). \quad (11)$$

To arrive at the counterfactual distribution in equation (10), the conditional distribution function of population j needs to be integrated with respect to the distribution of characteristics of population k , which makes the group characteristics comparable for the decomposition.

To identify the counterfactual distribution, I consider a finite number of populations $k \in \mathcal{K}$, for which covariates $X_k \in \mathbb{R}^d$ and the outcome $\tau_k \in \mathbb{R}$ are observed or estimated. Given the observability of covariates X_k , the covariate distribution F_{X_k} , $\forall k \in \mathcal{K}$, and the conditional distribution $F_{\tau_j|X_j}$, $\forall j \in \mathcal{K}$, can be identified. Let $\mathcal{X}_k \subseteq \mathbb{R}^d$ denote the support of X_k and $\mathcal{Y}_k \subseteq \mathbb{R}$ the support of τ_k .

The counterfactual distribution is well-defined if $\mathcal{X}_k \subseteq \mathcal{X}_j$, $\forall (j, k) \in \mathcal{K}$. To estimate the counterfactual distribution, one first has to estimate the covariate distribution $\hat{F}_{X_k}(x)$ using the empirical distribution function

$$\hat{F}_{X_k}(x) = \frac{1}{n_k} \sum_{i=1}^{n_k} 1_{\{X_{ki} \leq x\}}, \quad k \in \mathcal{K}. \quad (12)$$

For estimating the conditional distribution function $F_{\tau_j|X_j}(y|x)$, which describes the stochastic assignment of τ for individuals with characteristics X_j for population j , I make use of distribution regression. Chernozhukov et al. (2013) introduce distribution regression for estimating the conditional distribution function. It uses a binary regression framework to estimate the conditional distribution of an outcome Y given covariates. The proposed dis-

tribution regression model is defined as

$$F_{Y|X}(y|x) = \Lambda(P(x)' \beta(y)), \quad \forall y \in \mathcal{Y}, \quad (13)$$

where $P(x)$ is a vector of known transformations of X , $\beta(y)$ is parameter vector that may vary with y and Λ is a link function. Different link functions such as logit, probit, linear, and log-log are possible in this setting (Chernozhukov et al., 2013).

The estimator of the conditional distribution function takes the form

$$\hat{F}_{\tau_k, X_k} = \Lambda(P(x)' \hat{\beta}_k(y)), \quad (y, x) \in \mathcal{Y}_k \times \mathcal{X}_k, \quad k \in \mathcal{K}, \quad (14)$$

$$\hat{\beta}_k = \arg \max_{b \in \mathbb{R}^{d_p}} \sum_{i=1}^{n_k} [1_{\{\tau_{ki} \leq y\}} \ln [\Lambda(P(X_{ki})' b)] + 1_{\{\tau_{ki} > y\}} \ln [1 - \Lambda(P(X_{ki})' b)]] , \quad (15)$$

where $d_p = \dim(P(X_j))$. $\hat{\beta}_k$ is derived via Maximum-Likelihood estimation for fixed y . In practice, the counterfactual distribution is estimated on a fine mesh of points, resulting in the convenient form $\hat{F}_{\tau(j|k)}(y) = \frac{1}{n_k} \sum_{i=1}^{n_k} \Lambda(P(X_{ki})' \hat{\beta}_j(y))$.

The estimators are then plugged into (10), where the conditional distribution function is integrated with respect to the covariate distribution to arrive at $\hat{F}_{\tau(j|k)}(y)$. Chernozhukov et al. (2013) also construct confidence for the estimated counterfactual distribution. The construction of the pointwise and uniform confidence intervals over a set of quantile indexes relies on functional central limit theorems and bootstrap functional central limit theorems for the empirical counterfactual distribution.

For implementation, I make use of the **grf** R-package (Tibshirani et al., 2020) and the **Counterfactual** R-package (Chen et al., 2020), which provide a function for estimating causal forest and the counterfactual distribution.

The proposed procedure (Procedure 1) to decompose the treatment effect is based on two estimation steps in sequence (CATE and counterfactual distribution estimation). Thus, this procedure might suffer from over-fitting. To cope with this problem, I rely on a sample

splitting idea by Chernozhukov et al. (2018c), which splits the data into auxiliary (A) and main sample (M). The auxiliary sample is used for training the model to estimate CATE. The CATE is then predicted on the main sample, where the decomposition takes place. The sample split itself is a source of uncertainty, since the learned forest structure depends on the sample it is trained on. To control for uncertainty that is imposed by the splits, I use many splits and report the median over all results.

Procedure 1: Effect Decomposition

Repeat multiple times using different random splits:

1. Split the sample into auxiliary (A) and main sample (M)
 2. On Auxiliary sample: train the causal forest
 3. On Main sample:
 - Obtain estimates $\hat{\tau}(x)$ of the treatment effect via the causal forest trained on A
 - Define number of populations $k \in \mathcal{K}$ based on the partitioning variable $X^* \in X$ (continuous variables need to be categorized)
 - $\forall k \in \mathcal{K} \setminus \{0\}$: decompose total difference between the treatment effect in each group k and the reference group 0 into structural and compositional effect
 - Obtain estimates $\hat{F}_{X_k}$ of the covariate distributions F_{X_k} via (12)
 - Obtain estimates $\hat{F}_{\tau(x)^{(0)}|X_0}$ of the conditional distribution via (13)
 - Obtain estimates of the counterfactual distributions via (10)
 - Obtain decomposition via (9).
-

5 Results

Estimating the average treatment effect of smoking during pregnancy on the birth outcomes of interest, I find that smoking during pregnancy reduces birth weight on average by 155 grams. Further, smoking during pregnancy decreases gestation length by 0.1 weeks and the Apgar score and increases the risk of LBW (see table 4).

The average treatment effect might however mask heterogeneity that is present. Using the test for heterogeneity described in equation (7), table 3 shows that this is indeed the case. The test reveals that the mean forest prediction is indeed correct for all health outcomes of

Table 2: Average Treatment Effects (ATE) with Standard Errors

Outcome	ATE Estimate (SE)
Birth weight	-155.209 (6.274)
Standardized BW	-0.313 (0.012)
LBW	0.029 (0.003)
Gestation Length	-0.101 (0.026)
Apgar Score	-0.022 (0.011)

interest. Only for the Apgar score, there seems to be no detectable heterogeneity, whereas in all other cases, we can reject the null of no heterogeneity.

The main outcomes of interest is standardized birth weight, which is the most widely used proxy for health at birth. Additionally, I will present results on gestation length, non standardized birth weight (C.3) and 5-minute Apgar score (C.4). I will report the full decomposition for mother’s age, but will only present structural effects for other modifying factors.

Table 3: Calibration Estimates with Standard Errors

Outcome	Variable	Estimate (SE)
Birth weight	mean forest prediction	1.011 (0.021)
	differential forest prediction	1.122 (0.088)
Standardized BW	mean forest prediction	1.002 (0.019)
	differential forest prediction	0.954 (0.094)
LBW	mean forest prediction	1.008 (0.051)
	differential forest prediction	1.231 (0.098)
Gestation Length	mean forest prediction	1.032 (0.191)
	differential forest prediction	0.791 (0.142)
Apgar Score	mean forest prediction	0.930 (0.264)
	differential forest prediction	-0.139 (0.167)

5.1 Standardized Birth Weight

Figure 2 shows the conditional average treatment effect of smoking on standardized birth weight estimated using a causal forest and the ATE estimate including corresponding confidence intervals. The effect of smoking is negative, with a mean CATE of around -0.31

standard deviations. This roughly correspond to a mean reduction of -180 grams. Only considering the average treatment effect would mask the heterogeneity that is present. Some individuals are barely harmed by smoking, others face a reduction in standardized birth weight of -0.6 , corresponding to a birth weight of -310 grams. The estimated average treatment effect is clearly negative and we find significant heterogeneity.

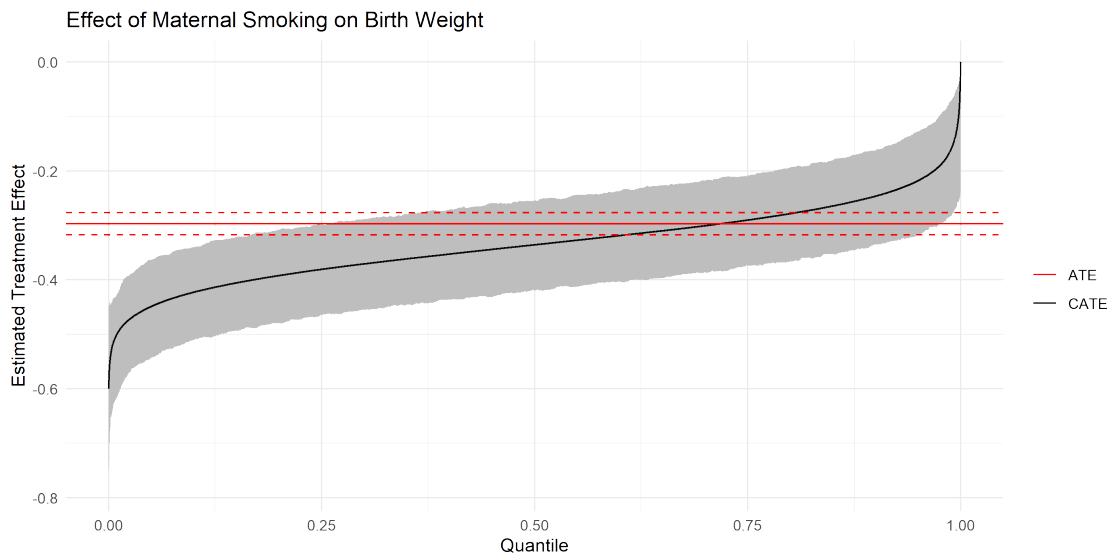


Figure 2: Sorted CATE and ATE estimates of smoking on standardized birth weight

Note: The solid black line shows the estimated CATE of smoking on standardized birth weight sorted in ascending order. The dashed black line shows the 95% confidence intervals derived via bootstrapping. CATE is estimated using a causal forest as described in 4.2 using the `grf` R-package. The solid red line shows the ATE and corresponding 95% confidence intervals.

The results in figure 3 follow the implementation steps as described in procedure 1, repeated 10 times using different random splits into main and auxiliary sample. I report confidence bands for $\alpha = 0.05$, meaning I use bootstrap confidence bands for confidence level 0.975 but discount for splitting uncertainty, resulting in confidence level of 0.95. The x -axis of the plot shows the quantile index. Since I am looking at (mostly) negative effects, the lower quantiles are the areas where the effect of smoking on the birth outcome of interest is strongest. The y -axis displays the differences in the outcome of interest. The plots for the total, structural and compositional effect always display the difference in treatment effect in the corresponding age group with the baseline, which is the youngest age group (mothers aged lower than 23). The top left plot shows the quantiles of the estimated CATE cumulative

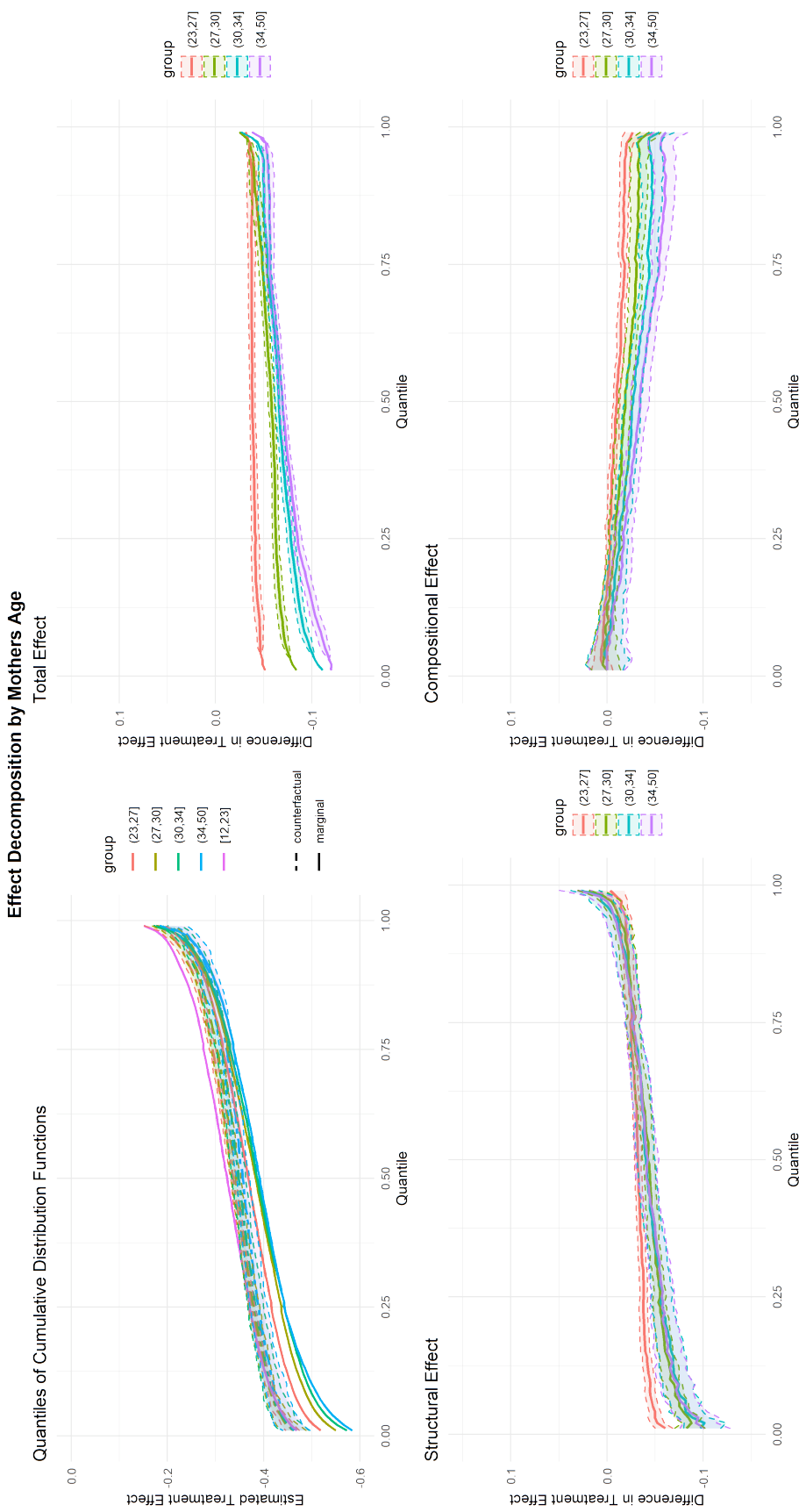


Figure 3: Standardized Birth Weight - Effect Decomposition by Mother's Age

Note: The figure displays the decomposition of the effect of maternal smoking during pregnancy on standardized birth weight by mother's age. It shows the quantiles of the cumulative distribution function of each group and their counterfactual distribution (top left), the total effect difference (top right), structural effect (bottom left) and compositional effect (bottom right) difference between each group and the reference group. The counterfactual distribution and decomposition are derived with respect to the reference group which are mothers aged 12 to 23 (the lowest quintile). Decomposition follows the procedure described in 1. Total, structural and compositional effect are defined as in 9. Positive difference corresponds to effect mitigation with increasing age, whereas negative difference corresponds to effect amplification with increasing age respectively. Shaded areas show 95% confidence intervals.

Structural Effect of Parity

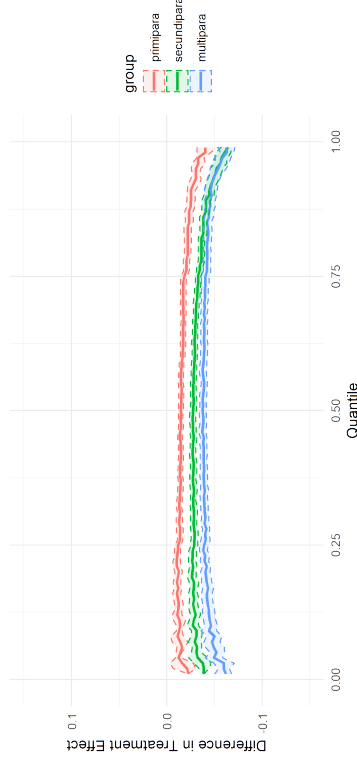


Figure 4: Standardized Birth Weight- Structural Effect by Parity

Structural Effect of BMI

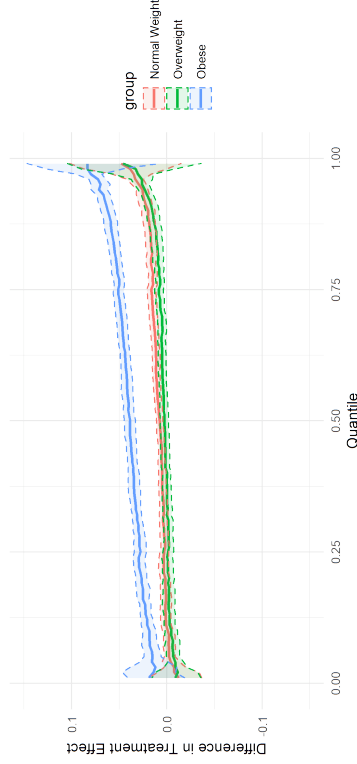


Figure 5: Standardized Birth Weight- Structural Effect of BMI

Structural Effect of Weight Gain

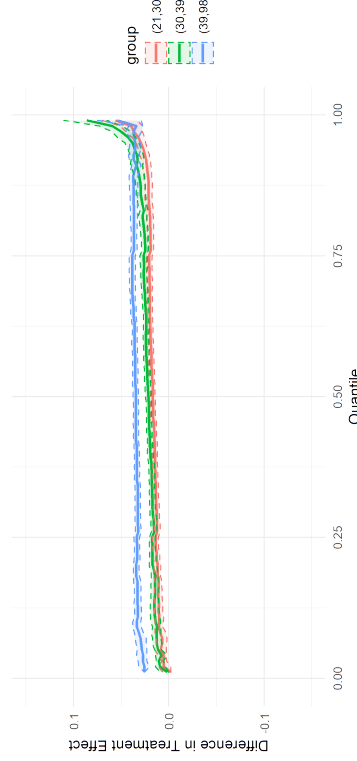


Figure 6: Standardized Birth Weight- Structural Effect of Weight Gain

Note: The figure shows the structural effect decomposition of the effect of maternal smoking on standardized birth weight for different modifying factors. Decomposition follows the procedure described in 1. The reference groups are nulliparous mothers 4, underweight mothers 5, mothers gaining less than 22 pounds (lowest quartile) 6, and mothers gaining less than recommended 7. Positive difference corresponds to effect mitigation, whereas negative difference corresponds to effect amplification respectively. Shaded areas show 95% confidence intervals.

Structural Effect of Weight Gain

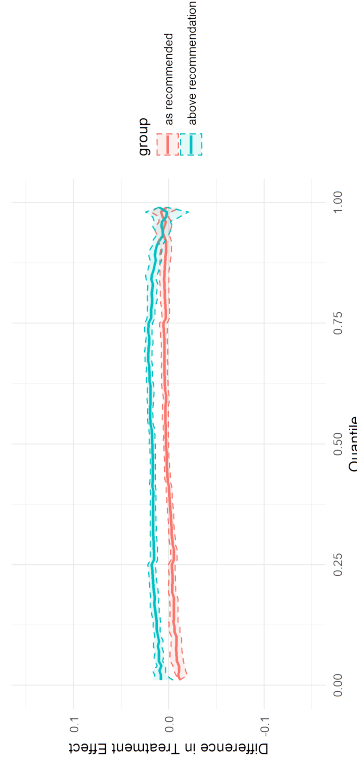


Figure 7: Standardized Birth Weight- Structural Effect of Weight Gain Recommendations

distribution (solid line) and counterfactual distribution that would have prevailed for other groups if they faced reference group’s distribution of characteristics (dashed line). The top right shows the total effect, which is the difference between each group’s CATE distribution and the CATE distribution of the baseline group. The bottom left plot shows the structural effect, which is the main effect of interest. It shows the difference only associated with mother’s age. Characteristics of groups are kept comparable by using counterfactual distribution. The plot on the bottom right shows the compositional effect, which is the difference associated with differences in group characteristics, leaving out decomposition dimension. I find strong modifying effect for mother’s age in the effect of smoking on standardized birth weight. The total effect indicates significant differences between mothers younger than 27 and older than 27. The structural effect confirms this observation. For mothers older than 27, the effect of smoking on standardized birth weight is nearly twice as strong as for younger mothers. For mothers older than 27, there is no significant difference between the three age groups considered. This profound difference fades for the last two quartiles, suggesting within group heterogeneity. For parity, I separately look at nullipara, primipara, secundipara, and mothers with 3 or more previous live births (multipara). The structural effect of the decomposition by parity (see Figure 4) reveals that increased number of previous children born to the mother amplify the effect of smoking on standardized birth weight. For multipara for example, this can explain up to -0.05 (around 27g) of the total difference in standardized birth weight. The difference is stable over the quantile indices, suggesting that each additional child the mother has given birth to is amplifying the effect of smoking on standardized birth weight by 0.025 standard deviations. Since the effect difference is stable over the quantile indices and clearly separated between the groups, there is no within group heterogeneity.

Figure 5 shows a strong mitigating effect of obesity on the effect of smoking on standardized birth weight. There seems to be no profound difference between overweight or normal weight women compared to underweight women, which are the reference group in this case.

For weight gain I look at the actual weight gain in pounds, but additionally relate weight gain to recommendations based on prepregnancy BMI published by CDC (2024)⁸.

I see similar results for weight gain in pounds and weight gain recommendations (Figures 6 and 7). Excessive weight gain is having a very small positive effect, as it helps in mitigating the effect of smoking on standardized birth weight. This effect is very small, there is only a detectable difference in standardized birth weight of 0.025, and there is no clear difference between low and moderate weight gain. When looking at absolute weight gain, the structural effects also point towards the hypothesis, that gaining excessive weight (above recommendation) has a very small positive effect. Gaining below recommendation and as recommended seem to have no modifying effect on the effect of smoking on standardized birth weight.

5.2 Gestation Length

Another widely used indicator for health at birth is gestation length. The observable mean difference in gestation length for babies born to smoker and non-smoker is only 0.17 weeks, and unlike a clear shift towards lower birth weights, smoking seems to be moving the mass of the distribution towards the tails for the gestation length for smoking mothers. The estimated ATE of smoking on gestation length is -0.101 weeks. The gestation length of some infants is more negatively affected (estimates up to -0.6 weeks), and some even experience a positive effect of smoking on gestation (see figure 8), which can be explained by growth retardation of very large babies caused by smoking.

In contrast to the results for birth weight, figure 9 reveals weak effect modification by

⁸Recommendations are as follows: For BMI less than 18.5 (underweight), recommended weight gain is 28 – 40 pounds, for BMI between 18.5 – 24.9 (normal weight) recommendation is 25 – 35 pounds, for BMI 25 – 29.9 (overweight) recommendation is 15 – 25 pounds, and for BMI above 30 (obese) recommendation is 11 – 20 pounds. I classify mothers into three categories which are 'below recommendation', 'as recommended', and 'above recommendation'.

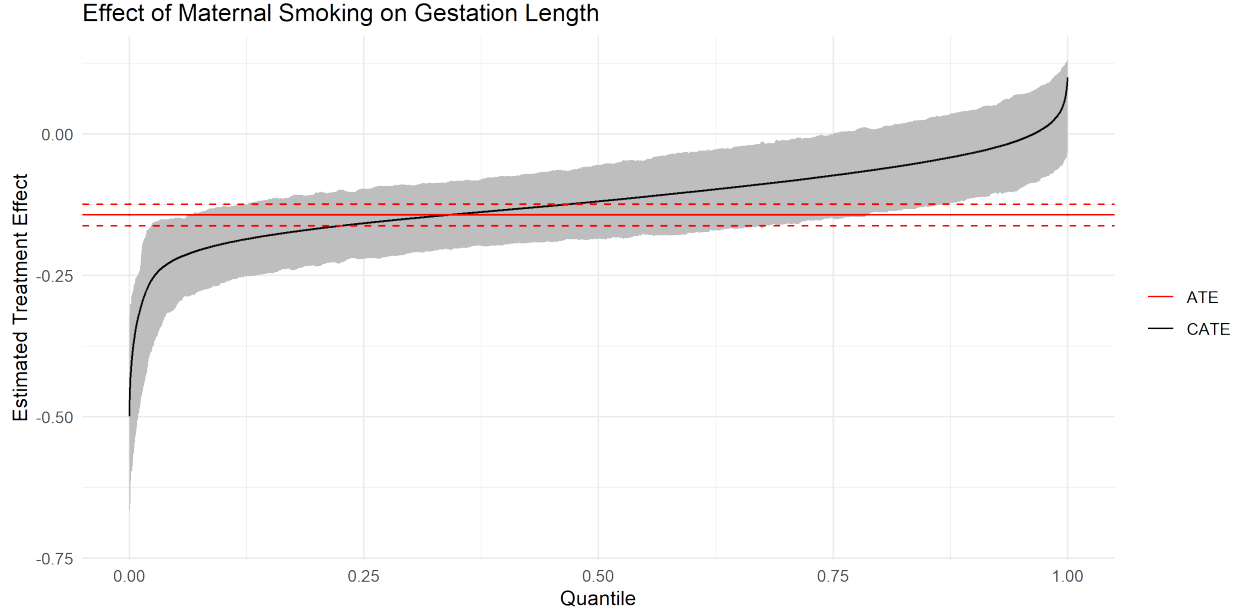


Figure 8: Sorted CATE and ATE estimates of smoking on Gestation Length

Note: The solid black line shows the estimated CATE of smoking on gestation length sorted in ascending order. The dashed black line shows the 95% confidence intervals derived via bootstrapping. CATE is estimated using a causal forest as described in 4.2 using the **grf** R-package. The solid red line shows the ATE and corresponding 95% confidence intervals.

maternal age. The structural effect suggests that there is an increase in treatment effects attributable to age. The difference in treatment effect that is solely based on mother's age is at most 0.2 weeks, corresponding to 1.4 days in difference of gestation length. Figure 10 shows the differences in effect magnitude that can be attributed to different levels of parity is very close to zero. Similarly, there seems to be no significant effect modification that can be attributed to pregnancy weight gain (see figure 12 and 13). Figure 11 shows large differences in treatment effects of smoking on gestation length by prepregnancy BMI. The structural effect reveals that 0.2 weeks difference in treatment effects between underweight and obese mothers can be attributed to their prepregnancy weight. The effect mitigation increases linearly with BMI.

Table 4: Overview: Robustness Checks

			Decomposition by		
	Mother's Age	Parity	Prepreg BMI	Weight Gain	Weight Gain Rec.
Dose dependency					
Birth Weight	+	+	- -	-	-
Pregnancy Smoking					
Birth Weight	+	+	- -	-	-
Propensity Trimming					
Birth Weight	+	+	-	-	-

++: strong amplification, +: amplification, 0: no effect, -: weak mitigation, - -: strong mitigation

See E in the supplemental material for detailed decompositions.

6 Robustness Checks

The identified drivers of heterogeneity are robust to different outcomes, choice of treatment definition and sample selection. Table 4 provides an overview of these checks.

Dose dependency - To ensure the findings are robust against heterogeneity in smoking levels across groups and to address dose dependency, I rerun the analysis on heavy smokers only. Heavy smoking is defined as smoking 20 or more cigarettes daily (Pierce et al., 2011). The effect size of smoking appears larger when considering only heavy smokers, indicating a clear dose dependency with stronger negative effects. Similar amplification effects are observed with increased maternal age and parity, consistent with the main analysis. However, weight gain shows less mitigation of smoking effects among heavy smokers. Overall, results closely mirror those of the main analysis, highlighting age-related amplification in smoking effects and consistent findings across parity and BMI categories.

Prepregnancy Smoking - Smoking during pregnancy faces strong social stigma, potentially leading to underreported smoking behaviors. Howland et al. (2015) found that birth certificates underestimated smoking by approximately 24.3% before pregnancy and 26.2% during pregnancy compared to medical records. To assess the robustness of our re-

sults against misreporting, we use smoking before pregnancy as an alternative treatment. For standardized birth weight, patterns in modifications do not change much compared to our main analysis. Effect size for those harming the baby most remains similar. However, the quantiles of the cumulative distribution function show, that some mothers barely harm their baby at all, which can be explained by mothers quitting smoking when pregnant.

Propensity Trimming - I adress concerns regarding limited overlap in figure C.1. To ensure sufficient overlap in the propensity score of treated and untreated observations, I apply asymmetric trimming methods as described by Stürmer et al. (2010)⁹. This trimming involves first, removing observations with no overlap in propensity scores and second, restricting the sample. The lower end of the propensity score distribution is restricted to the lowest 2.5th percentile of treated propensity scores and the upper end to the highest 97.5th percentile of untreated propensity scores.

7 Policy Implications

The results presented can serve as an effective mean to improve allocative efficiency by better targeting of intensified smoking cessation programs. These programs typically include intensified counseling session, thorough information material, follow-up calls, or medication to increase chances of successful smoking cessation (Agency for Healthcare Research and Quality, 2019). Intensified smoking cessation programs are scarce and usually not available for all pregnant smoking women. Using the case of low birth weight in a hypothetical scenario, where 10,000 places for intensified smoking cessation are available, I want to highlight the potential of my results by improving targeting using the identified drivers of heterogeneity.

Since my results suggest, that risk factors for amplifying the effect of smoking on birth weight are increased mothers age, increased parity, low prepregnancy BMI and low weight

⁹Crump et al. (2009) propose a symmetric trimming approach, but it would exclude most of our observations.

gain, I sort all smoking mothers by these four easy criteria and make smoking cessation available for the first 10,000 in our sorted list. So the smoking cessation treatment will be distributed to smoking mothers who are old, with high number of previous births, have low BMI and low weight gain. To compare results to an uninformed choice, I select 10,000 smoking mothers at random. For each selected smoking mother, I estimate the effect of smoking. Subtracting this reduction from the actual birth weight, I can get an estimate of the birth weight the infant would have had in case the mother did not smoke. In order to evaluate the costs saved, I multiply costs savings associated with each additional gram of birth weight with the estimated effect of smoking in grams. I use estimates provided by Almond et al. (2005)¹⁰, who estimate savings in hospital charges associated with each additional gram of birth weight in certain birth weight regions. For example costs savings associated with each additional gram of birth weight for babies born at 800-1000 grams is \$212.77, whereas it is only \$5.19 for babies born at 2500-3000 grams.

Given that not all mothers who are assigned to smoking cessation programs stop smoking, the resulting estimates would heavily overstate the cost savings in hospital charges. Following Balmumcu and Ünsal Atan (2021); Gadowski et al. (2011), I assume that 60% of mothers who receive intensified smoking cessation stop smoking. This simplification does not take into account any heterogeneous effects of smoking cessation programs on the success of smoking cessation.

Figure 14 shows birth weight distribution for children born to randomly targeted smokers and smokers targeted by using the identified drivers of heterogeneity. Clearly, the informed sorting captures more babies born with LBW. The sorting based on main modifying factors selects 1546 LBW babies, whereas the random sorting selects only 963.

This also translates into improved cost savings. While targeting at random saves \$8,987,994 in hospital charges, targeting using the results of this study saves \$16,108,357. The saved costs from the informed sorting are approximately 80% higher than in the random case.

¹⁰Details on costs can be found in D in the supplemental Appendix or Almond et al. (2005), Table IV

Considering medical inflation, these cost savings correspond to between \$56.1 million and \$31.2 million in year 2024 dollars. These estimates are already corrected for the probability of smoking cessation of mothers enrolled. Apart from saved costs for newborns, targeting those most at risk to harm their babies a lot, can also result in improved health of the mothers. So the estimated savings can be thought of as a lower bound for actual costs saved. This analysis underlines the strength of the main results derived in this study. Knowledge of factors amplifying the effect of smoking on health at birth is crucial and can help in better targeting intensified care.

8 Conclusion

To design targeted policies, policymakers must understand effects beyond averages because averages can mask significant variations and heterogeneity within populations. By understanding heterogeneity and its drivers, policymakers can design interventions that are more effective and suitable when facing scarce resources, addressing the specific needs and challenges of diverse groups. In this study, I uncover driving factors of heterogeneity in the effect of maternal smoking during pregnancy. When decomposing the treatment effect on several birth outcomes, I find the strongest modifying effects for mothers' age, which are robust to different specifications and outcomes.

As highlighted in section 7 the findings presented have important policy implications, especially regarding targeting of intensified smoking cessation programs to the most vulnerable. Babies harmed a lot by smoking in utero cause higher medical care costs throughout their life cycle, which could have been avoided in case the mother stopped smoking before or during pregnancy. By uncovering the drivers of heterogeneity, we can target intensified smoking cessation to those mothers potentially harming their babies the most. My findings suggest, that a simple targeted intervention of providing smoking cessation to the most vulnerable 10,000 smoking pregnant women, can save costs of more than \$17 million.

While the data abstracted from birth certificates is suited to evaluate newborns' health, I cannot capture the adverse health effects of smoking that result in fetal death. Especially the increased risk of stillbirth and miscarriage cannot be assessed using the data from birth certificates, since it only contains newborns that survived up until delivery at least. Consequently, the results might underestimate the adverse effects of smoking on health, as the most affected fetuses are not included in the sample. Further, the results do not account for dose dependency in the effect of smoking on infant health, since I look at smoking as a binary treatment. To rule out that results might only be driven by stronger smoking of more affected groups, I restrict the smoker sample to heavy smoking only. This way, smoking levels between different groups are comparable, and I can rule out possible effect differences related to differences in smoking intensity. In doing so, I find very similar results to those in our main analysis, confirming strong heterogeneity by the mother's age.

Despite these limitations, I can add new evidence for treatment effect heterogeneity in the effect of smoking on birth outcomes and identify mothers' characteristics that drive the heterogeneity found. In the future, this framework can potentially be used in other health-related contexts beyond health at birth, where identifying drivers of heterogeneity is an important mean to understand underlying mechanisms of treatment effects.

References

- Adams, E. Kathleen, Vincent P. Miller, Carla Ernst, Brenda K. Nishimura, Cathy Melvin, and Robert Merritt, "Neonatal health care costs related to smoking during pregnancy.," *Health economics*, Apr 2002, *11*, 193–206.
- Agency for Healthcare Research and Quality, "Research Protocol: Smoking Cessation Interventions During Pregnancy and the Postpartum Period.," <https://effectivehealthcare.ahrq.gov/products/smoking-pregnancy-infants/research-protocol> 2019.

- Almond, Douglas, Kenneth Y. Chay, and David S. Lee**, “The Costs of Low Birth Weight,” *The Quarterly Journal of Economics*, 08 2005, *120* (3), 1031–1083.
- American College of Obstetricians and Gynecologists.**, “Committee Opinion No. 644: The Apgar Score,” *Obstetrics and gynecology*, Oct 2015, *126*, e52–e55.
- Athey, Susan and Guido Imbens**, “Recursive partitioning for heterogeneous causal effects,” *Proceedings of the National Academy of Sciences*, 2016, *113* (27), 7353–7360.
- **and Guido W Imbens**, “Machine learning methods that economists should know about,” *Annual Review of Economics*, 2019, *11*, 685–725.
- **, Julie Tibshirani, and Stefan Wager**, “Generalized random forests,” *The Annals of Statistics*, 2019, *47* (2), 1148 – 1178.
- Baba, Sachiko, Anna-Karin Wikström, Olof Stephansson, and Sven Cnattingius**, “Influence of Snuff and Smoking Habits in Early Pregnancy on Risks for Stillbirth and Early Neonatal Mortality,” *Nicotine & Tobacco Research*, 08 2013, *16* (1), 78–83.
- Balmumcu, Ayça and Şenay Ünsal Atan**, “Smoking Cessation Programs for Pregnant Women: Utilizing WhatsApp Text Messaging,” *Journal of Addictions Nursing*, 2021, *32* (3).
- Bearth, Nora and Michael Lechner**, “Causal Machine Learning for Moderation Effects,” 2024.
- Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen**, “Inference on Treatment Effects after Selection among High-Dimensional Controls†,” *The Review of Economic Studies*, 11 2013, *81* (2), 608–650.
- Bharadwaj, Prashant, Julian V. Johnsen, and Katrine V. Løken**, “Smoking bans, maternal smoking and birth outcomes,” *Journal of Public Economics*, 2014, *115*, 72–93.
- Blinder, Alan S.**, “Wage Discrimination: Reduced Form and Structural Estimates,” *The Journal of Human Resources*, 1973, *8* (4), 436–455.
- Breiman, Leo**, “Random Forests,” *Machine Learning*, Oct 2001, *45* (1), 5–32.

- Cattaneo, Matias D.**, “Efficient semiparametric estimation of multi-valued treatment effects under ignorability,” *Journal of Econometrics*, 2010, *155* (2), 138–154.
- CDC**, “Weight Gain During Pregnancy,” <https://www.cdc.gov/maternal-infant-health/pregnancy-weight/index.html> May 2024. Accessed: 2024-07-11.
- Chen, Mingli, Victor Chernozhukov, Ivan Fernandez-Val, and Blaise Melly**, *Counterfactual: Estimation and Inference Methods for Counterfactual Analysis* 2020. R package version 1.2.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins**, “Double/debiased machine learning for treatment and structural parameters,” *The Econometrics Journal*, 01 2018, *21* (1), C1–C68.
- , **Iván Fernández-Val, and Blaise Melly**, “Inference on Counterfactual Distributions,” *Econometrica*, 2013, *81* (6), 2205–2268.
- , —, and **Ye Luo**, “The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages,” *Econometrica*, 2018, *86* (6), 1911–1938.
- , **Mert Demirer, Esther Duflo, and Iván Fernández-Val**, “Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India,” Working Paper 24678, National Bureau of Economic Research June 2018.
- Cnattingius, Sven**, “Does age potentiate the smoking-related risk of fetal growth retardation?,” *Early Human Development*, 1989, *20* (3), 203–211.
- , “The epidemiology of smoking during pregnancy: Smoking prevalence, maternal characteristics, and pregnancy outcomes,” *Nicotine & Tobacco Research*, 04 2004, *6* (Suppl.2), S125–S140.
- , **Michele R. Forman, Heinz W. Berendes, Barry I. Graubard, and Leena**

- Isotalo**, “Effect of age, parity, and smoking on pregnancy outcome: A population-based study,” *American Journal of Obstetrics and Gynecology*, 1993, *168* (1, Part 1), 16–21.
- Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik**, “Dealing with limited overlap in estimation of average treatment effects,” *Biometrika*, 01 2009, *96* (1), 187–199.
- Currie, Janet, Matthew Neidell, and Johannes F. Schmieder**, “Air pollution and infant health: Lessons from New Jersey,” *Journal of Health Economics*, 2009, *28* (3), 688–703.
- Deryugina, Tatyana, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif**, “The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction,” *American Economic Review*, December 2019, *109* (12), 4178–4219.
- D’Souza, S W, P Black, and B Richards**, “Smoking in pregnancy: associations with skinfold thickness, maternal weight gain, and fetal size at birth.,” *BMJ*, 1981, *282* (6277), 1661–1663.
- Elek, Péter and Anikó Bíró**, “Regional differences in diabetes across Europe – regression and causal forest analyses,” *Economics & Human Biology*, 2021, *40*, 100948.
- England, Lucinda J., Juliette S. Kendrick, Hoyt G. Wilson, Robert K. Merritt, Paul M. Gargiullo, and S. Christine Zahniser**, “Effects of Smoking Reduction during Pregnancy on the Birth Weight of Term Infants,” *American Journal of Epidemiology*, 10 2001, *154* (8), 694–701.
- Farbmacher, Helmut, Martin Huber, Lukáš Lafférs, Henrika Langen, and Martin Spindler**, “Causal mediation analysis with double machine learning,” *The Econometrics Journal*, 01 2022, *25* (2), 277–300.
- Gadomski, Anne, Laurie Adams, Nancy Tallman, Nicole Krupa, and Paul Jenkins**, “Effectiveness of a Combined Prenatal and Postpartum Smoking Cessation Program,” *Maternal and Child Health Journal*, 2011, *15* (2), 188–197.

- Gilman, Stephen E., Hannah Gardener, and Stephen L. Buka**, “Maternal Smoking during Pregnancy and Children’s Cognitive and Physical Development: A Causal Risk Factor?,” *American Journal of Epidemiology*, 07 2008, *168* (5), 522–531.
- Haworth, J.C., J.J. Ellestad-Sayed, Jean King, and Louise A. Dilling**, “Fetal growth retardation in cigarette-smoking mothers is not due to decreased maternal food intake,” *American Journal of Obstetrics and Gynecology*, 1980, *137* (6), 719–723.
- Howland, Renata E., Candace Mulready-Ward, Ann M. Madsen, Judith Sackoff, Michael Nyland-Funke, Jennifer M. Bombard, and Van T. Tong**, “Reliability of Reported Maternal Smoking: Comparing the Birth Certificate to Maternal Worksheets and Prenatal and Hospital Medical Records, New York City and Vermont, 2009.,” *Maternal and child health journal*, Sep 2015, *19*, 1916–24.
- Imbens, Guido W and Donald B Rubin**, *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press, 2015.
- J, Chamberlain C Dowswell T Oliver S Oakley L Lumley and L Watson**, “Interventions for promoting smoking cessation during pregnancy,” *Cochrane Database of Systematic Reviews*, 2009, (3).
- Kreif, Noemi, Karla DiazOrdaz, Rodrigo Moreno-Serra, Andrew Mirelman, Taufik Hidayat, and Marc Suhrcke**, “Estimating heterogeneous policy impacts using causal machine learning: a case study of health insurance reform in Indonesia,” *Health Services and Outcomes Research Methodology*, 2022, *22* (2), 192–227.
- La Merrill, Michele, Cheryl R. Stein, Philip Landrigan, Stephanie M. Engel, and David A. Savitz**, “Prepregnancy Body Mass Index, Smoking During Pregnancy, and Infant Birth Weight,” *Annals of Epidemiology*, 2011, *21* (6), 413–420.
- Lien, Diana S. and William N. Evans**, “Estimating the Impact of Large Cigarette Tax Hikes: The Case of Maternal Smoking and Infant Birth Weight,” *Journal of Human Resources*, 2005, *XL* (2), 373–392.

- Misra, Dawn P., Nan Astone, and Courtney D. Lynch**, “Maternal Smoking and Birth Weight: Interaction with Parity and Mother’s Own in Utero Exposure to Smoking,” *Epidemiology*, 2005, *16* (3), 288–293.
- National Center for Health Statistics**, “Vital Statistics Natality Birth Data, 2011-2018. Public-use data file and documentation,” 2011-2018.
- Nie, X and S Wager**, “Quasi-oracle estimation of heterogeneous treatment effects,” *Biometrika*, 09 2020, *108* (2), 299–319.
- Oaxaca, Ronald**, “Male-Female Wage Differentials in Urban Labor Markets,” *International Economic Review*, 1973, *14* (3), 693–709.
- Oster, Emily**, “Unobservable Selection and Coefficient Stability: Theory and Evidence,” *Journal of Business & Economic Statistics*, 2019, *37* (2), 187–204.
- Pierce, John P., Karen Messer, Martha M. White, David W. Cowling, and David P. Thomas**, “Prevalence of Heavy Smoking in California and the United States, 1965-2007,” *JAMA*, 03 2011, *305* (11), 1106–1112.
- Robinson, P. M.**, “Root-N-Consistent Semiparametric Regression,” *Econometrica*, 1988, *56* (4), 931–954.
- Simon, David**, “Does Early Life Exposure to Cigarette Smoke Permanently Harm Childhood Welfare? Evidence from Cigarette Tax Hikes,” *American Economic Journal: Applied Economics*, October 2016, *8* (4), 128–59.
- Smedberg, Janne, Angela Lupattelli, Ann-Charlotte Mårdby, and Hedvig Nordeng**, “Characteristics of women who continue smoking during pregnancy: a cross-sectional study of pregnant women and new mothers in 15 European countries,” *BMC Pregnancy and Childbirth*, 2014, *14* (1), 213.
- Spinillo, Arsenic, Ezio Capuzzo, Sabrina E. Nicola, Laura Colonna, Thomas O. Egbe, and Carlo Zara**, “Factors potentiating the smoking-related risk of fetal growth retardation,” *BJOG: An International Journal of Obstetrics & Gynaecology*, 1994, *101*

(11), 954–958.

Stürmer, Til, Kenneth J. Rothman, Jerry Avorn, and Robert J. Glynn, “Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study,” *American journal of epidemiology*, Oct 2010, *172*, 843–54.

Tibshirani, Julie, Susan Athey, and Stefan Wager, *grf: Generalized Random Forests* 2020. R package version 1.1.0.

US Surgeon General, “Health benefits of smoking cessation : a report of the Surgeon General,” 1990.

Wager, Stefan and Susan Athey, “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests,” *Journal of the American Statistical Association*, 2018, *113* (523), 1228–1242.

Wu Wen, Shi, Robert L. Goldenberg, Gary R. Cutter, Howard J. Hoffman, Suzanne P. Cliver, Richard O. Davis, and Mary B. DuBard, “Smoking, maternal age, fetal growth, and gestational age at delivery,” *American Journal of Obstetrics and Gynecology*, 1990, *162* (1), 53–58.

Zarén, Birgitta, Gunilla Lindmark, and Leiv Bakketeig, “Maternal smoking affects fetal growth more in the male fetus,” *Paediatric and Perinatal Epidemiology*, 2000, *14* (2), 118–126.

Zhang, Weihui and Tse-Chuan Yang, “Maternal Smoking and Infant Low Birth Weight: Exploring the Biological Mechanism Through the Mother’s Pre-pregnancy Weight Status,” *Population Research and Policy Review*, 2019, pp. 1–19.

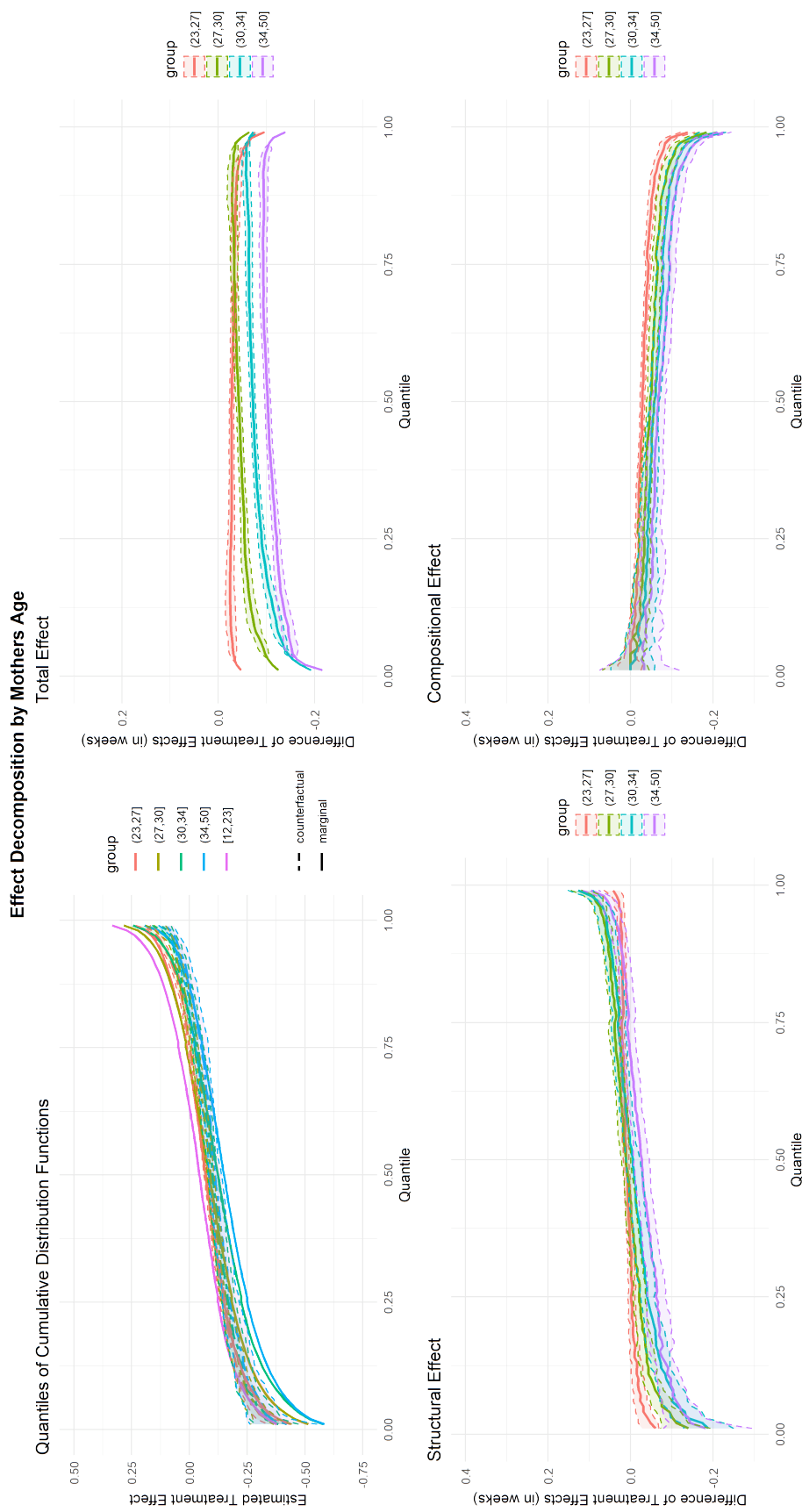


Figure 9: Gestation Length - Effect Decomposition by Mother's Age

Note: The figure displays the decomposition of the effect of maternal smoking during pregnancy on gestation length by mother's age. It shows the quantiles of the cumulative distribution function of each group and their counterfactual distribution (top left), the total effect difference (top right), structural effect (bottom left) and compositional effect (bottom right) difference between each group and the reference group. The counterfactual distribution and decomposition are derived with respect to the reference group which are mothers aged 12 to 23 (the lowest quintile). Decomposition follows the procedure described in 1. Total, structural and compositional effect are defined as in equation (9). Positive difference corresponds to effect mitigation with increasing age, whereas negative difference corresponds to effect amplification with increasing age respectively. Shaded areas show 95% confidence intervals.

Structural Effect of Parity

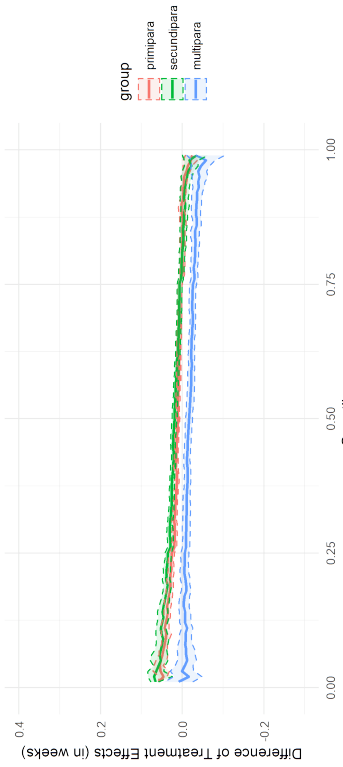


Figure 10: Gestation Length - Structural Effect of Parity

Structural Effect of Weight Gain

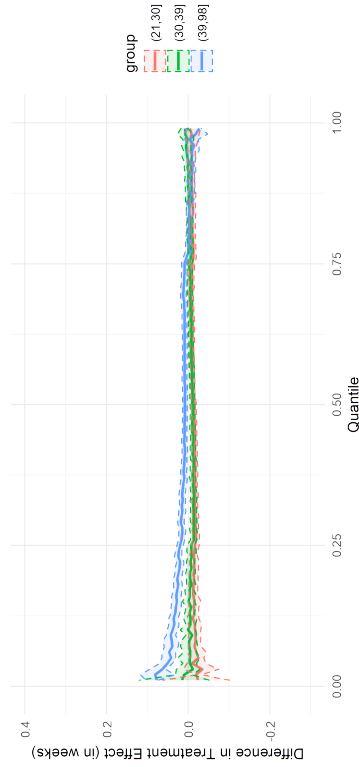


Figure 12: Gestation Length - Structural Effect of Weight Gain

Note: The figure shows the structural effect derived by decomposing the effect of maternal smoking during pregnancy on gestation length. Decomposition follows the procedure described in 1. Positive difference corresponds to effect mitigation, whereas negative difference corresponds to effect amplification respectively. Shaded areas show 95% confidence intervals.

Structural Effect of BMI

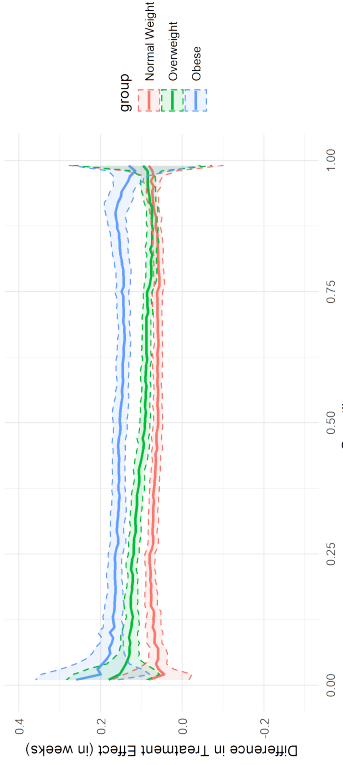


Figure 11: Gestation Length - Structural Effect of BMI

Structural Effect of Weight Gain

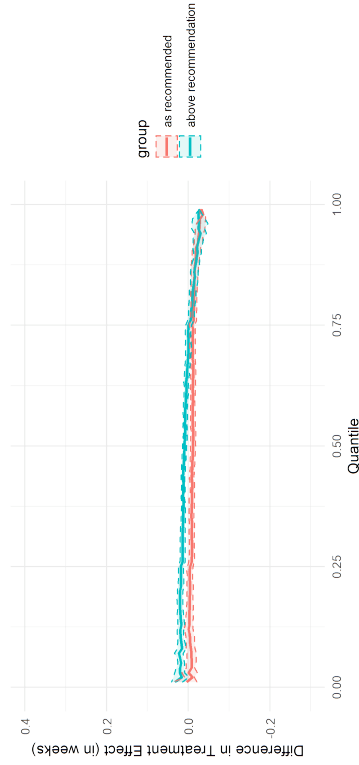


Figure 13: Gestation Length - Structural Effect of Weight Gain Recommendations

Note: The figure shows the structural effect derived by decomposing the effect of maternal smoking during pregnancy on gestation length. Decomposition follows the procedure described in 1. Positive difference corresponds to effect mitigation, whereas negative difference corresponds to effect amplification respectively. Shaded areas show 95% confidence intervals.

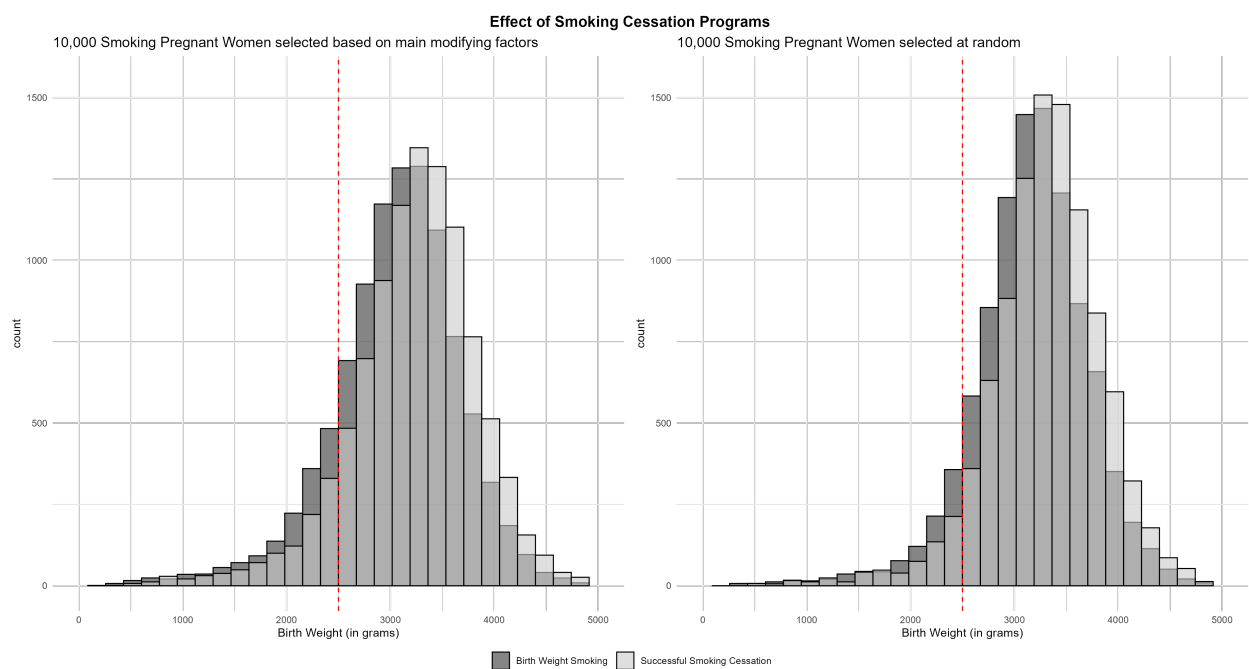


Figure 14: Birth Weight Distribution - Smoking Cessation

Note: Histogram of observed birth weight and estimated birth weight after smoking cessation for 10,000 smokers selected for smoking cessation program. The red line indicates LBW (2500 grams).