# Heterogeneous Treatment Effects of Behavioral and Environmental Risk Factors on Infants' Health at Birth: A Causal Machine Learning Approach

Inauguraldissertation zur Erlangung des Doktorgrades der

Wirtschafts- und Sozialwissenschaftlichen Fakultät der Unversität zu Köln

2023

vorgelegt von

Johanna Maria Zenzes

aus

Neuss

Referent: Prof. Dr. Tom Zimmermann
Korreferent: Prof. Dr. Daniel Wiesen
Tag der Promotion: 12.07.2023

*Für meine Eltern.*

## Acknowledgments

I'm extremely grateful to my main advisor Tom Zimmermann, for his invaluable mentorship and support over the last few years. I want to thank him for all the fruitful discussions, and for generously providing expertise and feedback on this thesis. His encouragement to pursue my research ideas and his advice have been invaluable to my growth as a researcher.

I would also like to express my gratitude to my second advisor Daniel Wiesen, who provided invaluable feedback and guidance on this thesis.

The last years would not have been the same without my colleagues at the Data Innovation Lab at AXA and at the Institute of Econometrics and Statistics at the University of Cologne. I am thankful for the great working atmosphere and moral support, which made these last years very special and enjoyable.

Above all, I am deeply indebted to my parents who always supported me unconditionally and encouraged me along the way. And special thanks to Kevin, for always having my back and believing in me.

# Contents

# List of Tables

# List of Figures

# Chapter 1.

# Introduction

Behavioral and environmental risk factors, such as maternal smoking during pregnancy and in-utero exposure to extreme weather events are significant threats to infants' health at birth. It is crucial to understand how these conditions and shocks during pregnancy affect the fetus, as these can have a large impact on health at birth, but also long-lasting effects on later life health outcomes and outcomes like educational attainment and adult earnings[1]. Effects of these risk factors on infant health have been extensively studied using traditional methods (Almond et al., 2005, Cattaneo, 2010, Lien and Evans, 2005, Deschênes et al., 2009, Andalón et al., 2016, Chen et al., 2020b), while mostly neglecting possible heterogeneity in the effects. Causal machine learning has emerged as a promising approach for estimating heterogeneous treatment effects and identifying the causal pathways that drive them (Athey and Imbens, 2016, Athey et al., 2019, Belloni et al., 2013, Chernozhukov et al., 2018b, Wager and Athey, 2018). Understanding sources of heterogeneity is crucial to identify the most vulnerable groups and possible underlying mechanisms, and offers valuable insights for policymakers working to address the health impacts of behavioral or environmental risk factors.

This dissertation consists of three essays that use causal machine learning techniques to study different aspects of behavioral and environmental risk factors that influence health at birth. Chapter 2 uncovers heterogeneity in the effects of maternal smoking during pregnancy on infant health at birth, while chapter 3 studies how smoking during pregnancy can be regulated using smoking bans. Chapter 4 studies the effects of in-utero heat shock exposure on health at birth and the possible heterogeneity in the effect. All of these are single-authored projects. They leverage machine learning and econometrics and apply and adapt recently developed causal estimation strategies to questions regarding the economics of smoking and climate change. They mark a significant contribution to the literature by introducing cutting-edge machine learning techniques to the economics of early human capital formation.

In Uncovering Sources of Heterogeneity in the Effects of Maternal Smoking on Infants' Health at Birth I study drivers of heterogeneity in the effects of maternal smoking during pregnancy. Maternal smoking during pregnancy is a substantial threat to infants' health at birth but beyond averages, its effect is not well understood. I study how the effects of maternal smoking during pregnancy on infants' health depend on mothers' characteristics,

[1] See Almond and Currie (2011) for a review of the literature.

using a comprehensive dataset of pregnancies in the United States. Using recent advances in the intersection of machine learning and econometrics, I provide a novel and structured framework to identify sources of heterogeneity. To estimate the heterogeneous treatment effects, I use the causal forest, a machine learning based algorithm. But from the estimation itself, the detection of driving factors of heterogeneity remains unclear. Therefore, I propose a novel decomposition approach that makes use of counterfactual distributions. This way, I can isolate differences in effects that are only driven by one single variable, while keeping other characteristics comparable. I find that especially increased mother's age is a robust amplification factor for the effect of smoking on health at birth and it can explain up to 75 grams of birth weight difference when comparing mothers younger than 23 to those that are 34 and older.

Apart from understanding the consequences of smoking on birth outcomes, it is crucial to get a better understanding of why women smoke during pregnancy and how this can be regulated. Public smoking bans are one instrument governments use to regulate smoking in public. In the third chapter "Effect of Smoking Bans on Smoking during pregnancy: Evidence from Germany", I investigate the consequences of a smoking ban introduction on the smoking behavior of pregnant women, which took place starting August 2007 – 2008 in all federal states of Germany. I exploit staggered implementation of state-level smoking ban legislation (differences over time and across states) using data on all births that occurred in German hospitals between 2004 and 2016. I estimate the effect of smoking bans on average cigarette consumption and smoking rate among pregnant women using a difference-in-differences approach. The introduction of smoking bans has a small but significant decreasing effect on the average number of cigarettes smoked by pregnant women (-0.3 daily cigarettes), whereas it does not affect the smoking rate. Considering regional differences in smoking ban implementation, especially strict smoking bans have strong effects on decreasing smoking intensity, however, partial smoking bans are less effective.

The last chapter "Effect of Temperature and Weather Shocks on Health at Birth: Evidence from the US" deals with the effect of in-utero exposure to extreme weather events on health at birth. Understanding in-utero exposure to extreme weather events is key to mitigating climate change's impact on health at birth. Using detailed historic weather records and data on infants born in the US between 1989-2004, I investigate how in-utero exposure to weather events, such as heat and cold waves or rainfall, impacts infants' health at birth. I focus on the effects of heat shocks on birth outcomes and systematically investigate heterogeneity therein using the causal forest, a recently developed causal machine learning technique. Exposure to a heat shock significantly reduces birth weight by around 6 grams on average and increases the small for gestational age (SGA) birth rate. There is substantial heterogeneity in the effect of heat shock exposure on birth weight. Especially infants born to black, Mexican, or low-educated mothers are disproportionately prone to health risks from extreme heat exposure.

# Chapter 2.

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

## 2.1. Introduction

Maternal smoking during pregnancy is strongly associated with birth weight reduction as well as fetal growth restriction, causing high neonatal health care costs. Mothers who smoke are at high risk of preterm delivery, stillbirth[1], and low birth weight, which are leading causes of death, disability, and disease among newborns (e.g. Almond et al., 2005, Baba et al., 2013a, Vogler and Kozlowski, 2002). One in five babies born to mothers who smoke during pregnancy has low birth weight (LBW) (U.S. Department of Health and Human Services, 2010), making smoking the main modifiable risk factor for LBW in the US and other developed countries (Almond et al., 2005). The neonatal costs attributable to smoking in the US are estimated to be almost $367,000,000 (in year 1996 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. The literature on factors influencing the effect of maternal smoking on birth weight is not consistent and studies oftentimes report contradictory results. Understanding sources of heterogeneity is crucial for personalized care and targeted support in smoking cessation for those mothers threatening the health of their newborn most. This would not only contribute to enhancing infants' health but also lead to savings of health care costs (Adams et al., 2002, Almond et al., 2005, Schwartz, 1989).

In this paper, we are identifying key factors of heterogeneity in the effect of smoking on infants' health at birth in a structured way by combining machine learning and econometrics. By decomposing the distribution of conditional average treatment effects (CATE), we can isolate driving factors of heterogeneity. This decomposition is making use of counterfactual distributions (Chernozhukov et al., 2013). We decompose differences in the distribution of CATE for groups into structural and compositional effects. The structural effect reveals the

[1]For a glossary of medical terminology used, see Appendix A.1.

effect differences solely associated with the variable of interest, whereas the compositional effect captures treatment effect differences based on differing group characteristics. This way we can learn about the observable factors that are strongly associated with the heterogeneity and offer a structured identification of effects. The decomposition helps in gaining insights into the estimated CATE function, which might be complex and hard to describe. Methods to estimate CATE based on machine learning offer flexibility in predicting heterogeneous effects but underlying sources of heterogeneity and driving factors for heterogeneity remain hidden. For estimation of the effect of smoking on birth weight conditional on the mother's characteristics, we use a causal forest (Athey et al., 2019, Wager and Athey, 2018).

In a comprehensive data set of infants born in the US we find strong modifying effects for mother's age on the effect of smoking on birth weight, Apgar score[2] and gestation length. Increased mother's age amplifies the effect of smoking on health at birth. It can explain up to 75 grams (when comparing mothers aged 23 or younger and mothers older than 34) of difference in treatment effects. For increased prepregnancy BMI and weight gain, we find no clear effect on health at birth. There are mitigating effects on birth weight, thus being overweight or obese and increased calorie intake can alleviate up to 50 grams of 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 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, we 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., 2018b, Wager and Athey, 2018), the estimated treatment effect function depending on individual characteristics may be very complex and hard to describe[3]. The proposed decomposition allows us to isolate the effect of a change in a single variable while keeping other characteristics comparable. To do so, we 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

[2]The Apgar score is a measure to quickly judge the health condition of the newborn right after birth. It is routinely measured 1 and 5 minutes after birth, for newborns with low scores, the measurement may be continued thereafter. It is derived by assessing the newborn on five simple criteria (appearance, pulse, grimace, activity, respiration), each evaluated on a scale from 0 to 2. The final Apgar score is the sum of the five criteria, ranging from 0 to 10. A newborn with an Apgar score of $7-10$ is considered healthy, a score of $4-6$ is considered moderately abnormal, whereas a score of $0-3$ is low (American Academy of Pediatrics, 2015).

[3]Despite the effectiveness and flexibility of machine learning approaches, very little application can be found. An application of the generic machine learning approach looks at heterogeneity in the effect of fine particulate matter exposure on life expectancy (Deryugina et al., 2019). An early application of causal forests is to predict treatment heterogeneity of summer jobs by Davis and Heller (2017). There are other applications in corporate finance (Gulen et al., 2020), environmental economics to estimate heterogeneity of environmental policy changes (Miller, 2020), and also application in health, where causal forests are used to examine regional differences in diabetes within Europe (Elek and Bíró, 2021) or heterogeneous policy impacts of health insurance reforms (Kreif et al., 2022).

population. 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 us to keep characteristics comparable in our decomposition. In contrast to Chernozhukov et al. (2013), we do not decompose an observed outcome distribution[4], but decompose the distribution of CATEs in order to find observables, for which large structural differences in treatment effects exist. We decompose group differences in the CATE distribution into structural and compositional effects. The structural effect is our main effect of interest, as it captures the effect solely attributable to a single variable while keeping all characteristics comparable.

Despite the popularity of birth weight as a proxy for infants' health at birth, we want to be able to capture additional aspects of health at birth by analyzing possible heterogeneity in the effect of smoking on the 5-minute Apgar score. The Apgar score is a simple measure to quickly evaluate infants' health right after birth. It is intended to predict neonatal survival, evaluate infants' condition immediately after birth, and determine their need for resuscitation (American College of Obstetricians and Gynecologists., 2015, Hegyi et al., 1998). Since the Apgar score is jointly evaluating heart rate, respiratory effort, muscle tone, response to stimulation, and skin coloration of the newborn, it measures indicators of health at birth that we cannot capture by birth weight. The possible modifying factors that we analyze mainly stem from the medical literature. The literature focuses on mother's age, number of previous births, prepregnancy BMI, and sex of the newborn. By using modifying factors from the medical literature, we want to ensure medically meaningful results. However, we also move beyond these factors and find modification by weight gain, but cannot find effects for other characteristics, such as race or education.

Considering the 5-minute Apgar score, the decomposition reveals that the mitigating effects of excessive weight gain and obesity are not persistent. While the decomposition for the mother's age shows similar patterns as for birth weight, increased number of children born to the mother in the past now shows weak mitigating effects instead of amplification effects. Increased mother's age amplifies the effect of smoking on the Apgar score and birth weight, even though the effect difference explained by the mother's age is smaller regarding the Apgar score. In both cases, the age difference is especially profound between mothers younger than 27 and those older than 27. For the weight-related characteristics, such as BMI and weight gain, we cannot find any positive effect of increased weight or calorie intake on the Apgar score. Excessive weight gain and an increased prepregnancy BMI can alleviate the harmful effects of maternal smoking regarding birth weight, but not regarding the Apgar score. This suggests that birth weight increased through overweight and excessive weight gain does not result in better infant health.

Our analysis implies that improved allocation of enhanced smoking cessation based on

[4]Ideas 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 key factors of heterogeneity has the potential to generate huge healthcare expenditure savings. In a scenario, where 10,000 smoking pregnant women can be assigned to intensified assistance with smoking cessation an assignment based on the drivers of heterogeneity could increase savings by nearly 80% compared to random targeting. This proves that our results can be used to identify mothers most at risk to harm their babies. In case these mothers would stop smoking, this could lead to large cost savings and improved infants' health.

There are, however, important caveats to the estimation of the effect of smoking on our outcomes of interest and in uncovering the heterogeneity therein. We address these in several additional robustness checks. First, we analyze heterogeneity in the effects of smoking on gestation length and non-standardized birth weight to capture additional aspects of health at birth. Results for gestational age are very similar to the ones recorded for Apgar score and standardized birth weight. Second, we try to capture the dose dependency of smoking in the heterogeneous treatment effects by rerunning the analysis for heavy smokers only. We now find larger overall treatment effects, but drivers of heterogeneity remain similar to the main analysis. Third, we evaluate whether effect modification is still apparent at the lower end of the birth weight distribution (birth weight below 2800 grams), where adverse health effects are concentrated. Effect size decreases but modifying factors remain, even though the magnitude of effect modification decreases. Additionally, we want to overcome possible concerns regarding underreporting of smoking during pregnancy due to stigmatization, by looking at the effect of pre-pregnancy smoking instead. Lastly, we address possible concerns of limited overlap in the characteristics of smokers and non-smokers by asymmetrically trimming the propensity score as proposed by Stürmer et al. (2010). This way we ensure a larger overlap in characteristics of individuals considered for treatment effect estimation. The 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 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 an outcome of interest (e.g., England et al., 2001, Almond et al., 2005, Lien and Evans, 2005, Cattaneo, 2010, Bharadwaj et al., 2014)[5]. Possible heterogeneity in the effect of smoking is typically neglected in this stream of literature. Only a few studies in the medical literature have examined the possibility that the effect of maternal smoking depends on the underlying characteristics of the mothers (e.g., Cnattingius et al., 1985, Haworth et al., 1980a, La Merrill et al., 2011, Misra et al., 2005, Spinillo et al., 1994b). These mainly focus on birth weight or small for gestational age births as outcomes of interest[6]. Typically, these studies consider small samples, and their estimates provide unclear causal interpretations, suggesting a need for improvement. This paper thus contributes to, first, overcoming problems of existing studies and providing insights into

[5]Figure A.1 provides an overview of treatment effect estimates for the effect of maternal smoking during pregnancy on birth weight in the literature.

[6]See table A.3 for an overview of heterogeneity in the medical literature.

drivers of heterogeneity. Leveraging advances in the machine learning literature, we are able to provide causal interpretations based on a much more comprehensive data set than previous studies. Second, we go beyond the effect of smoking on birth weight and widen the range of possible proxies for health at birth. By also looking at outcome measures, we are able to capture additional details of the effect of smoking on newborn health. Third, we introduce new, cutting-edge machine learning techniques to the economics of smoking and the economics of early human capital formation.

This paper also contributes to the emerging literature of applying causal machine learning to health-related questions (Deryugina et al., 2019, Elek and Bíró, 2021, Kreif et al., 2022), however focusing more on methodological advances to make results of machine learning methods interpretable. We try to describe 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 identification of CATE, albeit the interpretation of driving factors of heterogeneity is not always clear. Therefore, we propose using a decomposition technique to uncover driving factors of heterogeneity.

In the next section, we provide background on smoking and its effect on birth outcomes and related literature regarding heterogeneity from medical research. Next, we describe the data used in the study and describe the method, where we clarify the set-up and discuss theoretical details of the procedure and estimators used. Then we present the empirical results and end with a discussion of implications of the findings.

## 2.2. Background

### 2.2.1. Smoking and Health at Birth

The most commonly used proxy measure for health at birth in economics is birth weight[7], as low birth weight is an indicator for poor health at birth (Almond et al., 2005, Currie and Schwandt, 2013). We want to capture health at birth beyond birth weight, which only reflects a limited weight related aspect of health at birth. To do so, we will additionally use the Apgar Score as an outcome of interest. In the robustness section, we also look at the effect of maternal smoking on gestation length.

Poor health at birth has important (economic) short- and long-term implications. Short term implications include higher infant mortality, poor childhood health and higher medical care costs (Almond et al., 2005, Lightwood et al., 1999). Medical care costs for children born at low birth weight, which is birth weight below 2500 grams, exceed those of children born at normal weight. They account for only 9% of hospital case load in the US, but for 57%

[7]The ability of birth weight as a proxy measure for health at birth has been questioned by Conti et al. (2020). They find that birth weight mainly proxies abdominal circumference. A fetus with larger abdominal circumference usually has higher birth weight, but also shorter lengths of gestation and lower Apgar scores. Thus, birth weight is capturing both positive and negative aspects of fetal health.

of costs for neonatal hospital care (Schwartz, 1989). Thus, the costs that low birth weight poses on health care systems are huge. Considering long term implications, health at birth is predictive of educational attainment and adult earnings. Additionally, it is predictive of adult health and intellectual and social development (Black et al., 2007, Currie and Almond, 2011, Conley and Bennett, 2000).

Health at birth heavily depends on genetic endowment but there are many modifiable factors that influence health at birth significantly. A major risk factor for poor health at birth is smoking. In the US, smoking is the leading modifiable risk factor for LBW (Almond et al., 2005), and it is a public health goal to reduce smoking during pregnancy. Cigarettes and other tobacco products contain several substances, such as nicotine, which can be harmful to the fetus. During pregnancy, the developing fetus is exposed to higher nicotine levels than is the smoking mother (Luck et al., 1985). Harm of maternal smoking increases with later gestational age at exposure, the third trimester being most sensitive (Cohen et al., 2005). In utero exposure to tobacco products is responsible for several complications of pregnancy and birth (Cnattingius, 2004), but also causes long term complications. Smoking is found strongly associated with birth weight reduction and fetal growth retardation (Almond et al., 2005, Baba et al., 2013b, England et al., 2001, Harrod et al., 2014), even after controlling for gestational age. The negative effect of smoking on birth weight remains after controlling for genetic factors. Using a large-sample of sibling pairs for which the mother smoked during one pregnancy but not the other, Currie et al. (2009) show that the risk of LBW is only increased for the pregnancy in which the mother smoked. Almond et al. (2005) use a twin based approach and Knopik et al. (2016) controls for genetic and familial confounding factors and document similar results. Additionally, smoking increases the risk of early pregnancy loss, stillbirth, preterm birth and infant mortality (Baba et al., 2013a, Cohen et al., 2005, Holbrook, 2016).

While the effect of smoking on birth weight and gestation length are well established, the effect of smoking on Apgar score is unclear. Garn et al. (1981) find that the proportion of low Apgar scores is related to maternal smoking during pregnancy, Thorngren-Jerneck and Herbst (2001) state that smoking is a significant risk factor for Apgar score below 7. Contrary, Straube et al. (2010), Gilman et al. (2008) find no significant association between smoking and Apgar score.

### 2.2.2. Heterogeneity in the Smoking Effect

Even though the effect of smoking on several outcomes is well established, only few studies have examined the possibility that mothers' characteristics influence the effects of maternal smoking. Those studies focus on birth weight, birth weight standardized for gestational age, preterm birth, and small for gestational age birth as outcomes of interest, Apgar score is mainly not considered. Table A.3 in Appendix A.9 gives a detailed overview of studies focusing on heterogeneity in the effect of smoking on birth weight and other outcomes.

The literature on factors influencing the effect of maternal smoking on birth weight is not consistent and studies oftentimes report contradictory results. Spinillo et al. (1994a) analyze several factors possibly potentiating effect of smoking, suggesting that nulliparous mothers at lower ages are most in danger to harm their fetus by smoking. This is contradictory to the results of other studies on maternal age and parity, where stronger effects of smoking are found for multiparous, older mothers (see for example Cnattingius, 2004, Misra et al., 2005). Unreliable or contradictory findings mostly arise from small sample sizes or imbalanced samples. Additionally, most studies do not have clear causal identification, and some only look at correlations or perform mean comparisons. Further, these studies only partially control for mothers' characteristics. See Appendix A.9 for an overview on sample size, methods and controls considered in the studies.

Of the factors studied to amplify the risk of strong effects of maternal smoking on birth weight, maternal age is studied best. There are several studies looking at the effect of maternal smoking in different age groups, most finding that advancing maternal age is increasing the effect of smoking on birth weight or the risk of small for gestational age birth (SGA). Even though maternal age in itself has no effect on birth weight (Cnattingius et al., 1985), studies like Cnattingius (1997, 1989), Wu Wen et al. (1990) suggest that smoking actually influences fetal growth more among older smokers. The relative risk of small for gestational age births for smoker versus non-smoker is 3.4 among women between 40-44, and only 1.9 among women aged 15-19 (Cnattingius, 1989). Wu Wen et al. (1990) find similar results when comparing smokers and non-smokers older than 35 and younger than 17. Birth weight reduction for young women is found to be 134 grams, whereas smoking reduces birth weight by 301 grams for older women.

Also, parity is influencing the effect of smoking on birth weight and impaired fetal growth. When considering the interaction with smoking, Cnattingius et al. (1993) document an interaction of parity and maternal age. They find parous smokers being at an especially high risk for low birth weight and preterm delivery, even though the age effect was greater among nulliparas than multiparas. Similarly, Misra et al. (2005) report increasing risks of smoking related birth weight reduction for multiparas, even though no age effect can be found. However, Nabet et al. (2007) argues, that the effect of smoking (on preterm delivery) might only seem stronger among multipara than among primipara, since there is a higher rate of gestational hypertension in primipara than multipara, with smoking tending to decrease the risk of gestational hypertension.

For non-smokers, increased weight gain during pregnancy increases the birth weight of the infant (Ludwig and Currie, 2010). But smoking is found to reduce appetite, thus leading to decreased maternal weight gain during pregnancy (Muscati et al., 1996, D'Souza et al., 1981). Regardless of these correlations, there seems to be no evidence, that the effect of maternal smoking is affected by maternal weight gain. Studies by Muscati et al. (1996), D'Souza et al. (1981), Haworth et al. (1980a) suggest, that the negative effect of smoking on fetal growth

retardation and birth weight is not caused or mediated by nutritional deficiencies or smaller energy intake.

It is a matter of debate, whether high prepregnancy BMI can help to mitigate strong adverse outcomes of smoking during pregnancy. A study by Haworth et al. (1980b) focusing on obesity argues that maternal smoking and maternal obesity act independently of each other. This suggests that maternal overweight does not protect the fetus from growth retardation. However, a more recent study by La Merrill et al. (2011) suggests that the effect of smoking on birth weight and SGA is markedly reduced among obese and overweight women. Similarly, Zhang and Yang (2019) find that prepregnancy BMI is a biologically plausible mediating factor for adverse effect of smoking on birth outcomes.

The 'fragile male' hypothesis suggests, that female fetuses are more robust than their male counterparts. However, it is unclear, whether gender of the fetus affects birth weight reduction or growth retardation due to smoking. Zarén et al. (2000), Spinillo et al. (1994a) and Varvarigou et al. (2009) suggest that growth retardation is significantly stronger for male fetuses, thus suggesting presence of sex dimorphism. However, studies like Suzuki et al. (2011) do not find gender differences in birth weight reduction.

## 2.3. Data

For this study we use the annual natality micro data by the National Center for Health Statistics (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 such as race, age, educational attainment, marital status, place of residence, childbearing history, prenatal care, and information on medical risk factors. It also contains information on the father and detailed health information on the infant such as sex, gestational age, birth weight, Apgar score, and plurality. Starting from 1989, the data contains self-reported information on the mother's smoking behavior. We classify every woman as smoker, who reported to have smoked at least one cigarette per week in any of the three trimesters of pregnancy.

As the smoking indicator is self-reported, underreporting might be an issue. After the revision of the birth certificate in 2003, numerous states observed an increase in self-reported smoking prevalence, suggesting that the revised birth certificate question captures more smoking mothers (Curtin and Matthews, 2016). But there is evidence, that underreporting is still an issue, even after revision of the birth certificate (Howland et al., 2015, Searles Nielsen et al., 2014). This mainly concerns women with high education, who might be ashamed admitting to smoking before and during pregnancy.

The main outcome of interest in this study is health at birth, which we cannot measure directly. Hence, we use proxy measures for health at birth. A widely used proxy measure for

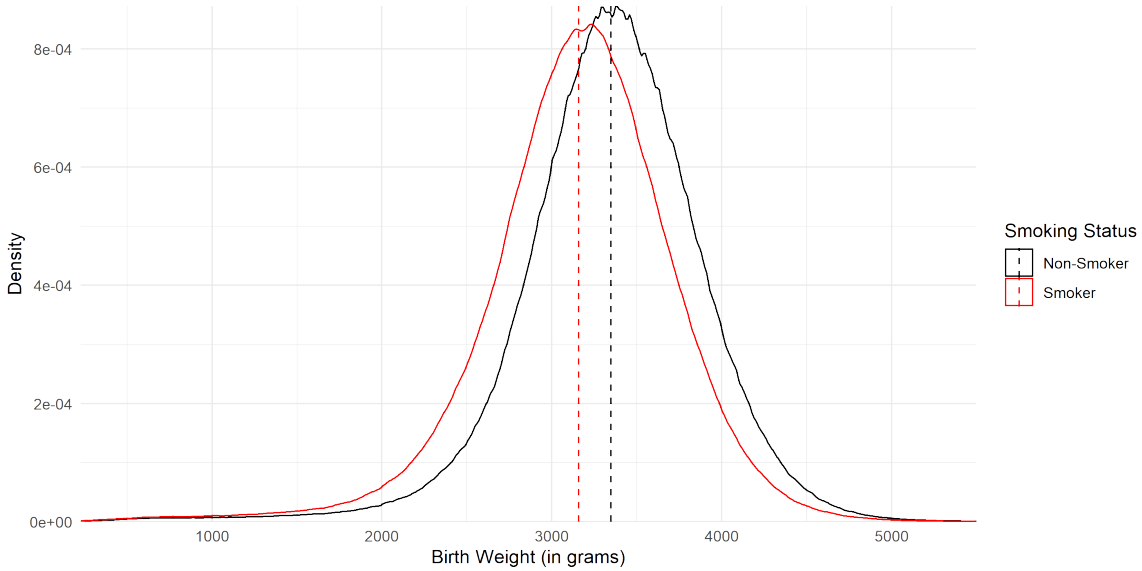


Figure 2.1.: Density of Birth Weight for Smokers and Non-Smokers for 2011-2018
*Note:* Solid lines show density estimates of birth weight for smokers and non-smokers for births between 2011 and 2018. The dashed lines indicate the mean birth weight of each group. Smokers (smoking status 1) are defined as mothers who reported to have smoked at least one cigarette in any of the three trimester of pregnancy.

health at birth is birth weight. Especially low birth weight is strongly associated with adverse health outcomes. Hence, birth weight is most predictive of infant health at the tails of the distribution, which both have different adverse consequences for infant health. Low birth weight is an important predictor of low infant health, similarly extremely high birth weight leads to adverse birth outcomes such as labor problems, diabetes, or metabolic syndrome (Ju et al., 2009). Further, there exist large differences in birth weight by gestational age, which can also be influenced by smoking. To address possible concerns regarding birth weight as a measure for newborn health, we standardize birth weight for gestational age and sex of newborn. Additionally, we examine possible heterogeneity in the effect of smoking on 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). But since it is evaluated by doctors and nurses, reported measures might be biased. Further, Apgar Score scale ranges from 0 to 10, which makes detectable differences caused by smoking very small. None of the used proxy measures is a perfect proxy for health at birth, but evaluating a wide range of outcomes enables us to capture several aspects of health at birth.

Following Cnattingius (1989), standardized birth weight is calculated as follows: First, we calculate mean reference birth weight and its standard deviation for given gestation length and sex. The reference population contains singletons, non-smoker, and pregnancies without complications. We now standardize each individual birth weight by subtracting the reference

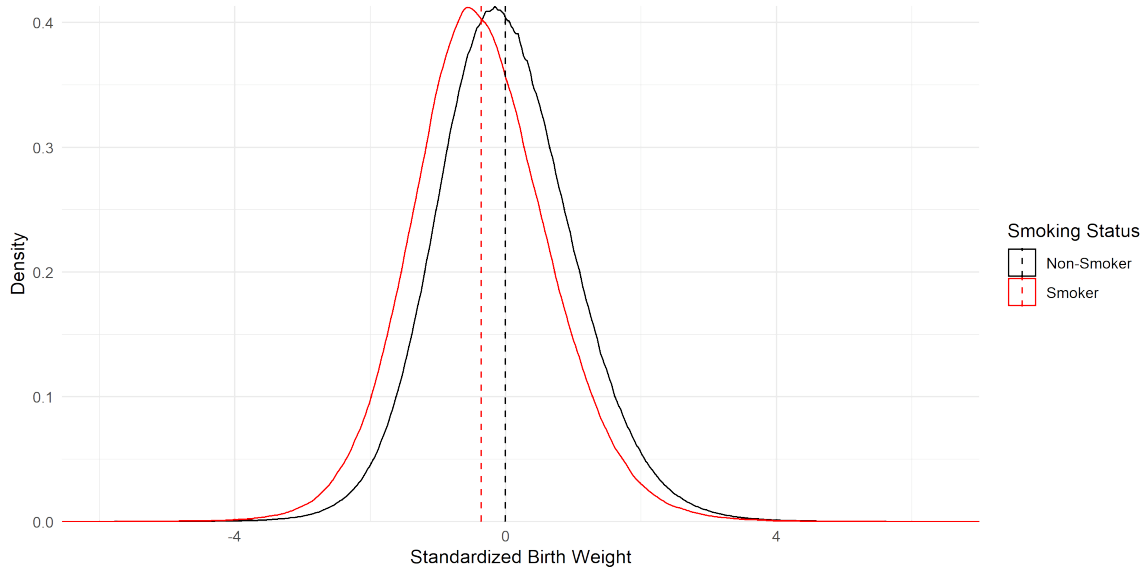


Figure 2.2.: Density of Standardized Birth Weight for Smokers and Non-Smokers for 2011-2018

*Note:* Solid lines show density estimates of standardized birth weight for smokers and non-smokers for births between 2011 and 2018. The dashed lines indicate the mean birth weight of each group. Smokers (smoking status 1) are defined as mothers who reported to have smoked at least one cigarette in any of the three trimester of pregnancy. Standardized birth weight is calculated by subtracting mean reference birth weight and dividing by its standard deviation for given gestation length and sex. Reference birth weight and standard deviation are calculated from non-smoker, singleton pregnancies without complications.

weight given individuals gestation length and sex and dividing by the corresponding standard deviation. The standardized birth weight now fully controls for gestation effects and possible gender effects.

For this analysis, we restrict ourselves to singletons born in the US. This is because multiple pregnancies are generally associated with significant medical risks and complications for the mother and children (Kogan et al., 2000). Multiples are usually born at lower birth weight and at lower gestational age than singletons. Further, prenatal care for multiple pregnancies will be more intensive than for singletons. Additionally, we discard all observations with missing information in the variables of interest.

We have a wide range of possible variables available. All variables, that might be post pregnancy related outcomes (i.e., complication during birth) need to be discarded, as well as variables that can be an outcome of smoking during pregnancy themselves, i.e., gestation length. Thus, we restrict ourselves to general demographic information of the mother and father, as well as her pregnancy history, if applicable[8]. Further, we consider information on

[8]Some variables, such as education, Hispanic origin, and race, are recoded to different levels than in the original birth certificate. Since the reported levels of father's race changed over the seven years of interest, we recoded it to match the format of mother's race. We further combine Cuban, Puerto Rican and other Central and South American Hispanic origin of both, mother, and father. Additionally, we combine several groups of educational attainment. We combine educational attainment of 8th grade or less and

prenatal care, medical and other risk factors, infections and characteristics of the newborn[9]. The resulting sample contains observations on 17,398,750 pregnancies which occurred in the US between 2011-2018.

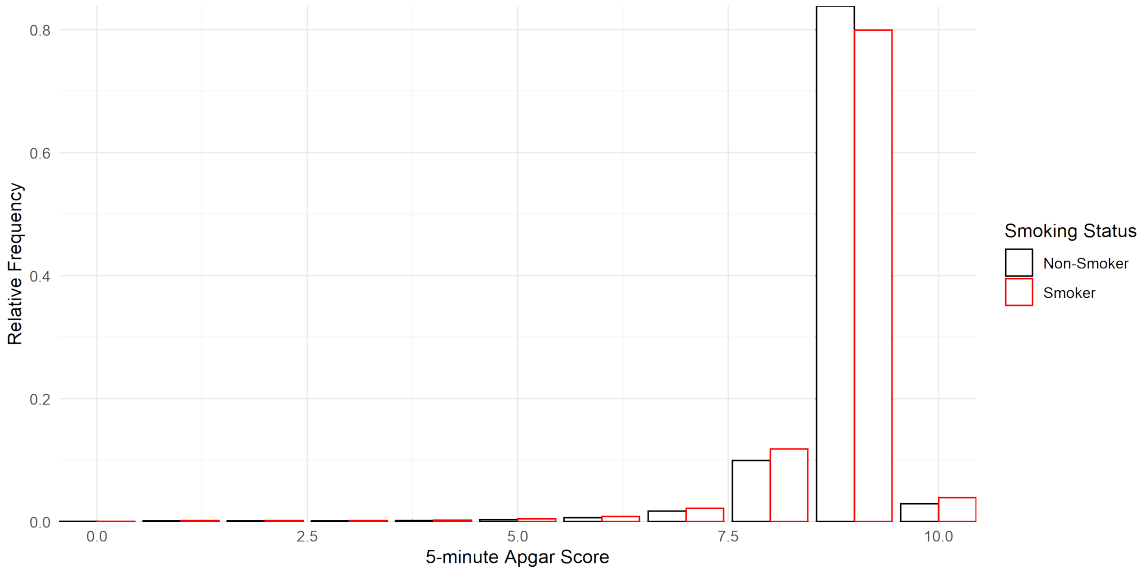


Figure 2.3.: Relative Frequency of 5-minute Apgar Score for Smokers and Non-Smokers for 2011-2018

*Note:* The barplot shows the relative frequency of different Apgar score values for smokers and non-smokers for births between 2011 and 2018. Smokers (smoking status 1) are defined as mothers who reported to have smoked at least one cigarette in any of the three trimester of pregnancy. Apgar score of 10 corresponds to perfect health, whereas 0 corresponds to bad health.

The density plot in Figure 2.1 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. Figure 2.2 shows the density plot for birth weight standardized for gestational age and sex of newborn. Since the reference group for birth weight standardization are non-smokers with pregnancies without complications, the mean standardized birth weight for non-smokers is zero by construction. Again, there is a clear shift towards lower birth weight for smokers of around $-0.375$ standard deviations. For Apgar score, the mean difference is

9th through 12th grade with no diploma to one group and combine educational attainment that results in a Master's degree or Doctorate. We also discard the small fraction of foreign residents from the sample.

[9] Variables considered: Mother's age, mother's education, mother's race, mother's Hispanic origin, mother's marital status, mother's residence status, smoking status, father's age, father's education, father's race, father's Hispanic origin, acknowledgment of paternity, month prenatal care began, number of prenatal care visits, prior other terminations, total delivery order, live birth order, sex of child, birth weight, 5-minute Apgar score, weight gain during pregnancy, gestation length (in weeks), prepregnancy diabetes, gestational diabetes, prepregnancy hypertension, gestational hypertension, eclampsia, former preterm delivery, infertility treatment, use of fertility drugs, reproductive assistance, number of past cesarean, chlamydia, syphilis, gonorrhea, hepatitis b, hepatitis c, prepregnancy BMI, interval since last live birth, intervals since last other birth, mothers height, month of birth, place of birth, WIC receipt, payment type.

very small, as Figure 2.3 shows. Interestingly, the relative frequency for Apgar score 10 is slightly higher among smokers than non-smokers. Same holds for Apgar score 5,6,7, and 8. The relative frequency for very low Apgar scores is higher among smokers than non-smokers.

Table 2.1.: Summary Statistics for Births in 2011, 2013, 2015, 2018

| | **2011** | | **2013** | | **2015** | | **2018** | |
|---|---|---|---|---|---|---|---|---|
| Variables | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| **Infant Characteristics** | | | | | | | | |
| Birth Weight in grams | 3336.32 | 536.83 | 3340.39 | 538.73 | 3336.99 | 540.95 | 3326.16 | 543.88 |
| Low Birth Weight | 0.05 | 0.22 | 0.05 | 0.22 | 0.05 | 0.23 | 0.06 | 0.23 |
| 5-min. APGAR Score | 8.84 | 0.73 | 8.82 | 0.76 | 8.81 | 0.75 | 8.81 | 0.75 |
| Gestation Length | 38.85 | 2.17 | 38.88 | 2.16 | 38.87 | 2.17 | 38.80 | 2.21 |
| Fraction Male | 0.51 | 0.50 | 0.51 | 0.50 | 0.51 | 0.50 | 0.51 | 0.50 |
| Cesarean | 0.14 | 0.34 | 0.14 | 0.35 | 0.15 | 0.35 | 0.15 | 0.36 |
| Birth Place Hospital | 0.98 | 0.13 | 0.98 | 0.13 | 0.98 | 0.13 | 0.98 | 0.14 |
| **Risk Factors** | | | | | | | | |
| Smoking during Pregnancy | 0.07 | 0.26 | 0.07 | 0.25 | 0.06 | 0.24 | 0.06 | 0.23 |
| Prepregnancy Smoking | 0.10 | 0.30 | 0.09 | 0.29 | 0.08 | 0.28 | 0.07 | 0.26 |
| Cigarettes Before Pregnancy | 1.32 | 5.15 | 1.22 | 4.83 | 1.08 | 4.53 | 0.98 | 4.41 |
| Cigarettes 1st Trimester | 0.75 | 3.55 | 0.70 | 3.34 | 0.62 | 3.15 | 0.58 | 3.08 |
| Cigarettes 2nd Trimester | 0.56 | 2.90 | 0.52 | 2.72 | 0.46 | 2.57 | 0.43 | 2.52 |
| Cigarettes 3rd Trimester | 0.50 | 2.71 | 0.46 | 2.55 | 0.41 | 2.41 | 0.38 | 2.37 |
| Weight Gain in pounds | 30.77 | 14.36 | 30.62 | 14.43 | 30.27 | 14.53 | 29.79 | 14.70 |
| Prenatal Care Visits | 11.56 | 3.69 | 11.59 | 3.71 | 11.58 | 3.84 | 11.58 | 3.87 |
| Month Prenatal Care Began | 2.94 | 1.41 | 2.94 | 1.42 | 2.80 | 1.41 | 2.86 | 1.41 |
| Total Delivery Order | 2.25 | 1.42 | 2.27 | 1.43 | 2.30 | 1.45 | 2.36 | 1.50 |
| Live Birth Order | 1.99 | 1.18 | 2.00 | 1.18 | 2.02 | 1.19 | 2.05 | 1.22 |
| Prepregnancy Diabetes | 0.01 | 0.08 | 0.01 | 0.08 | 0.01 | 0.08 | 0.01 | 0.09 |
| Gestational Diabetes | 0.05 | 0.21 | 0.05 | 0.22 | 0.06 | 0.23 | 0.07 | 0.25 |
| Gestational Hypertension | 0.04 | 0.20 | 0.05 | 0.21 | 0.05 | 0.23 | 0.07 | 0.26 |
| Eclampsia | 0.00 | 0.04 | 0.00 | 0.04 | 0.00 | 0.05 | 0.00 | 0.05 |
| BMI (prepregnancy) | 25.99 | 6.20 | 26.16 | 6.29 | 26.43 | 6.43 | 26.99 | 6.65 |
| Prior other Terminations | 0.27 | 0.71 | 0.28 | 0.72 | 0.29 | 0.73 | 0.32 | 0.78 |
| Previous Preterm Birth | 0.02 | 0.14 | 0.02 | 0.15 | 0.02 | 0.16 | 0.03 | 0.18 |
| Interval since last Livebirth | 47.77 | 35.03 | 48.38 | 35.20 | 48.46 | 35.30 | 47.84 | 35.22 |
| **Mothers Demographic Information** | | | | | | | | |
| Mother's Age | 28.17 | 5.79 | 28.41 | 5.71 | 28.68 | 5.63 | 29.06 | 5.54 |
| Married | 0.70 | 0.46 | 0.69 | 0.46 | 0.69 | 0.46 | 0.70 | 0.46 |
| Race - White | 0.82 | 0.38 | 0.80 | 0.40 | 0.80 | 0.40 | 0.79 | 0.41 |
| Race - Black | 0.10 | 0.30 | 0.11 | 0.31 | 0.12 | 0.32 | 0.14 | 0.34 |
| Race - American Indian / Eskimos | 0.01 | 0.09 | 0.01 | 0.09 | 0.01 | 0.10 | 0.01 | 0.10 |
| Race - Asian / Pacific Islander | 0.07 | 0.26 | 0.08 | 0.26 | 0.08 | 0.27 | 0.07 | 0.25 |
| Hispanic Origin - None | 0.77 | 0.42 | 0.79 | 0.41 | 0.79 | 0.41 | 0.83 | 0.37 |
| Hispanic Origin - Mexico | 0.16 | 0.36 | 0.14 | 0.35 | 0.14 | 0.34 | 0.09 | 0.29 |
| Hispanic Origin - Puerto Rico | 0.01 | 0.10 | 0.01 | 0.11 | 0.01 | 0.11 | 0.01 | 0.12 |
| Hispanic Origin - Cuba | 0.00 | 0.07 | 0.00 | 0.07 | 0.00 | 0.07 | 0.01 | 0.08 |
| Hispanic Orig. - C./S. America | 0.02 | 0.15 | 0.02 | 0.15 | 0.02 | 0.15 | 0.03 | 0.16 |
| Education - up to 12th grade | 0.04 | 0.19 | 0.03 | 0.17 | 0.03 | 0.16 | 0.02 | 0.15 |
| Education - Highschool | 0.10 | 0.31 | 0.09 | 0.29 | 0.08 | 0.27 | 0.07 | 0.25 |
| Education - College without degree | 0.23 | 0.42 | 0.23 | 0.42 | 0.23 | 0.42 | 0.23 | 0.42 |
| Education - Associate degree | 0.21 | 0.41 | 0.21 | 0.41 | 0.21 | 0.41 | 0.20 | 0.40 |
| Education - Bachelor's degree | 0.08 | 0.28 | 0.09 | 0.28 | 0.09 | 0.29 | 0.09 | 0.29 |
| Education - Master's degree and PhD | 0.22 | 0.41 | 0.23 | 0.42 | 0.23 | 0.42 | 0.24 | 0.43 |
| Resident | 0.74 | 0.44 | 0.72 | 0.45 | 0.71 | 0.45 | 0.66 | 0.47 |
| Intrastate Nonresident | 0.24 | 0.43 | 0.26 | 0.44 | 0.26 | 0.44 | 0.31 | 0.46 |
| Interstate Nonresident | 0.02 | 0.14 | 0.02 | 0.14 | 0.02 | 0.15 | 0.03 | 0.16 |
| Foreign Resident | 0.00 | 0.05 | 0.00 | 0.05 | 0.00 | 0.06 | 0.00 | 0.05 |
| Payment - Medicaid | 0.37 | 0.48 | 0.37 | 0.48 | 0.36 | 0.48 | 0.35 | 0.48 |
| Payment - Private Insurance | 0.54 | 0.50 | 0.54 | 0.50 | 0.56 | 0.50 | 0.57 | 0.49 |
| Payment - Self-Pay | 0.04 | 0.19 | 0.04 | 0.20 | 0.04 | 0.19 | 0.04 | 0.20 |
| Number of Observations | 1,886,927 | | 2,050,584 | | 2,439,253 | | 2,077,663 | |

Source: National Center for Health Statistics (2011, 2013, 2015, 2018)

In Table 2.1 sample means and standard deviation of some relevant variables for a collection of the years of interest (2011, 2013, 2015, 2018) are given. Overall, means do not differ a lot between the years and are comparable. The smoking prevalence declined from 7% cigarette smokers in the 2011 sample to 6% in the 2018 sample. Prepregnancy smoking steadily declined from 10% in 2011 to 7% in 2018. The number of daily cigarettes smoked among the whole population before and during pregnancy also declined, which could just be due to the decrease in smoking mothers and does not necessarily need to be caused by an actual decline in daily cigarettes smoked by smoker. Smoking prevalence among all adults in the United States follows a similar pattern. Cigarette smoking prevalence declined from 20.9% in 2005 to 15.1% in 2015 (Jamal et al., 2016) and reached an all-time low of 13.7% in 2018 (Creamer et al., 2019). Interestingly, the decline is sharper for the total adult population than for pregnant women, who not only expose themselves to the harmful effects of cigarettes, but their baby. Birth weight and low-birth-weight rate remain stable, which also holds for the second outcome of interest, 5-minute Apgar score and average gestation length.

Table 2.2 shows an overview of relevant variables, grouped by mother's demographic information, pregnancy related risk factors and infant characteristics for smoker and non-smoker. The sample includes observations from 2011 - 2018. 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 this 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.

## 2.4. Methods

### 2.4.1. Setup

In our setup we observe $N$ mothers that gave birth in the time period observed, indexed by $i = 1, \ldots, N$. For each individual, we observe $(Y_i^{obs}, T_i, X_i)$ for $i = 1, \ldots, N$, where $Y_i^{obs}$ denotes the observed birth outcome of the newborn child, $T_i$ the binary smoking indicator and $X_i$ the vector of mother's characteristics. We define causal effects via the potential outcomes model by Imbens and Rubin (2015). For each observation $i$ we define the outcome under potential treatment as $Y_i^{(T_i)}$, $T_i \in \{0, 1\}$, 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. Never can both potential outcomes be observed together, we can only observe the realized outcome $Y_i^{obs} = T_i Y_i^{(1)} + (1 - T_i) Y_i^{(0)}$. Thus, the "ground truth" for a causal effect can never be observed for any individual.

Table 2.2.: Summary Statistics for Smoker and Non-Smoker for Births between 2011-2018

| | **Non-Smoker** | | **Smoker** | | **Normalized** |
|---|---|---|---|---|---|
| Variables | Mean | SD | Mean | SD | **Difference** |
| **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 |
| Cesarean | 0.14 | 0.35 | 0.16 | 0.36 | 0.03 |
| Birth Place Hospital | 0.98 | 0.13 | 1.00 | 0.06 | 0.10 |
| **Risk Factors** | | | | | |
| Smoking during Pregnancy | 0.00 | 0.00 | 1.00 | 0.00 | - |
| Prepregnancy Smoking | 0.03 | 0.16 | 0.98 | 0.12 | 4.75 |
| 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 |
| Live Birth Order | 2.00 | 1.18 | 2.24 | 1.30 | 0.13 |
| Prepregnancy Diabetes | 0.01 | 0.09 | 0.01 | 0.10 | 0.02 |
| Gestational Diabetes | 0.06 | 0.23 | 0.06 | 0.23 | 0.00 |
| Gestational Hypertension | 0.05 | 0.23 | 0.05 | 0.22 | -0.00 |
| Eclampsia | 0.00 | 0.05 | 0.00 | 0.05 | 0.00 |
| BMI (prepregnancy) | 26.39 | 6.37 | 26.88 | 7.07 | 0.05 |
| Prior other Terminations | 0.28 | 0.72 | 0.45 | 0.98 | 0.14 |
| Previous Preterm Birth | 0.02 | 0.15 | 0.05 | 0.21 | 0.09 |
| Interval since last Livebirth | 47.82 | 34.85 | 52.80 | 39.40 | 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. - Central/South 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 |
| Resident | 0.71 | 0.45 | 0.65 | 0.48 | -0.09 |
| Intrastate Nonresident | 0.26 | 0.44 | 0.32 | 0.47 | 0.09 |
| Interstate Nonresident | 0.02 | 0.15 | 0.03 | 0.16 | 0.01 |
| Foreign Resident | 0.00 | 0.06 | 0.00 | 0.01 | -0.05 |
| 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)

The average treatment effect is defined as $\tau_{ATE} = \mathrm{E}[Y_i^{(1)} - Y_i^{(0)}]$ which is the expected difference between the potential outcomes. The main interest in this analysis are treatment effects which differ across individuals by their characteristics. Therefore, we focus on the conditional average treatment effect, defined as

$$\tau(x) = \mathrm{E}[Y_i^{(1)} - Y_i^{(0)} | X_i = x]. \tag{2.1}$$

To ensure causal interpretation we assume common support (a discussion of the common support assumption can be found in section 2.5)

$$0 < P(T_i = 1|X_i) < 1 \tag{2.2}$$

and unconfoundedness, that is, that the treatment assignment $T_i$ is as good as randomly assigned conditional on the covariates $X_i$:

$$\{Y_i^{(0)}, Y_i^{(1)}\} \perp T_i | X_i, \tag{2.3}$$

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

Unconfoundedness is indeed a very strong assumption. After controlling for the available observables, treatment status needs to be as good as random. We do have a rich dataset available with many factors, that the literature has found to be associated with smoking decision and birth outcomes (Smedberg et al., 2014). But only relying on rich data does not ensure unconfoundedness. While we cannot test for unconfoundedness directly, we can check whether our estimated ATE is in line with the literature. Parts of the literature are able to instrument for smoking (Evans and Ringel, 1999, Lien and Evans, 2005, Bharadwaj et al., 2014), others are able to control for unobserved maternal factors by comparing the same mothers in different pregnancies (Currie et al., 2009). Figure A.1 shows that our estimated average treatment effect is well in line with the literature. Since our ATE is in line with the literature, unconfoundedness is likely to hold.

### 2.4.2. Causal Forest

For estimation of the conditional average treatment effect (4.1), we 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, but adjusting split criterion to be able to cope with unobserved counterfactuals.

The starting point for Athey et al.'s idea behind the causal forest is a partially linear model as in Robinson (1988), in which the treatment effect is considered to be constant $\tau(x) = \tau$:

$$Y_i = g(X_i) + T_i\tau + \epsilon_i. \tag{2.4}$$

Writing $p(x) = \mathrm{E}[T_i|X_i = x]$ for the propensity score and $m(x) = \mathrm{E}[Y_i|X_i = x]$ for the expected outcome marginalizing over treatment, referred to as nuisance parameters, this

model can be rewritten in a centered form, following Robinson (1988):

$$Y_i - m(X_i) = (T_i - p(X_i))\tau + \epsilon_i. \tag{2.5}$$

The causal effect $\tau$ can be estimated by plugging in estimates of $m(X_i)$ and $p(X_i)$ and using a residual on residual regression as

$$\hat{\tau} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}(X_i)\right)\left(T_i - \hat{p}(X_i)\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(T_i - \hat{p}(X_i)\right)^2}, \tag{2.6}$$

which is a semiparametrically efficient estimator of $\tau$ when assuming unconfoundedness (Chernozhukov et al., 2018a, Robinson, 1988). Similarly, in a setting where we allow for treatment effects to differ with characteristics, but assuming that the treatment effect is constant in a sufficiently small neighborhood $N(x)$, the heterogeneous treatment effect can be estimated as (Kreif et al., 2022)

$$\hat{\tau}(x) = \frac{\sum_{\{i:X_i \in N(x)\}}\left(Y_i - \hat{m}(X_i)\right)\left(T_i - \hat{p}(X_i)\right)}{\sum_{\{i:X_i \in N(x)\}}\left(T_i - \hat{p}(X_i)\right)^2}. \tag{2.7}$$

To be able to estimate $\hat{\tau}(x)$ as in (2.7), it is crucial to find the neighborhood $N(x)$. To do so, Athey et al. (2019) propose the generalized random forest, where these neighborhoods are derived as a locally weighted set of neighboring observations for each test point $x$, estimated by an adaptation of the random forest (Breiman, 2001). Random forests are a combination of many full-grown regression trees (Breiman, 2001). Regression trees split the given feature space into non-overlapping rectangular regions, also referred to as leaves. A tree is grown by greedy recursive partitioning, where we recursively choose axis aligned splits, which minimize the prediction error in the sample. For every observation that falls into a certain leaf $L$, a tree returns the mean of all response values of training observation that fall into the same leaf (Hastie et al., 2009). To minimize variance of the estimates, each tree $b = 1, \ldots, n$ is grown on a random subset sampled from the original feature space $\mathcal{S}_b \subseteq 1, \ldots, n$ and averaging over all those trees' predictions leads to the prediction of the random forest for $\mu(x) = \mathrm{E}[Y_i | X_i = x]$:

$$\hat{\mu}(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{Y_i 1(\{X_i \in L_b(x), i \in \mathcal{S}_b\})}{|\{X_i \in L_b(x), i \in \mathcal{S}_b\}|}, \tag{2.8}$$

where $L_b$ is the leaf containing $x$ in tree $b$ (Athey and Wager, 2019). Athey et al. (2019) build on this algorithm, but alternatively think of the random forest as an adaptive kernel method. The representation in (2.8) can equivalently be written as

$$\hat{\mu}(x) = \sum_{i=1}^{n}\alpha_i(x)Y_i, \quad \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{1(\{X_i \in L_b(x), i \in \mathcal{S}_b\})}{|\{X_i \in L_b(x), i \in \mathcal{S}_b\}|}, \tag{2.9}$$

where the weights $\alpha_i(x)$ measures how often the $i$-th observation falls into the same leaf as

$x$.

Thinking of these weights $\alpha_i(x)$ as characterization of the neighborhood $N(x)$, one can modify (2.7) to estimate the CATE function as follows:

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x) \left(Y_i - \hat{m}(X_i)\right) \left(T_i - \hat{p}(X_i)\right)}{\sum_{i=1}^{n} \alpha_i(x) \left(T_i - \hat{p}(X_i)\right)^2}. \tag{2.10}$$

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. 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). Several trees are grown on many subsamples of the data and averaged to build a causal forest, which is then used to derive the weights for each observation $\alpha_i(x)$. The weights measure how often an observation was used to estimate treatment effect at $x$.

### 2.4.3. Effect Decomposition

We are interested in characteristics that are the potentially driving factors for treatment effect heterogeneity. Thus, we want to detect mothers' characteristics for which we can find large differences in treatment effects, while keeping all other characteristics fixed. To do so, we split the sample into non-overlapping populations based on a variable of interest (partitioning characteristic, i.e., sex: male and female population) and decompose the treatment effect for individuals in these populations. We 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. Unlike them, we do not apply the decomposition with focus on changes in the distribution of outcome $Y$, but 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.

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

$$\underbrace{F_{\tau\langle k|k\rangle}(y) - F_{\tau\langle j|j\rangle}(y)}_{\textit{Total Effect}} = \underbrace{\left[F_{\tau\langle k|k\rangle}(y) - F_{\tau\langle j|k\rangle}(y)\right]}_{\textit{Structural Effect}} + \underbrace{\left[F_{\tau\langle j|k\rangle}(y) - F_{\tau\langle j|j\rangle}(y)\right]}_{\textit{Compositional Effect}}. \tag{2.11}$$

$F_{\tau\langle k|k\rangle}(y)$ is the observed distribution function of $\tau(x)$ for population $k$. $F_{\tau\langle j|k\rangle}(y)$ represents the counterfactual distribution function of $\tau$ that would have prevailed for population $k$, if

they faced populations $j$'s distribution of $\tau$. It can also be interpreted as the distribution function of $\tau$ 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. We can also look at the decomposition of quantile effects. Quantile effect decomposition is easier to interpret and directly shows the magnitude of effect differences. The resulting decomposition looks as follows:

$$\underbrace{Q_{\tau\langle k|k\rangle}(p) - Q_{\tau\langle j|j\rangle}(p)}_{\textit{Total Effect}} = \underbrace{[Q_{\tau\langle k|k\rangle}(p) - Q_{\tau\langle j|k\rangle}(p)]}_{\textit{Structural Effect}} + \underbrace{[Q_{\tau\langle j|k\rangle}(p) - Q_{\tau\langle j|j\rangle}(p)]}_{\textit{Compositional Effect}}, \tag{2.12}$$

$$\forall k, j \in \mathcal{K}, p \in (0,1), \tag{2.13}$$

where $Q_{\tau\langle j|k\rangle}(p) := F^{-1}_{\tau\langle j|k\rangle}(p), \;\; p \in (0,1)$ 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 rather constructed 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.

$$F_{\tau\langle j|k\rangle}(y) := \int_{\mathcal{X}_k} F_{\tau_j|X_j}(y|x) dF_{X_k}(x), \tag{2.14}$$

$$Q_{\tau\langle j|k\rangle}(y) := F^{-1}_{\tau\langle j|k\rangle}(p), \;\; p \in (0,1). \tag{2.15}$$

To arrive at the counterfactual distribution in equation (2.14), 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. For ease of notation, we will denote the outcome of interest $Y$, instead of $\tau(x)$ for the following details on estimation. To be able to identify the counterfactual distribution, we consider a finite number of populations $k \in \mathcal{K}$, for which covariates $X_k \in \mathbb{R}^d$ and the outcome $Y_k \in \mathbb{R}$ are observed. Given the observability of covariates $X_k$ and outcome $Y_k$, the covariate distribution $F_{X_k}$, $\forall k \in \mathcal{K}$ and the conditional distribution $F_{Y_j|X_j}$, $\forall j \in \mathcal{K}$ 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 $Y_k$. The counterfactual distribution is well-defined if $\mathcal{X}_k \subseteq \mathcal{X}_j$, $\forall (j,k) \in \mathcal{K}$. For estimation, we assume that observe samples $\{(Y_{ki}, X_{ki}), i = 1, \ldots, n_k\}$. To estimate the counterfactual distribution, one first has to estimate the covariate distribution $\hat{F}_{X_k}(x)$ using the empirical distribution function

$$\hat{F}_{X_k}(x) = \frac{1}{n_k} \sum_{i=1}^{n_k} 1_{\{X_{ki} \leq x\}}, \quad k \in \mathcal{K}. \tag{2.16}$$

For estimating the conditional distribution function $F_{Y_k|X_k}(y|x)$, which describes the stochastic assignment of $Y$ for individuals with characteristics $X_j$ for population $j$, we 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 distribution regression model is defined as

$$F_{Y|X}(y|x) = \Lambda(P(x)'\beta(y)), \quad \forall y \in \mathcal{Y}, \tag{2.17}$$

where $P(x)$ is a vector of transformations of X, $\beta(y)$ is parameter vector that may vary with $y$ and $\Lambda$ 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

$$\hat{F}_{Y_k,X_k} = \Lambda(P(x)'\hat{\beta}_k(y)), \quad (y,x) \in \mathcal{Y}_k, \mathcal{X}_k,\, k \in \mathcal{K}, \tag{2.18}$$

$$\hat{\beta}_k = arg \max_{b \in \mathbb{R}^{d_p}} \sum_{i=1}^{n_k} \left[1_{\{Y_{ki} \leq y\}} \ln\left[\Lambda\left(P(X_{ki})'b\right)\right] + 1_{\{Y_{ki} > y\}} \ln\left[1 - \Lambda\left(P(X_{ki})'b\right)\right]\right], \tag{2.19}$$

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 $\hat{F}_{Y\langle j|k\rangle}(y) = \frac{1}{n_k}\sum_{i=1}^{n_k} \Lambda(P(X_{ki})'\hat{\beta}_j(y))$.

The estimators are then plugged into (2.14), where the conditional distribution function is integrated with respect to the covariate distribution to arrive at $\hat{F}_{\tau\langle j|k\rangle}(y)$. Chernozhukov et al. (2013) also construct confidence for the estimated counterfactual distribution. The construction of the intervals pointwise and uniform confidence intervals over a prespecified set of quantile indexes rely on functional central limit theorems and bootstrap functional central limit theorems for the empirical counterfactual distribution.

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

**Procedure 1:** Effect Decomposition

Repeat multiple times using different random splits into auxiliary and main sample:

- Split the sample into auxiliary (A) and main sample (M)
- On Auxiliary sample: train the causal forest
- 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 (2.16)
    - Obtain estimates $\hat{F}_{\tau(x)^{(0)}|X_0}$ of the conditional distribution via (2.17)
    - Obtain estimates of the counterfactual distributions via (2.14)
    - Obtain decomposition via (2.12):

$$\underbrace{Q_{\tau\langle k|k\rangle}(y) - Q_{\tau\langle 0|0\rangle}(y)}_{\textit{Total Effect}} = \underbrace{[Q_{\tau\langle k|k\rangle}(y) - Q_{\tau\langle 0|k\rangle}(y)]}_{\textit{Structural Effect}} + \underbrace{[Q_{\tau\langle 0|k\rangle}(y) - Q_{\tau\langle 0|0\rangle}(y)]}_{\textit{Compositional Effect}}$$

Other than traditional methods, the causal forest and other machine learning based approaches provide more flexibility and a very structured way of identifying heterogeneity. Using a simple interaction model, one faces the issue of multiple hypothesis testing, since many variables and threshold would need to be tested. This is not the case when using machine learning based methods. Prediction of heterogeneous effects is more precise when using those advanced procedures, as they are able to search over high-dimensional functions of covariates, rather than only searching over a small subgroup by using a limited number of interaction terms (Davis and Heller, 2017). Through this, researchers are able to search for heterogeneity over the whole population and not just based on a limited number of preselected covariates. Additionally, the researcher does not need to impose parametric assumptions about covariate interactions. Further, the causal forest offers statistical tests to test, whether there is significant heterogeneity present that can be explained by observable covariates.

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, we rely on a sample splitting idea by Chernozhukov et al. (2018b), 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, we use many splits and report the median over all results. For confidence intervals derived on the main sample the confidence level needs to be discounted.

## 2.5. Empirical Results

To be able to identify CATE, we assume unconfoundedness (2.3) and overlap (2.2) in the covariate distributions. While we cannot test for unconfoundedness, we can assess overlap by assessing overlap of propensity score estimates of treated and control group. The propensity score $p(x) = \mathrm{E}[T_i|X_i = x]$ is one of the nuisance parameters estimated using a regression forest in the causal forest algorithm `grf`, see section 2.4.2 for details. Figure 4.4 shows propensity estimates for both smoker and non-smokers estimated by a standard regression forest. It shows overlap for the whole range of estimated propensity scores, however especially in the non-smoker group, individuals with a propensity score estimate of larger than 0.4 are rare. Since there might be concerns due to the limited overlap, we also run our analysis after asymmetrically trimming the propensity as proposed in Stürmer et al. (2010). See Section 2.6.4 for details.

Table 2.3 provides an overview of the findings of the empirical analysis for standardized birth weight and Apgar score, the main outcomes of interest. We will report the full decomposition for mother's age, but will only present structural effects for toher modifying factors. For the full decomposition, refer to Appendix A.8. For results on additional outcomes, non standardized birth weight and gestation length, refer to Appendix A.3.1 and A.3.2.

Table 2.3.: Overview: Empirical Results

| | **Decomposition by** | | | | | |
|---|---|---|---|---|---|---|
| **Effect of Smoking on** | Mother's Age | Parity | Prepregnancy BMI | Weight Gain | Weight Gain Recommendation | Sex |
| **Standardized Birth Weight** | ++<br>strong amplifying effects of increased mother's age | +<br>amplifying effects of increased parity | **- -**<br>strong mitigating effects of obesity, only weak for overweight | **-**<br>weak mitigating effects of increased weight gain | **-**<br>weak mitigating effects of gaining above recommendation | **0**<br>no effect difference |
| **Apgar Score** | +<br>amplifying effects of increased mother's age | **-**<br>weak mitigation of increased parity | **0**<br>no modifying effect of increased BMI | +<br>weak amplifying effects of increased weight gain | **0**<br>no modifying effect of weight gain by recommendations | **0**<br>no effect difference |

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

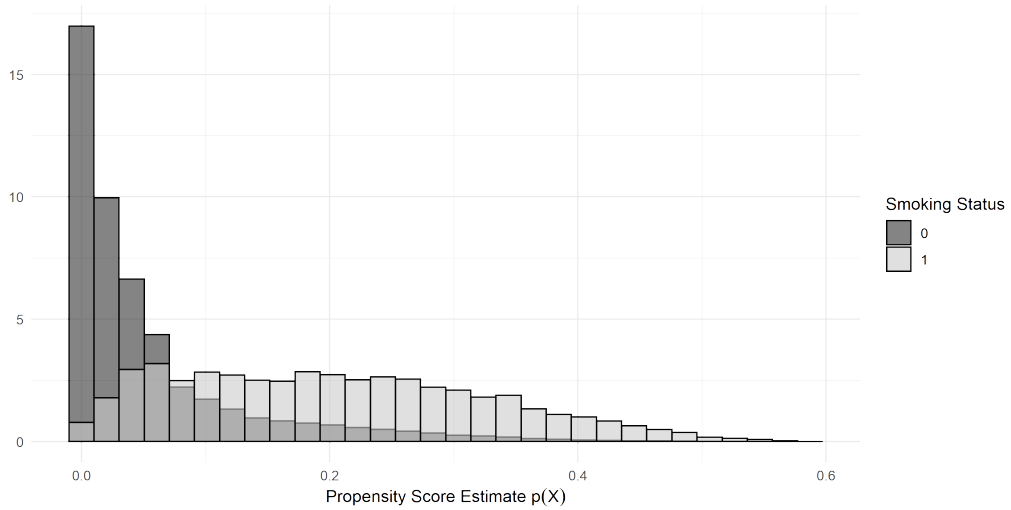


Figure 2.4.: Propensity Score estimates for Smokers and Non-Smokers
*Note:* Histogram of estimated propensity score estimates for smokers (smoking status 0) and non-smokers (smoking status 1). Propensity score is estimated using a regression forest. For propensity score smaller than 0.4 there is large overlap between the two groups. For propensity score estimates above 0.4 overlap is smaller, since untreated individuals are rare.

### 2.5.1. Standardized Birth Weight

Figure 2.5 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 $-0.35$ standard deviations. This roughly correspond to a mean reduction of $-180$ grams, in case we consider the standard deviation in birth weight of the reference group used for standardization. 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 og $-310$ grams. Estimated average treatment effect is clearly negative, but also confidence intervals do not account for the strong heterogeneity present.

The results in figure 2.7 follow the implementation steps as described in procedure 1, repeated 10 times using different random splits into main and auxiliary sample. The differences that we see, always refer to the same baseline group, which is the youngest age group for mother's age. We report confidence bands for $\alpha = 0.05$, meaning we use bootstrap confidence bands for confidence level 0.975 but discount for splitting uncertainty, resulting

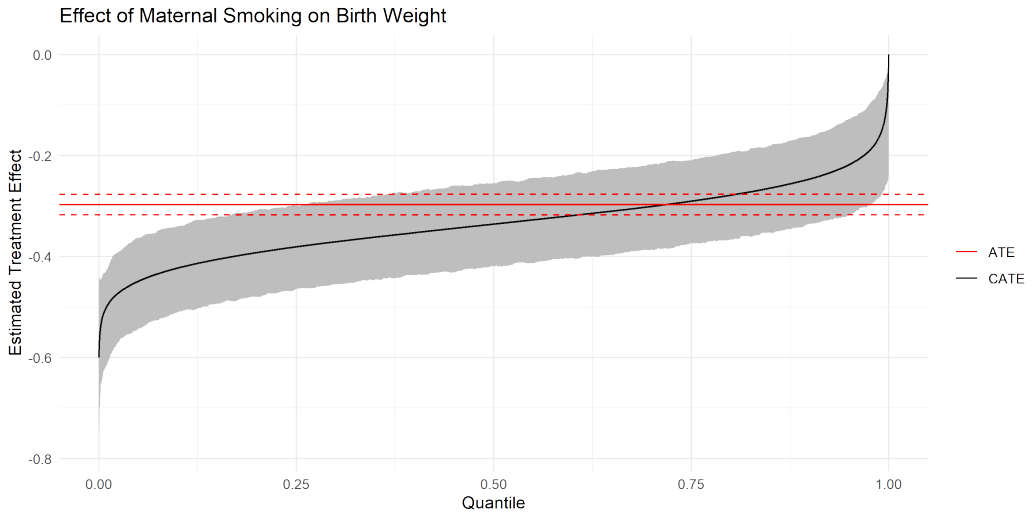


Figure 2.5.: Sorted CATE and ATE estimates of smoking on standardized birth weight
*Note:* The solid black line shows the estimated CATE of smoking on standardized birth weight sorted in ascending order. The dashed black line shows the 95% confidence intervals derived via bootstrapping. CATE is estimated using a causal forest as described in 2.4.2 using the `grf` R-package. The solid red line shows the ATE and corresponding 95% confidence intervals. These are estimated using the `grf` R-package by doubly robust augmented inverse propensity weighting.

in confidence level of 0.95. Figure 2.7 shows the decomposition for mother's age. The $x$-axis of the plot shows the quantile index. Since we are 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 top left plot shows the quantiles of the observed CATE cumulative 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 observed difference between each group's CATE distribution and the CATE distribution of the baseline group. On the bottom left we see the structural effect, which is our main effect of interest. It shows the difference only associated with mother's age. Characteristics of groups are kept comparable by using counterfactual distribution. This difference is always relative to the reference group. The plot on the bottom right shows the compositional effect, which is the difference associated with differences in group characteristics, leaving out decomposition dimension. We 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, we 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 2.8) 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 2.9 shows a strong mitigating effect of obesity on the effect of smoking on standardized birth weight. However, the positive effect of overweight, for example reported by La Merrill et al. (2011), cannot be observed here. 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 we look at the actual weight gain in pounds, but additionally relate weight gain to recommendations based on prepregnancy BMI published by CDC – National Center for Health Statistics (2021). Recommendations 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. We classify mothers into three categories which are 'below recommendation', 'as recommended', and 'above recommendation'.

We see similar results for weight gain in pounds and weight gain recommendations (Figures 2.10 and 2.11). 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.

### 2.5.2. Apgar Score

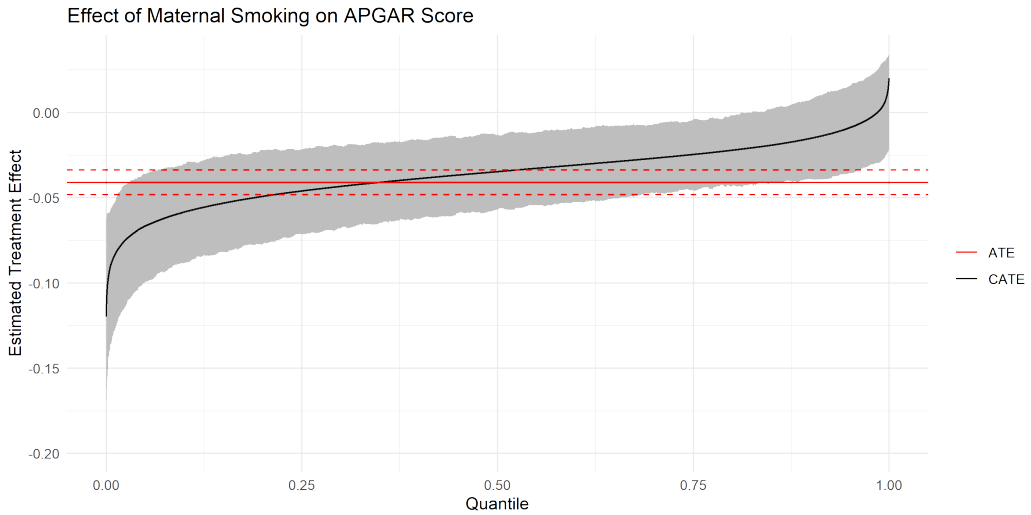


Figure 2.6.: Sorted CATE and ATE estimates of smoking on 5-minute Apgar score
*Note:* The solid black line shows the estimated CATE of smoking on 5-minute Apgar score sorted in ascending order. The dashed black line shows the 95% confidence intervals derived via bootstrapping. CATE is estimated using a causal forest as described in 2.4.2 using the `grf` R-package. The solid red line shows the ATE and corresponding 95% confidence intervals. These are estimated using the `grf` R-package by doubly robust augmented inverse propensity weighting.

Since the Apgar score ranges from 0 to 10, we cannot expect to see large differences in treatment effect for the decomposition of interest. The observable mean difference in Apgar score for smoker and non-smoker is only 0.04, whereas it is 189.72g for birth weight and 0.375 for standardized birth weight for example. The average treatment effect of smoking on Apgar score is $-0.025$, which is quite small. However, Figure 2.6 reveals some heterogeneity in the effect. Further, the figure shows that some women potentially positively affect the Apgar score of their newborn when smoking. This might seem surprising at first, since smoking is known to be harmful. But the positive effect can be explained by the weight reduction caused by smoking. Birth weight proxies for the abdominal circumference of the

fetus, which is negatively correlated with the Apgar score. A large abdominal circumference may be associated with labor dystocia, hence reducing the Apgar score (Conti et al., 2020). Thus, a decrease of birth weight caused by smoking also reduces abdominal circumference, which reduces the risks of labor dystocia and increases Apgar score, especially for those mothers who give birth to very large and heavy children.

For mother's age, we can see a similar pattern as for the effect on standardized birth weight (Figure 2.12). However, the total effect indicates smaller differences than for decomposing the CATE on birth weight standardized. Increased mothers age amplifies the effect of smoking on Apgar score. At the lower quantile indices, it can be responsible for up to a difference of $-0.075$ in Apgar Score when smoking. As for the decomposition for standardized birth weight, this difference fades for the upper quantiles, indicating large within group heterogeneity. The compositional effect captures differences for mothers at the lower quantile end of CATE distribution, but fades towards 0 elsewhere, thus mother's age explains most of the observable total difference.

Figure 2.13 shows no clear effect of parity on the effect of smoking on Apgar score. The structural effect captures both small mitigating and amplifying effects for parity. Increased number of previous pregnancies is mitigating the effect of smoking on birth weight, but only at the lower quartile of the estimated distribution and we only observe a significant difference between nullipara and multipara. For all other quantiles, the effect is close to 0. Thus, the modifying influence of parity on the effect of smoking on Apgar score seems to be mostly non-existent.

The structural effect of the decomposition for Apgar score suggests, that strong weight gain during pregnancy is slightly amplifying the effect of smoking on Apgar score. The effect difference is very small, so that only a decrease up to $-0.01$ in Apgar score when smoking can be explained by a weight gain of more than 39 pounds, see Figure 2.15. The effect is stable over the quantile indices, suggesting no within group heterogeneity. Additionally, there is no modifying effect of weight gain classified by recommendations depending on prepregnancy BMI (Figure 2.16). Similar to the findings on weight gain, BMI does not show any modifying effect on the effect of smoking on Apgar score, as shown in Figure 2.14. Structural effect is very small and close to 0 for all three groups observed. As for the decomposition of the effect

of smoking on standardized birth weight, the decomposition shown here does not indicate any sex differences (Appendix A.8, Figure A.59).

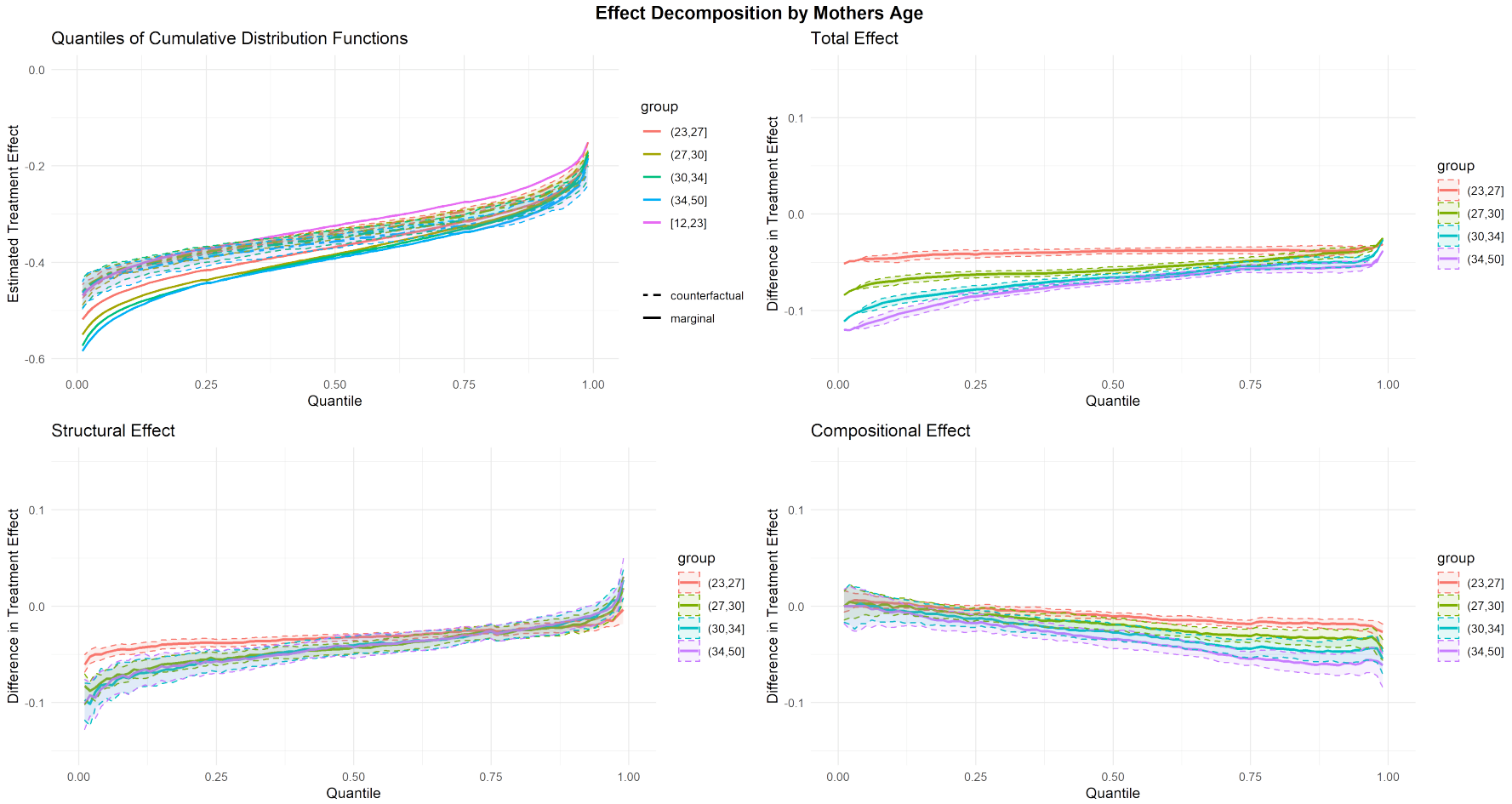


Figure 2.7.: Standardized Birth Weight - Effect Decomposition by Mother's Age

*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 2.12. 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.

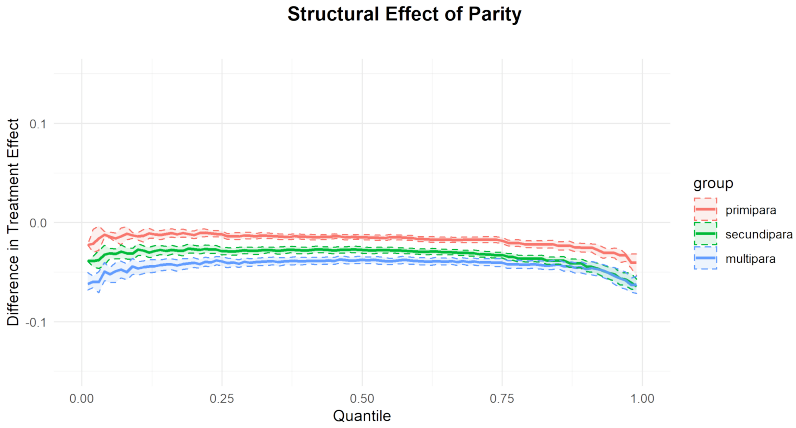


Figure 2.8.: Standardized Birth Weight - Structural Effect by Parity

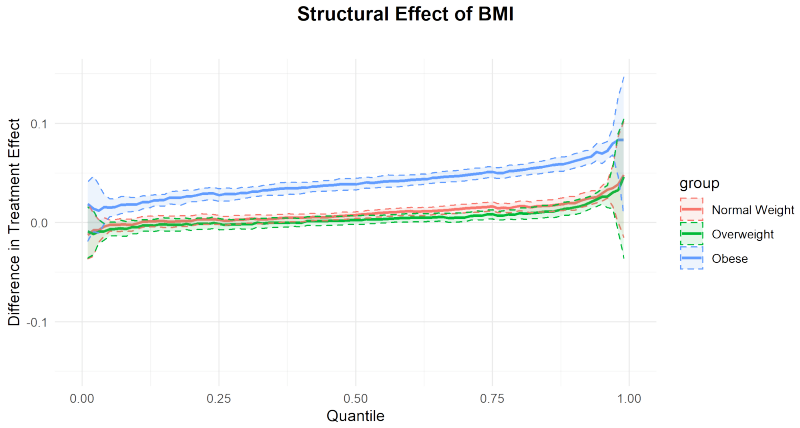


Figure 2.9.: Standardized Birth Weight - Structural Effect of BMI

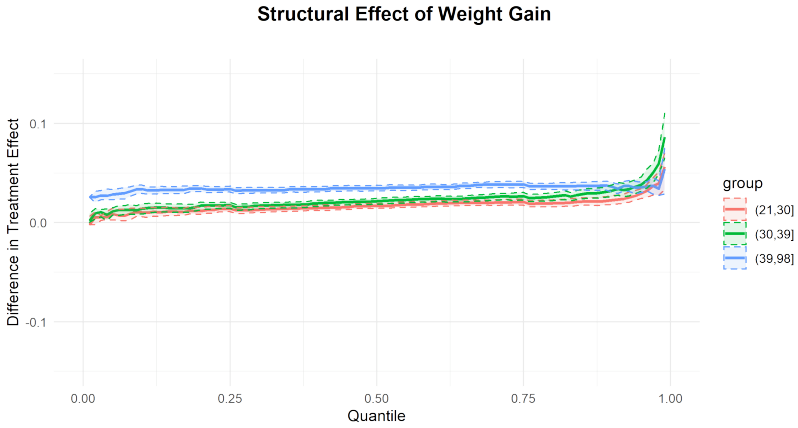


Figure 2.10.: Standardized Birth Weight - Structural Effect of Weight Gain

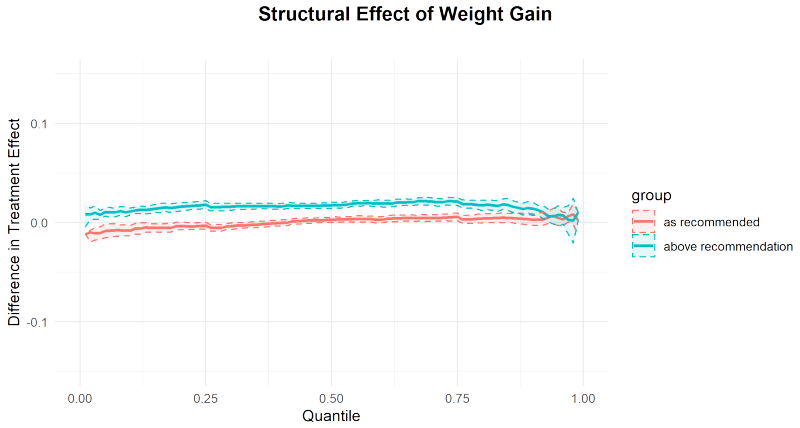


Figure 2.11.: Standardized Birth Weight - Structural Effect of Weight Gain Recommendations

*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. Positive difference corresponds to effect mitigation, whereas negative difference corresponds to effect amplification respectively. Shaded areas show 95% confidence intervals. For full decomposition results, refer to Appendix A.8.

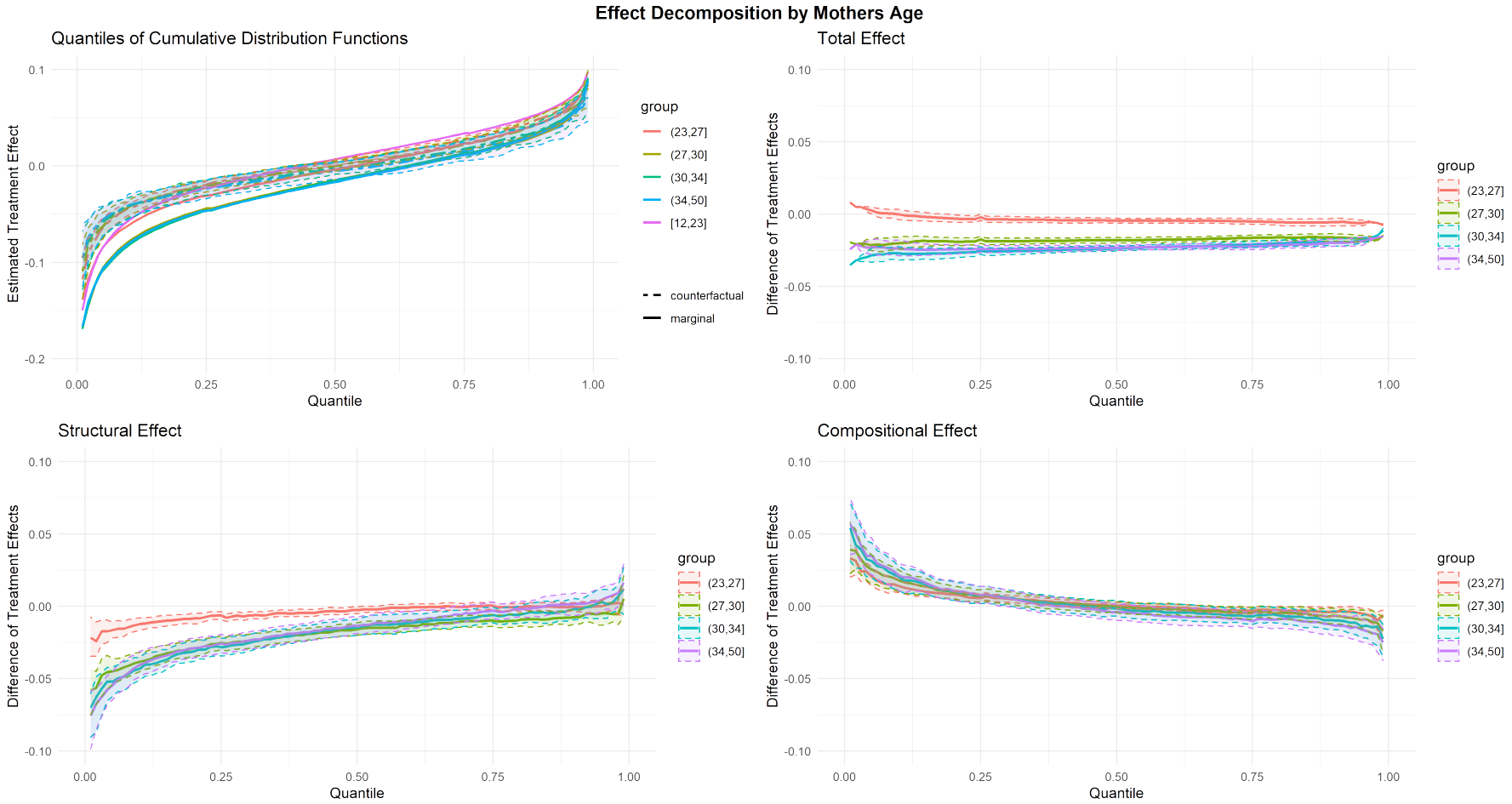


Figure 2.12.: Apgar Score - Effect Decomposition by Mother's Age

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on 5-minute Apgar score 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 2.12. 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.

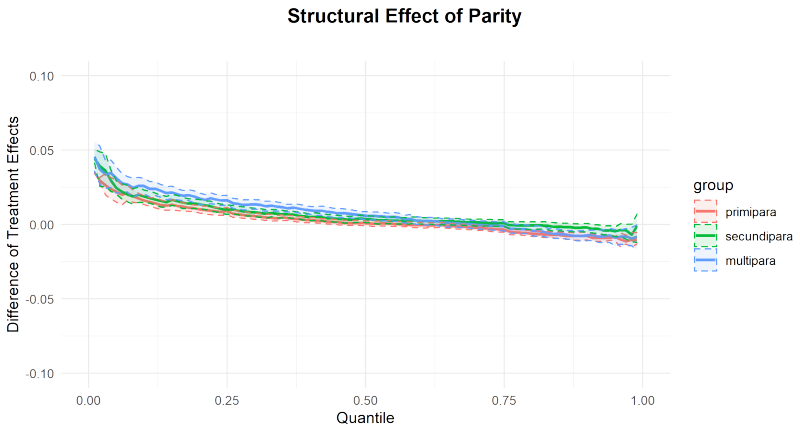


Figure 2.13.: Apgar Score - Structural Effect of Parity

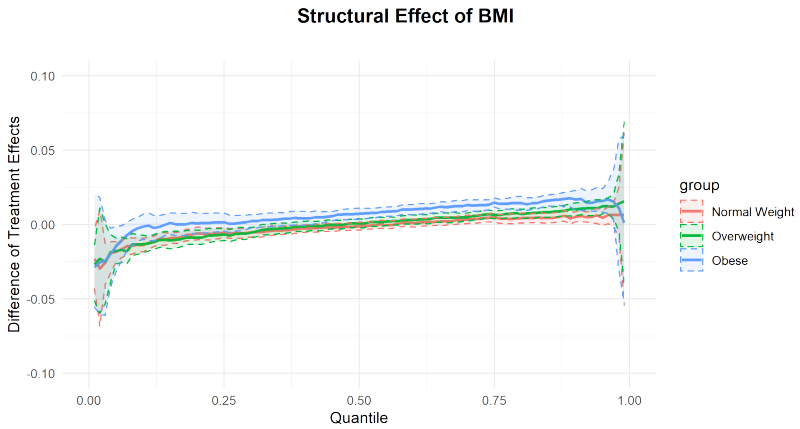


Figure 2.14.: Apgar Score - Structural Effect of BMI

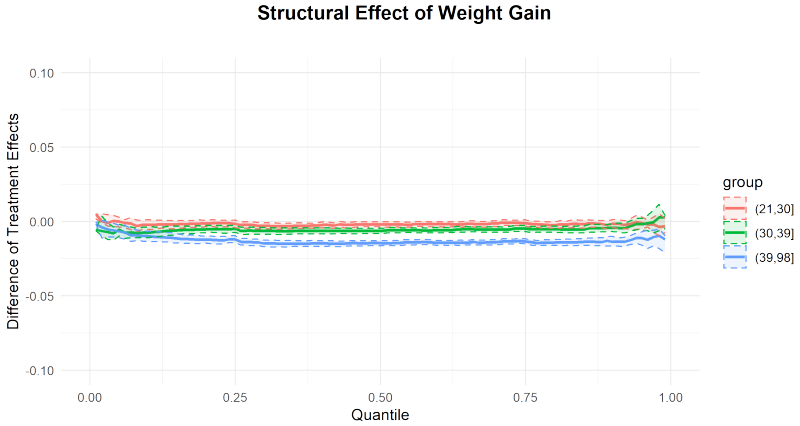


Figure 2.15.: Apgar Score - Structural Effect of Weight Gain

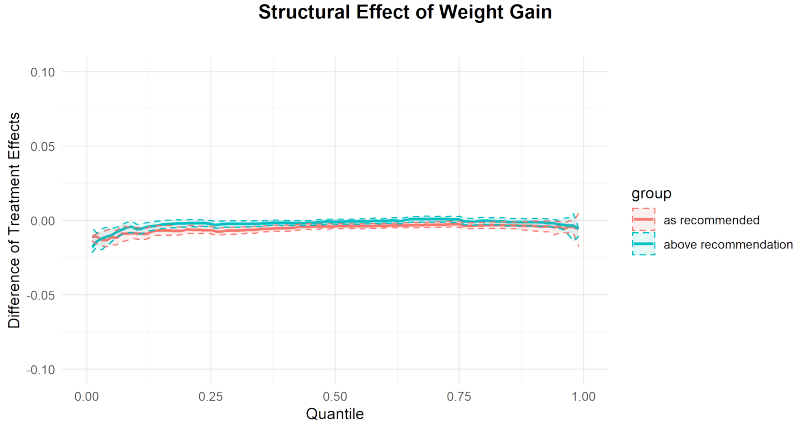


Figure 2.16.: Apgar Score - Structural Effect of Weight Gain Recommendations

*Note:* The figure shows the structural effect decomposition of the effect of maternal smoking on Apgar score for different modifying factors. 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.

## 2.6. Robustness Checks

### 2.6.1. Heavy Smokers

Regarding the groups of interest used for the decomposition, we find significant difference in smoking behavior (Figures 2.17 and 2.18). To verify that our findings are not driven by the strong heterogeneity in smoking levels of different groups and to account for dose dependency in the effect of smoking, we will rerun the analysis with focus on heavy smokers. There is no consensus over the definition of heavy smoking. Definitions usually vary between more than 20 daily cigarettes or more than 25 daily cigarettes[10]. We follow the definition of heavy smoking as 20 or more daily cigarettes, in order to not loose too many observations. Therefore, we only include those smokers, who smoke more than 20 daily cigarettes on average in the first trimester and do not quit smoking during pregnancy.

Restricting the analysis to heavy smokers only, removes the significant differences in smoking intensity between mother's age groups and prepregnancy BMI (see Figures 2.17 and 2.18).

When looking at the effect size of smoking (i.e., figure A.16), we observe larger estimates when only including heavy smokers in the analysis. This points towards a clear dose dependency of the effect, with stronger negative effects for heavy smokers. As in the main analysis, we find strong amplification by increased mothers age (Figure A.16) and parity (Figure A.17). While figure A.18 shows that the effect modification by obesity is similar in size to the one in the main analysis, weight gain does not have strong mitigating effects regarding heavy smoking (Figure A.19). Overall, results are very similar to those in the main analysis.

The analysis of the effect of heavy smoking on Apgar score reveals similar effect size as for the overall smoker population. This might stem from the overall very small effect that smoking has on the Apgar score and dose dependency might be limited in this relationship. As in the main analysis, Figure A.21 shows a clear age difference in the structural effect,

[10]The office for national statistics in the UK defines heavy smoking as smoking more than 20 daily cigarettes (`https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/compendium/generallifestylesurvey/2013-03-07/chapter1smokinggenerallifestylesurveyoverviewareportonthe2011generallifestylesurvey`), so does the Canadian government (`https://www.canada.ca/en/health-canada/services/health-concerns/tobacco/research/tobacco-use-statistics/terminology.html`). A study by Pierce et al. (2011) also defines heavy smoking as 20 or more daily cigarettes, however Wilson et al. (1992) uses the definition of 25 or more cigarettes per day.

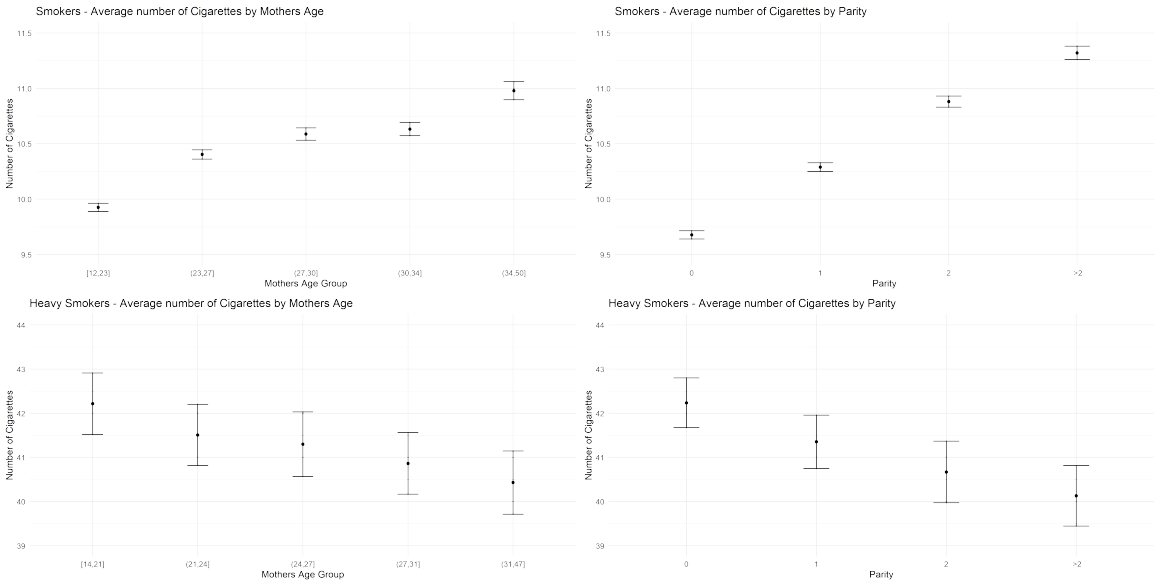


Figure 2.17.: Average daily cigarette consumption for mother's age and parity decomposition groups of interest

*Note:* The figure shows the mean number of cigarettes in the groups used for decomposition, using data from 2011, 2013, 2015, and 2018. For each group, the figure displays the mean of reported average cigarette smoked in the first trimester of pregnancy and corresponding 95% confidence intervals in the first row. The second row displays reported average cigarette smoked in the first trimester of pregnancy and corresponding 95% confidence intervals after filtering for heavy smoking (more than 20 daily cigarettes).

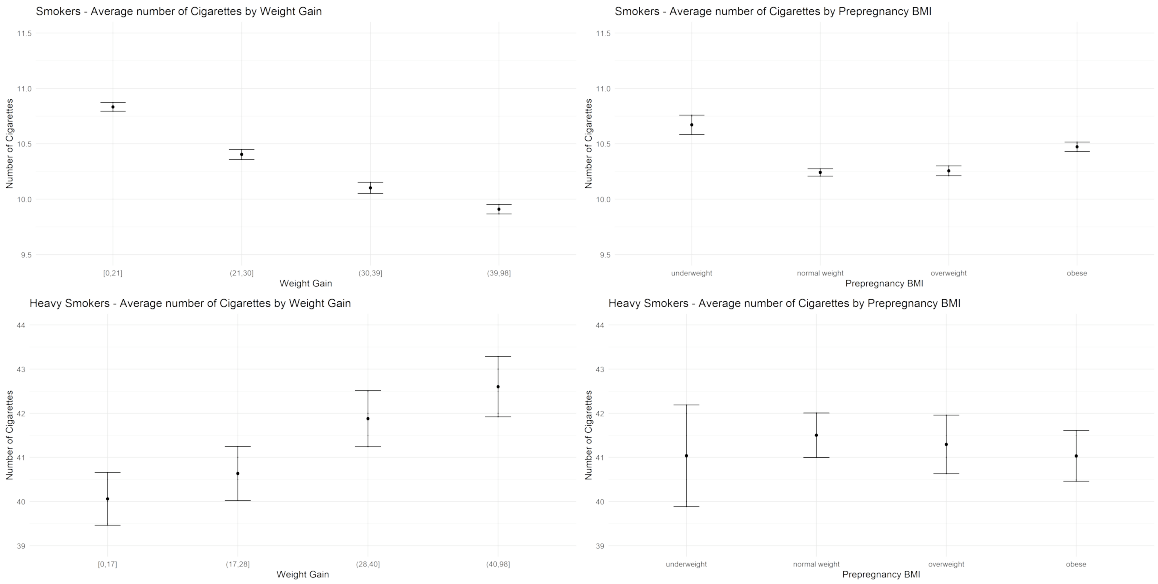


Figure 2.18.: Average daily cigarette consumption in the weight gain and BMI decomposition groups of interest

*Note:* The figure shows the mean number of cigarettes in the groups used for decomposition, using data from 2011, 2013, 2015, and 2018. For each group, the figure displays the mean of reported average cigarette smoked in the first trimester of pregnancy and corresponding 95% confidence intervals in the first row. The second row displays reported average cigarette smoked in the first trimester of pregnancy and corresponding 95% confidence intervals after filtering for heavy smoking (more than 20 daily cigarettes).

which are more profound when only looking at heavy smokers. An increase in age results in amplification of effect size. Contrary to our findings in the main analysis, all prepregnancy BMI groups show at least weak mitigating effect compared to the underweight group (Figure A.23), whereas we could only observe mitigating effects for obesity in the main analysis. For weight gain (see Figure A.24 and A.25), there is no significant difference in treatment effect magnitude for any of the groups considered, which is partially in line with the main analysis, where we found weak amplification regarding excessive weight gain.

### 2.6.2. Low Birth Weight

The birth weight distribution has large dispersion and weight loss caused by smoking is especially relevant in the lower tail of the distribution, as strong adverse effects are concentrated there. For example, a decrease in birth weight of 200 grams is less serious in case the baby would weight 3500 grams without mother's smoking during pregnancy compared to babies that would be born at 2000 grams without smoking of the mother. In order to understand whether the driving factors of heterogeneity identified in the main analysis are also relevant in the group with the largest adverse health consequences, we rerun the analysis for standardized birth weight on a subsample with all births below 2800 grams. This way, we include all LBW births while also including numerous normal weight babies. These normal weight babies are close to LBW regarding the effect size of smoking during pregnancy ranging from $-300$ grams to 0 grams. As in the robustness check on heavy smoking, the distribution regression is evaluated on a less comprehensive grid, resulting in less smooth estimates and confidence intervals.

Overall, estimated effect size of smoking on standardized birth weight is significantly smaller, when only considering babies born at below 2800 grams. The effect is nearly cut in half (Figure A.26). This indicates, that especially at higher birth weights, where a weight reduction might not be as critical as in the subsample considered here, effect of smoking is larger. However, the decomposition reveals similar risk factors, as the main analysis. For increased mother's age, the structural effect (Fiogure A.26) shows a significant but small amplification for older age groups (older than 30) in the first quartile. In all other quartiles the effect is not different from 0. The decomposition in figure A.27 does not reveal any

difference in effect size related to parity. Considering only babies born at lower birth weight (Figure A.28), there is still a small mitigating effect of obesity present. For weight gain (Figure A.29, A.30) we can again observe clear mitigating effects of increased weight gain or gaining weight as recommended or above recommendation.

### 2.6.3. Prepregnancy Smoking

Smoking during pregnancy is strongly stigmatized. Therefore, misreporting of self-reported measures of smoking behavior might be a problem. In a sample of mothers in New York City and Vermont, Howland et al. (2015) find birth certificates underestimated smoking by 24.3% before pregnancy and 26.2% during pregnancy compared to medical records. Since smoking before pregnancy is less misreported than smoking during pregnancy, we want to use it as a way to check how robust our results are against misreporting. As in preceding robustness checks, the distribution regression is evaluated on a less comprehensive grid, resulting in less smooth estimates and confidence intervals.

Appendix A.6 shows the decomposition for the effect of prepregnancy smoking on standardized birth weight and 5-minute Apgar Score. For both standardized birth weight and Apgar score, patterns in modifications do not change much compared to our main analysis. Effect size for those harming the baby most remains similar, see figure A.31. However, the quantiles of the cumulative distribution function show, that some mothers barely harm their baby at all. This can be explained by mothers quitting smoking when they find out they are pregnant.

Figure A.31 shows, that results for mother's age change slightly. In the first two quartiles, effect size, as well as patterns of modification with increasing age remain. For the two upper quartiles however, we now observe mitigating effects of age. This can be explained by the noisy signal of the prepregnancy smoking treatment. Since some mothers stop smoking before pregnancy, the effect of prepregnancy smoking is close to zero for some, and might also show mitigating effects regarding mother's age. For both standardized birth weight and Apgar score, we see changes in effect modification by BMI. For the Apgar score (Figure A.38), we now find effect modification by BMI, which we did not find when looking at smoking during pregnancy. For standardized birth weight (Figure A.33), the strong modification by obesity

remains, but we now also find mitigating effects of overweight and normal weight compared to underweight mothers.

### 2.6.4. Propensity Trimming

In order to ensure sufficient overlap in the propensity score of treated and untreated observations, we trim the propensity score as described in Stürmer et al. (2010)[11]. The proposed trimming is asymmetric. In a first step they cut all observations for which there is no overlap in propensity scores. In a second step, which we apply to our data, they add a range restriction to the upper and the lower end of the overlapping propensity scores. For the lower end, they use the lowest 2.5-th percentile of the propensity score estimated for treated observations. At the upper end of the propensity score distribution, they use a restriction based on the highest 97.5-th percentile of propensity scores in the untreated population. Adopting this asymmetric propensity score trimming, we trim the propensity score as indicated in figure 2.19. Our results do not change much compared to the results using the entire range of observations. Regarding standardized birth weight, effect modification by age (Figure A.41) is not as strong as without trimming and for parity, there is no clear distinction between the effect of one child or two children previously born to the mother (Figure A.42). Concerning the 5-minute Apgar score, we now find weak effect mitigation by increased BMI (Figure A.48) and effect amplification by multiparity (Figure A.47), which we did not find before.

## 2.7. Discussion and 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). These intensified smoking cessation programs are scarce and usually not available for all pregnant smoking women.

[11] Crump et al. (2009) propose to trim propensity scores in order to ensure sufficient overlap. However, their approach uses a symmetric trimming rule, which we cannot use in our setting, since we would lose most of our observations.

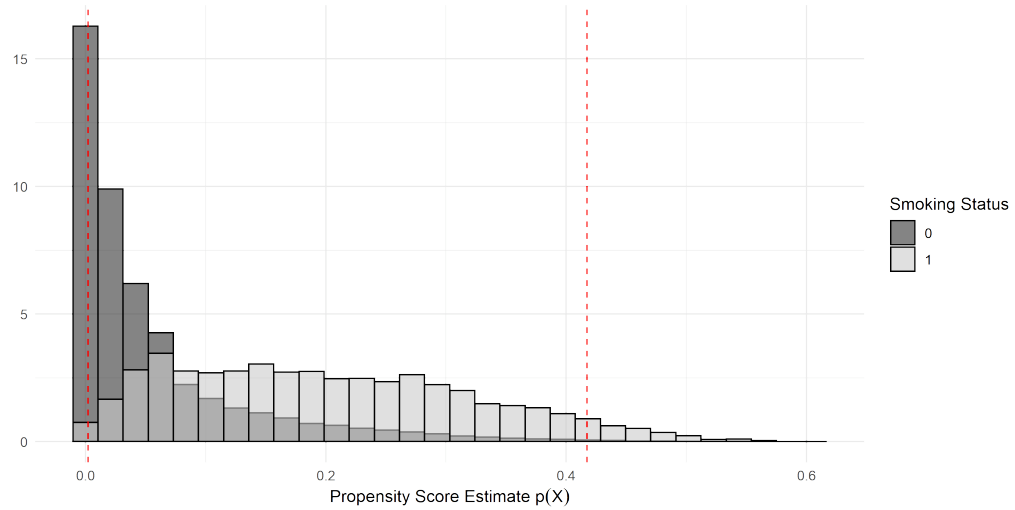


Figure 2.19.: Propensity Score estimates for Smokers and Non-Smokers

*Note:* Histogram of estimated propensity score estimates for smokers (smoking status 0) and non-smokers (smoking status 1). Propensity score is estimated using a regression forest. Red lines indicate trimming.

Using the case of low birth weight, we want to highlight the potential of our results. In a hypothetical scenario, we have 10,000 places for intensified smoking cessation available.[12]. For simplicity, we assume that all mothers who receive intensified smoking cessation stop smoking. Since our 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 gain, we want to sort all smoking mothers by these four easy criteria and make smoking cessation available for the first 10,000 in our sorted list. To compare results to an uninformed choice, we select 10,000 smoking mothers at random.

In order to evaluate the costs saved, we multiply costs savings associated with each additional gram of birth weight with the estimated effect of smoking in grams. We use estimates provided by Almond et al. (2005) (for details on costs, see Almond et al. (2005), Table IV), who estimate savings in hospital charges associated with each additional gram of birth weight in certain birth weight regions using data from 1995-2000. They provide pooled cross-section estimates and estimates using mother's fixed effects. 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. For babies born at 600-800

[12]Cost estimates for smoking cessation programs range from \$42 (Drouin et al., 2021),\$45 (Pollack, 2001), to up to \$1482 Levy et al. (2022)

grams, Almond et al. (2005) estimate additional costs of $186.59 per gram of birth weight, which they explain by a selection effect in which babies born at low birth weights are likely to die soon after birth, and therefore accumulate fewer charges for hospital charges. Since we want to estimate costs saved by smoking cessation through which the babies should get heavier, healthier at birth, and therefore more likely to survive, we will assume that the same cost effects per additional gram of birth weight holds for babies born at 600-800 grams and 800-1000 grams.

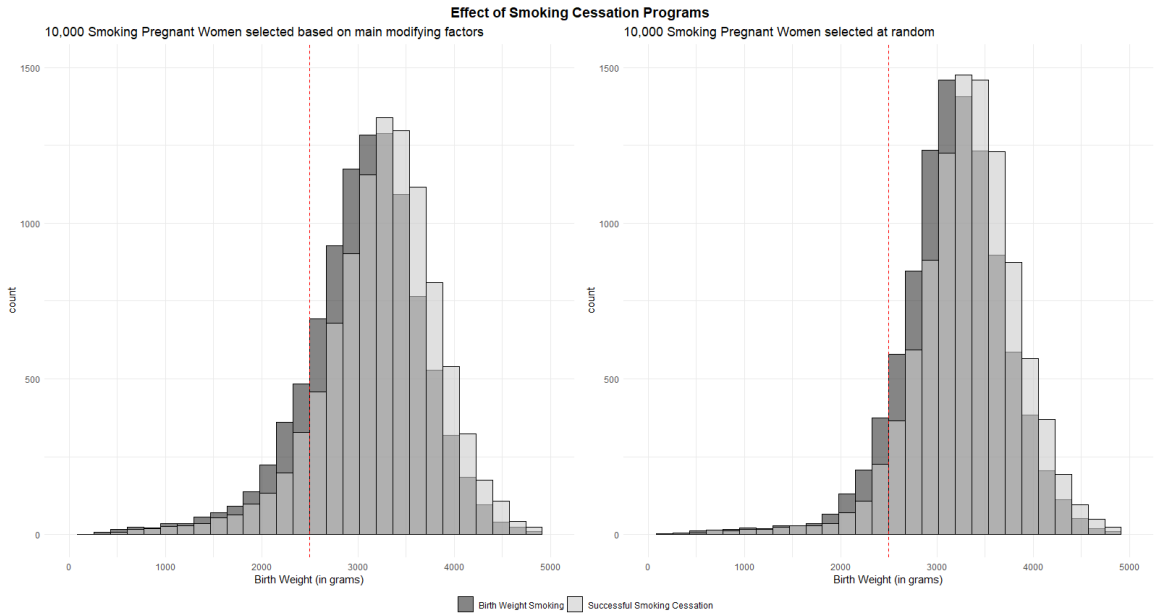


Figure 2.20.: Birth Weight Distribution - Smoking Cessation

*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).

Figure 2.20 shows birth weight distribution for children born to randomly targeted smokers and smokers targeted by using results of this study. Clearly, the informed sorting captures more babies born at LBW or below. The sorting based on main modifying factors selects 1546 LBW babies, whereas the random sorting selects only 954. This also translates into improved cost savings. While targeting at random saves costs between $9,637,395 (pooled cross-section estimates) and $3,242,219 (mother's fixed effects), targeting using the results of this study saves costs between $17,004,792 and $5,819,760. The saved costs from the informed sorting are approximately 80% higher than in the random case. Considering medical inflation, these cost savings correspond to between $37 million and $12.5 million in year 2022 dollars. Apart

from saved costs, targeting those most a risk to harm their babies a lot, can also result in improved health. While the assumption that all mothers enrolled in smoking cessation programs is not likely to hold, 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.

## 2.8. Conclusion

It is a matter of debate whether smoking is more damaging in some pregnancies than others. Few studies exist in the medical literature that try to shed a light on mother's characteristics driving the heterogeneity. However, these studies suffer several weaknesses and provide ambiguous results. But what are the driving factors of the heterogeneity that is present? When decomposing the treatment effect on several birth outcomes, we find strongest modifying effects for mothers age, which are robust to different specifications and outcomes.

As highlighted in section 2.7 the findings presented have important policy implications, especially regarding targeting of intensified smoking cessation programs. Smoking cessation is not easy, and most smokers who try to stop smoking without support fail. Effective means of smoking cessation, which comprise a combination of medication and therapy, are costly and time intensive. Babies harmed a lot by smoking in utero cause higher medical care costs in their first years of life, which could have been avoided in case the mother would have been able to stop smoking before or during pregnancy. Our findings suggest, that a simple targeted intervention of providing smoking cessation to 10,000 smoking pregnant women, can save costs of more than $17 million.

Our results are subject to several limitations. One might argue, that the estimates is biased due to the self-reported smoking indicator. Self-reported measures will never be accurate, and there is no way to check for the reliability of smoking levels indicated by the mothers themselves in the data used. Using prepregnancy smoking as a less misreported treatment, we find very similar results to our main analysis. Additionally, the smoking indicator does not thoroughly capture e-cigarette smoking or vaping, which is more and more popular among younger Americans. An investigation on the effect of e-cigarette smoking or vaping

on birth outcomes seems like a useful direction for future research.

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

Despite these limitations, we can add new evidence for treatment effect heterogeneity in the effect of smoking on birth outcomes and identify mother's 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.

# Chapter 3.

# Effect of Smoking Bans on Smoking during Pregnancy: Evidence from Germany

## 3.1. Introduction

Maternal smoking during pregnancy harms the unborn child's health and is strongly associated with birth weight reduction as well as fetal growth restriction. Mothers who smoke are at high risk of preterm delivery, stillbirth, and low birth weight birth, which are leading causes of death, disability, and disease among newborns (Almond et al., 2005). These adverse effects pose excessive costs on health care systems (Jacob et al., 2017, Kathleen Adams et al., 2002). It is therefore one of the key public health priorities of the WHO and many governments to reduce smoking prevalence and exposure to secondhand smoke, for example by smoke-free legislation and smoking bans in public places. However, most studies find no firm evidence on the effectiveness of smoking bans on active smoking (e.g., Adda and Cornaglia, 2010, Anger et al., 2011, Jones et al., 2015).

This study[1] tries to shed a light on the smoking habits of pregnant women in Germany and evaluates how smoke-free legislation impacts their smoking behavior. Pregnant women are especially at risk when smoking, as they are not only harming themselves but substantially threatening the health of their unborn babies. Therefore, it is important to evaluate who is smoking during pregnancy, whether there are differences by federal states, and especially

[1] Data from quality assurance procedures pursuant to Section 136 of the German Social Code, Book Five (Sozialgesetzbuch, SGB V) of the Federal Joint Committee (Gemeinsamer Bundesausschuss, G-BA) were used for this study.

focus on how different implementations of smoking bans across federal states in Germany influence smoking among pregnant women.

We exploit staggered implementation of state-level smoking ban legislation to estimate its effect on pregnant women's smoking behavior, using recent data from the German quality assurance procedure perinatal medicine, including all births that occurred in German hospitals between 2004-2016. We pursue a difference in differences (DiD) approach and make use of variations in smoke-free policy details by federal states over time. For smoking bans, we focus on bans in restaurants and bars, which differ substantially in terms of strictness across states. Therefore, we classify smoking bans into two groups, strict and partial bans, to account for different levels of strictness. Partial smoking bans include all bans with exceptions, i.e. smoking pubs, and separate smoker rooms, whereas strict smoking bans do not allow for exceptions and prohibit smoking in all restaurants and bars.

Our findings suggest a small but significant decreasing effect of smoking bans on smoking intensity, but no effect on prevalence. Controlling for years and state fixed effects, we find a decrease in average daily cigarette consumption among smoking pregnant women, which is stronger for strict bans than for partial bans. For strict smoking bans, we find a decrease of around 0.3 in daily cigarettes, corresponding to a reduction of 4 packs of cigarettes during pregnancy. For partial bans, we only find a reduction of 0.05 daily cigarettes. Regarding smoking prevalence among pregnant women, smoke-free legislation shows no robust effect. Overall, smoke-free legislation seems to be an effective mean to reduce the number of cigarettes smoked, but do not affect smoking prevalence.

Since the analysis of smoking behavior trends in federal states suggests that groups of states behave similarly over time, we want to relax the time-constant heterogeneity assumption imposed by the DiD framework and allow for unobserved group heterogeneity to vary over time. Using a recently introduced grouped fixed effects estimator by Bonhomme and Manresa (2015), we find very similar effects of smoking bans on smoking intensity as in our main specification. Again, we find a decrease of around 0.3 daily cigarettes due to strict smoking bans. Regarding the smoking prevalence results suggest, that unobserved factors within certain groups of states explain the overall smoking prevalence reduction.

Overall, smoking prevalence in Germany is declining, but since prevalence is especially

increasing for younger women of childbearing age (aged 15-45) (Bergmann et al., 2008, Lampert et al., 2013), there is a need to evaluate trends in smoking during pregnancy and underlying mechanisms. Most Western countries show declining smoking prevalence (Cnattingius, 2004), however, there are still a lot of smoking pregnant women, and especially Germany shows high smoking prevalence (World Health Organization, 2015). For example, 10.7% of pregnant women are smoking in the US in 2010 (Tong et al., 2013), 12.5% in Denmark (2010), 16.5% in Norway in 2009, and 6.9% in Sweden in 2008 (Ekblad et al., 2013). Compared to other countries, data availability on smoking during pregnancy in Germany is poor, since only a few surveys exist, that elicit smoking behavior (Kuntz and Lampert, 2016). Nevertheless, the prevalence of maternal smoking during pregnancy in Germany has been studied using different data sources (Kuntz and Lampert, 2016, Kuntz et al., 2018, Scholz et al., 2013, Schneider et al., 2008) and studies agree on declining prevalence. We also find declining smoking prevalence among pregnant women, which drops from 13% (2004) to 7.6% (2016). Additionally, we analyze smoking intensity. For average cigarette consumption, we find a reduction of 1.6 daily cigarettes, from 10.4 daily cigarettes in 2004 to 8.8 in 2016. However, both smoking prevalence and intensity differ substantially across states. For smoking prevalence, there is a profound North/South disparity, where Northern and Central German states show higher prevalence. Looking at smoking intensity, a West/East disparity is visible, where pregnant smokers smoke on average more cigarettes in West Germany.

The literature on the effects of smoking bans on smoking behavior shows limited evidence for their effectiveness on active smoking (e.g. Adda and Cornaglia (2010) for the US, Anger et al. (2011) for Germany, Carpenter et al. (2011) for Canada, Jones et al. (2015) for the UK). To our knowledge, this is the first study with emphasis on the effects on smoking behavior of pregnant women in Germany, but the effect on the entire German population has been studied by Anger et al. (2011). Using a difference-in-differences approach, Anger et al. (2011) study the effect of smoking bans in Germany shortly after the introduction of federal laws and focus on short-term effects. They find that the introduction of smoke-free legislation in Germany did not change average smoking behavior within the whole population. However, they find effects on individuals that go out to restaurants and bars often, where individuals are both less likely to smoke and reduce smoking intensity after the introduction of the ban.

Regarding differences by federal state, smokers who live in states with stricter bans in place show a stronger reduction, which is also supported by our findings. A study by Kvasnicka et al. (2018) examines the effect of public smoking bans in Germany on hospitalization. They find smoking bans to be effective in preventing hospital admissions due to cardiovascular diseases and asthma but do not evaluate active smoking behavior in detail. Hankins and Tarasenko (2016) study effects of smoking bans on neonatal health outcomes and maternal smoking behavior during pregnancy in the US. They find no effect of smoking bans on maternal smoking behavior or neonatal health outcomes. Jones et al. (2015) examine the effect of public smoking bans on smoking behavior in the UK, where no firm evidence on the effects of smoking bans on smoking can be found. Similar to most studies in the literature, we also find mixed evidence for the effectiveness of bans on active smoking.

Since our primary focus is on pregnant women and their smoking behavior, this study adds to the existing body literature on the effects of smoking bans by moving the focus to a narrower, but very important subgroup of the population. Further, we evaluate long-term trends on the most comprehensive data set including all hospital births in Germany and shed a light on differences in smoking ban enforcement by federal states. Studying smoking in Germany is of special interest, since Germany is one of the high-income countries with the highest smoking prevalence (World Health Organization, 2015) and smoking behavior in Germany has not been researched extensively.

In the next section, we give more details on smoke-free legislation in Germany. Section 3 describes the main data source and discusses the empirical strategy. In section 4, we present our findings and additional robustness checks, followed by a discussion of the results in section 5.

## 3.2. Smoking Bans in Germany

Smoke-free policies were introduced starting from August 2007 to ban smoking from several public places (i.e., bars and restaurants, schools, hospitals) in all 16 German federal states in order to protect non-smokers from adverse effects of second-hand tobacco smoke. Baden-Wuerttemberg was the first state to introduce smoke-free legislation in August 2007 and by

the end of August 2008, all states had introduced corresponding laws. However, policy details differ in terms of several exemptions by federal states (DEHOGA, 2008). A federal law introduced on September 01, 2007 regulates strict smoking bans in federal institutions and public transport, additionally the federal government raised minimum legal age for buying cigarettes from 16 years to 18 years in all of Germany (September 01, 2007). Details on smoking bans in all other areas lie within responsibility of each federal state, causing differences in introduction dates and strictness of smoking bans over time and federal state. Overall, German smoking bans were less comprehensive than smoking bans introduced in other countries and several exemptions applied. In states like Baden-Wuerttemberg, Berlin, Lower Saxony, and Rhineland-Palatinate, pubs can self-declare as "smoker pub", allowing people to legally smoke inside. There are only three states where currently a strict smoking-ban applies in restaurant and bars. In Bavaria the at that time most comprehensive smoke-free legislation of all German states was introduced on January 01, 2008. The enforcement of this strict ban, among other reasons, was made responsible for poor election results of the ruling party (CSU) in the 2008 Bavaria state election. Therefore, following the election, strict smoking bans were relaxed on August 01, 2009, leading to massive criticism by smoke-free initiatives, which eventually led to a referendum in 2010. The referendum was a success for the smoke-free initiatives and the comprehensive smoke-free legislation from 2008 was reintroduced on August 01, 2010. The strict smoking ban prohibits any smoking in restaurant, bars, pubs, and beer tents on local fairs. The state of Saarland changed their smoking-bans on July 01, 2010, which now excludes any exemptions for smoking bans in restaurants and bars, taking place after transitional arrangements at latest in December 2011. Lastly, North Rhine-Westphalia which had one of the loosest smoking-ban legislation of all federal states, enforced stricter rules starting from May 01, 2013.

For smoking bans, we focus on bans in restaurants and bars, since they differ markedly between states (see Table 3.1 for details). We classify smoking bans into partial smoking bans and strict smoking bans, where strict smoking bans means smoking ban without exceptions in place (after updated legislation in Bavaria, North Rhine-Westphalia, Saarland). Partial smoking bans include exceptions like separate rooms dedicated to smoking inside (i.e., Brandenburg, Hamburg) or smoking pubs (i.e. Baden-Wuerttemberg, Berlin). Data on

Table 3.1.: Enforcement Dates of Smoking Bans in Germany

| Federal State | Enforcement of smoking ban | Updated (Restaurants & Bars) |
|---|---|---|
| Baden-Wuerttemberg | August 01, 2007 | - |
| Bavaria | January 01, 2008 | August 01, 2009 and August 01, 2010 |
| Berlin | July 01, 2008 | - |
| Brandenburg | July 01, 2008 | - |
| Bremen | July 01, 2008 | - |
| Hamburg | January 01, 2008 | - |
| Hesse | October 01, 2007 | - |
| Lower Saxony | November 01, 2007 | - |
| Mecklenburg-West Pomerania | August 01, 2008 | - |
| North Rhine-Westphalia | July 01, 2008 | May 01, 2013 |
| Rhineland-Palatinate | February 15, 2008 | - |
| Saarland | June 01, 2008 | December 01, 2011* |
| Saxony | February 01, 2008 | - |
| Saxony-Anhalt | July 01, 2008 | - |
| Schleswig-Holstein | January 01, 2008 | - |
| Thuringia | July 01, 2008 | - |
| Germany | September 2007 | - |

* Stricter smoking ban was enforced July 01, 2010, but transition period for certain exceptions ended December 01, 2011.
*Note:* Information on smoking ban introductions are based on authors personal research in current and archived federal state laws and overview provided by DEHOGA (2008).

smoking ban introductions are based on authors personal research in current and archived federal state laws and overview provided by DEHOGA (2008). More detailed information on smoking bans in Germany can be found for example in Anger et al. (2011) and Kvasnicka et al. (2018).

## 3.3. Data and Method

### 3.3.1. Data Basis

Our main data source is data collected on behalf of the Common Federal Commission of Germany (Gemeinsamer Bundesausschuss, GBA) for the purpose of quality assurance. Data driven quality assurance is routinely conducted in all of Germany to ensure transparency in care, medical and nursing quality (IQTIG, Institut für Qualitätssicherung und Transparenz im Gesundheitswesen, 2016). One area covered by the healthcare quality assurance system is obstetrics and neonatology. Data in the area of obstetrics and neonatology comprises all inpatient births in German hospitals. We analyze data between 2004 and 2016. By law, all deliveries in hospitals need to be documented by hospital staff. Overall, the data comprises

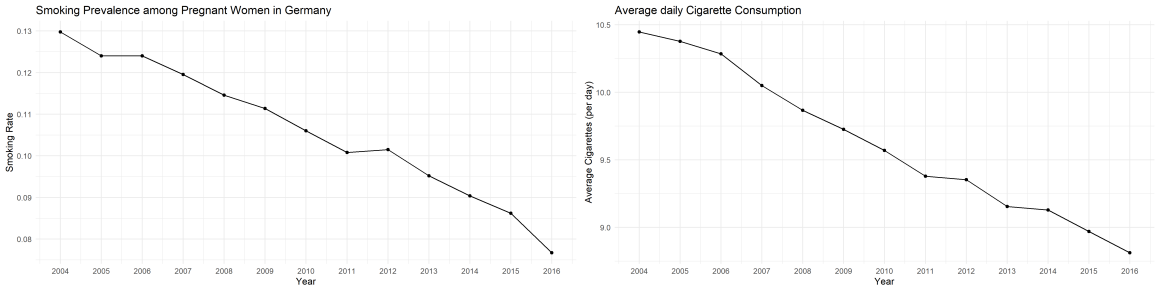


Figure 3.1.: Smoking prevalence (left) and average daily cigarette consumption (right) of pregnant women in Germany between 2004-2016

nearly all births that occurred in Germany between 2004-2016.

The data contains socio-demographic information on the mother (i.e., age, nationality, employment status), detailed information on pregnancy care (i.e. number of prenatal examinations), and detailed information about the birth and health information on the newborn (i.e. birth weight, APGAR Score, crown-heel length). See Table 3.2 for an overview of mothers demographic information, pregnancy risk factors and infant characteristics of the study population. Information on smoking during pregnancy is elicited by obstetrician and mothers are asked to recall their average daily cigarette consumption during pregnancy. The availability of information on smoking is limited to data from years 2004-2016. We classify each mother as smoker, who reported smoking at least one cigarette per day during pregnancy.

Overall, the data of 2004-2016 comprises 8,844,029 births. Of those, 85% reported their smoking behavior in 2004, whereas it declines to a reporting rate of only 79% in 2016. The study population comprises all pregnancies with smoking information available, resulting in 6,915,824 births. Mean and standard deviations of characteristics of interest remain comparable to the overall population after filtering for availability of smoking information.

For analyzing effects of smoking bans on smoking behavior among pregnant women, we use additional data sources. We link the female population of each federal state[2] to the observations on federal state level and additionally use data on smoking ban enforcement dates (for details see Table 3.1).

We observe a clear downward trend in both smoking prevalence and smoking intensity (see

[2]Source: Fortschreibung des Bevölkerungsstandes (EVAS-Nr. 12411), Bevölkerung nach Geschlecht - Stichtag 31.12. - regionale Ebenen [2004-2016]. Statistische Ämter des Bundes und der Länder, 2021.

Table 3.2.: Sample Means and Standard Deviation for Smoker and Non-Smoker in quality assurance procedure Perinatal Medicine (2004-2016)

| | **Whole Population** | | **Smoker** | | **Non-Smoker** | |
|---|---|---|---|---|---|---|
| Variables | Mean | SD | Mean | SD | Mean | SD |
| **Mothers Demographic Information** | | | | | | |
| age | 30.19 | 5.56 | 27.34 | 6.06 | 30.46 | 5.44 |
| unmarried | 0.14 | 0.35 | 0.26 | 0.44 | 0.14 | 0.34 |
| employed | 0.54 | 0.50 | 0.33 | 0.47 | 0.55 | 0.50 |
| Country of origin: Germany | 0.81 | 0.39 | 0.85 | 0.36 | 0.82 | 0.39 |
| housewife | 0.27 | 0.44 | 0.45 | 0.50 | 0.28 | 0.45 |
| in training/ studying | 0.03 | 0.17 | 0.05 | 0.22 | 0.03 | 0.16 |
| unskilled worker | 0.03 | 0.17 | 0.05 | 0.22 | 0.03 | 0.17 |
| skilled worker etc. | 0.32 | 0.47 | 0.23 | 0.42 | 0.36 | 0.48 |
| highly skilled worker etc. | 0.12 | 0.33 | 0.04 | 0.18 | 0.14 | 0.34 |
| **Pregnancy Risk Factors** | | | | | | |
| previous pregnancies | 1.10 | 1.34 | 1.46 | 1.67 | 1.06 | 1.29 |
| previous stillbirths | 0.01 | 0.11 | 0.01 | 0.12 | 0.01 | 0.10 |
| live-births | 1.16 | 1.02 | 1.37 | 1.24 | 1.12 | 0.99 |
| prenatal care visits | 11.48 | 3.46 | 10.79 | 3.62 | 11.57 | 3.43 |
| SSW first care visit | 9.32 | 4.04 | 10.06 | 5.15 | 9.24 | 3.85 |
| inpatient stay (days) | 1.73 | 10.31 | 2.09 | 10.37 | 1.50 | 10.28 |
| weight before pregnancy | 68.45 | 15.05 | 69.39 | 16.82 | 68.28 | 14.79 |
| weight before birth | 82.30 | 15.32 | 82.52 | 16.88 | 82.29 | 15.07 |
| high risk pregnancy | 0.33 | 0.48 | 0.39 | 0.50 | 0.31 | 0.47 |
| gestation length | 39.28 | 2.17 | 39.06 | 2.08 | 39.34 | 2.17 |
| cigarettes per day | 1.01 | 3.59 | 9.53 | 6.37 | 0.00 | 0.00 |
| **Infant Characteristics** | | | | | | |
| male | 0.51 | 0.50 | 0.51 | 0.50 | 0.51 | 0.50 |
| birth weight | 3324.89 | 599.75 | 3133.35 | 585.27 | 3345.99 | 594.50 |
| 5-min APGAR score | 9.63 | 0.96 | 9.59 | 1.03 | 9.64 | 0.93 |
| head circumference | 34.69 | 1.94 | 34.22 | 1.85 | 34.74 | 1.91 |
| length child | 51.08 | 3.40 | 50.15 | 3.32 | 51.16 | 3.31 |
| gestation length | 39.28 | 2.17 | 39.06 | 2.08 | 39.34 | 2.17 |
| Number of Observations | 8,844,029 | | 736,260 | | 6,179,564 | |

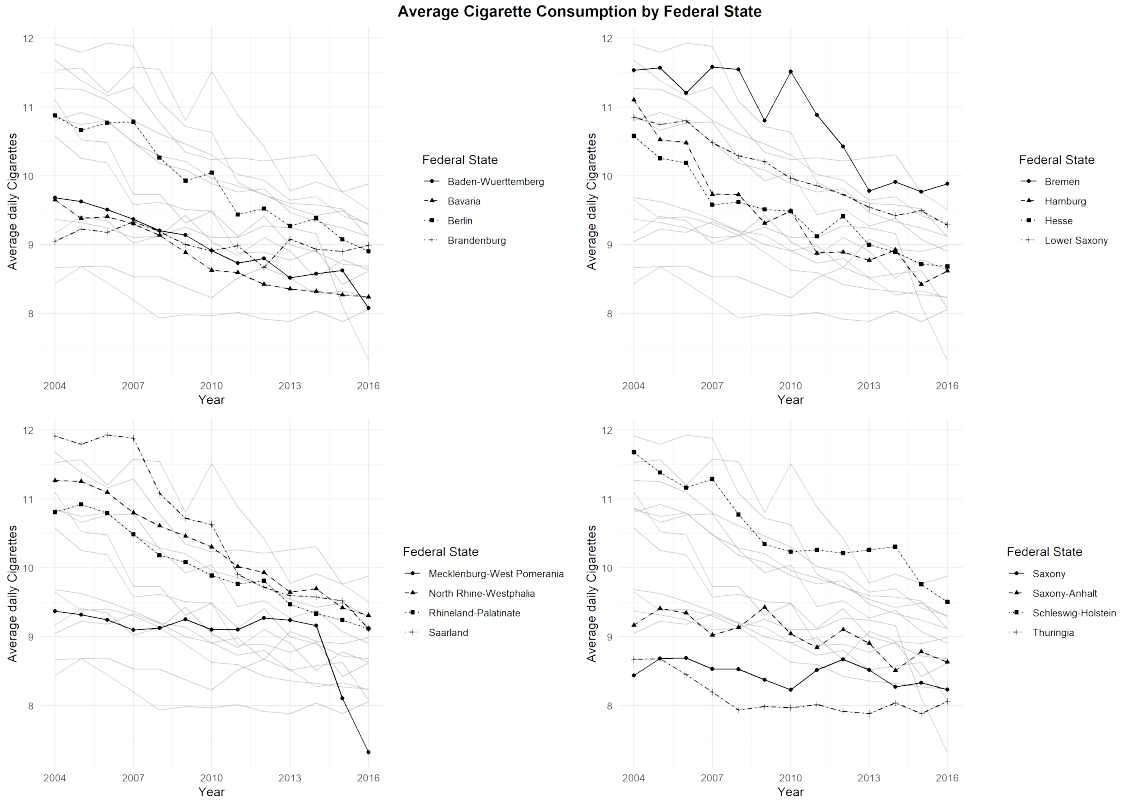


Figure 3.2.: Average daily cigarette consumption by Federal State

Figure 3.1). In 2004, 13% of mothers who reported smoking information did report smoking more than one cigarette per day, whereas in 2016, only 7.6% of all reporting mothers smoked. The average number of cigarettes smoked by smoking women during pregnancy declined from 10.4 in 2004 to 8.81 in 2016 (see Figure 3.1).

Comparing smokers to non-smokers, we find that smoking pregnant women are on average younger than non-smoker, more likely to be single and more likely to be of German origin than non-smokers (see Table 3.2). Smoking pregnant women are usually less educated and less likely to be employed than non-smoking mothers. Women who smoke during pregnancy have on average 0.5 pregnancies more than non-smoking women. Further, there are differences in utilization of prenatal care. Smoking pregnant women start their pregnancy care on average later than non-smokers and attend less prenatal care visits throughout their pregnancy, even though the percentage of high-risk pregnancies is higher for smokers than non-smokers. Their inpatient stay at the hospital after birth is on average 0.59 days longer than for non-smoker.

There are strong regional differences by federal state in smoking prevalence among preg-

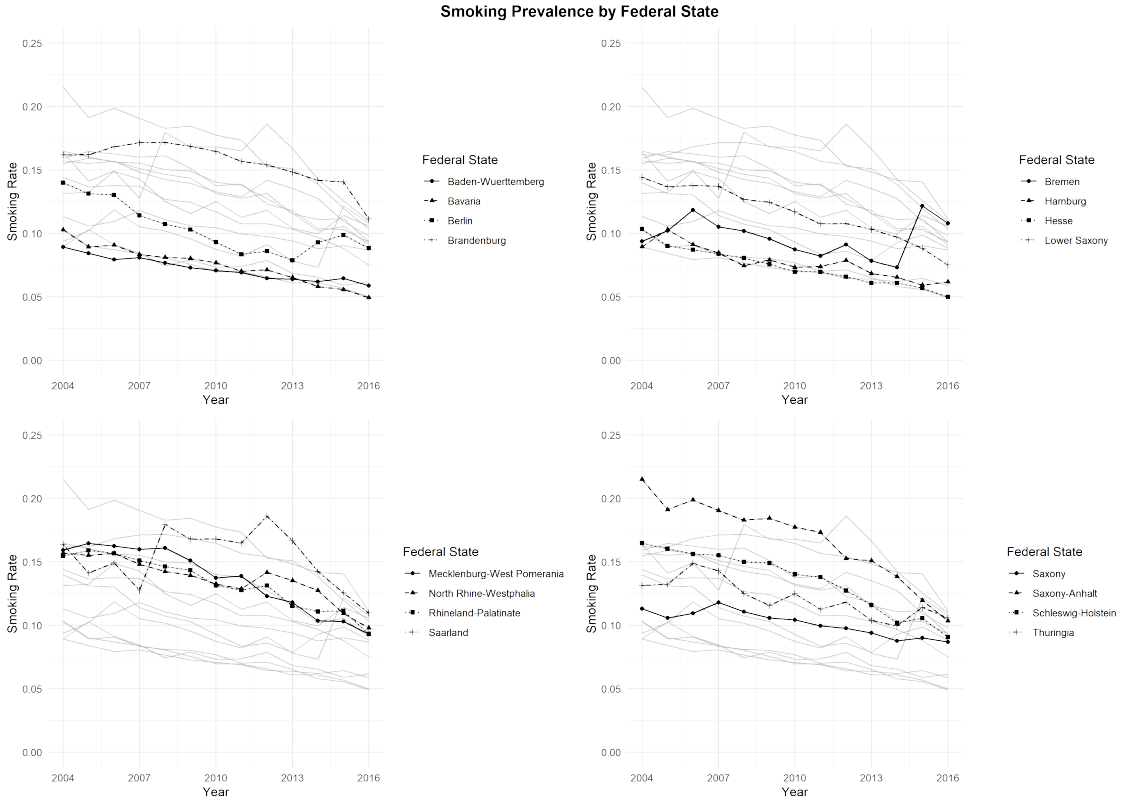


Figure 3.3.: Smoking Rate by Federal State

nant women in Germany (see Figure 3.4). There seems to be a North/ South separation when looking at smoking prevalence. In northern and central Germany, the smoking rate among pregnant women is higher than in southern Germany. In 2004, especially for Saxony-Anhalt there is a high smoking prevalence of 21.5%. Other northern regions show prevalence of around 15%. Only in southern Germany, there are lower smoking rates among pregnant women of below 10%. This difference fades over the 13 observed years but is still observable in 2016. This is especially interesting, since studies like Cnattingius (2004) on the overall smoking prevalence of young women report a West/East difference, whereas we find a North/South separation.

Regarding the average number of cigarettes smoked per day, one can see a profound West/East disparity (see Figure 3.5). Smoking pregnant women in West Germany smoke more daily cigarettes than those in the former East German regions. For Bavaria and Baden-Wuerttemberg we see lower average cigarette consumption, too. In 2004, the highest average daily cigarette consumption was reported in the Saarland (11.9 cigarettes per day), whereas

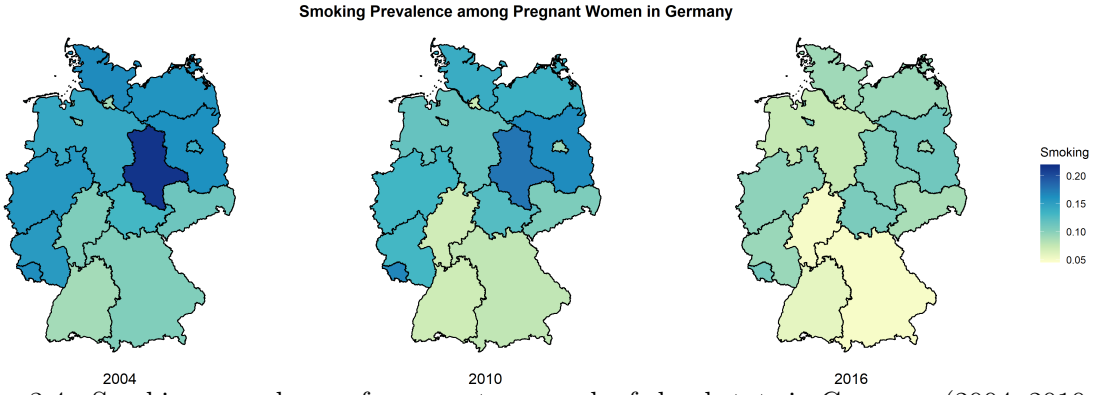


Figure 3.4.: Smoking prevalence of pregnant women by federal state in Germany (2004, 2010, 2016)

the lowest average daily cigarette consumption was reported in Saxony (8.4 cigarettes per day). Over the 13 years of interest, average daily cigarettes smoked decline and difference between federal states fades.

Figure 3.2 shows, that average number of daily cigarettes smoked by smoking pregnant women follows the same downward trend in most federal states of Germany. For Bremen, a federal state with relatively small population, we observe unstable trends. But also for Brandenburg we see an increase in average cigarette consumption after 2012, getting close to the all-time high in 2007. For Mecklenburg-West Pomerania, we observe a sharp decrease in average cigarette consumption after 2014.

Figure 3.3 shows strong differences in smoking rate between states. Especially for states with small population, like Saarland or Bremen, we observe unstable trends. For Saarland, we see a sharp increase in smoking rate in 2008, for Bremen we observe a sharp increase in 2015, despite the steady downward trend observed in the years before. Therefore, weights of population share are needed to evaluate effects of smoking bans in the next section.

### 3.3.2. Method

In order to estimate effects of smoking bans on average cigarette consumption and smoking rate among pregnant women, we exploit staggered implementation of smoking bans over time and over the 16 federal states using a difference-in-differences approach. Since we observe smoking behavior on a yearly basis, we approximate the introduction date of a new smoking ban regulation with the actual year of introduction. So, for the introduction of the smoking

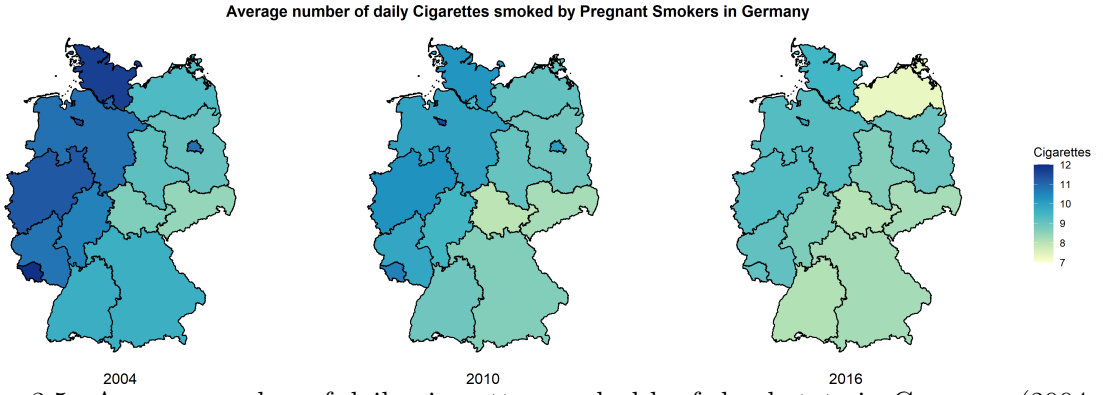


Figure 3.5.: Average number of daily cigarettes smoked by federal state in Germany (2004, 2010, 2016)

ban legislation on federal level (Nichtraucherschutzgesetz) on September 01, 2007, we would assume introduction starting in 2007. For legislation on federal state level, we focus on smoking bans related to restaurants and bars, since those differ strongest between the 16 states over time. We model smoking behavior as

$$\text{Smoking}_{st} = \beta_0 + \beta_1 \text{StrictBan}_{st} + \beta_2 \text{PartialBan}_{st} + \gamma_s + \mu_t + \epsilon_{st}, \tag{3.1}$$

where $\text{Smoking}_{st}$ is either smoking rate among pregnant women or average cigarette consumption among smoking pregnant women in year $t$ and federal state $s$. The parameters of interest are $\beta_1$ and $\beta_2$, the effect of introduction of strict smoke-free legislation $\text{StrictBan}_{st}$ or partial smoke-free legislation $\text{PartialBan}_{st}$ in federal state $s$ at time $t$, respectively. In our identification, we include fixed effects to absorb confounding variation. $\gamma_s$, federal state fixed effects, control for unobserved heterogeneity in smoking behavior in federal states of Germany. Year fixed effects $\mu_t$ eliminate unobserved differences in smoking behavior in the years of interest and control for price changes. Since we observe smoking behavior on federal state level and not on individual level, we weight our observations by share of female population in each federal state to control for unstable trends in federal states with very small population (i.e., Saarland, Bremen).

## 3.4. Results

### 3.4.1. Smoking Ban and Smoking Behavior

We focus on smoking bans in restaurants and bars and estimate their effect on smoking rate and average cigarette consumption among pregnant women. Since smoking bans are implemented by federal states independently (Table 3.1) and legislation differs not only in enforcement date, but also in terms of strictness, we can estimate the effect of the introduction of state level legislation using a difference in differences approach.

The results for the effect of smoking ban introduction on smoking rate are presented in Table 3.3. Controlling for state fixed effects only, we find similar effects of partial smoking bans and strict smoking bans. Strict smoking bans seem to reduce smoking rate by 2 percentage points, the reduction of partial bans is only 1 percentage point. Both estimates are significant at the 1% level. Including years fixed effects, estimates are cut in half and appear insignificant. In our richest specification, including both state and year fixed effects, estimates on both ban types change sign and are positive, but insignificant. This change in sign suggests, that the estimates are not robust to different model specifications. Further, we cannot find enough evidence that decline in smoking rate is driven by smoking ban introduction in any way.

Table 3.4 shows the results for average number of cigarettes as outcome of interest. Regardless of the specification, we find a negative effects of both strict and partial smoking bans on average number of cigarettes smoked. Only including state fixed effects, the estimated effects are quite large and highly significant for both bans. Strict smoking bans reduce number of cigarettes smoked by more than one daily cigarette, the effect of partial ban is $-0.63$. Including years fixed effects, the size of the estimates reduces sharply and estimates are not significant. The estimate for strict bans reduces to $-0.33$ and the one for partial ban reduces even further to $-0.09$. In our richest specification, including both state and year fixed effects, the effect size is comparable to results when only including year fixed effects. This time, only the estimate on strict smoking bans is significantly different from zero.

Our results suggest that smoking bans mainly have an effect on the intensive margin and do not succeed in reducing the extensive margin, as they seem to have no effect on

smoking rate, but significantly reduce the number of daily cigarettes pregnant women smoke. This mechanism is intuitive, since smoking bans reduce occasions to smoke in everyday life, particularly when going out, which might not lead to people quitting smoking overall, but reducing their consumption. Reduction in smoking rate, however, seems like a long-run trend, independent of the introduction of smoking bans.

Table 3.3.: Effect of Federal State Smoking Ban Introduction in Restaurant/Bars (starting from 2007) on Smoking Rate among Pregnant Women

| Dependent Variable: | Smoking Rate | | |
|---|---|---|---|
| Model: | (1) | (2) | (3) |
| *Variables* | | | |
| Strict Smoking Ban | -0.0267*** | -0.0148 | 0.0042 |
| | (0.0041) | (0.0168) | (0.0032) |
| Partial Smoking Ban | -0.0159*** | -0.0058 | 0.0042 |
| | (0.0049) | (0.0114) | (0.0030) |
| *Fixed-effects* | | | |
| state | Yes | | Yes |
| year | | Yes | Yes |
| *Fit statistics* | | | |
| Observations | 208 | 208 | 208 |
| $R^2$ | 0.84620 | 0.21031 | 0.95469 |
| Within $R^2$ | 0.35511 | 0.01551 | 0.01530 |

*Clustered (state) standard-errors in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

### 3.4.2. Robustness Checks

To further evaluate treatment dynamic and check pre-treatment periods for balance between treatment and control group, we conduct an event study. We interact time dummies for the years before and after smoking ban introduction with the treatment indicators. Formally, we introduce several leads and lags of treatment in our main specification (3.1) and estimate

$$\text{Smoking}_{st} = \sum_{\substack{\tau=-3 \\ \tau\neq-1}}^{6} \alpha_\tau \mathbb{1}_{[t-\text{Strict}_s=\tau]} + \sum_{\substack{\tau=-3 \\ \tau\neq-1}}^{6} \beta_\tau \mathbb{1}_{[t-\text{Partial}_s=\tau]} + \gamma_s + \mu_t + \epsilon_{st}, \tag{3.2}$$

where $\text{Strict}_s$, $\text{Partial}_s$ are the event dates, on which treatment status switches from 0 to 1 for the strict treatment or the partial treatment, respectively.

Table 3.4.: Effect of Federal State Smoking Ban Introduction in Restaurant/Bars (starting from 2007) on Cigarette Consumption among Pregnant Women

| Dependent Variable: | Average Cigarettes | | |
|---|---|---|---|
| Model: | (1) | (2) | (3) |
| *Variables* | | | |
| Strict Smoking Ban | -1.119*** | -0.3302 | -0.2865** |
| | (0.1969) | (0.3368) | (0.1047) |
| Partial Smoking Ban | -0.6263*** | -0.0851 | -0.0524 |
| | (0.1348) | (0.1970) | (0.0577) |
| *Fixed-effects* | | | |
| state | Yes | | Yes |
| year | | Yes | Yes |
| *Fit statistics* | | | |
| Observations | 208 | 208 | 208 |
| $R^2$ | 0.82955 | 0.33631 | 0.94540 |
| Within $R^2$ | 0.55751 | 0.01713 | 0.08267 |

*Clustered (state) standard-errors in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Focusing on smoking rate among pregnant women, no estimates in the two pre-treatment periods of interest appear to be significant (see Figure 3.6). Similarly, after introduction of treatment effects remain insignificant and close to 0. Effects appear positive in some periods after treatment introduction for both partial and strict bans. The event study shows that even after treatment introduction there is no significant effect on smoking rate.

Figure 3.7 shows the event study for the effect of smoking bans on average cigarette consumption. We find slightly positive, but insignificant estimates in the pre-treatment periods for strict smoking bans. Partial smoking bans show negative and insignificant effects in the pre-treatment periods. After treatment introduction, estimates for strict bans are considerably below zero and negative effects are significant in at least some years of interest, especially in the short run. The point estimates for partial bans are positive and increasing after treatment introduction, but insignificant.

Both event studies support plausibility of the common trends assumption and show negative effects of strict smoking bans on average cigarette consumption, especially shortly after treatment introduction. Effect of bans on smoking rate seems to be non-existent.

Since two-way fixed effects estimation might not be optimal to capture effects of smoking

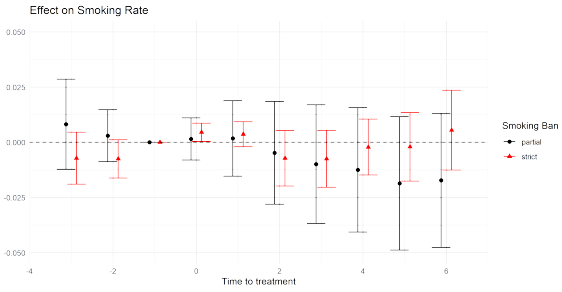


Figure 3.6.: Event study for smoking rate

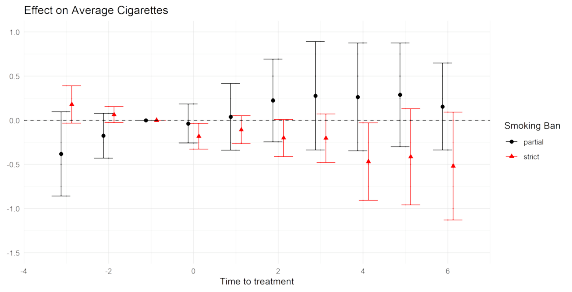


Figure 3.7.: Event study for average cigarette consumption

bans on smoking behavior due to the restrictive assumption of time-constant unobserved heterogeneity, we want to allow for unobserved group heterogeneity varying over time in our specification. The analysis of trends of smoking in federal states revealed differences in the evolution of smoking behavior over time across certain groups of states, which we hope to capture by group specific time trends. Additionally, the assumption of time varying unobserved group heterogeneity seems especially plausible in the case of Germany, where we have certain groups of states, that historically evolve similarly over time. For example, even years after German reunification, structural differences between federal states in the former GDR and West Germany remain.

We will make use of a novel grouped fixed effects (GFEs) estimator proposed by Bonhomme and Manresa (2015). GFEs cluster individuals with similar unobserved characteristics into a finite number of groups. Group assignments are not picked by the researcher. Rather, group assignment is data driven, by minimizing a least squares criterion over all possible groupings. GFEs assume that states within the same group share the same time profile of group-specific unobserved heterogeneity,

$$\text{Smoking}_{st} = \beta_1 \text{StrictBan}_{st} + \beta_2 \text{PartialBan}_{st} + \alpha_{g_s t} + \epsilon_{st}, \tag{3.3}$$

where $\alpha_{g_s t}$ refers to the time profile of group $g_s$ for $g_s \in \{1, \ldots, G\}$. Since $\alpha_{g_s t}$ captures the groups' time trajectories, we exclude time fixed effects from the model. By defining a parameter $\theta = (\beta_1, \beta_2)$ and a vector of regressors $x_{st} = (\text{StrictBan}_{st}, \text{PartialBan}_{st})$, we can rewrite equation 3.3 more compactly

$$\text{Smoking}_{st} = x'_{st}\theta + \alpha_{g_s t} + \gamma_s + \epsilon_{st}. \tag{3.4}$$

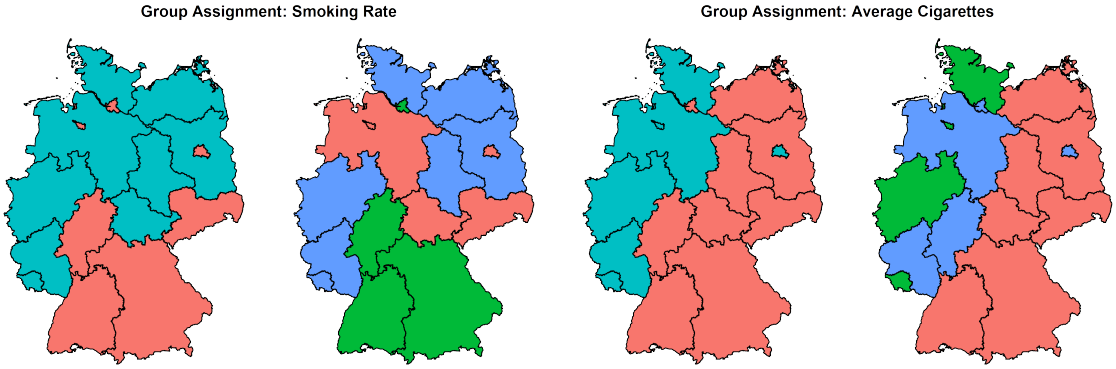


Figure 3.8.: GFE group assignment for smoking prevalence (left) and average daily cigarette consumption (right).

The grouped fixed effects estimator is the solution of the following minimization problem

$$(\hat{\theta}, \hat{\alpha}, \hat{\gamma}) = \underset{\theta,\alpha,\gamma}{\operatorname{argmin}} \sum_{s=1}^{N} \sum_{t=1}^{T} (\text{Smoking}_{st} - x'_{st}\theta - \alpha_{g_s t})^2, \tag{3.5}$$

where we search for the minimum over all possible groupings $\gamma = \{g_1, \ldots, g_N\}$, common parameter $\theta$ and group-specific time effects $\alpha$. For details on computation refer to Appendix B.1. GFEs also allow for individual specific time invariant fixed effects. In our setting, we will therefore add federal state fixed effects and estimate

$$\text{Smoking}_{st} = \beta_1 \text{StrictBan}_{st} + \beta_2 \text{PartialBan}_{st} + \alpha_{g_s t} + \gamma_s + \epsilon_{st}. \tag{3.6}$$

To not impose too many restrictions on the model, we will make use of either 2 or 3 groups for the GFEs. In practice, we first estimate the group assignment for each state and outcome for a given number of groups and interact this group assignment with time.

Figure 3.8 shows the group assignment for both outcomes of interest using two or three groups, respectively. The group assignment captures the North/South disparity in smoking rate when using two groups, where smoking prevalence is higher in the northern part of Germany (blue), than in the South (red). For three groups, we find a low prevalence group (green), a mid-prevalence group (red) and states with high prevalence (blue). Groups for average cigarettes smoked also reflect East/West disparity, as observed before. For two groups, we find a high-intensity smoking group for western parts of Germany (blue) and a low intensity smoking group for east and southern Germany (red). Considering three groups,

high intensity states are marked in green, mid-intensity is marked in blue and low intensity in red, which again comprises most of the former East Germany and southern Germany states. Groups loosely reflect former GDR and West Germany differences as hypothesized before, especially with regard to average number of cigarettes smoked.

Table 3.5.: Grouped Fixed Effects: Effect of Introduction of Federal State Smoking Bans

| Dependent Variables: | Smoking Rate | | Average Cigarettes | |
|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) |
| *Variables* | | | | |
| Strict Smoking Ban | 0.0085* | 0.0092* | -0.2151*** | -0.3110** |
| | (0.0041) | (0.0050) | (0.0650) | (0.1105) |
| Partial Smoking Ban | 0.0045*** | 0.0041** | -0.0013 | 0.0246 |
| | (0.0014) | (0.0015) | (0.0588) | (0.0772) |
| *Fixed-effects* | | | | |
| state | Yes | Yes | Yes | Yes |
| sr_GFE_2-year | Yes | | | |
| sr_GFE_3-year | | Yes | | |
| ac_GFE_2-year | | | Yes | |
| ac_GFE_3-year | | | | Yes |
| *Fit statistics* | | | | |
| Observations | 208 | 208 | 208 | 208 |
| $R^2$ | 0.96955 | 0.96901 | 0.95980 | 0.96893 |
| Within $R^2$ | 0.05721 | 0.05598 | 0.06609 | 0.11791 |

*Clustered (state) standard-errors in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Allowing for unobserved patterns of heterogeneity, we find that the effect of smoking bans on average number of cigarettes smoked among pregnant women is comparable to the one found in our main specification. Strict smoking bans significantly reduce average number of cigarettes smoked by $-0.22$ or $-0.31$ daily cigarettes, depending on the number of groups used for clustering. Effect size does not differ a lot between different model specifications. Partial bans do not seem to have a robust effect of smoking intensity, since estimates vary in sign and are not significant. Using GFEs, we find positive significant effects of bans on smoking rate. The estimates for the effect of smoking bans on smoking rate, suggest that partial as well as strict bans increase smoking rate among pregnant women by around 0.9 percentage points. Given a baseline smoking rate in 2004 of around 13, this increase is fairly

large. Reduction in smoking rate over time is therefore not attributable to introduction of smoking bans, but to other unobservable factors common to certain groups of states.

## 3.5. Discussion

Data from the quality assurance procedure obstetrics and quality assurance procedure neonatology suggests, that there is a declining trend in smoking during pregnancy. However, underlying mechanism that lead to smoking reduction remain unclear. Therefore, we evaluate the effect of smoke-free legislation in Germany that differ across state and over time to estimate their effect on smoking behavior of pregnant women. Pregnant women are a group of special interest, since they do not only harm themselves, but their unborn baby while smoking.

Starting from 2007, German federal states introduced laws to protect non-smokers, which ban smoking from public transport, restaurants, and public places. Those laws differ in strictness by federal state. We exploit these time and state varying differences, to provide causal estimates of the effect of the introduction of smoking bans on smoking behavior of pregnant women.

We find significant but small reduction effects on average cigarette consumption among pregnant women due to the introduction of smoking bans, but no such effect on smoking rate. Considering the effect of smoking bans on smoking intensity, we find that especially strict bans decrease smoking throughout pregnancy by 4 packs. This effect proves robust to different assumptions and model specifications. However, the effect of bans on smoking prevalence remains unclear, as different specifications and model assumptions lead to substantially different estimates. Therefore, this study suggests that smoking bans mainly work on the intensive margin and not the extensive margin, as they seem to have no clear effect on smoking rate, but significantly reduce the number of daily cigarettes pregnant women smoke.

To our knowledge this study is the first to evaluate the effect of smoking bans or smoke-free legislation on the smoking behavior of pregnant women. Previous studies assessing the effects of smoking bans on smoking behavior, mostly with focus on the entire population,

also find mixed results of smoking bans on active smoking (Anger et al., 2011, Jones et al., 2015). However, certain subgroups seem to respond differently to smoke-free legislation than others, i.e. Anger et al. (2011) find decreasing effects on both smoking rate and cigarette consumption for individuals that go out often.

Since adverse effects of smoking are widely known and especially pregnant women are informed about harmful effects of smoking on their child, they might have an intrinsic willingness to stop smoking or at least reduce their cigarette consumption. This willingness is most probably higher than among the normal population, and therefore pregnant women might respond differently to smoking bans than other population groups. Especially strict smoking bans, which are currently in place in Saarland, North Rhine-Westphalia and Bavaria prove to be effective in reducing smoking intensity among pregnant women. Similarly, Anger et al. (2011) find that smokers living in states with stricter smoke-free legislation reduce smoking more than those living in states with more relaxed bans.

However, the results on declining smoking prevalence and also our estimates need to be considered with care. For most studies including this one, smoking during pregnancy is a self-reported measure, meaning that mothers do not need to admit to smoking during pregnancy. In Germany, smoking during pregnancy is socially unaccepted, therefore mothers might fear negative consequences when admitting to smoking during pregnancy. Bergmann et al. (2008) find overall decreasing trends in smoking, however it is increasing among young women. Therefore, Bergmann et al. (2008) assume strong underreporting for smoking during pregnancy, since most studies report significantly lower smoking rates even at the beginning of pregnancy, where they should be at least comparable to those in the childbearing age group. Similarly, Fleitmann et al. (2010) assume strong underreporting when it comes to smoking during pregnancy. Therefore, our results may only act as a lower bound for the effect of smoking bans on smoking during pregnancy due to underreporting issues.

As this study has shown, smoking prevalence during pregnancy is declining, and especially strict smoking bans enforce lower cigarette consumption among pregnant women. This effect of bans on smoking behavior found, might only hold true for the special subgroup of interest, but it indicates, that in order to enforce smoking cessation or reduction in cigarette consumption in the general population, stricter smoking bans in all of Germany might need

to be considered.

# Chapter 4.

# Effect of Temperature and Weather Shocks on Health at Birth: Evidence from the US

## 4.1. Introduction

Climate change is the biggest health threat humanity faces in the 21st century (World Health Organization, 2018). In recent years, extreme weather events such as heat waves, cold spells, and heavy rainfall have become increasingly frequent and severe due to climate change. These extreme weather events threaten health in different ways, for example through the stress associated with the shock, food, and water insecurities, undernutrition, and forced displacements. Especially vulnerable groups, including older populations, women, and children are disproportionately prone to climate-sensitive health risks. While the impact of these events on human health has been extensively studied with a focus on average effects, less attention has been paid to possible heterogeneity in their effects on the health of infants born to mothers who were exposed to them during pregnancy. Pregnant women are considered a particularly vulnerable group when it comes to exposure to extreme weather events due to the changes in their bodies physiology (Samuels et al., 2022). It is crucial to understand how exposure to weather in utero affects the fetus and to be able to identify the most vulnerable groups, as it may affect the infant's long-term outcomes, in line with the fetal origins

hypothesis[1].

This paper investigates the effects of in-utero exposure to several extreme weather events in the US between 1989 and 2004 and systematically explores potential heterogeneity in these effects. Weather events considered include heat and cold events, and rainfall shocks. We analyze their effects on multiple outcomes of health at birth, like standardized birth weight, small for gestational age birth (SGA)[2], gestation length, and 5-minute Apgar score[3]. Following the literature, we first analyze the effects of these shocks on the birth outcomes of interest using a fixed effect regression model, counting the days of exposure in certain temperature bins. Our main interest, however, lies in understanding whether there is heterogeneity in the effect of extreme heat on health at birth. Therefore, we study the conditional average treatment effect (CATE) of heat shocks on health at birth, which we estimate using a causal forest (Athey et al., 2019). We uncover heterogeneity in the effect and describe the most vulnerable groups and attempt to learn about possible mechanisms underlying the effect of heat shocks on health at birth by employing a recently developed decomposition approach.

We show that in-utero exposure to extreme heat events generally has adverse effects on infants' health at birth, where heat shocks can reduce birth weight by up to 6 grams. Given that exposure to extreme heat shows the strongest effects, the main analysis focuses on the effect of heat shocks on weight-related birth outcomes exclusively. We find a significant reduction in standardized birth weight due to heat shock exposure. The heat shocks considered do particularly affect the most vulnerable infants, as they significantly increase the SGA rate. We find strong heterogeneity, especially in the effect on standardized birth weight. Analysis of heterogeneity in the effects reveals that especially infants born to black, Mexican, and low-educated mothers are disproportionately prone to heat-related health issues. The decomposition, however, reveals that none of these factors on their own can explain the heterogeneity found.

The analysis of heat shocks is motivated by our evaluation of exposure to the entire

[1]The fetal origins hypothesis posits that conditions and shocks in utero can have long-lasting impacts on human capital, by influencing, for example, health, educational attainment, and adult earnings. See for example Almond and Currie (2011) for a review of this literature.

[2]A birth weight of less than 10th percentile for gestational age is considered small for gestational age.

[3]The Apgar Score is a key measure of infant health right after birth, focusing on activity and respiration of the newborn (American College of Obstetricians and Gynecologists., 2015).

temperature distribution, where especially exposure to extremely hot temperatures is found to significantly decrease birth weight. Exposure to days with temperatures exceeding 24 °C has adverse effects on infants' health at birth. The strongest effect can be found for the exposure to a day with temperatures of more than 32° C, which decreases birth weight by 0.57 grams, compared to a day in a temperature range of $0 - 24°$ C. Exposure to cold days shows smaller, yet significant increases in standardized birth weight, so exposure to temperatures below 0 tends to have mitigating effects. Exposure to extreme rainfall does not significantly affect any of the measures of health at birth used. Concerning timing, especially the second and third trimesters show a strong response to extreme temperature exposure, whereas the first trimester does not seem very vulnerable.

Even though the analysis of the average effects of extreme weather events on health has found much attention, heterogeneity in these effects is not well understood. So far, analysis of heterogeneity has mainly been limited to infant's sex, mother's race, and mother's education (Chen et al., 2020b, Deschênes et al., 2009, Le and Nguyen, 2021). Understanding the heterogeneity in the effects of extreme weather shocks on health at birth is crucial for several reasons. It can help in identifying the most vulnerable groups and provide insight into the underlying mechanisms through which these shocks affect fetal development and health outcomes at birth. This eventually leads to improved strategies and targeted interventions to prevent damage from extreme weather events.

The analysis of heterogeneity in the effects of heat shocks on health at birth reveals strong heterogeneity, where the most negative impact is concentrated among black, Mexican, and low-educated mothers. We define the shock as the average temperature of 5 consecutive days exceeding at least 30 °C and the temperature of the 9th hottest day per county. We estimate the conditional average treatment effect of a temperature shock on birth outcomes of interest using a causal forest (Athey et al., 2019), which we then analyze further to uncover drivers of heterogeneity. To gain an understanding of how different maternal characteristics might affect treatment outcomes, the study compares the average characteristics of the 10% most and 10% least affected groups. This analysis reveals that the most affected group comprises primarily black, Mexican, and low-educated mothers, suggesting that they may have limited access to protective measures against heat exposure. Additionally, we employ a newly devel-

oped decomposition approach to uncover drivers of heterogeneity. Using this decomposition, we are able to isolate the effects of a single variable, while keeping the others comparable. Our analysis suggests that the mother's race and Hispanic origin alone are not significant drivers of heterogeneity. This observation may be due to the fact that these variables serve as proxies for more complex factors such as living conditions and socioeconomic status, which could be the true drivers of heterogeneity. We can nonetheless detect weak heterogeneity by mother's age, where increased age amplifies the effect of heat shock exposure and mitigating effects of excessive weight gain.

A large body of literature analyzes how extreme weather events affect health and well-being. Focusing on mortality rates and hospital admissions Karlsson and Ziebarth (2018) investigate effects of high ambient temperature in Germany. They find that extreme heat immediately increases hospitalizations and deaths. Similarly, Deschênes and Greenstone (2011) analyze temperature effects on mortality in the US, finding that extremes on both ends of the temperature scale lead to increased mortality rates. Following the fetal origins hypothesis by David J. Barker, a growing part of the economic literature analyzes how in utero exposure to extreme weather events affects later life outcomes (Wilde et al., 2017, Isen et al., 2017, Chang et al., 2022). Evidence on the effect of temperature exposure on later life outcomes is mixed. Wilde et al. (2017) find exposure to increased temperature leads to higher educational attainment and literacy, whereas Isen et al. (2017) find such exposure leads to reduced adult earnings.

Most closely related to this study is a strand of literature that analyzes how extreme weather events during pregnancy affect birth outcomes. Studies focus on different countries of interest, like the US (Currie and Rossin-Slater, 2013, Deschênes et al., 2009), China (Chen et al., 2020b), Vietnam (Le and Nguyen, 2021), and sub-Saharan Africa (Bratti et al., 2021). These studies use different ways to measure the exposure to extreme temperature events, i.e., Deschênes et al. (2009), Chen et al. (2020b) use the count of days in certain temperature bins, Andalón et al. (2016), Le and Nguyen (2021), Molina and Saldarriaga (2017) use deviations from the long-term mean. All studies share the same conclusion: extreme weather events can have detrimental effects on health at birth. Especially for birth weight, studies find significant and robust reductions in birth weight and increases in low birth rates caused

by extreme heat events. Similar to this study, Deschênes et al. (2009) analyze the effects of extreme heat on birth weight and low birth weight in the US for births between 1972-1988. They find that exposure to extreme heat events during pregnancy reduces birth weight and increases the prevalence of low birth weight (LBW) birth. Considering climate change predictions, the overall reductions correspond to losses of 7.5 to 11.5 grams. The effect is especially profound in the second and third trimesters, which can explain up to 95% of the effect found. Overall, the literature lacks a good understanding of possible heterogeneity in the effects of temperature on health at birth. Heterogeneity in the effects has mainly been studied with regard to the mother's race, education, and sex of the child (Deschênes et al., 2009, Le and Nguyen, 2021).

While a large body of literature assesses the effects of in utero exposure to weather events (Andalón et al., 2016, Deschênes et al., 2009, Rocha and Soares, 2015, Zhang et al., 2020), they mostly neglect a fundamental problem regarding a mechanical correlation between gestation length and probability of exposure. Only Currie and Rossin-Slater (2013) discuss this problem when analyzing the effect of hurricane exposure on birth outcomes. Measuring temperature exposure during the gestational period leads to a problem since a longer gestation length increases the probability of exposure to an extreme event. This simultaneous interaction between the exposure and the mediator results in challenges with the identification of the causal effect of interest and in most cases leads to biased estimates if not handled carefully. To eliminate the simultaneity of temperature exposure and gestation length, the literature mostly imposes full 9 months of pregnancy for each birth. These fixed pregnancy lengths are either counted backward from the date of birth or forward from the date of conception. Each of these strategies leads to biases in the effect estimates, which in most cases are not discussed further. To overcome this problem, we analyze the effects on outcomes standardized for gestational age.

This paper contributes to the literature on in-utero exposure to weather shocks in three ways. First, it marks a significant contribution to the literature by introducing cutting-edge machine learning techniques to uncover the heterogeneity in the effect of environmental factors on infant health. The paper highlights the potential of machine learning approaches for advancing our knowledge in the field of environmental health research. By leveraging

recent causal machine learning techniques, we can identify the drivers of heterogeneity and characterize the most vulnerable groups, which may offer valuable insights for policymakers.

Second, this study is the first to systematically explore heterogeneity in the effect of in-utero exposure to heat waves on infant health. To the best of our knowledge, heterogeneity has been studied only by mother's race, mother's education, and sex of child (Deschênes et al., 2009, Le and Nguyen, 2021). This paper addresses an important gap in our understanding of the complex relationship between heat exposure during pregnancy and infant health outcomes. Exploiting heterogeneity by several mothers' characteristics allows us to understand the effects of extreme weather events beyond averages and helps to identify patterns of vulnerability.

Additionally, we provide new evidence for a range of birth outcomes, which control for differences in gestation length, and discuss problems arising from the mechanical correlation of gestation length and shock exposure. Only a few of the aforementioned studies consider outcomes other than birth weight and gestation length (Andalón et al., 2016, Molina and Saldarriaga, 2017). We additionally estimate the effect on standardized birth weight, standardized 5-minute Apgar score, and small for gestational age birth. We highlight the mechanical correlation between gestation length and the exposure measure and discuss issues arising from solutions usually employed in the literature.

The remainder of the paper is structured as follows: The next section provides an overview of the birth data and weather data sources. We then describe the empirical strategy and go on to describe the results. The last section discusses the results and concludes with a discussion of possible implications of the findings.

## 4.2. Data

### 4.2.1. Birth Data

Microdata on births in the US between 1989 and 2004 are taken from the National Vital Statistics System of the NCHS (National Center for Health Statistics, 2018). The data provides information on births in the United States, based on information abstracted from birth certificates. The public use files contain information on the mother's county of residence only

for counties with a population of at least 100,000 residents. The analysis therefore mainly focuses on highly populated areas on the US coasts, including a total of 525 distinct counties. See figure C.1 for an overview of counties considered. The data contains information on the socioeconomic and demographic background of the mothers, such as race, age, educational attainment, marital status, childbearing history, prenatal care, and information on medical risk factors. Further, it contains health information on the infant such as sex, gestational age, birth weight, Apgar score, plurality, and month of birth.

Details for trimester dates are calculated using the mother's last reported menses and gestation length in weeks. Since the exact date of birth is not given in the data, we need to approximate pregnancy start, trimester borders, and end dates. To determine the start of the pregnancy, we count forward from the mother's last reported menses. As Currie and Rossin-Slater (2013) point out, counting backward from the date of birth shifts trimester borders and therefore induces bias. We set the start date of the pregnancy as the month after mother's last reported menses. From this, we set the first trimester as the following three months, and the second trimester as the three months thereafter. The last trimester is determined by the month of birth and the gestation length. We, therefore, allow for differing lengths in the third trimester depending on the gestation length.

To measure infants' health at birth, we use several proxy measures. A widely used proxy measure for health at birth is birth weight. Especially low birth weight is strongly associated with adverse health outcomes Almond et al. (2005). But given the mechanical correlation between gestation length and exposure to extreme weather events, we mainly concentrate on outcomes, that account for differences in gestation length. Our main outcomes are therefore standardized birth weight[4], standardized 5-minute Apgar score and a measure for small-for-gestational age (SGA) birth. Additionally, we consider gestation length but concentrate on effects in the first and second trimesters.

We focus our analysis exclusively on singletons born in the United States. The rationale

[4]Standardized birth weight is calculated as follows: First, we calculate mean reference birth weight and its standard deviation for given gestation length and sex. The reference population contains singletons, non-smokers, and pregnancies without complications. We standardize each birth weight by subtracting the reference weight given individuals gestation length and sex and dividing by the corresponding standard deviation. The standardized birth weight now fully controls for gestation effects and possible gender effects. We use the same procedure for the standardized 5-minute Apgar score.

behind this decision is that multiple pregnancies tend to carry greater medical risks and complications for both the mother and the children involved (Kramer, 1987). We also remove the 1% most extreme outliers at both ends of the distribution for birth weight and gestation length. Moreover, we exclude all observations with missing data in the relevant variables of interest. Table 4.1 provides an overview of the birth data that we consider. Our sample comprises information on $19,865,677$ pregnancies that occurred in the US between 1989 and 2004. On average, the infants in our sample have a birth weight of 3401.86 grams and are born after 39.11 weeks of gestation. Approximately 15% of the births are preterm, and 4% are low birth weight. The average age of mothers in our sample is 28.22 years, and most of them are married and white. More than half of the mothers attended college, while 11% smoked during pregnancy.

### 4.2.2. Weather Data

Meteorological data were obtained from two sources: the North America Land Data Assimilation System (NLDAS) available on CDC WONDER (`https://wonder.cdc.gov/`) and the National Centers for Environmental Information (NCEI, `https://www.ncei.noaa.gov/`). We gather information on ambient temperature, rainfall and sunshine exposure from the North America Land Data Assimilation System, which can be accessed on county level[5] from CDC WONDER. Information on snowfall and snow density is obtained from NCEI for single weather stations. For required aggregation on the county level, we average data from all weather stations in the corresponding county. Information on snowfall is not recorded for all days in some counties. Therefore, we miss information on snowfall for 1,377,089 observed pregnancies.

We use measures for average ambient temperature and information on rainfall (in mm) to describe our main (extreme) weather events. We additionally gather monthly averages for each county and year for temperature, rainfall, snowfall, and sunshine (in KJ/$m^2$). Several studies have found effects of either of these weather conditions on health or health at birth

[5]The North America Land Data Assimilation System provides the following information on county aggregation: The county coded represents the spatial average of data observations from 14x14 kilometer square (1/8 degree) geographic-area grids. Grids are coded to the county that includes the grid centroid. For small counties where no grid centroid fell within county boundaries, county data are aggregated from grids where most of the grid area fell within county boundaries.

Table 4.1.: Summary Statistics for Birth Data

| Variables | Mean | SD |
|---|---|---|
| **Infant Characteristics** | | |
| Birth Weight in grams | 3401.86 | 482.59 |
| Standardized Birth Weight | -0.04 | 1.02 |
| Low Birth Weight | 0.04 | 0.19 |
| Small for Gestational Age | 0.11 | 0.31 |
| 5-min. APGAR Score | 8.97 | 0.55 |
| Gestation Length | 39.11 | 1.74 |
| Preterm Birth | 0.15 | 0.36 |
| Assisted Ventilation | 0.02 | 0.13 |
| Male | 0.51 | 0.50 |
| **Mothers Characteristics** | | |
| Mother's Age | 28.22 | 5.73 |
| Married | 0.81 | 0.39 |
| Race - White | 0.82 | 0.38 |
| Race - Black | 0.13 | 0.34 |
| Race - Other | 0.05 | 0.21 |
| Non-Hispanic | 0.88 | 0.32 |
| Hispanic - Mexican | 0.05 | 0.23 |
| Hispanic - Other | 0.06 | 0.24 |
| Education - <HS | 0.03 | 0.18 |
| Education - Highschool | 0.10 | 0.30 |
| Education - some College | 0.55 | 0.50 |
| Education - College + | 0.32 | 0.47 |
| Smoking during Pregnancy | 0.11 | 0.31 |
| Weight Gain in pounds | 30.95 | 12.29 |
| Prenatal Care Visits | 11.88 | 3.56 |
| Month Prenatal Care Began | 2.32 | 1.27 |
| Birth Order | 2.38 | 1.44 |
| **Weather** | | |
| Average Temperature | 12.98 | 5.04 |
| Average Rainfall | 2.90 | 1.01 |
| Average Snowfall | 1.81 | 2.20 |
| Average Sunlight | 15803.66 | 2065.49 |
| **Average Days in Temperature Bins** | | |
| (-Inf,-8] | 6.11 | 10.08 |
| (-8,-4] | 9.50 | 10.15 |
| (-4,0] | 19.28 | 15.86 |
| (0,24] | 195.43 | 40.96 |
| (24,28] | 32.64 | 29.01 |
| (28,32] | 12.12 | 21.79 |
| (32, Inf] | 1.29 | 7.70 |
| **Average Days in Rainfall (in mm) Bins** | | |
| [0,10] | 251.39 | 18.36 |
| (10,20] | 16.02 | 6.82 |
| (20,50] | 8.23 | 4.80 |
| (50,100] | 0.70 | 1.02 |
| (100, Inf] | 0.03 | 0.19 |
| Number of Observations | | 19,865,677 |

Source: NCHS, NLDAS, NCEI (1989-2004)

Table 4.2.: Summary Statistics for Birth Data and Heat Shocks

| | No Shock | | Heat Shock | | |
|---|---|---|---|---|---|
| Variables | Mean | SD | Mean | SD | Normalized Difference |
| **Infant Characteristics** | | | | | |
| Birth Weight in grams | 3403.15 | 483.16 | 3390.19 | 477.19 | -0.02 |
| Standardized Birth Weight | -0.04 | 1.02 | -0.06 | 1.01 | -0.01 |
| Low Birth Weight | 0.04 | 0.19 | 0.04 | 0.19 | -0.00 |
| Small for Gestational Age | 0.11 | 0.31 | 0.11 | 0.31 | 0.01 |
| 5-min. APGAR Score | 8.97 | 0.55 | 8.96 | 0.56 | -0.02 |
| Gestation Length | 39.12 | 1.74 | 39.09 | 1.75 | -0.01 |
| Preterm Birth | 0.15 | 0.35 | 0.15 | 0.36 | 0.01 |
| Assisted Ventilation | 0.02 | 0.13 | 0.02 | 0.13 | -0.01 |
| Male | 0.51 | 0.50 | 0.51 | 0.50 | -0.00 |
| **Mothers Characteristics** | | | | | |
| Mother's Age | 28.29 | 5.73 | 27.51 | 5.73 | -0.10 |
| Married | 0.81 | 0.39 | 0.80 | 0.40 | -0.01 |
| Race - White | 0.82 | 0.38 | 0.81 | 0.40 | -0.03 |
| Race - Black | 0.13 | 0.34 | 0.15 | 0.36 | 0.05 |
| Race - Other | 0.05 | 0.21 | 0.04 | 0.20 | -0.02 |
| Non-Hispanic | 0.89 | 0.32 | 0.84 | 0.36 | -0.09 |
| Hispanic - Mexican | 0.05 | 0.22 | 0.10 | 0.30 | 0.14 |
| Hispanic - Other | 0.06 | 0.24 | 0.05 | 0.23 | -0.03 |
| Education - <HS | 0.03 | 0.18 | 0.04 | 0.19 | 0.01 |
| Education - Highschool | 0.09 | 0.29 | 0.11 | 0.32 | 0.04 |
| Education - some College | 0.55 | 0.50 | 0.57 | 0.49 | 0.03 |
| Education - College + | 0.32 | 0.47 | 0.28 | 0.45 | -0.07 |
| Smoking during Pregnancy | 0.11 | 0.31 | 0.10 | 0.30 | -0.03 |
| Weight Gain in pounds | 30.92 | 12.24 | 31.28 | 12.66 | 0.02 |
| Prenatal Care Visits | 11.85 | 3.55 | 12.09 | 3.72 | 0.05 |
| Month Prenatal Care Began | 2.33 | 1.27 | 2.27 | 1.32 | -0.03 |
| Birth Order | 2.38 | 1.44 | 2.36 | 1.42 | -0.01 |
| **Weather** | | | | | |
| Average Temperature | 12.36 | 4.70 | 18.64 | 4.47 | 0.97 |
| Average Rainfall | 2.95 | 0.94 | 2.43 | 1.38 | -0.31 |
| Average Snowfall | 1.95 | 2.25 | 0.48 | 0.97 | -0.60 |
| Average Sunlight | 15560.66 | 1892.58 | 18019.19 | 2249.17 | 0.84 |
| Number of Observations | 17,902,118 | | 1,963,559 | | |

Source: NCHS, NLDAS, NCEI (1989-2004)

(i.e., Andalón et al., 2016, Rocha and Soares, 2015, Zhang et al., 2020).

Over the 25 years that we observe, there is a slight increase in average temperatures, see figure 4.1. Particularly average temperatures in the winter months seem to have increased. But also summer averages show a subtle increase. To motivate our shock definition, we are first interested in the effects of hot and cold temperatures simultaneously. We use the mean daily temperature and divide it into the following segments: $< 8°C$, $(-8°C, -4°C]$, $(-4°C, 0°C]$, $(0°C, 24°C]$, $(24°C, 28°C]$, $(28°C, 32°C]$, $> 32°C$. For the analysis, we set $(0°C, 24°C]$ as the reference temperature bin, since most days lie within these boundaries (see figure C.3). For each mother, we calculate the number of days in each exposure bin during pregnancy. To further examine sensitivity by the timing of the shock, we calculate the number of days of exposure in each trimester of the pregnancy.

In the main part of the analysis, we want to analyze the effect of exposure to extreme temperature events. Since especially long-term exposure to heat or extended heat periods are found to have negative effects, we define a heat shock as the average temperature of 5 consecutive days exceeding the 0.85 percentile of historic July and August temperatures and at least 30 degrees Celsius. Specifically, we assess whether the 5-day average temperature exceeds that of the 9th hottest day recorded in each county between 1979 and 1988. In addition, we require that the temperature surpasses a threshold of at least 30 degrees, which serves to exclude counties that do not typically experience extremely hot days.

Table 4.2 shows that mothers who experience a heat shock during pregnancy are on average very similar to those who do not experience such a shock. The normalized difference indicates no substantial difference in any of the infant characteristics or most of the maternal characteristics. There is slightly higher education for those mother's not experiencing a shock. Black mothers are more likely to experience a heat shock, the same holds for mothers of Mexican origin. There are substantial differences regarding average weather. The average temperature in pregnancies by mothers experiencing a shock is nearly 8 $°C$ higher, whereas average rainfall is substantially lower than for mothers not experiencing a shock. In particular, most mothers who are pregnant during summer are exposed to heat shocks.

The second weather condition we consider is daily rainfall (in mm). Similar to temperature, we first investigate the non-linear effects of rainfall. We count the number of days per

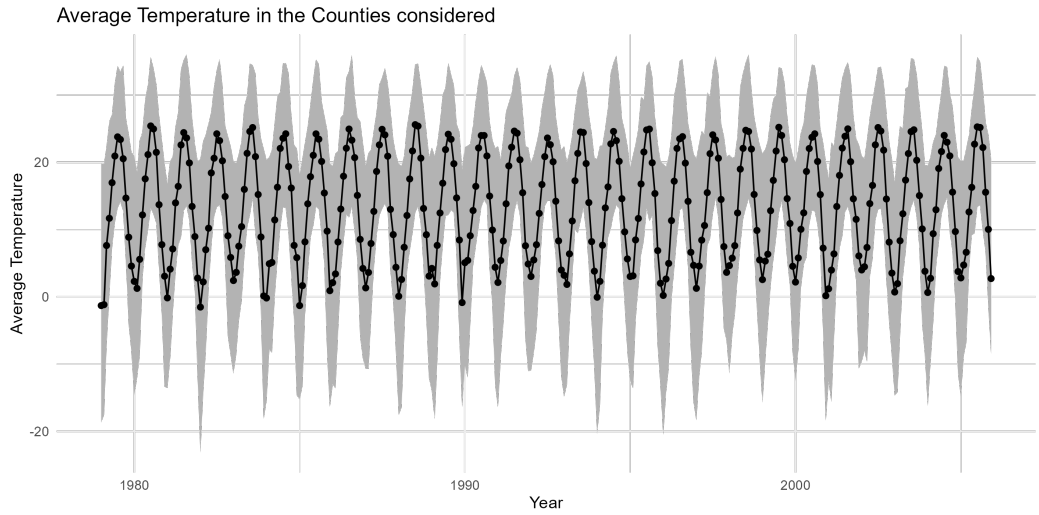


Figure 4.1.: Average Temperature

*Note:* This plot shows the average daily temperature per month in the counties considered. The grey area indicates the range for the maximum and minimum value for average temperature in a certain month.

pregnancy or trimester that each mother is exposed to the following rainfall bins: $0-10$mm, $10-20$mm, $20-50$mm, $50-100$mm, $>100$mm. These reflect a range from very light rain to very heavy rainfall. Overall, average rainfall does not change during the 25 years of interest and shows expected seasonality. As expected, most days lie in the low precipitation bin of $0-10mm$, which serves as a reference category for our analysis. As figure 4.2 shows, there is quite some deviation from the mean average rainfall. Other than for temperature, we do not see strong seasonality. While the average daily rainfall mostly lies below 5 mm, there are large outliers. Figure C.4 shows the average distribution of days in a certain rainfall bin. On average, counties experience less than a day in the most extreme rainfall bin (more than 100mm per day), indicating that this is a very rare event. Usually, counties mostly experience days of no or very light rainfall.

## 4.3. Empirical Strategy

### 4.3.1. Setup and Notation

For the analysis of heterogeneity, we focus on extreme weather shocks. We want to estimate the causal effects of these binary weather shocks in the potential outcomes framework by Imbens and Rubin (2015). For each observation $i$ we define the outcome under potential

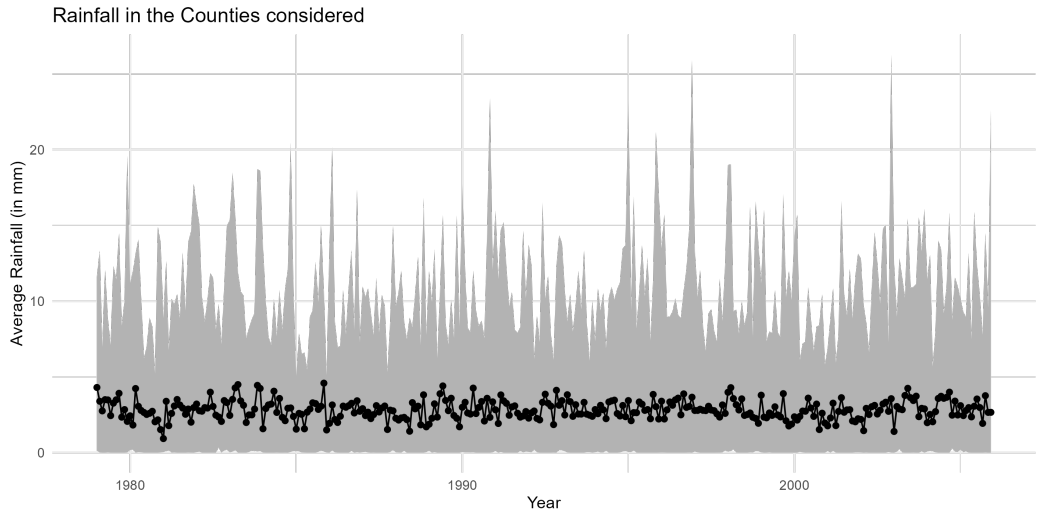


Figure 4.2.: Average Rainfall (in mm)

*Note:* This plot shows the average daily rainfall per month in the counties considered. The grey area indicates the range for the maximum and minimum value for average rainfall in a certain month.

treatment as $Y_i^{(T_i)}$, $T_i \in \{0, 1\}$, 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. In any case, we can only observe the realized outcome $Y_i^{obs} = T_i Y_i^{(1)} + (1 - T_i) Y_i^{(0)}$, since both potential outcomes can never be observed together. The propensity score $p(X_i) = P(T_i = 1|X_i)$ denotes the probability of receiving the treatment, giving the observable covariates of individual $i$. The average treatment effect (ATE) is defined as $\tau_{ATE} = \mathrm{E}[Y_i^{(1)} - Y_i^{(0)}]$, which is the expected difference between the potential outcomes. The main interest in this analysis are treatment effects that differ across individuals by their observable characteristics. Therefore, we focus on the conditional average treatment effect (CATE), defined as

$$\tau(x) = \mathrm{E}[Y_i^{(1)} - Y_i^{(0)}|X_i = x]. \tag{4.1}$$

Given that the heat shock is not a random shock, due to its correlation with the mothers' location of residence and the minimum required temperature of 30 °C, we assume unconfoundedness $\{Y_i^{(0)}, Y_i^{(1)}\} \perp T_i|X_i$ and overlap $0 < P(T_i = 1|X_i) < 1$ in the covariate distributions to be able to identify the ATE and CATE.

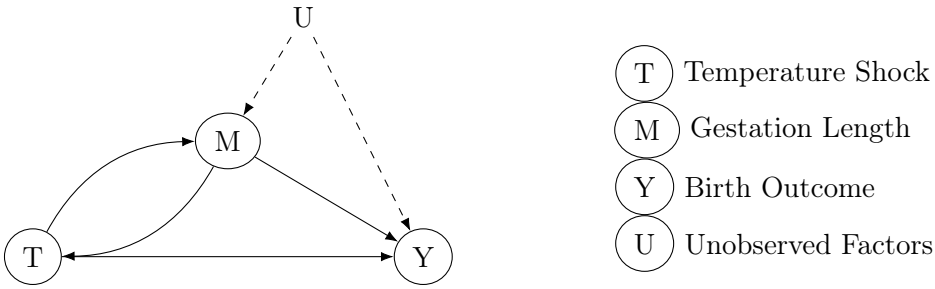


Figure 4.3.: DAG representation of the relation between Outcomes, Gestation and Temperature Exposure

### 4.3.2. Identification Challenges

A problem we face in the identification of causal effects of shocks during pregnancy on health at birth is the mechanical correlation between gestation length and our temperature exposure measures. For a longer gestation length, the probability of exposure to a temperature shock increases. Figure 4.3 is a simplified representation of the problem at hand. Temperature exposure has a direct effect on the birth outcome and an indirect effect via gestation length. There is a simultaneous interaction between temperature exposure and gestation length. Gestation acts as a mediator on the path $T \rightarrow M \rightarrow Y$ and as a confounder on the path $T \leftarrow M \rightarrow Y$. In case of gestation length only acting as a confounder affecting both the treatment and the outcome, one would want to control for it. However, given that gestation is also a possible outcome of the treatment it is a bad control. Additionally, there might be other unobserved factors that are common causes of gestation length and birth outcomes. We denote this unobserved factor as U in figure 4.3. U could for example be genetic factors or birth defects. In the literature, $M$ is referred to as a collider (Imbens, 2020). Controlling for gestation length would induce some sort of selection bias, referred to as collider stratification bias. When conditioning on gestation length, we open the causal path from $T \rightarrow M \leftarrow U \rightarrow Y$, which is a source of bias in the estimates (Imbens, 2020)[6]. So in the given setup, both conditioning and not conditioning on gestation length induces bias.

[6] A prominent example for collider bias is the low birth weight paradox. Several studies find that infants born to smokers have higher risks for LBW and infant mortality than infants born to non-smokers. However, among infants born at LBW, mortality is lower for those babies born to smokers, which would imply some beneficial effects of smoking (Hernández-Díaz et al., 2006). Similar to gestation length in figure 4.3, LBW is a collider. Smoking is not beneficial, but LBW infants born to smokers have lower mortality than LBW infants born to smokers since LBW of babies born to non-smokers is due to more detrimental causes associated with higher infant mortality, like birth defects.

The literature usually attempts to solve this problem by assuming fixed pregnancy length, i.e., 40 weeks or 9 months of pregnancy (Bratti et al., 2021, Chen et al., 2020b, Deschênes et al., 2009, Le and Nguyen, 2021). Only Currie et al. (2013) discuss the problem of the mechanical correlation between exposure and gestation length. Assuming fixed pregnancy length imposes bias, depending on how the gestation period is measured. As Currie and Rossin-Slater (2013) point out, counting backward from the date of birth shifts trimester borders, where exposure in the third trimester is most likely overstated and the strategy measures exposure that happened before the pregnancy even started. While the second strategy does not shift trimester borders, it measures exposure to extreme events after pregnancy, therefore after measurement of the outcomes of interest. Some individuals with shorter gestation lengths are therefore part of the treatment group, even though they should be in the control group, given that exposure happened after birth. This will lead to an overestimation of the treatment effect. To handle the problem of this mechanical correlation, we confine outcomes to birth weight standardized for gestational age, SGA, and 5-minute Apgar score. This way, the outcome itself attempts to control for gestational age. Appendix C.2 illustrates the problem of the mechanical correlation between gestation length and temperature exposure.

### 4.3.3. Treatment Effect Estimation using Causal Forests

To estimate heterogeneous treatment effects of the binary shocks, we use the causal forest by Athey et al. (2019). The causal forest builds on a partially linear model by Robinson (1988) with constant treatment effect $\tau(x) = \tau$ for all $x \in \mathcal{X}$:

$$Y_i = g(X_i) + T_i\tau + \epsilon_i. \tag{4.2}$$

Let $m(X_i) = E[Y_i|X_i]$ be the expected outcome and $p(X_i)$ the propensity score as defined before, then the model can be rewritten as

$$Y_i - m(X_i) = (T_i - p(X_i))\tau + \epsilon_i \tag{4.3}$$

and one can estimate $\tau$ using the following estimator, which is a semiparametrically efficient estimator under unconfoundedness (Robinson, 1988, Chernozhukov et al., 2018b)

$$\hat{\tau} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}(X_i)\right)\left(T_i - \hat{p}(X_i)\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(T_i - \hat{p}(X_i)\right)^2}. \tag{4.4}$$

Athey et al. (2019) build on this idea, but allow the treatment effect to differ with observables. To estimate $\tau(x)$ they rely on equation (4.4), but assume the treatment effect to be constant in a sufficiently small neighborhood $N(x)$. They use the random forest (Breiman, 2001) to find these neighborhoods where the treatment effect is assumed to be constant.

Random forests are a combination of many full-grown regression trees (Breiman, 2001). Regression trees split the given feature space into non-overlapping rectangular regions, also referred to as leaves. Each tree $b = 1, \ldots, n$ is grown on a random subset sampled from the original feature space $\mathcal{S}_b \subseteq 1, \ldots, n$, by greedy recursive partitioning, where we recursively choose splits, which minimize the prediction error in the sample. For every observation that falls into a certain leaf $L$, a tree returns the mean of all response values of training observations that fall into the same leaf. Averaging over all those trees' predictions leads to the prediction of the random forest for $\mu(x) = \mathrm{E}[Y_i | X_i = x]$:

$$\hat{\mu}(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{Y_i 1(\{X_i \in L_b(x), i \in \mathcal{S}_b\})}{|\{X_i \in L_b(x), i \in \mathcal{S}_b\}|}, \tag{4.5}$$

where $L_b$ is the leaf containing $x$ in tree $b$.

Thinking of the random forest as an adaptive kernel method, with data-adaptive kernel $\alpha_i(x)$, we can equivalently write

$$\hat{\mu}(x) = \sum_{i=1}^{n}\alpha_i(x) Y_i, \quad \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{1(\{X_i \in L_b(x), i \in \mathcal{S}_b\})}{|\{X_i \in L_b(x), i \in \mathcal{S}_b\}|}, \tag{4.6}$$

where the weights $\alpha_i(x)$ measures how often the $i$-th observation falls into the same leaf as $x$. These data-adaptive kernel weights can be used to characterize the neighborhood $N(x)$, and estimate the CATE function as follows:

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{n}\alpha_i(x)\left(Y_i - \hat{m}(X_i)\right)\left(T_i - \hat{p}(X_i)\right)}{\sum_{i=1}^{n}\alpha_i(x)\left(T_i - \hat{p}(X_i)\right)^2}. \tag{4.7}$$

In practice, the `grf` implementation of the causal forest fits two separate regression forests

to estimate unknown propensity scores $\hat{p}(x)$ and marginal outcomes $\hat{m}(x)$ on and makes out-of-bag predictions. With these estimates it residualizes outcomes and treatments to grow causal trees. The splits of the trees are chosen, so that the treatment effect within each leaf is homogeneous, while there is heterogeneity in treatment effects between leaves. For details see Athey et al. (2019). Several trees are grown on many subsamples of the data and averaged to build a causal forest, which is then used to derive the weights for each observation $\alpha_i(x)$.

The average treatment effect (ATE) can be estimated by a variant of doubly robust augmented inverse-propensity weighting (AIPW):

$$\begin{aligned}\hat{\tau}_{ATE} &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tau}(X_i) + \frac{T_i - \hat{p}(X_i)}{\hat{p}(X_i)(1-\hat{p}(X_i))}\left(Y_i - \hat{m}(X_i) - (T - \hat{p}(X_i))\hat{\tau}(X_i)\right)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + \frac{T_i}{p(\hat{X}_i)}(Y_i - \hat{\mu}_1(X_i)) - \frac{1-T_i}{1-p(\hat{X}_i)}(Y_i - \hat{\mu}_0(X_i))\right),\end{aligned} \tag{4.8}$$

where $\mu_{(T)}(X_i) = E[Y_i|X_i, T_i] = m(X_i) + (T_i - p(X_i))\tau(X_i)$. Cross-fitting is used to avoid overfitting, for both the ATE and CATE prediction (see Chernozhukov et al. (2018a) for details on cross-fitting.). The propensity score and outcomes models are trained on one part of the data (auxiliary sample), whereas the causal forest is estimated on another part of the data (main sample). To not loose half of the observations, we can switch the roles of the main and the auxiliary samples and report averaged results.

### 4.3.4. Assessing Treatment Effect Heterogeneity

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

$$Y - \hat{m}(x) = \beta_1(T_i - \hat{p}(X_i))\bar{\hat{\tau}}(x) + \beta_2(T_i - \hat{p}(X_i)) * (\hat{\tau}(x) - \bar{\hat{\tau}}(x)), \tag{4.9}$$

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 $\beta_1$ close to 1 suggests that the mean forest prediction is correct. The p-value of the $\beta_2$ coefficient acts as an omnibus test for the

presence of heterogeneity. If $\beta_2$ is significantly different from 0, we can reject the null of no heterogeneity.

Following chapter 2, we will additionally use an effect decomposition to uncover differences in treatment effects solely attributable to a single variable. The decomposition allows us to detect mothers' characteristics for which we can find large differences in treatment effects while keeping all other characteristics fixed. This decomposition is based on a decomposition approach by Chernozhukov et al. (2013) that makes use of the counterfactual distribution. We split the sample into non-overlapping populations based on a variable of interest and decompose the treatment effect for individuals in these populations. We decompose the treatment effect difference into the structural effect, which arises due to differences in treatment effect for individuals with the same characteristics, and the compositional effect arising from differences in the characteristics between the groups. 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.

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

$$\underbrace{F_{\tau\langle k|k\rangle}(y) - F_{\tau\langle j|j\rangle}(y)}_{\textit{Total Effect}} = \underbrace{[F_{\tau\langle k|k\rangle}(y) - F_{\tau\langle j|k\rangle}(y)]}_{\textit{Structural Effect}} + \underbrace{[F_{\tau\langle j|k\rangle}(y) - F_{\tau\langle j|j\rangle}(y)]}_{\textit{Compositional Effect}}. \tag{4.10}$$

$F_{\tau\langle k|k\rangle}(y)$ is the observed distribution function of $\tau(x)$ for population $k$. $F_{\tau\langle j|k\rangle}(y)$ represents the counterfactual distribution function of $\tau$ that would have prevailed for population $k$, if they faced populations $j$'s distribution of $\tau$. It can also be interpreted as the distribution function of $\tau$ 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. For details on the procedure, see algorithm 3 in appendix C.3 and section 2.4.3.

### 4.3.5. Fixed Effects Regression Analysis

To be able to compare estimates with previous studies and to get a better understanding of how different temperature bins affect health at birth to determine meaningful binary weather

shocks, we model the effect of temperature and rain exposure during pregnancy as follows:

$$Y_{icmy} = \alpha + \sum_{k=1}^{n} \beta_k T_{icmy} + \gamma W_{cmy} + \theta X_{icmy} + \eta_c + \mu_m + \nu_y + \sigma_s + \mu_m x \eta_c + \epsilon_{icmy}, \quad (4.11)$$

where the dependent variable $Y_{icmy}$ is the birth outcome of infant $i$, in county $c$ and month $m$ and year $y$. The key variable of interest is $T_{icmy}$, which is the count of days during pregnancy, where the temperature or rainfall lies in bin $k$. The vector $W_{cmy}$ contains several weather controls, such as average rainfall, snowfall, and sunshine in county $c$, month $m$, and year $y$. $X_{icmy}$ contains a set of maternal and child characteristics, such as mother's age, educational attainment, marital status, sex of child, prenatal care, and birth order. Additionally, we control for several fixed effects, on the county of residence $\eta_c$, month of birth $\mu_m$, year of birth $\nu_y$, and state level $\sigma_s$. Standard errors are clustered at the county level. By controlling for these fixed effects, we control for seasonality in birth outcomes, structural differences between places of residence and the selection into places of residence and hence climate of residence.

To further assess how different timing of certain shocks during pregnancy alter the effect, we estimate the following specification:

$$Y_{icmy} = \alpha + \sum_{k=1}^{n} \beta_k^{TR} T_{icmy}^{TR} + \gamma W_{cmy} + \theta X_{icmy} + \eta_c + \mu_m + \nu_y + \sigma_s + \mu_m x \eta_c + \epsilon_{icmy}, \quad (4.12)$$

where we now look at the effects in the different trimesters of pregnancy. The variables of interest $T_{icmy}^{TR}$ either corresponds to $T_{icmy}^{TR1}$, $T_{icmy}^{TR2}$, or $T_{icmy}^{TR3}$ which are counting the days the temperature (or rainfall) falls in bin $k$ in each trimester of pregnancy. $TR1$, $TR2$, and $TR3$ are the indicators for the first, second, and third trimesters of pregnancy respectively. All other variables are defined as in equation (4.11).

## 4.4. Results

### 4.4.1. Temperature Effects

Table 4.3 presents estimates of the effect of temperature on several birth outcomes of interest. The results displayed follow equation (4.11), so the estimates show the effect of an additional

day in a certain temperature bin compared to a day in the reference bin, which is 0-24°C. Column (1) shows the effect of temperature on standardized birth weight. Consistent with the literature, we find significant, but economically small negative effects of an additional day in hot temperatures on standardized birth weight, compared to a day in the reference temperature. Especially very hot temperatures show a significant reduction of $-0.001$ (corresponding to $-0.48$ grams), whereas the effect of exposure to less hot temperatures is smaller. On the other extreme of the temperature distribution, we observe positive effects, but these do not increase with lower temperatures. Controlling for mother's characteristics (column (2)) and average weather in the pregnancy months does not change the estimates much. Regarding the effect on SGA birth (columns (3) and (4)), we observe effects very close to zero. However, the same pattern as for birth weight is visible. Hot temperatures increase SGA births, whereas lower temperatures decrease SGA births. For the standardized Apgar score, we see mostly insignificant and very small effects of exposure to temperature extremes. Only extreme heat (above 32°C) shows significant positive effects on the Apgar score in both specifications.

In summary, there are negative effects of exposure to hot temperatures and small positive effects of cold exposure on birth weight. Effects are particularly strong for extreme temperatures above 32 °C. Overall, effects most profound in the second and third trimesters (see appendix C.4). This is well in line with findings from the literature. Comparing the effect size found to other studies, we find very similar economically small effects as previous studies. Deschênes et al. (2009) also study births in the US and find reductions in birth weight of around 0.003 to 0.009 percent for exposure to temperatures higher than 30 °C. This corresponds to changes of 0.102 to 0.3 grams. Chen et al. (2020b) study births in rural China and find reductions of birth weight by 1.66 grams for temperatures above 28 °C. Estimates from both studies perfectly align with our average effects found. We see slightly different estimates, given that our temperature bins are not exactly comparable.

### 4.4.2. Heat Shock Effects

We want to understand possible heterogeneity in the effect of temperature on birth outcomes. This analysis focuses on temperature exposure only, given that rainfall exposure has no

Table 4.3.: Temperature Effects

| Dependent Variables: | Standardized Birth Weight | | SGA | | Standardized Apgar Score | |
|---|---|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) | (5) | (6) |
| *Variables* | | | | | | |
| Avg. Temp. <-8°C | 0.4083*** | 0.1418 | -0.0659*** | -0.0027 | -0.2954 | -0.1987 |
| | (0.0840) | (0.1174) | (0.0243) | (0.0252) | (0.2112) | (0.2436) |
| Avg. Temp. -8 - -4 °C | 0.5469*** | 0.3567*** | -0.1315*** | -0.0970*** | 0.1979 | 0.3773 |
| | (0.0973) | (0.0890) | (0.0238) | (0.0249) | (0.2700) | (0.2736) |
| Avg. Temp. -4 - 0 °C | 0.6056*** | 0.2629*** | -0.1264*** | -0.0499*** | -0.0059 | 0.1224 |
| | (0.0703) | (0.0968) | (0.0156) | (0.0192) | (0.1449) | (0.1592) |
| Avg. Temp. 24 - 28 °C | -0.1439** | -0.3211*** | -0.0127 | 0.0357** | -0.1129* | -0.1012 |
| | (0.0644) | (0.0787) | (0.0123) | (0.0149) | (0.0625) | (0.0708) |
| Avg. Temp. 28 - 32 °C | -0.2833*** | -0.4122*** | 0.0123 | 0.0436** | -0.0714 | -0.0196 |
| | (0.0823) | (0.0932) | (0.0163) | (0.0177) | (0.0914) | (0.1175) |
| Avg. Temp. >32 °C | -1.0404*** | -1.1810*** | 0.1349*** | 0.1758*** | 0.4340** | 0.5180*** |
| | (0.1109) | (0.1142) | (0.0263) | (0.0217) | (0.1734) | (0.1806) |
| Weather Controls | | Yes | | Yes | | Yes |
| Mother's Characteristics | | Yes | | Yes | | Yes |
| *Fixed-effects* | | | | | | |
| Year of Birth | Yes | Yes | Yes | Yes | Yes | Yes |
| Month of Birth | Yes | Yes | Yes | Yes | Yes | Yes |
| County | Yes | Yes | Yes | Yes | Yes | Yes |
| State | Yes | Yes | Yes | Yes | Yes | Yes |
| Month of Birth-County | Yes | Yes | Yes | Yes | Yes | Yes |
| *Fit statistics* | | | | | | |
| Observations | 19,865,677 | 18,595,704 | 19,865,677 | 18,595,704 | 19,865,677 | 18,595,704 |
| $R^2$ | 0.0092 | 0.0900 | 0.0039 | 0.0376 | 0.0265 | 0.0283 |
| Within $R^2$ | 0.0000 | 0.0815 | 0.0000 | 0.0338 | 0.0000 | 0.0025 |

*Notes:* Entries show the coefficient on the relevant Temp. exposure measure, scaled by 1000 for ease of reading. This means that a one-unit change in *Avg. Temperature <-8° C* is associated with a 0.0004 unit change in standardized birth weight in column (1), holding all other variables constant. Samples uses all birth in counties with more than 100,000 inhabitants and no missing information in the variables of interest. Weather controls include average rainfall, average sunlight, and average snowfall in the 9 month after pregnancy start. Mother's controls include mother's age, mother's race, mother's Hispanic origin, mother's education, marital status, sex of child, month prenatal care began, number of prenatal visits, birth order, smoking during pregnancy, diabetes, and hypertension.
Standard errors clustered by County of residence in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

significant effect on the outcomes of interest (see appendix C.5). The results in table 4.3 motivate our shock definition. Since exposure to hot temperatures showed the largest, robust effects on birth weight, we will focus on the heterogeneity in the effect of extreme temperature shocks on birth weight-related outcomes.

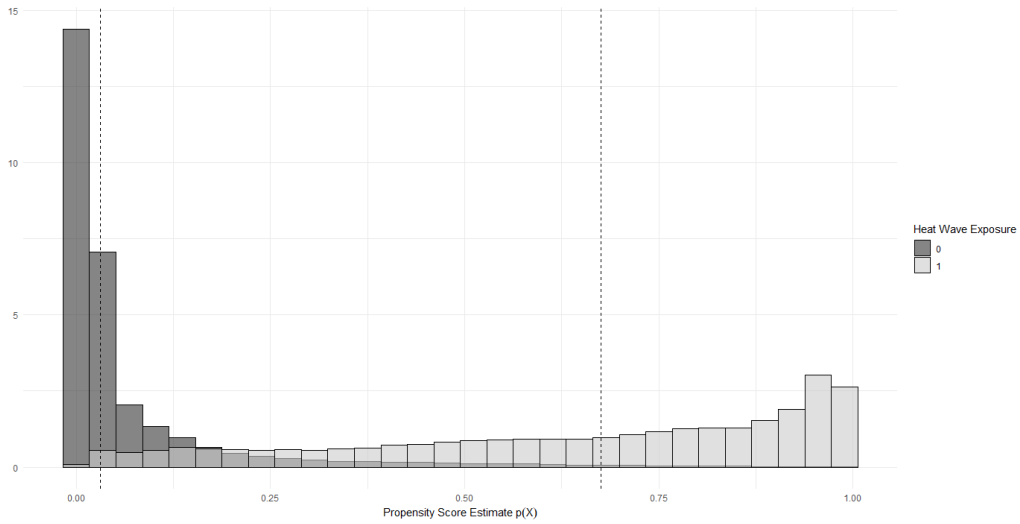


Figure 4.4.: Propensity Score estimate for heat shock exposure
*Note:* The propensity of heat shock exposure was estimated using a random forest. Dashed lines indicate the borders to trim the propensity score for sufficient overlap.

When estimating the propensity score for the entire sample we encounter problems of weak overlap. A lot of mass is concentrated at zero. This is because the temperature shock is not exogenous, given its strong correlation with the location and some locations never experience a heat shock. To have better overlap, we therefore exclude states, that either show no exposure at all or very weak overlap. The excluded states are Connecticut, Delaware, Maine, Massachusetts, New Hampshire, New York, Rhode Island, Pennsylvania, Vermont, and Virginia. So, we exclude all states in the North East. To further ensure sufficient overlap, we employ an asymmetric propensity score trimming approach by Stürmer et al. (2010). They first discard all observations in regions without overlap in estimated propensities of treated and untreated observations. In the second step, they derive an upper and lower bound to trim the propensity score estimates. For the lower end, they use the lowest 0.01 percentile of the propensity score estimated for treated observations. At the upper end of the propensity score distribution, they use a restriction based on the highest 0.99 percentile of propensity scores in the untreated population. Following this approach results in the borders

indicated in figure 4.4.

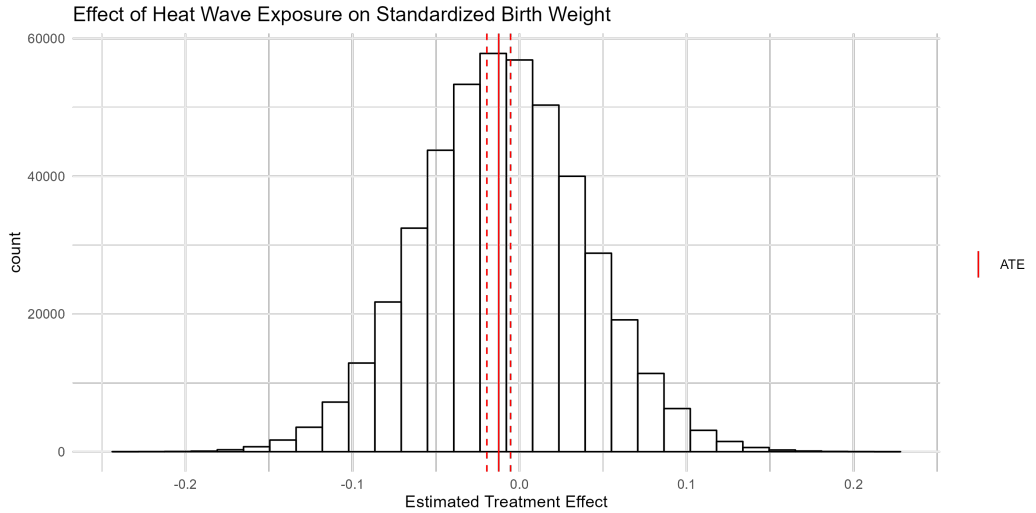


Figure 4.5.: CATE of heat shock on standardized birth weight

*Note:* The histogram shows the estimated CATE of heat shock on standardized birth weight. CATE is estimated using a causal forest using the `grf` R-package. The solid red line shows the ATE and corresponding 95% confidence intervals. These are estimated using the `grf` R-package by doubly robust augmented inverse propensity weighting.

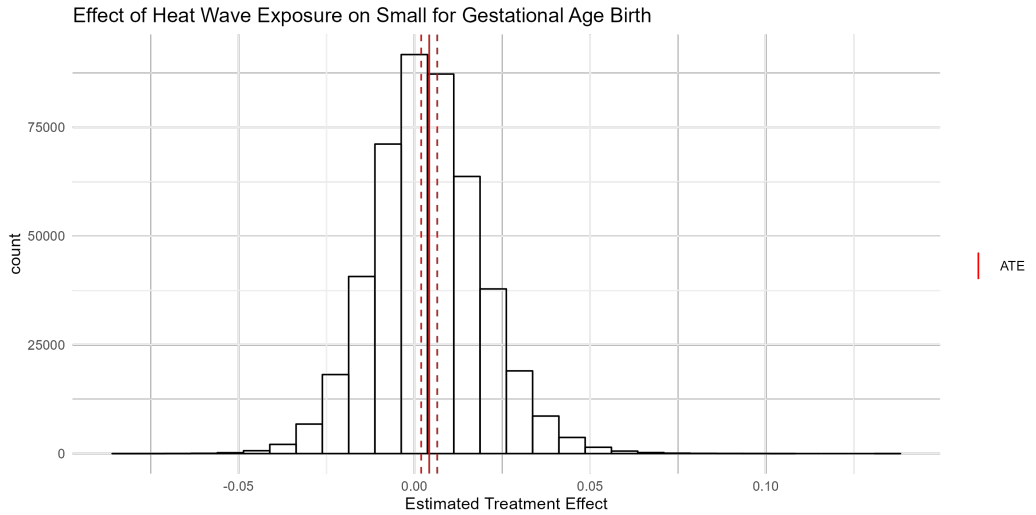


Figure 4.6.: CATE of heat shock on SGA birth

*Note:* The histogram shows the estimated CATE of heat shock on SGA birth. CATE is estimated using a causal forest using the `grf` R-package. The solid red line shows the ATE and corresponding 95% confidence intervals. These are estimated using the `grf` R-package by doubly robust augmented inverse propensity weighting.

Figure 4.5 shows a histogram of the estimated CATE of the heat shock on standardized birth weight and the ATE with corresponding 95% confidence intervals. The average treatment effect is $-0.0124$ (see table 4.4) and is significantly different from zero. It corre-

Table 4.4.: Average Treatment Effect of Heat Shock during pregnancy

| | ATE | Std. Error |
|---|---|---|
| Standardized Birth Weight | -0.0124*** | 0.0042 |
| SGA | 0.0042*** | 0.0014 |

*Notes: ATE is estimated using the inverse-propensity weighting in equation* (4.8).
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

sponds to a reduction of around 6 grams of birth weight. The calibration test in table 4.5 indicates heterogeneity being present. The coefficient for 'mean.forest.prediction' is close to 1, suggesting, that the mean forest prediction is correct. We can additionally reject the null of no heterogeneity being present since the coefficient of 'differential.forest.prediction' is significantly greater than 0.

Figure 4.6 displays the CATE estimates for the effect of a heat shock in small for gestational age birth. We observe a small but significant increase in SGA rate of 0.0042 (see table 4.4). So, on average, the heat shock affects infants on the lower end of the birth weight distribution, which are most vulnerable. The results of the calibration test in table 4.5 show that the average treatment effect estimate is correct, however the test cannot detect substantial heterogeneity in the effect.

Table 4.5.: Calibration Test Results

| Birth Outcome | Variables | Estimate | Std. Error | t value | Pr(>t) |
|---|---|---|---|---|---|
| Std. Birth Weight | mean.forest.prediction | 0.9520 | 0.4352 | 2.1877 | 0.0143 |
| | differential.forest.prediction | 0.2056 | 0.0871 | 2.3598 | 0.0091 |
| SGA Birth | mean.forest.prediction | 0.9653 | 0.3581 | 2.6953 | 0.0035 |
| | differential.forest.prediction | -0.0965 | 0.0974 | -0.9906 | 0.8391 |

*Notes: Best linear fit using forest predictions (on held-out data) as well as the mean forest prediction as regressors, along with standard errors, as described in equation* (4.9).

Given that the literature agrees on negative effects of in-utero exposure to heat events on health at birth, the results displayed in figure 4.5 are questionable. While it makes sense, that some mothers can cope better with heat wave exposure than others, questions arise from the positive effect found. These results were robust to different propensity score estimation and trimming approaches and inclusion and exclusion of several characteristics of the mother.

While we do not support the idea of positive effects of heat waves exposure, we still think it is appropriate to analyze differences in groups with strong negative (most affected) and strong positive (least affected) groups to uncover structural differences between them.

To be able to characterize the group most vulnerable to heat shocks, we will compare the average characteristics of 10% most and 10% least affected groups. Most affected refers to mothers with the strongest negative effect and least affected to mothers with estimated positive effect. Table 4.6 shows this comparison. Most and least affected groups differ a lot by their characteristics, especially when it comes to mother's race, Hispanic origin, and education. The 10% most affected group comprises a lot more black and Mexican mothers than the least affected group. Also, the education level of the most affected group is lower than that of the least affected group. On average, most affected mothers are 2 years younger and less likely to be married. We don't observe a large difference in birth order or smoking during pregnancy, but lower prenatal visits and lower weight gain in the group of most affected mothers. Comparing average exposure to different weather measures, we do not find a large difference between the two groups.

To further evaluate drivers of heterogeneity, we decompose differences in estimated treatment effects for groups into the effect of a single variable, while keeping the other characteristics comparable. So the decomposition follows the procedure in 3. Given that the differences found between most and least affected groups (table 4.6) were mostly regarding the mother's age, race, Hispanic origin, and weight gain, we restrict the decomposition analysis to these factors. Figure 4.7 shows the full decomposition for mother's age. For all other possibly modifying factors, we will only display the structural effect, as it is the main effect of interest.

Figure 4.7 follows the implementation steps as described in procedure 3, repeated 5 times using different random splits into a main and auxiliary sample. The differences refer to the baseline group of mothers aged 10-23. We report confidence bands for $\alpha = 0.05$, meaning we use bootstrap confidence bands for confidence level 0.975 but discount for splitting uncertainty, resulting in a confidence level of 0.95. The $x$-axis of the plot shows the quantile index, lower quantiles show the most harmful effects. The $y$-axis displays the differences in treatment effects. The top left plot shows the quantiles of the observed CATE cumulative

Table 4.6.: 10% Most and Least affected Mothers

| | 10% most affected | | 10% least affected | | |
|---|---|---|---|---|---|
| Variables | Mean | SD | Mean | SD | Normalized Difference |
| **Mothers Characteristics** | | | | | |
| Mother's Age | 26.61 | 6.03 | 28.61 | 6.43 | -0.23 |
| Married | 0.73 | 0.45 | 0.85 | 0.36 | -0.22 |
| Race - White | 0.69 | 0.46 | 0.85 | 0.35 | -0.28 |
| Race - Black | 0.28 | 0.45 | 0.11 | 0.32 | 0.30 |
| Race - Other | 0.03 | 0.17 | 0.03 | 0.18 | -0.01 |
| Non-Hispanic | 0.81 | 0.39 | 0.90 | 0.30 | -0.19 |
| Hispanic - Mexican | 0.15 | 0.36 | 0.03 | 0.17 | 0.30 |
| Hispanic - Other | 0.04 | 0.20 | 0.07 | 0.25 | -0.09 |
| Education - <HS | 0.07 | 0.25 | 0.02 | 0.14 | 0.16 |
| Education - Highschool | 0.15 | 0.36 | 0.08 | 0.26 | 0.17 |
| Education - some College | 0.53 | 0.50 | 0.56 | 0.50 | -0.04 |
| Education - College + | 0.25 | 0.43 | 0.35 | 0.48 | -0.15 |
| Smoking during Pregnancy | 0.07 | 0.26 | 0.10 | 0.29 | -0.05 |
| Cigarettes during Pregnancy | 1.03 | 6.02 | 1.47 | 7.38 | -0.05 |
| Weight Gain in pounds | 29.01 | 12.06 | 32.89 | 14.33 | -0.21 |
| Prenatal Care Visits | 11.45 | 3.77 | 12.79 | 3.90 | -0.25 |
| Month Prenatal Care Began | 2.51 | 1.53 | 2.16 | 1.21 | 0.18 |
| Birth Order | 2.30 | 1.34 | 2.29 | 1.44 | 0.00 |
| **Infant Characteristics** | | | | | |
| Birth Weight in grams | 3339.78 | 480.77 | 3414.37 | 480.37 | -0.11 |
| Standardized Birth Weight | -0.14 | 1.01 | -0.01 | 1.01 | -0.09 |
| Low Birth Weight | 0.05 | 0.21 | 0.04 | 0.19 | 0.03 |
| Small for Gestational Age | 0.13 | 0.33 | 0.10 | 0.30 | 0.05 |
| 5-min. APGAR Score | 8.94 | 0.55 | 8.96 | 0.54 | -0.02 |
| Gestation Length | 38.95 | 1.81 | 39.10 | 1.75 | -0.06 |
| Preterm Birth | 0.17 | 0.38 | 0.15 | 0.36 | 0.05 |
| Assisted Ventilation | 0.02 | 0.14 | 0.02 | 0.13 | 0.00 |
| Male | 0.51 | 0.50 | 0.51 | 0.50 | -0.00 |
| **Weather** | | | | | |
| Average Temperature | 16.46 | 4.78 | 17.24 | 4.75 | -0.12 |
| Average Rainfall | 2.82 | 1.22 | 3.02 | 1.13 | -0.12 |
| Average Snowfall | 0.93 | 1.92 | 0.72 | 1.50 | 0.09 |
| Average Sunlight | 16902.66 | 1997.23 | 17159.06 | 1894.14 | -0.09 |

distribution (solid line) and counterfactual distribution that would have prevailed for other groups if they faced the reference group's distribution of characteristics (dashed line). The top right shows the total effect, which is the observed difference between each group's CATE distribution and the CATE distribution of the baseline group. The bottom left displays the structural effect, which is our main effect of interest. It shows the difference only associated with the mother's age. Characteristics of groups are kept comparable by using the counterfactual distribution. This difference is always relative to the reference group. The bottom right shows the compositional effect, which is the difference associated with differences in group characteristics, leaving out mother's age.

We find a weak modifying effect for mother's age in the effect of a heat shock on standardized birth weight. The total effect indicates significant differences between mothers younger than 27 and older than 27. Similarly, the compositional effect shows that there are strong differences in treatment effects given differences in group composition, especially for older mothers. However, the structural effect does not reveal significant differences between mothers in different age groups. For mothers aged 27 to 32, we can detect weakly significant mitigating effects associated with age. At lower quantiles, there are significant amplifying effects for older mothers, but these fade with increasing quantile index. For mothers aged 23 to 27, we cannot find any difference compared to our reference group.

While we detected significant differences between most and least affected groups regarding their race, Hispanic origin, and education, the decomposition does not confirm this observation. We do observe that differences in race tend to amplify effects, but these are not significantly different from zero (figure 4.8). Similarly, we cannot detect significant modification by Mexican or other Hispanic origins (figure 4.9) or any level of mother's education (figure 4.10). Only the decomposition for weight gain during pregnancy reveals weak mitigating effects for excessive weight gain (figure 4.11).

In summary, this means that while we can detect differences in most and least affected groups, we are unable to clearly describe the drivers of heterogeneity found, apart from a weak effect modification by age and weight gain. This means that heterogeneity is not driven by race, Hispanic origin, education, or corresponding associated factors on their own. Keeping other characteristics comparable fully removes any modification effect attributable

to race or other strongly associated factors. The actual drivers might be a combination of certain factors or circumstances that are more complex to measure. The decomposition thus shows that we can rule out the mother's race, Hispanic origin, and education as single driving factors of heterogeneity.

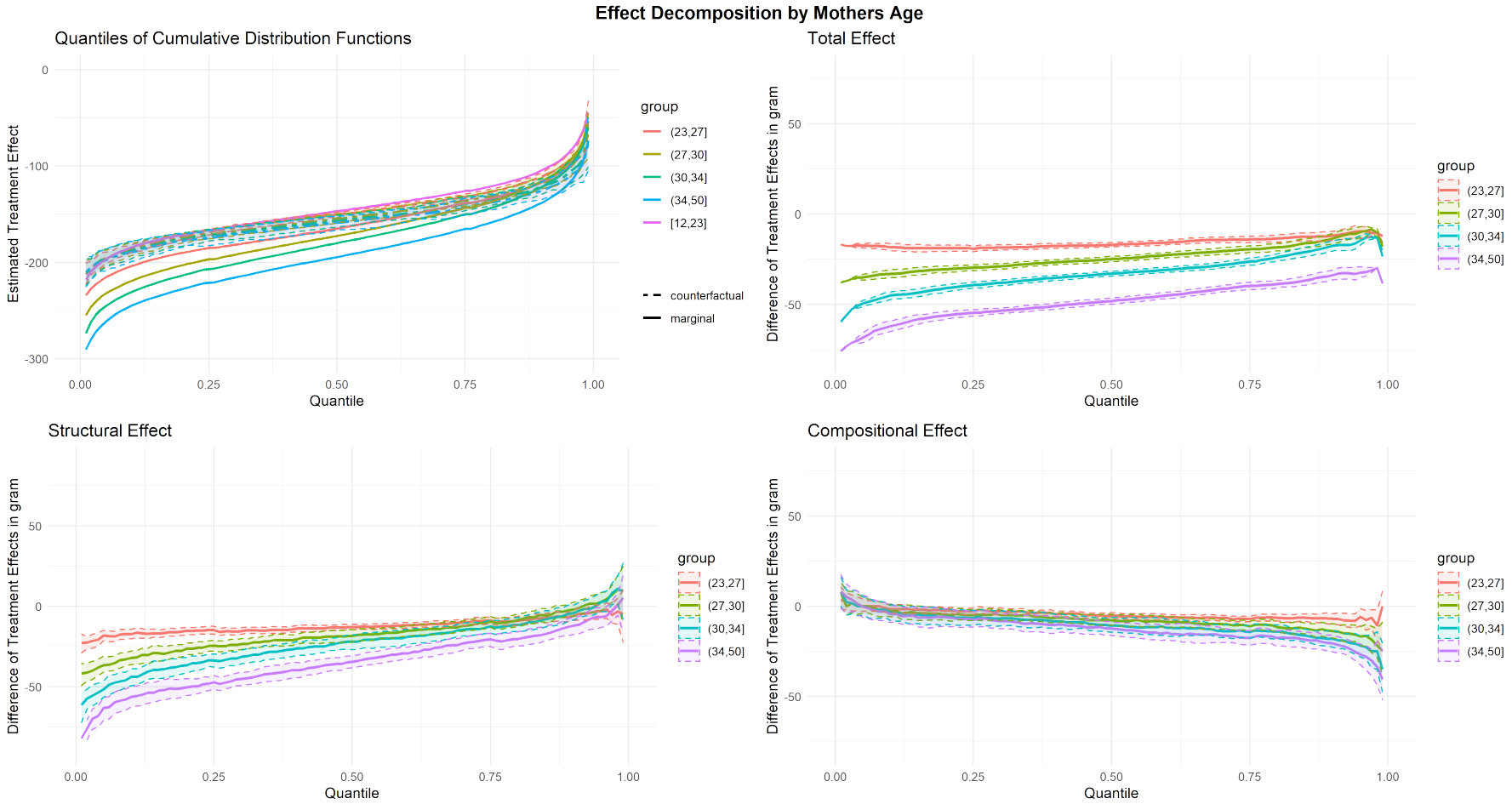


Figure 4.7.: Standardized Birth Weight - Effect Decomposition by Mother's Age

*Note:* The figure displays the decomposition of the effect of a heat shock 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 10 to 23 (the lowest age quartile). Decomposition follows the procedure described in 3. Total, structural and compositional effect are defined as in 4.10. 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.

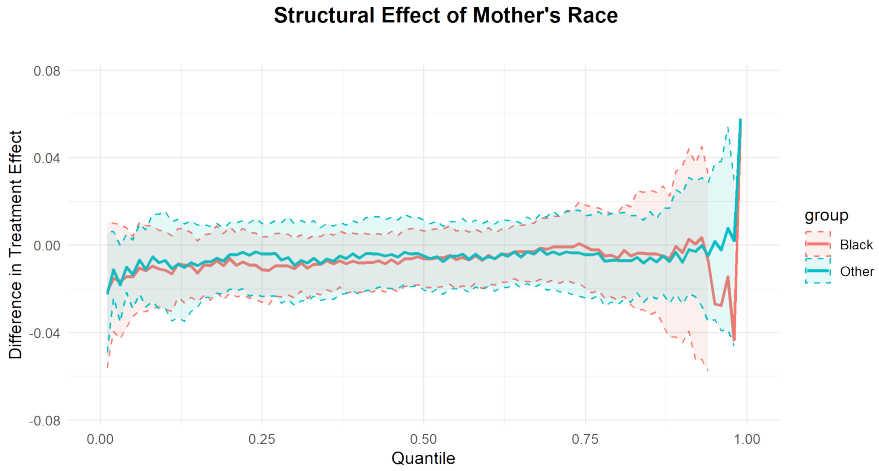


Figure 4.8.: Standardized Birth Weight - Structural Effect of Mother's Race

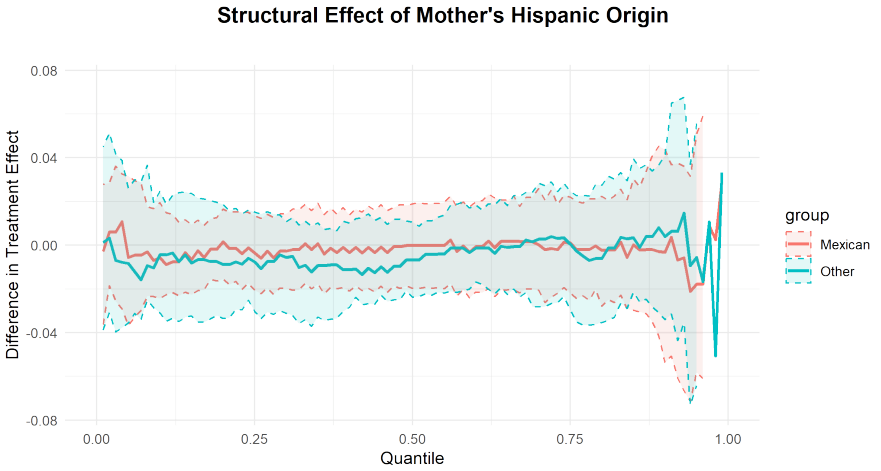


Figure 4.9.: Standardized Birth Weight - Structural Effect of Mother's Hispanic Origin

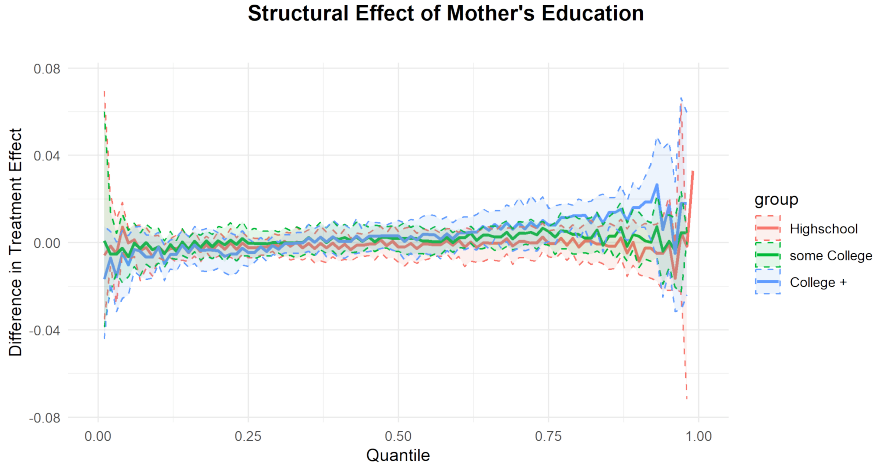


Figure 4.10.: Standardized Birth Weight - Structural Effect of Mother's Education

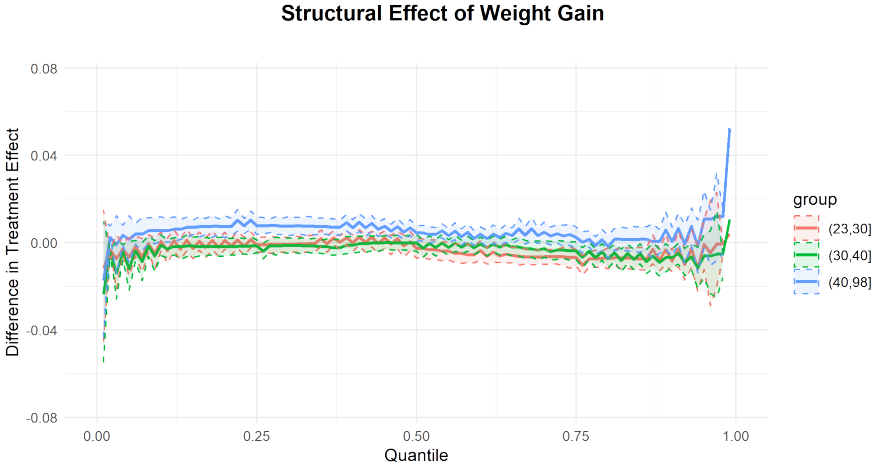


Figure 4.11.: Standardized Birth Weight - Structural Effect of Weight Gain

*Note:* The figure shows the structural effect decomposition of the effect of a heat shock during pregnancy on standardized birth weight for different potentially modifying factors. Decomposition follows the procedure described in 3. Positive difference corresponds to effect mitigation, whereas negative difference corresponds to effect amplification respectively. Shaded areas show 95% confidence intervals.

### 4.4.3. Placebo Test

Table 4.7 presents the results of a "placebo test" in which we estimate the effect of a heat shock that occurred six months after birth. A heat shock after the date of birth is known to have no causal effect on birth outcomes. But if our estimates of the treatment effect of heat shocks during pregnancy only reflect trends in the birth outcomes or an omitted variable, we might as well see significant effects of the placebo shock. In addition, this serves as an indirect test to assess plausibility of the unconfoundedness assumption. Finding a significant effect of the heat shock happening after birth would make the unconfoundedness assumption less plausible.

Table 4.7.: Placebo Test: Average Treatment Effect of shock 6 months after birth

| | ATE | Std. Error |
|---|---|---|
| Standardized Birth Weight | -0.0049 | 0.0053 |
| SGA | 0.0012 | 0.0017 |

*Notes: ATE is estimated using the inverse-propensity weighting in equation* (4.8).
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

We find no significant average treatment effect of the heat shock 6 months after birth on either standardized birth weight or small for gestational age birth. As expected, the magnitude of effects is smaller than in the main analysis. Overall, the placebo test supports the effectiveness of our strategy in identifying causal impacts of heat shock exposure on health at birth.

## 4.5. Discussion

How does temperature exposure during pregnancy influence infants' health at birth? And who is most prone to climate change related health risks? This study evaluates the effects of temperatures and rain exposure on several outcomes for health at birth and study heterogeneity in these effects. It is the first to analyze heterogeneity in the effects of temperature shocks on infants' health at birth in a structured way and to shed a light on the possible mechanism of how these exposures work.

We show that there is strong heterogeneity in the effect of heat shocks on health at birth, which has not been studied in detail before. Analysis reveals that especially babies born to black, Mexican, and low-educated mothers are disproportionately prone to climate-related health risks. Given the combination of race, Hispanic origin, and lower education, it seems like disadvantaged populations are most prone to health risks of in-utero exposure to health at birth. A possible explanation might be the lack of protection. These mothers might not have the possibility to stay inside, since they might need to work outside and do not have access to air conditioning in their homes. Unfortunately, we do not have data available to explore this mechanism further. These racial health disparities have not been explored in detail in the literature, but given the diversity in the US population, it is imperative to understand how climate change impacts infants born to mothers of different racial and ethnic backgrounds. This can aid in identifying whom to target for the prevention of heat-associated morbidity.

For the fixed effects regression estimation approach, which is often used in the literature, we highlight problems arising from a mechanical correlation of gestation length and the exposure measures and discuss possible solutions. We find a significant effect on both sides of the temperature distribution, where cold temperatures positively affect standardized birth weight-related measures and gestation, whereas hot days negatively affect birth outcomes. Especially in the second and third trimesters, where the fetus grows most, the effects are strongest. Given the mechanical correlation between gestation length and the exposure, it is hard to gather insights into the effects of temperature exposure on gestation length. However, effects in the first and second trimesters, which are not prone to this mechanical correlation, show that extreme heat negatively affects gestation length.

There are however some limitations to our empirical approach. First, we only observe effects on live births in our data. Since the data does not include any information on miscarriage and stillbirth, we cannot measure the effects on most vulnerable infants, who did not survive up until the time of birth. Therefore, we expect our estimates for average effects to be a lower bound for the actual effect of temperature and weather exposure on health at birth. Second, we cannot control for mothers' mobility and actual temperature exposure. While we have information on the mother's county of residence, we cannot track

the mother's mobility patterns. Additionally, we do not know whether she experienced any hot temperatures or if she was able to protect herself from extreme temperatures through heating, air conditioning, or location changes. From this perspective, our exposure measure is only a proxy for actual temperature and weather exposure. Third, the heterogeneity analysis unexpectedly reveals possible positive effects of heat wave exposure. As the literature agrees on negative effects of extreme heat on infants' health at birth and human health in general, we question these positive treatment effects found. This might point toward a possible adaptation the causal forest algorithm. In situations where the researcher has a strong prior about the effect to be estimated, solving the estimation problem under the constraint of negative treatment effects would be a possible extension of the algorithm.

Despite these limitations, this study provides important evidence of how temperature and extreme heat events affect health at birth and shed a light on heterogeneity in these effects. It introduces cutting-edge machine learning techniques to uncover heterogeneity in the effect of environmental factors on infant health and highlights the potential of machine learning approaches. Given the growing number of extreme weather events due to climate change and overall global warming, our results point out the most vulnerable groups which should be protected further from temperature and weather-related shocks.

# Appendix A.

# Appendix of Chapter 2

## A.1. Glossary: Medical Terminology

| | |
|---|---|
| **Apgar Score** | The Apgar score is a measure to quickly judge the health condition of the newborn right after birth. It is routinely measured 1 and 5 minutes after birth, for newborn with low scores, measurement may be continued thereafter. It is derived by assessing the newborn on five simple criteria (appearance, pulse, grimace, activity, respiration), each evaluated on a scale from 0 to 2. The final Apgar score is the sum of the five criteria, ranging from 0 to 10. A newborn with Apgar score of $7-10$ is considered healthy, a score of $4-6$ is considered moderately abnormal, whereas a score of $0-3$ as low (American Academy of Pediatrics, 2015). |
| **Body Mass Index (BMI)** | Body Mass Index (BMI) is a person's weight in kilograms (or pounds) divided by the square of height in meters (or feet). A high BMI can indicate high body fatness. BMI screens for weight categories that may lead to health problems, but it does not diagnose the body fatness or health of an individual. (`https://www.cdc.gov/healthyweight/assessing/bmi/index.html`) |
| **Fetal growth retardation** | Fetal growth retardation (FGR) is a condition in which an unborn baby (fetus) is smaller than expected for the number of weeks of pregnancy (gestational age). It is often described as an estimated weight less than the 10th percentile. (`https://www.stanfordchildrens.org/en/topic/default?id=intrauterine-growth-restriction-iugr-90-P02462`) |
| **Gestation length** | Fetal development period from the time of conception until birth. For humans, the full gestation period is normally 9 months, or 40 weeks. (`https://www.rxlist.com/gestation_period/definition.htm`) |

**Gestational hypertension** Gestational hypertension is a form of high blood pressure in pregnancy. It occurs in about 6 percent of all pregnancies. Another type of high blood pressure is chronic hypertension–high blood pressure that is present before pregnancy begins. (https://www.chop.edu/conditions-diseases/gestational-hypertension)

**Low birth weight (LBW)** Birth weight below 2500 grams.

**Miscarriage** A miscarriage is usually defined as loss of a baby before the 20th week of pregnancy. (https://www.cdc.gov/ncbddd/stillbirth/facts.html)

**Multiparous** Multiparous – the mother has previously given birth more than once (http://www.datadictionary.wales.nhs.uk/index.html#!WordDocuments/parity.htm)

**Neonate** A newborn infant, or neonate, is a child under 28 days of age. During these first 28 days of life, the child is at highest risk of dying. (https://www.who.int/westernpacific/health-topics/newborn-health)

**Nulliparous** Nulliparous – the mother has never previously given birth (http://www.datadictionary.wales.nhs.uk/index.html#!WordDocuments/parity.htm)

**Obesity** If your BMI is 30.0 or higher, it falls within the obesity range. Obesity is frequently subdivided into categories:

- Class 1: BMI of 30 to $< 35$
- Class 2: BMI of 35 to $< 40$
- Class 3: BMI of 40 or higher.

Class 3 obesity is sometimes categorized as "severe" obesity. (https://www.cdc.gov/obesity/basics/adult-defining.html)

**Overweight** If your BMI is 25.0 to $<30$, it falls within the overweight range. (https://www.cdc.gov/obesity/basics/adult-defining.html)

**Parity** Parity is the number of times a woman has given birth to a live neonate (any gestation) or at 24 weeks or more, regardless of whether the child was viable or non-viable (i.e. stillbirths). (http://www.datadictionary.wales.nhs.uk/index.html#!WordDocuments/parity.htm)

**Plurality** The number of fetuses delivered live or dead at any time in the pregnancy regardless of gestational age, or if the fetuses were delivered at different dates in the pregnancy. (https://www.cdc.gov/nchs/nvss/facility-worksheets-guide/33.htm?Sort=URL%3A%3Aasc&Categories=Newborn%20Information)

| | |
|---|---|
| **Preterm Birth** | Preterm is defined as babies born alive before 37 weeks of pregnancy are completed. There are sub-categories of preterm birth, based on gestational age:<br>• extremely preterm (less than 28 weeks)<br>• very preterm (28 to 32 weeks)<br>• moderate to late preterm (32 to 37 weeks).<br>(`https://www.who.int/news-room/fact-sheets/detail/preterm-birth`) |
| **Primiparous** | Primiparous – the mother has previously given birth once only (`http://www.datadictionary.wales.nhs.uk/index.html#!WordDocuments/parity.htm`) |
| **Resuscitation** | Neonatal resuscitation is defined as the set of interventions at the time of birth to support the establishment of breathing and circulation. (`https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-11-S3-S12`) |
| **Small for gestational age (SGA)** | Small for gestational age is a term used to describe a baby who is smaller than the usual amount for the number of weeks of pregnancy. SGA babies usually have birthweights below the 10th percentile for babies of the same gestational age. This means that they are smaller than many other babies of the same gestational age. (`https://www.chop.edu/conditions-diseases/small-gestational-age`) |
| **Stillbirth** | A stillbirth is the death or loss of a baby before or during delivery. It is defined as loss of a baby at or after 20 weeks of pregnancy. (`https://www.cdc.gov/ncbddd/stillbirth/facts.html`) |

## A.2. Overview: Effect of Smoking on Birth Weight

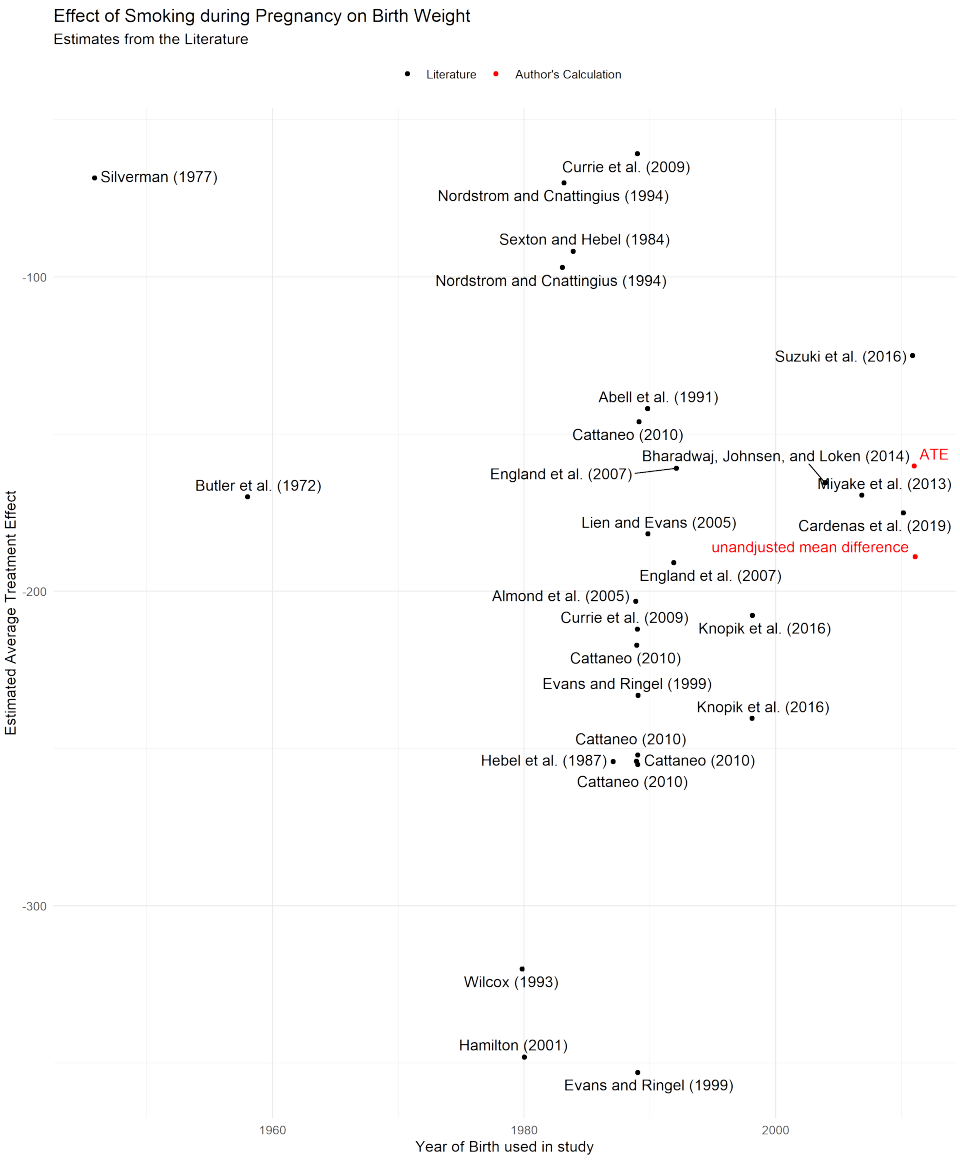


Figure A.1.: Overview of Estimates of the effect of smoking on birth weight

*Note:* The figure provides an overview of treatment effect estimates for the effect of smoking on birth weight in the economics and medical literature (Sexton and Hebel, 1984, Hebel et al., 1988, Abell et al., 1991, Wilcox, 1993, Nordström and Cnattingius, 1994, Evans and Ringel, 1999, Hamilton, 2001, Almond et al., 2005, Lien and Evans, 2005, England et al., 2007, Currie et al., 2009, Cattaneo, 2010, Miyake et al., 2013, Bharadwaj et al., 2014, Knopik et al., 2016, Suzuki et al., 2016, Cardenas et al., 2019). Some studies provide several point estimates for different reasons. Cattaneo (2010) and Nordström and Cnattingius (1994) provide dose dependent estimates, Currie et al. (2009) provide estimates with and without controlling for mothers fixed effects, Evans and Ringel (1999) provide OLS and IV estimates, England et al. (2007) use different smoking indicators (self-reported and based on cotinine levels), and Knopik et al. (2016) use models with smoking information provided by different family members.

## A.3. Decomposition: Alternative Health Outcomes

Table A.2.: Overview: Empirical Results alternative Health Outcomes

| **Effect of Smoking on** | **Decomposition by** Mother's Age | Parity | Prepregnancy BMI | Weight Gain | Weight Gain Recommendation | Sex |
|---|---|---|---|---|---|---|
| **Birth Weight** | ++ strong amplifying effects of increased mother's age | ++ strong amplifying effects of increased parity | **- -** strong mitigating effects of increased BMI | **- -** strong mitigating effects of increased weight gain | **-** mitigating effects of gaining above recommendation | **0** no effect difference |
| **Gestation Length** | + amplifying effects of increased mother's age | **0** no modifying effect of increased parity | **- -** strong mitigating effects of increased BMI | **0** no modifying effect of weight gain | **0** no modifying effect of weight gain | **0** no effect difference |

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

### A.3.1. Birth Weight

In order to catch absolute differences in the effect of smoking on birth weight related to mediating factors of interest, we additionally present result for non-standardized birth weight. This decomposition cannot fully control for gestation length.

Figure A.2 shows the conditional average treatment effect of smoking on birth weight. The figure shows substantial heterogeneity, estimates range from $< -300g$ to no effect at around 0. The average treatment effect is around $-160g$.

We find strong modifying effect for mother's age in the effect of smoking on birth weight (Figure A.3). The total effect indicates significant differences between the age groups, which the structural effect confirms. There is a monotone increase in the effect of smoking on birth weight via mother's age, making older mothers at great risk for harming their fetus severely when smoking. There is $-75$ grams difference in the effect of smoking on birth weight can be

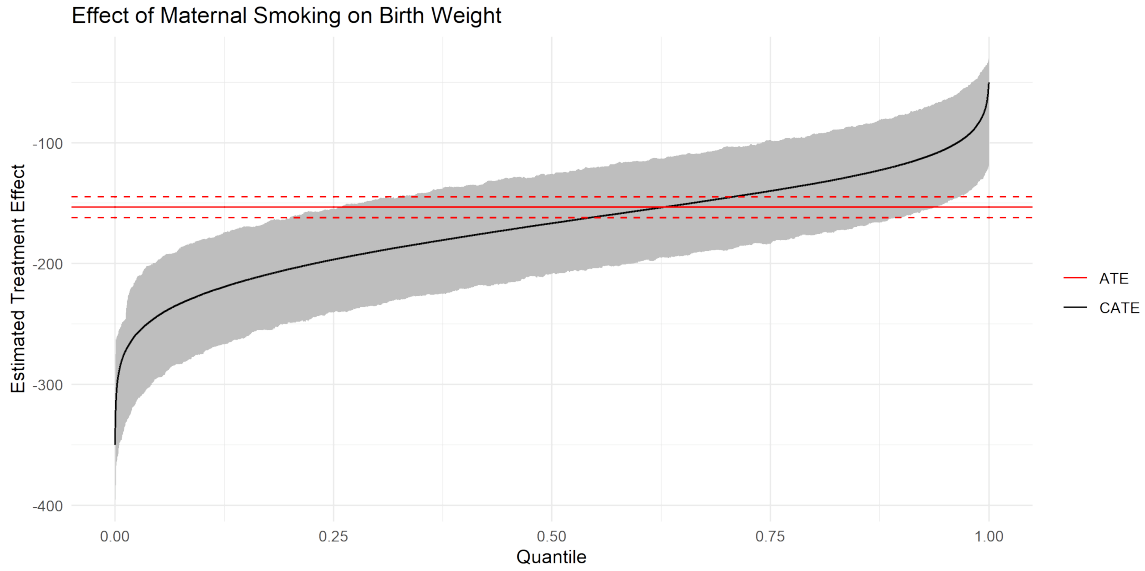


Figure A.2.: Sorted CATE and ATE estimates of smoking on birth weight

*Note:* The solid black line shows the estimated CATE of smoking on 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 2.4.2 using the `grf` R-package. The solid red line shows the ATE and corresponding 95% confidence intervals. These are estimated using the `grf` R-package by doubly robust augmented inverse propensity weighting.

attributed to mother's age in the group of mothers being 34 or older compared to mothers younger than 23.

We find strong amplification with increased parity (Figure A.5). For primipara and secundipara, no large difference can be found. The strongest modification of the effect of smoking on birth weight can be found for multipara, where around $-30$ grams of the total difference can be attributed to multiparity. For weight gain during pregnancy (Figure A.6), one can see that increased weight gain has mitigating effects up to 50g for the fetus of smoking mothers. The mitigation is stable of the quantile indices. Mitigation seems to increase linearly with weight gain.

Similarly, the prepregnancy BMI shows strong mitigating effects (Figure A.7). Compared to being underweight, a BMI considered normal weight can reduce the effect of smoking on birth weight by up to 40 grams. The effect is similar to the mitigating effect of being overweight. The structural effect is stable over the quantile indices for both groups. The decomposition in the dimension of sex of newborn does not show significant differences between male and female fetuses (Figure A.9).

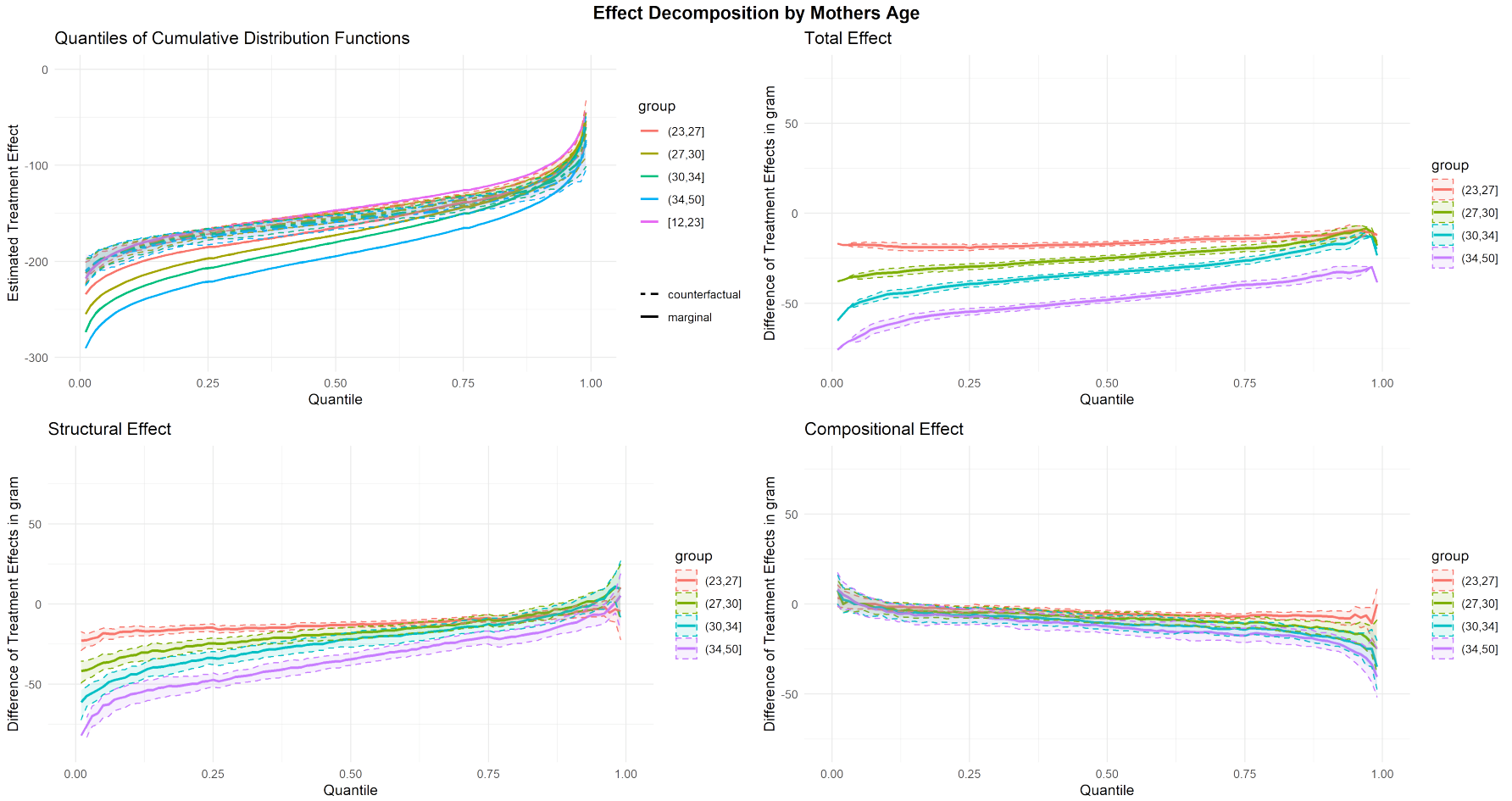


Figure A.3.: Birth Weight - Effect Decomposition by Mother's Age

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on 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 equation (2.12). 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.

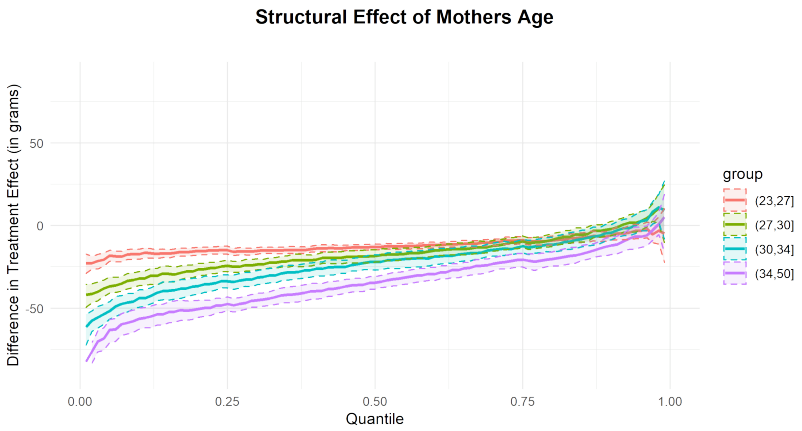


Figure A.4.: Birth Weight - Structural Effect by Mother's Age

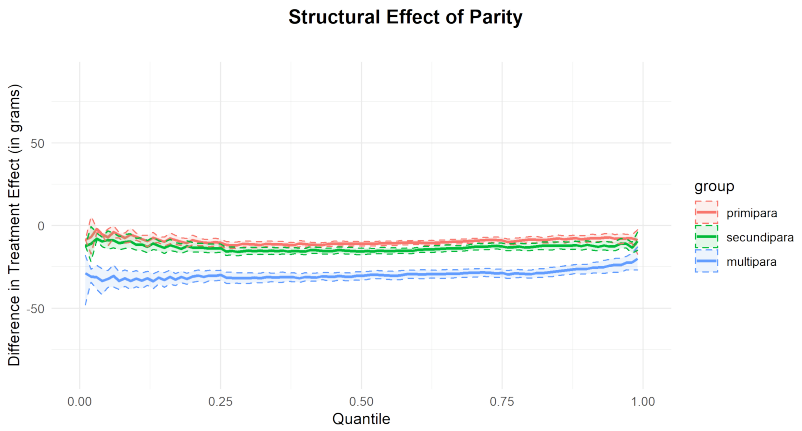


Figure A.5.: Birth Weight - Structural Effect by Parity

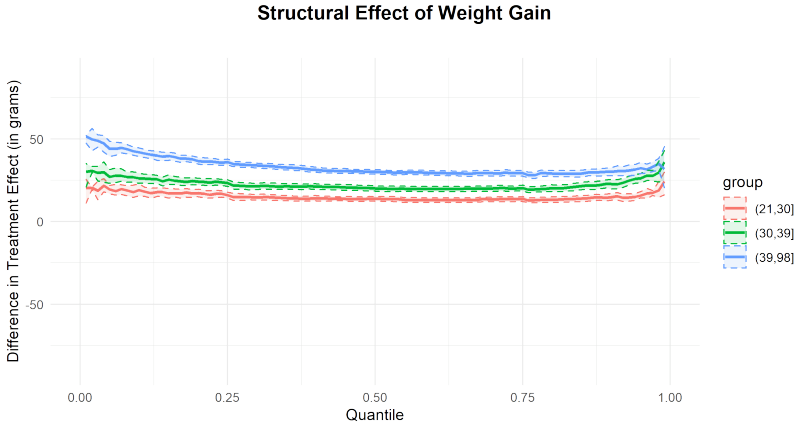


Figure A.6.: Birth Weight - Structural Effect by Weight Gain

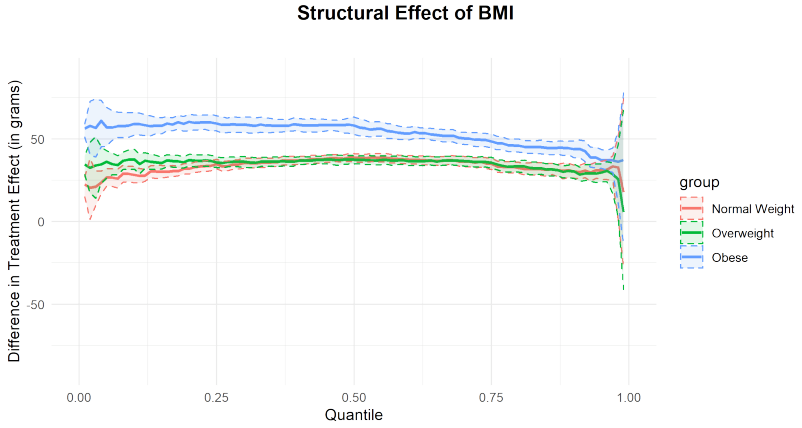


Figure A.7.: Birth Weight - Structural Effect by BMI

*Note:* The figure shows the structural effect derived by decomposing the effect of maternal smoking during pregnancy on birth weight. 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.

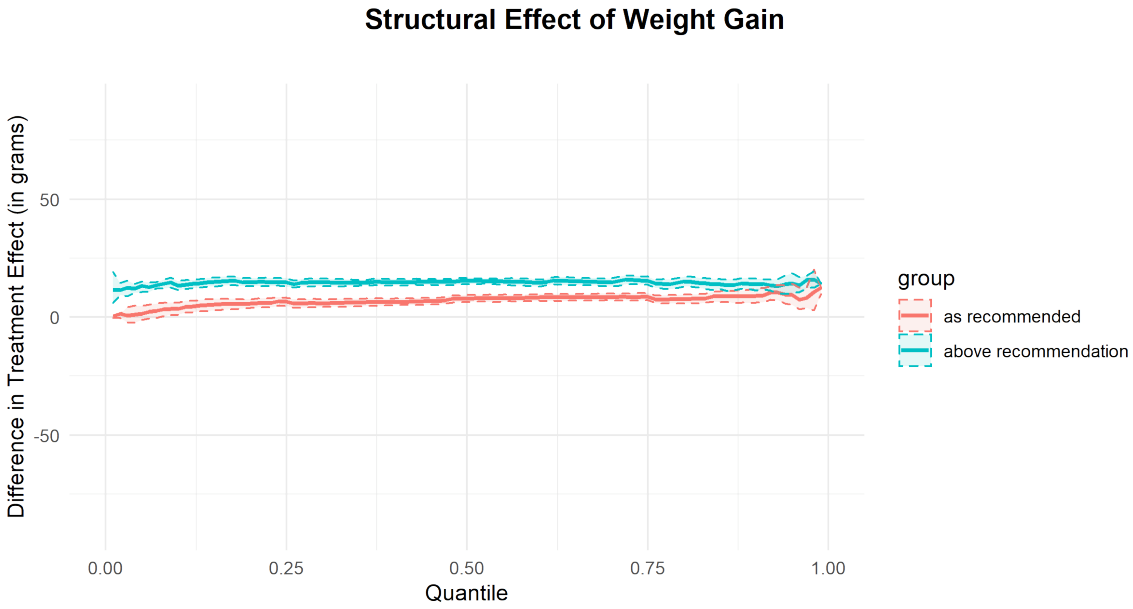


Figure A.8.: Birth Weight - Structural Effect of Weight Gain Recommendations

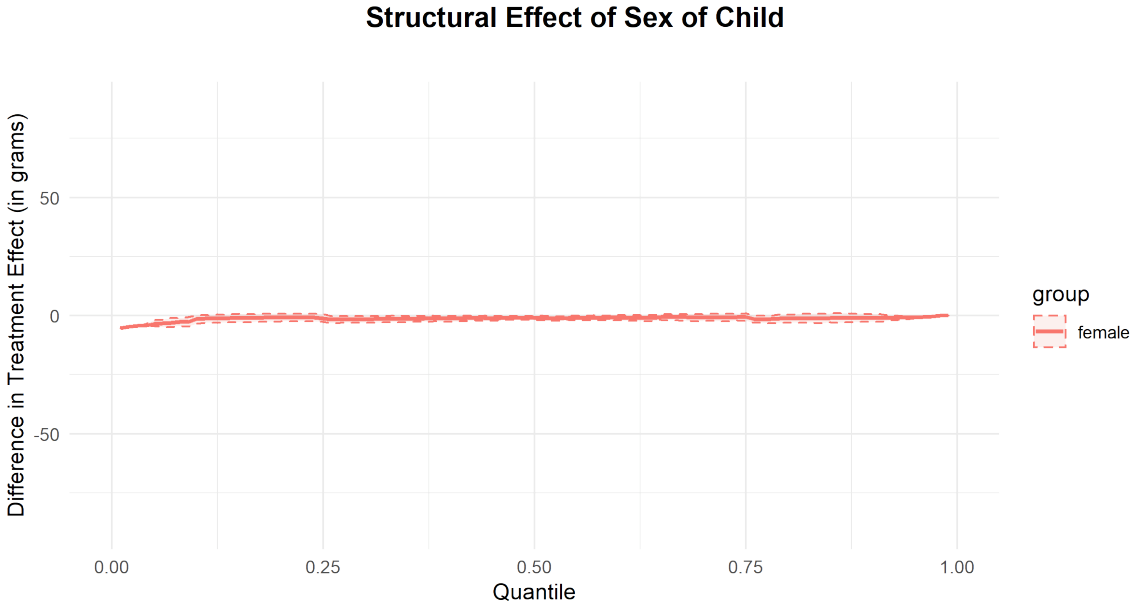


Figure A.9.: Birth Weight - Structural Effect of Sex of Newborn

*Note:* The figure shows the structural effect derived by decomposing the effect of maternal smoking during pregnancy on birth weight. 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.

### A.3.2. Gestation Length

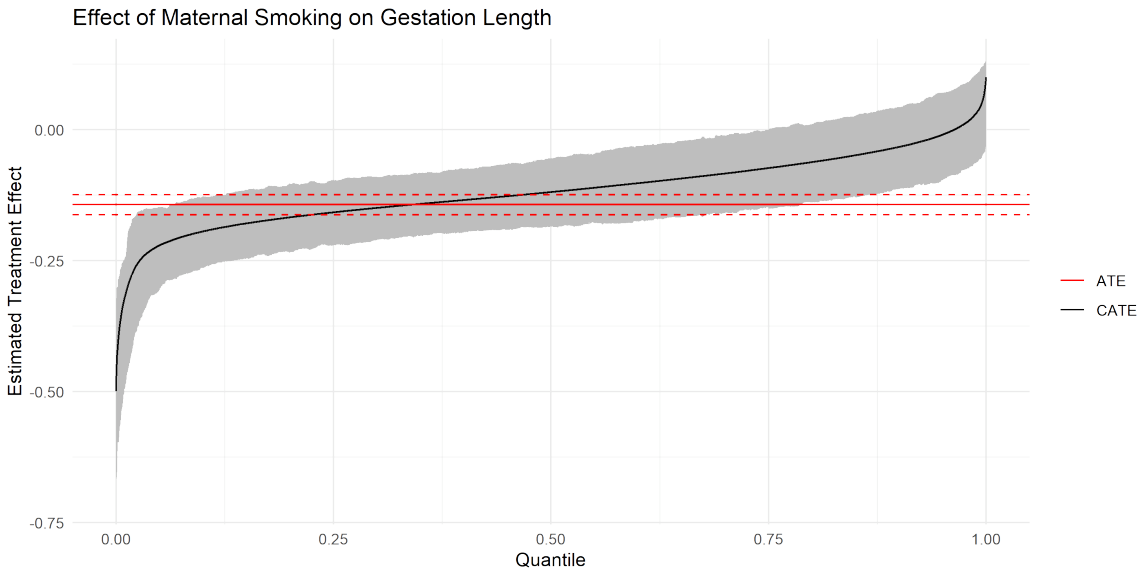


Figure A.10.: Sorted CATE and ATE estimates of smoking on Gestation Length
*Note:* The solid black line shows the estimated CATE of smoking gestation length sorted in ascending order. The dashed black line shows the 95% confidence intervals derived via bootstrapping. CATE is estimated using a causal forest as described in 2.4.2 using the `grf` R-package. The solid red line shows the ATE and corresponding 95% confidence intervals. These are estimated using the `grf` R-package by doubly robust augmented inverse propensity weighting.

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 among smoking mothers. The estimated average treatment effect of smoking on gestation length is $-0.08$ 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 A.10), which can be explained by growth retardation of very large babies caused by smoking.

For mother's age, Figure A.11 reveals weak effect modification by maternal age. The structural effect reveals that there is an increase in treatment effects caused by 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 A.12 shows the differences in effect magnitude that can be attributed to differ-

ent 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 A.56 and A.57). Figure A.58 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 prepregnancy BMI.

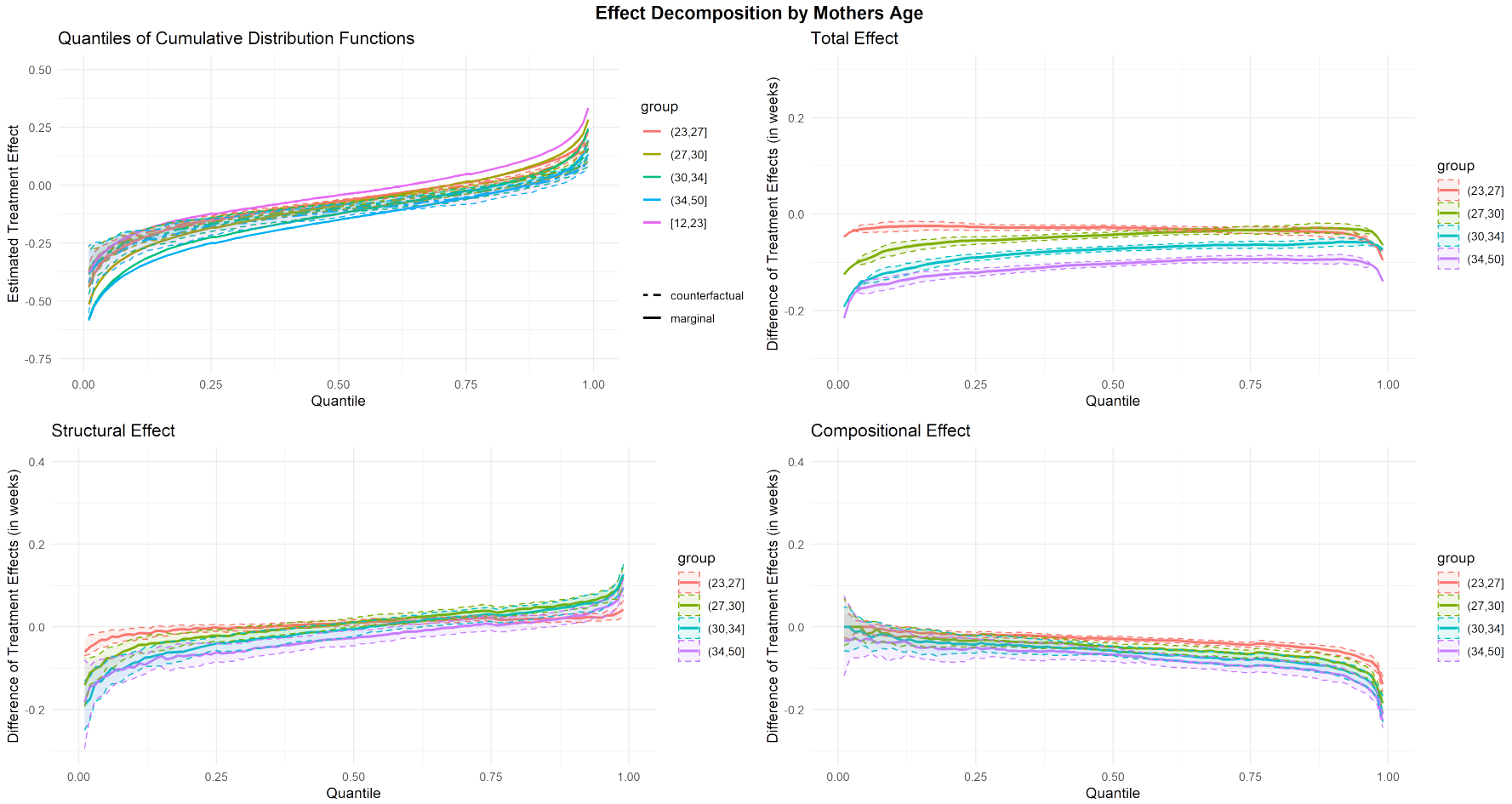


Figure A.11.: Gestation Length - Effect Decomposition by Mother's Age

*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 (2.12). 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.

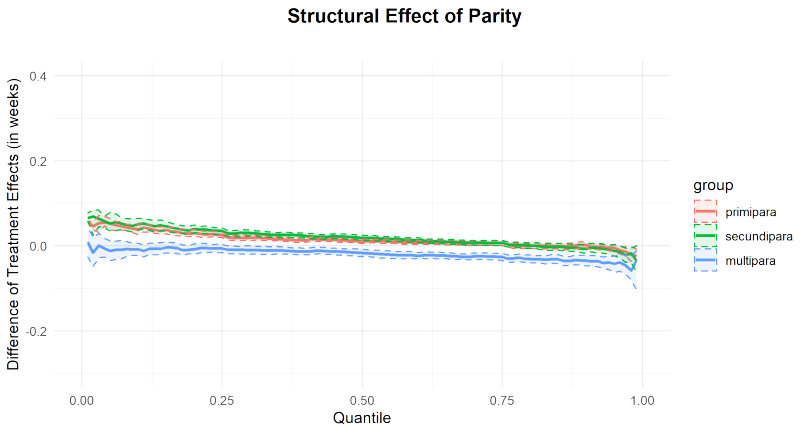


Figure A.12.: Gestation Length - Structural Effect of Parity

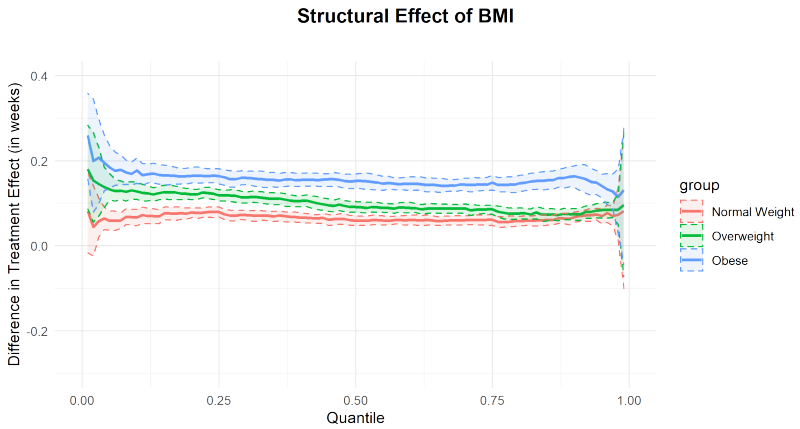


Figure A.13.: Gestation Length - Structural Effect of BMI

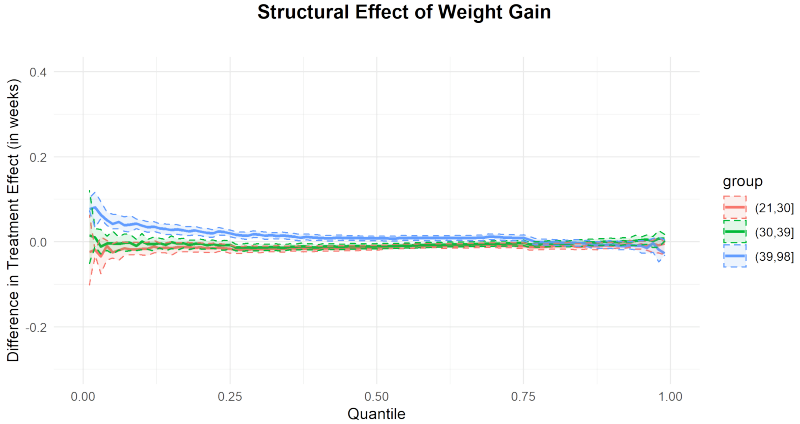


Figure A.14.: Gestation Length - Structural Effect of Weight Gain

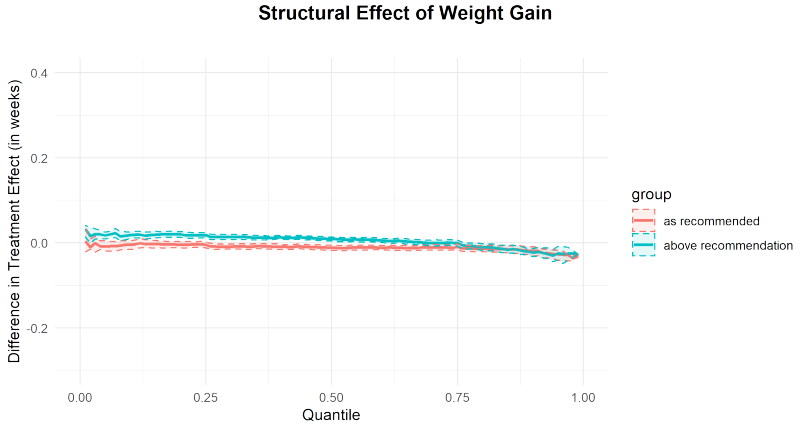


Figure A.15.: Gestation Length - Structural Effect of Weight Gain Recommendations

*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.

## A.4. Robustness Check: Heavy Smokers

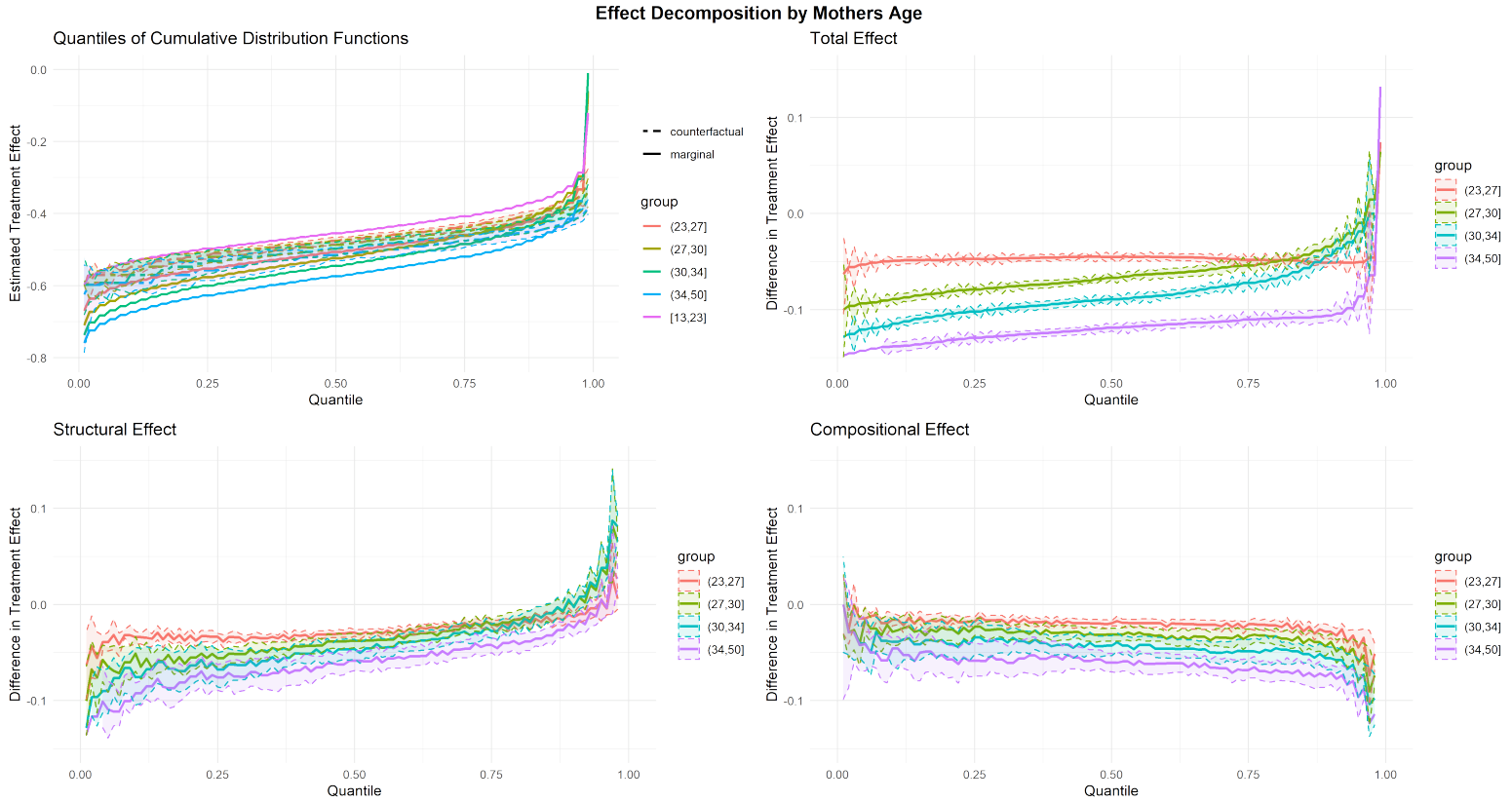


Figure A.16.: Heavy Smoking: Standardized Birth Weight - Effect Decomposition by Mother's Age

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on standardized birth weight by mother's age for heavy smokers (more than 20 daily cigarettes). 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 reference group are mothers aged 12 to 23 (the lowest quintile). Decomposition follows the procedure described in 1. 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.

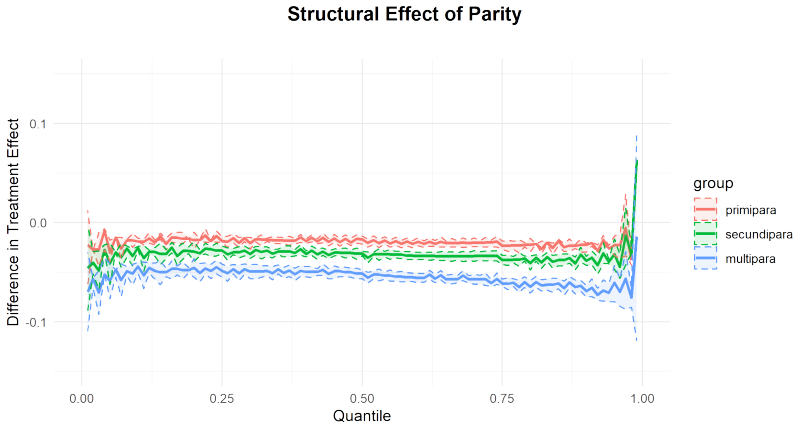


Figure A.17.: Heavy Smoking: Standardized Birth Weight - Structural Effect of Parity

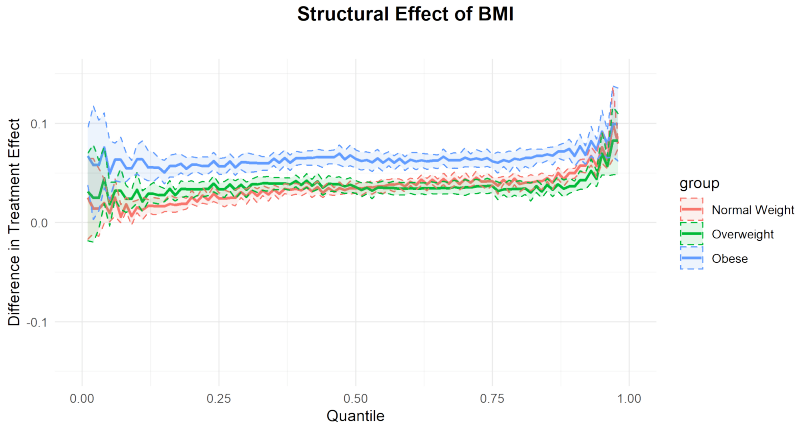


Figure A.18.: Heavy Smoking: Standardized Birth Weight - Structural Effect of Prepregnancy BMI

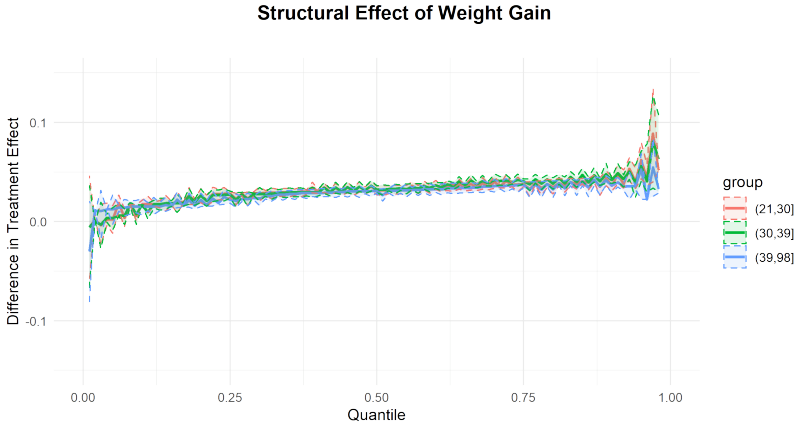


Figure A.19.: Heavy Smoking: Standardized Birth Weight - Structural Effect of Weight Gain

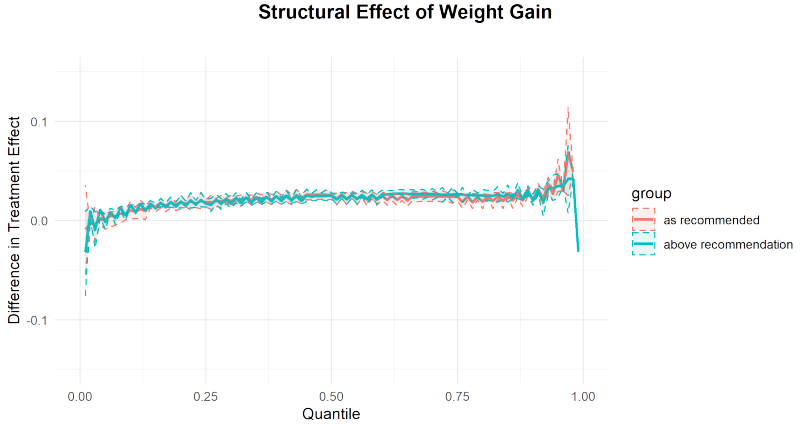


Figure A.20.: Heavy Smoking: Standardized Birth Weight - Structural Effect of Weight Gain (Recommendations)

*Note:* The figure shows the structural effect derived by decomposing the effect of heavy maternal smoking during pregnancy on standardized birth weight. 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.

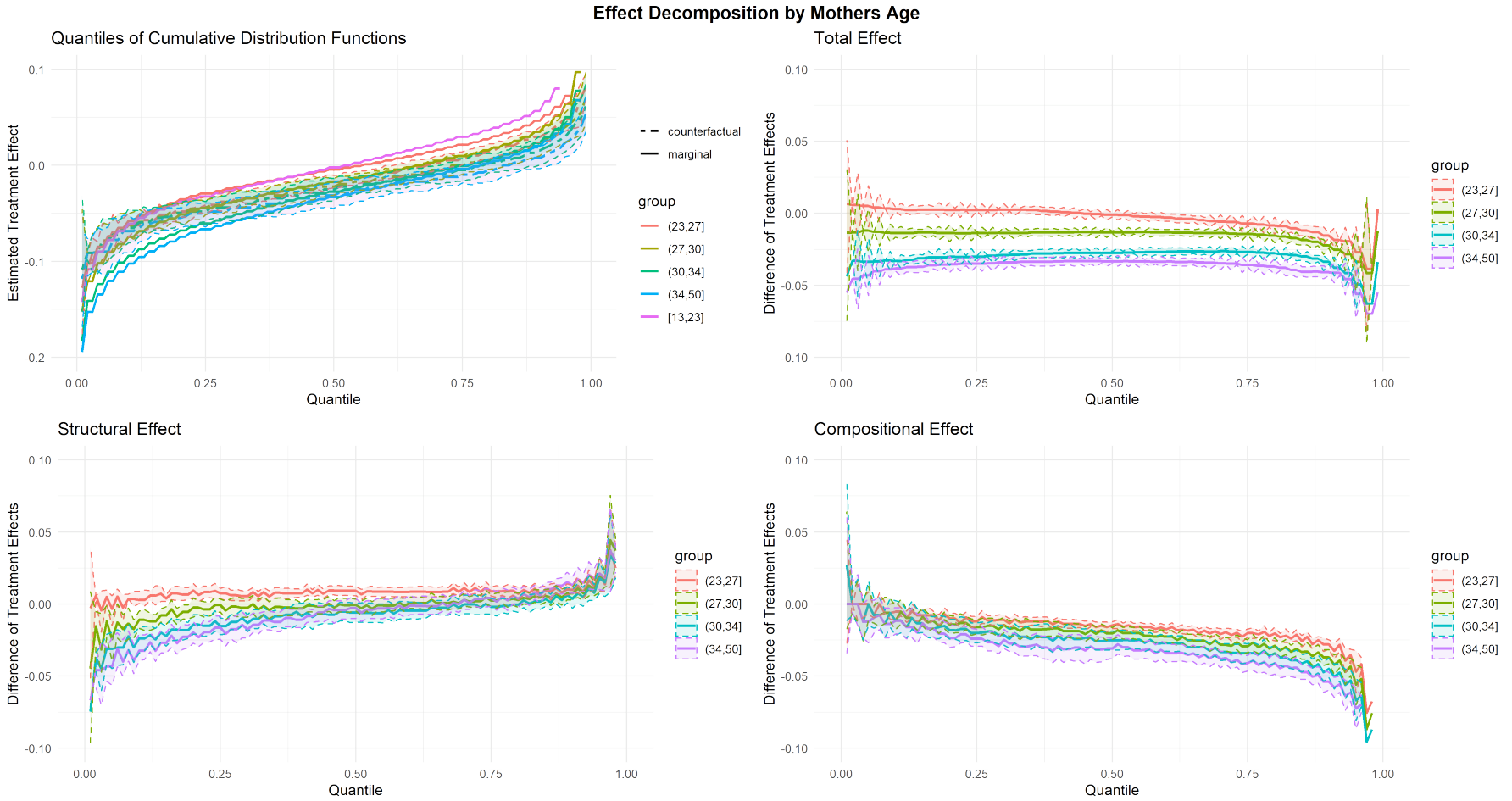


Figure A.21.: Heavy Smoking: Apgar Score - Effect Decomposition by Mother's Age

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on 5-minute Apgar score by mother's age for heavy smokers (more than 20 daily cigarettes). 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 (2.12). 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.

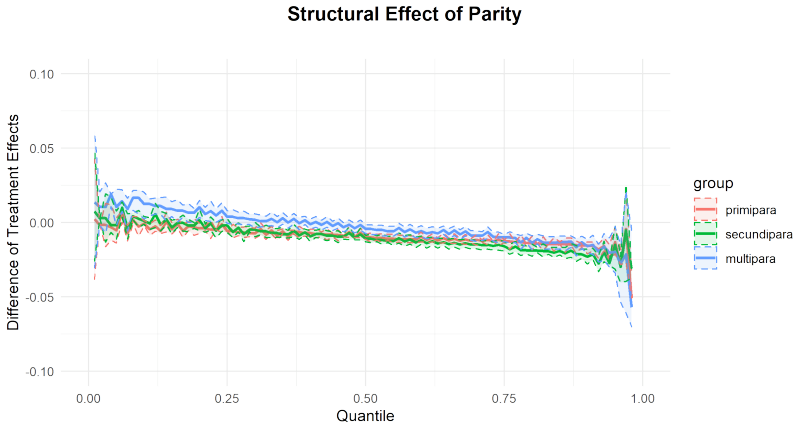


Figure A.22.: Heavy Smoking: Apgar Score - Structural Effect of Parity

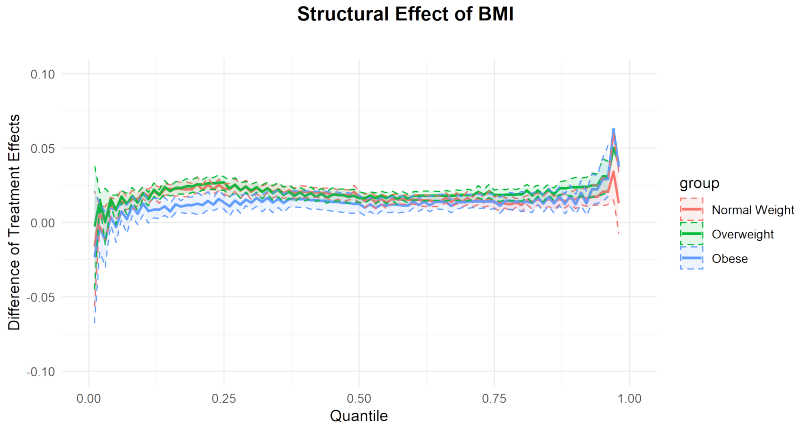


Figure A.23.: Heavy Smoking: Apgar Score - Structural Effect of Prepregnancy BMI

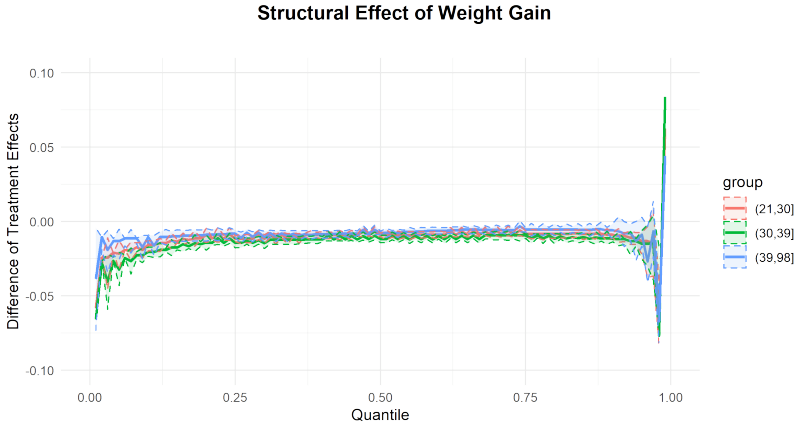


Figure A.24.: Heavy Smoking: Apgar Score - Structural Effect of Weight Gain

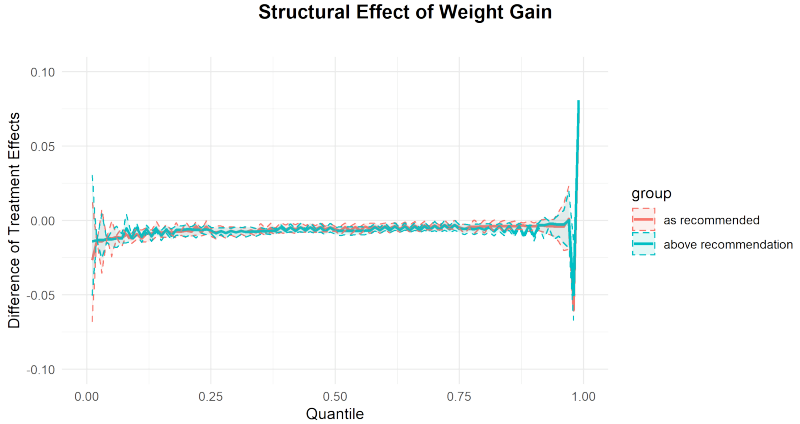


Figure A.25.: Heavy Smoking: Apgar Score - Structural Effect of Weight Gain (Recommendations)

*Note:* The figure shows the structural effect derived by decomposing the effect of heavy maternal smoking during pregnancy on 5-minute Apgar score. 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.

## A.5. Robustness Check: Low Birth Weight

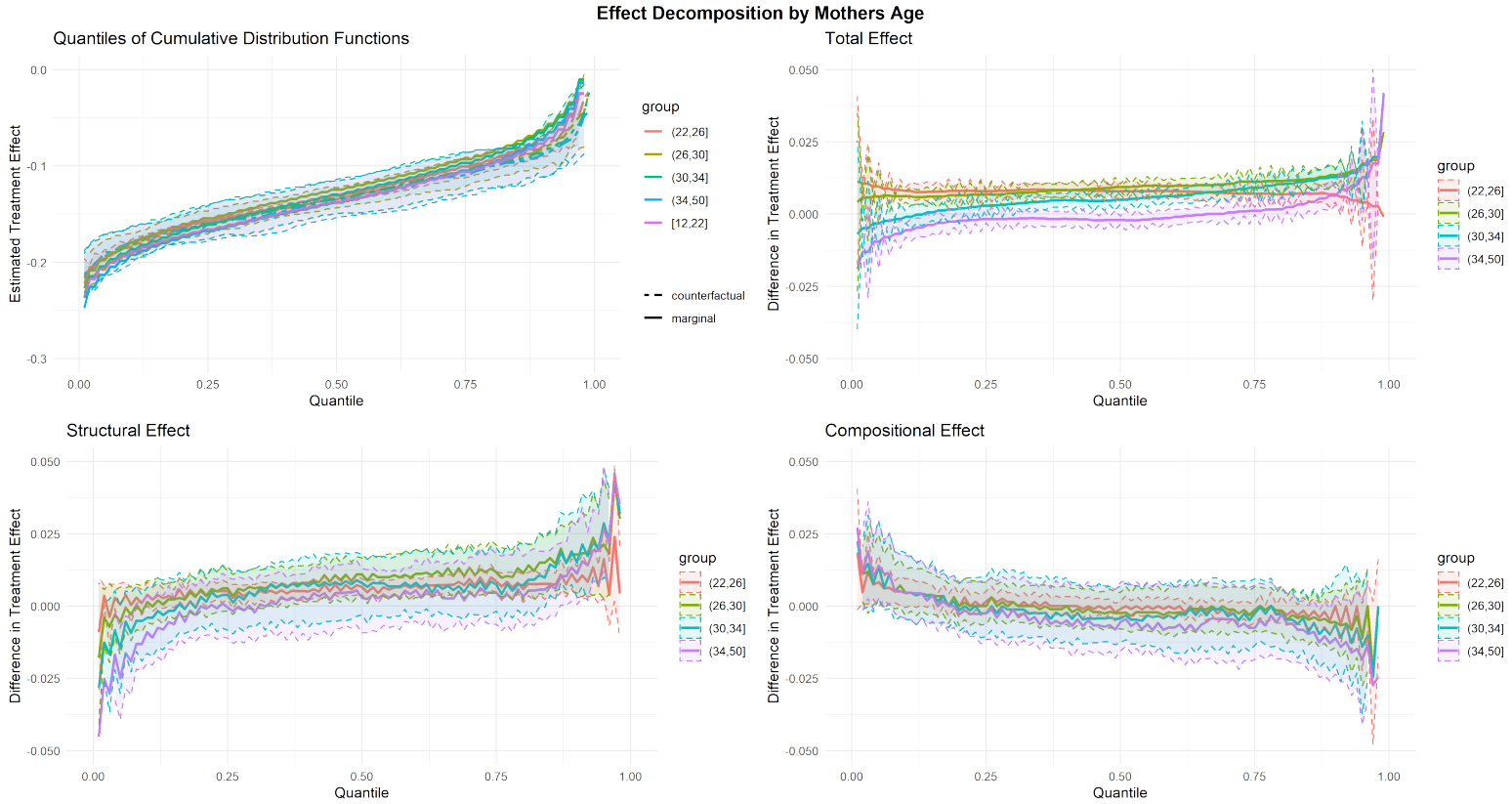


Figure A.26.: Standardized Birth Weight (<2800g) - Effect Decomposition by Mother's Age

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on standardized birth weight by mother's age, only considering birth weight below 2800 grams. 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 reference group are mothers aged 12 to 23 (the lowest quintile). Decomposition follows the procedure described in 1. 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.

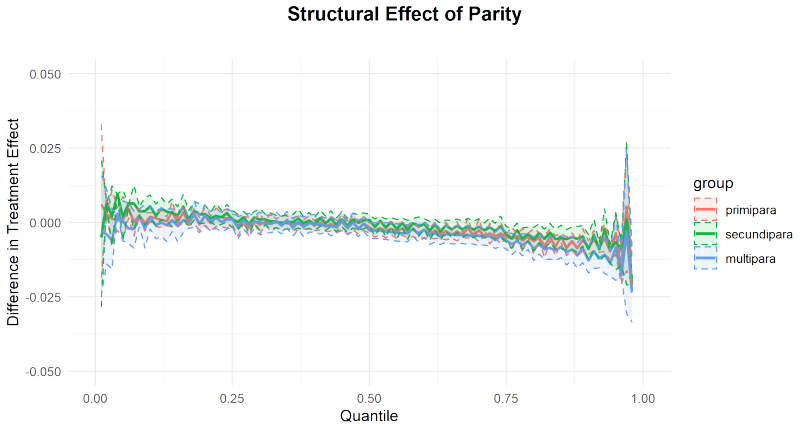


Figure A.27.: Standardized Birth Weight (<2800g) - Structural Effect of Parity

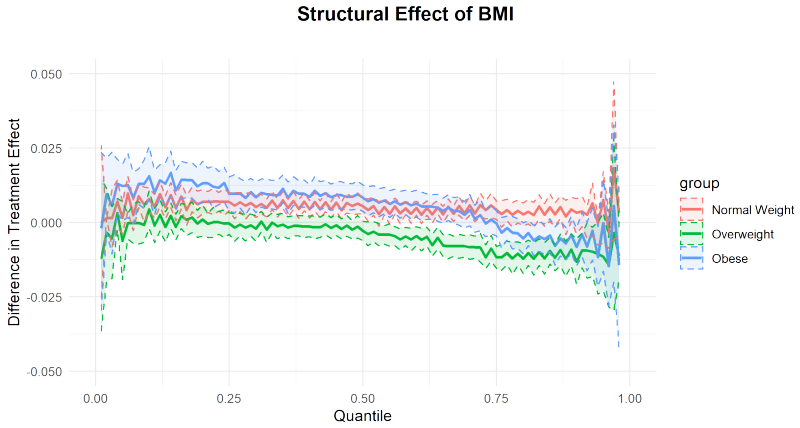


Figure A.28.: Standardized Birth Weight (<2800g) - Structural Effect of Prepregnancy BMI


Figure A.29.: Standardized Birth Weight (<2800g) - Structural Effect of Weight Gain

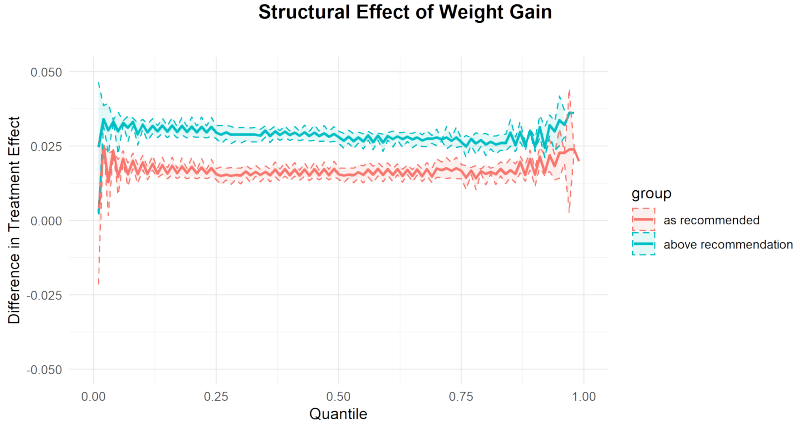


Figure A.30.: Standardized Birth Weight (<2800g) - Structural Effect of Weight Gain (Recommendations)

*Note:* The figure shows the structural effect derived by decomposing the effect of maternal smoking during pregnancy on standardized birth weight, only considering birth weight below 2800 grams. 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.

## A.6. Robustness Check: Prepregnancy Smoking

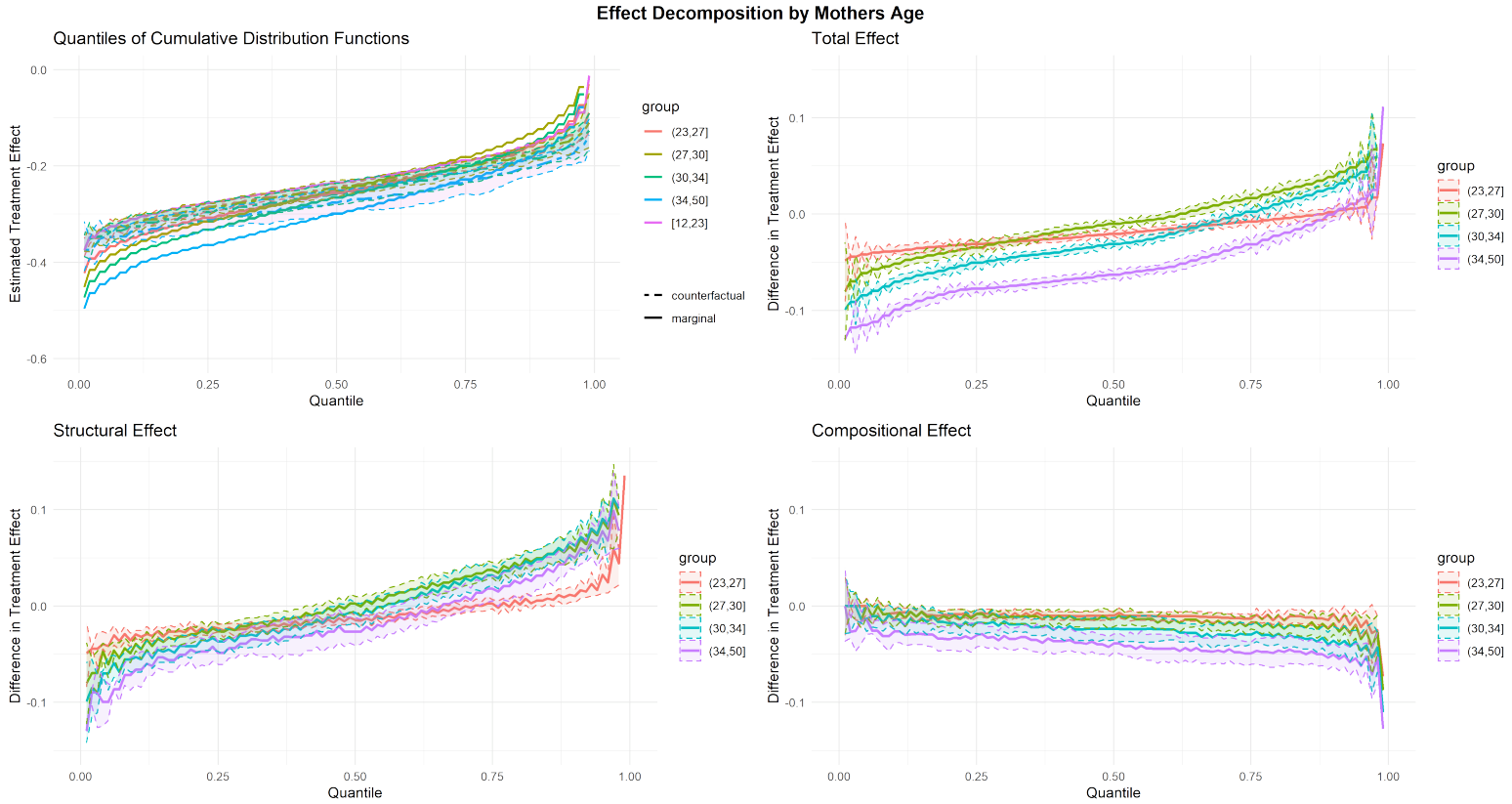


Figure A.31.: Prepregnancy Smoking: Standardized Birth Weight - Effect Decomposition by Mother's Age

*Note:* The figure displays the decomposition of the effect of prepregnancy smoking 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 reference group are mothers aged 12 to 23 (the lowest quintile). Decomposition follows the procedure described in 1. 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.

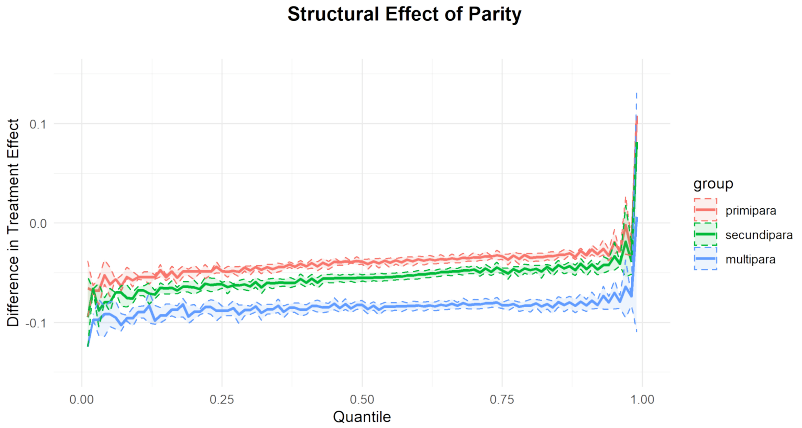


Figure A.32.: Prepregnancy Smoking: Standardized Birth Weight - Structural Effect of Parity

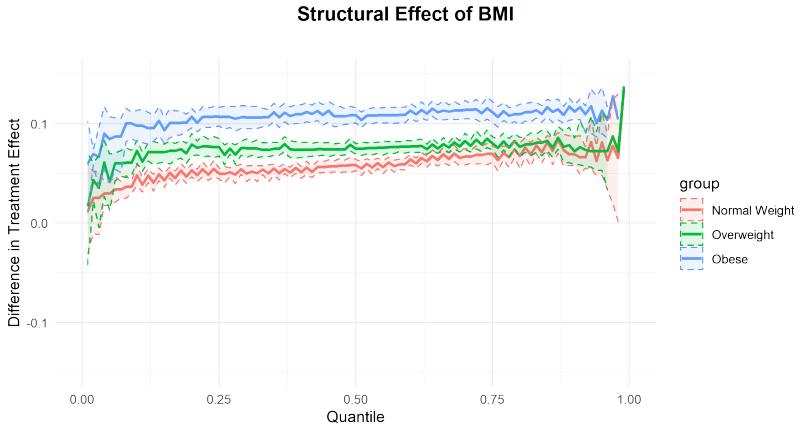


Figure A.33.: Prepregnancy Smoking: Standardized Birth Weight - Structural Effect of Prepregnancy BMI

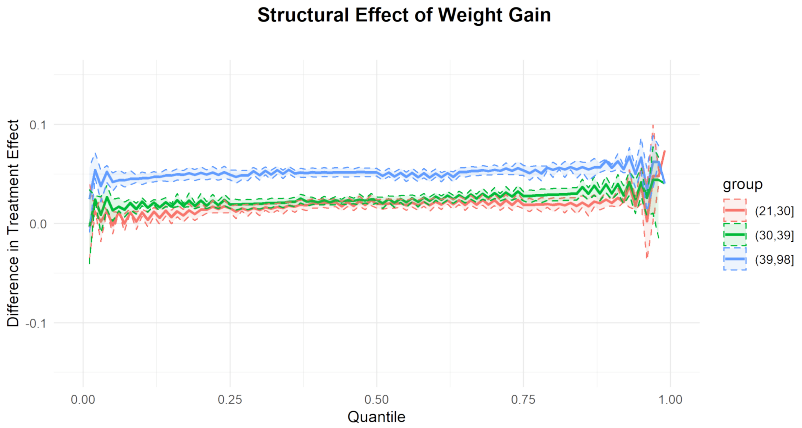


Figure A.34.: Prepregnancy Smoking: Standardized Birth Weight - Structural Effect of Weight Gain

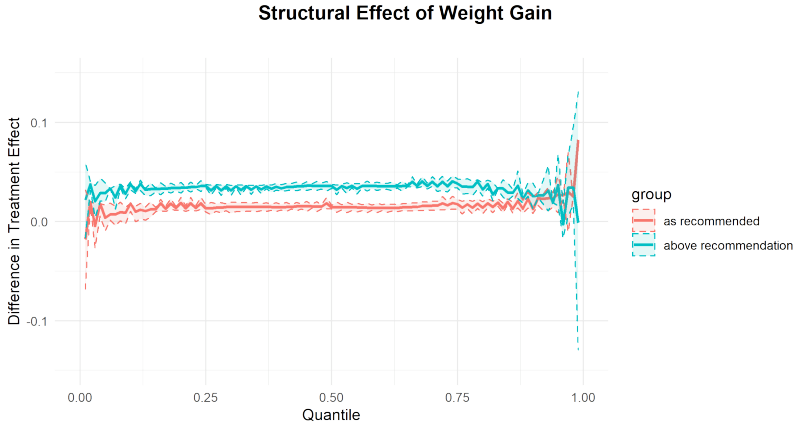


Figure A.35.: Prepregnancy Smoking: Standardized Birth Weight - Structural Effect of Weight Gain (Recommendations)

*Note:* The figure shows the structural effect derived by decomposing the effect of prepregnancy smoking on standardized birth weight. 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.

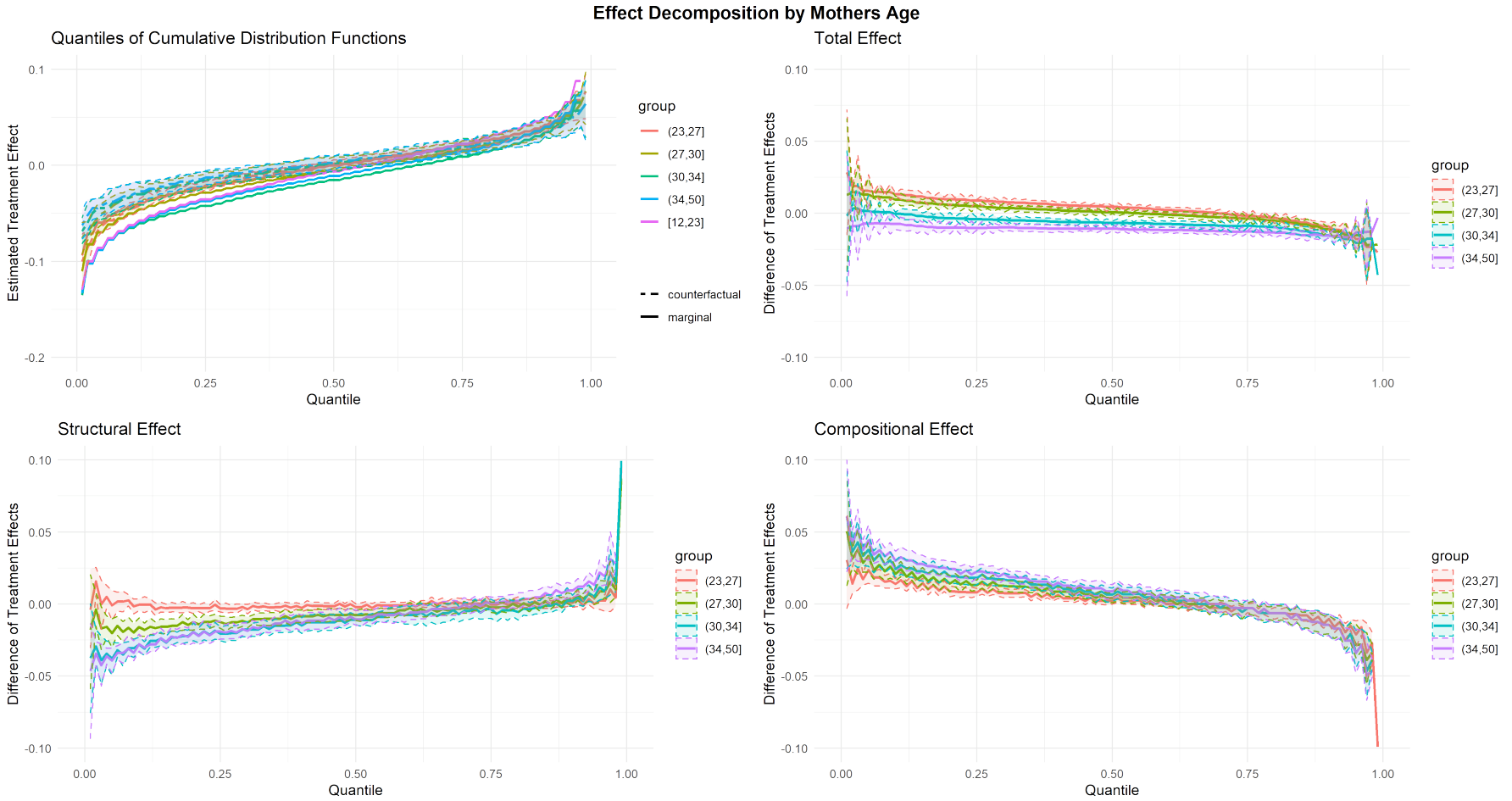


Figure A.36.: Prepregnancy Smoking: Apgar Score - Effect Decomposition by Mother's Age

*Note:* The figure displays the decomposition of the effect of prepregnancy smoking on Apgar score 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 (2.12). 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.

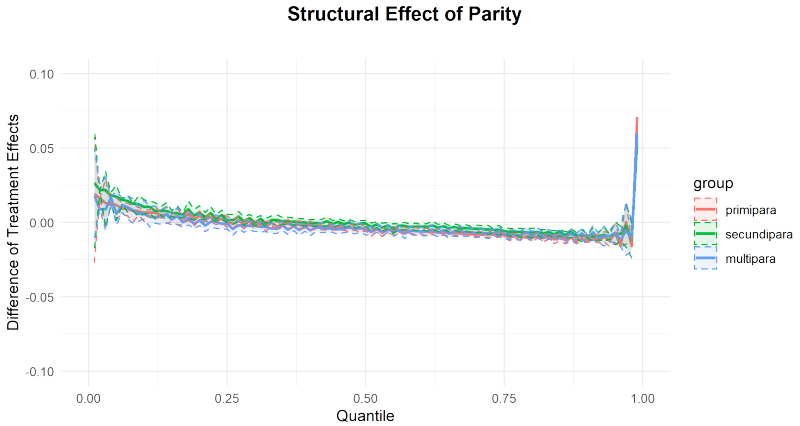


Figure A.37.: Prepregnancy Smoking: Apgar Score - Structural Effect of Parity

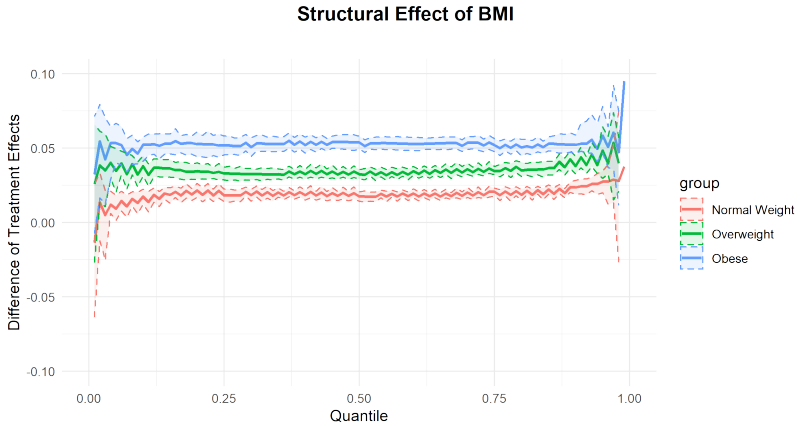


Figure A.38.: Prepregnancy Smoking: Apgar Score - Structural Effect of Prepregnancy BMI

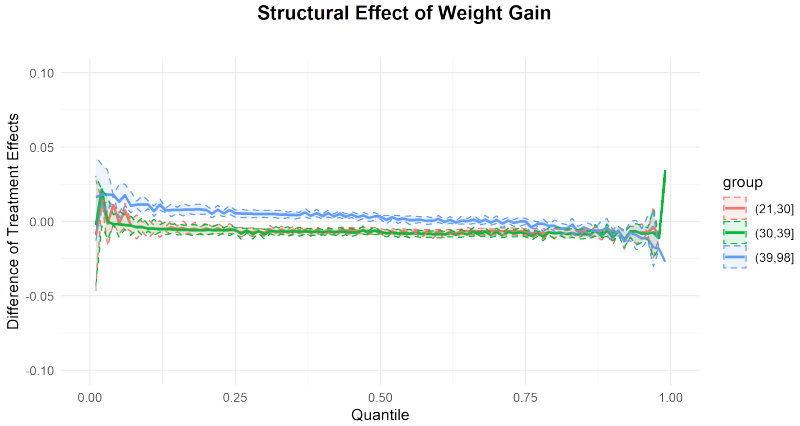


Figure A.39.: Prepregnancy Smoking: Apgar Score - Structural Effect of Weight Gain

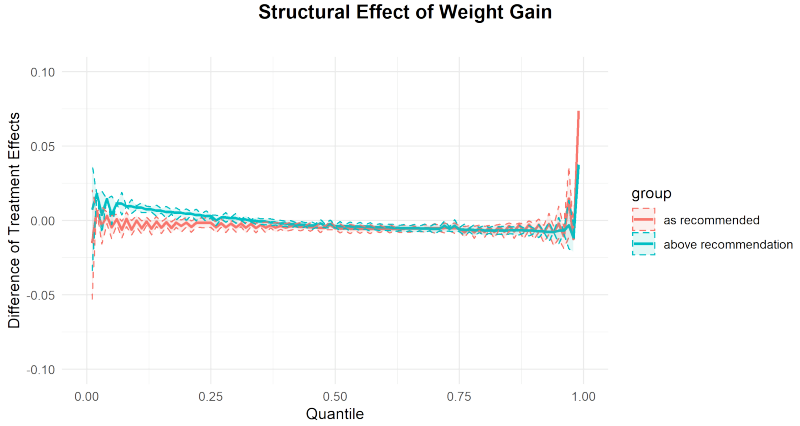


Figure A.40.: Prepregnancy Smoking: Apgar Score - Structural Effect of Weight Gain (Recommendations)

*Note:* The figure shows the structural effect derived by decomposing the effect of prepregnancy smoking on the 5-minute Apgar score. 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.

## A.7. Robustness Check: Propensity Trimming

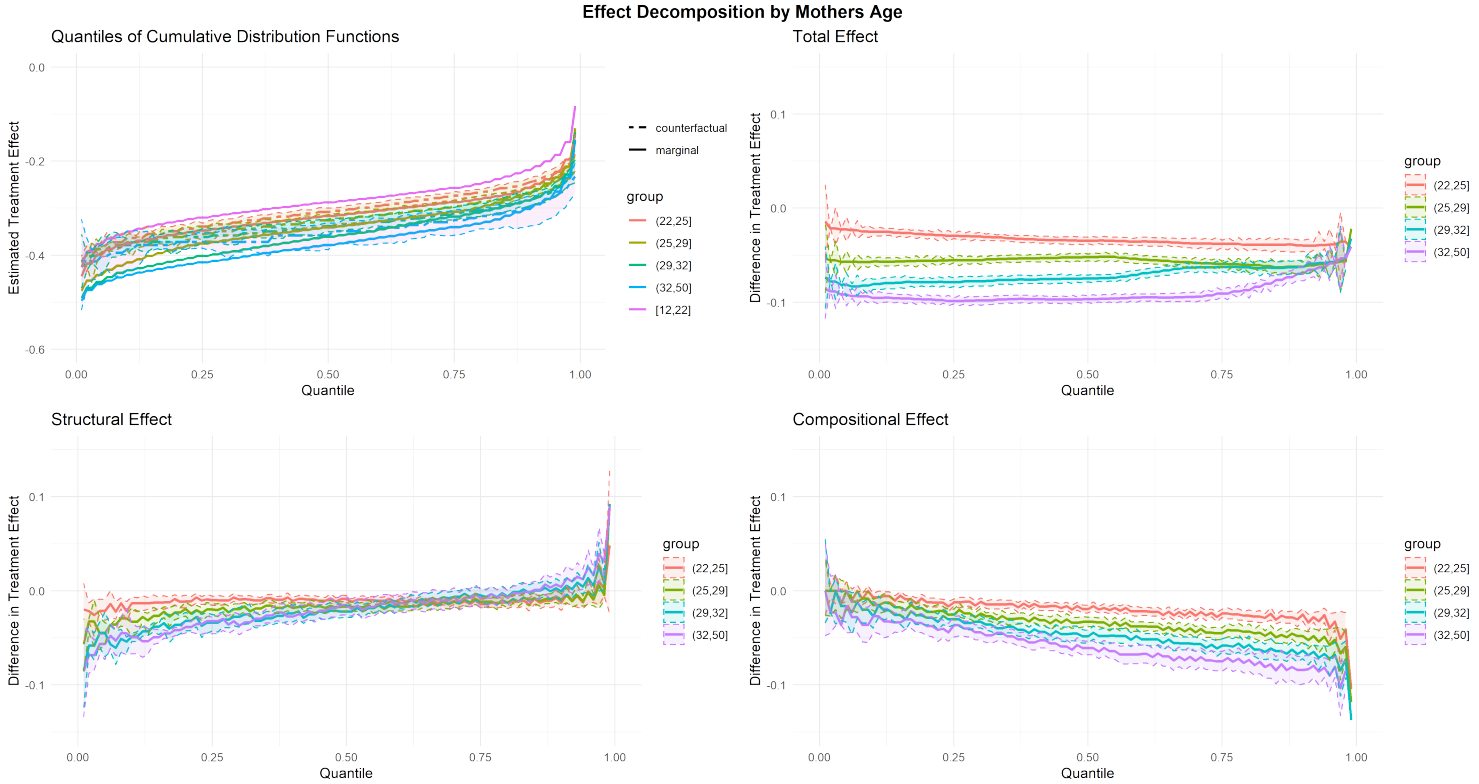


Figure A.41.: Propensity Trimming: Standardized Birth Weight - Effect Decomposition by Mother's Age

*Note:* The figure displays the decomposition of the effect of smoking during pregnancy on standardized birth weight by mother's age. We only consider observations with a propensity of smoking falling between the 2.5-th percentile for treated observations and the 97.5-th percentile for untreated observations. The figure 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 reference group are mothers aged 12 to 23 (the lowest quintile). Decomposition follows the procedure described in 1. 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.

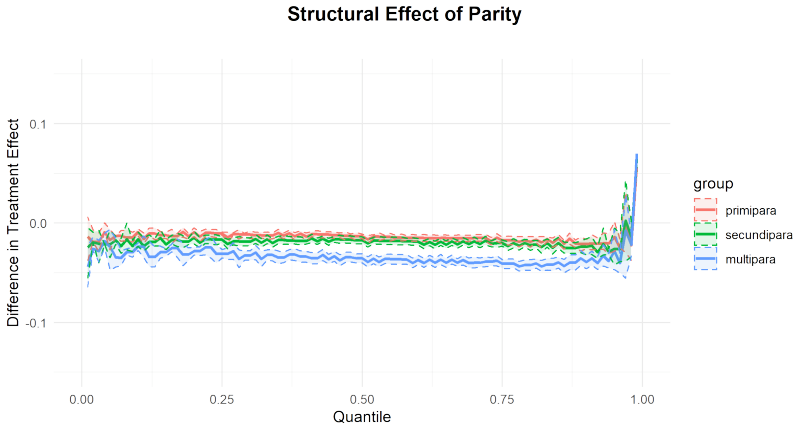


Figure A.42.: Propensity Trimming: Standardized Birth Weight - Structural Effect of Parity

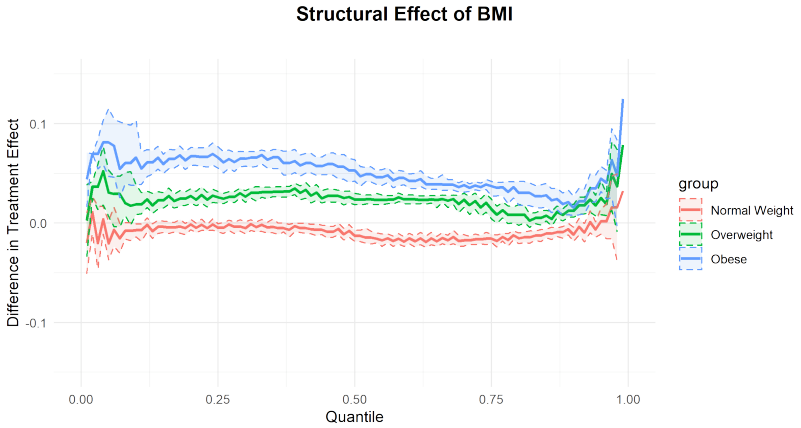


Figure A.43.: Propensity Trimming: Standardized Birth Weight - Structural Effect of prepregnancy BMI

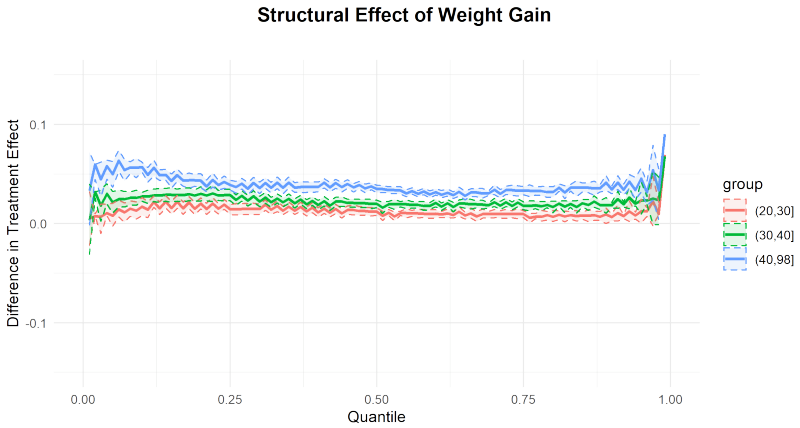


Figure A.44.: Propensity Trimming: Standardized Birth Weight - Structural Effect of Weight Gain

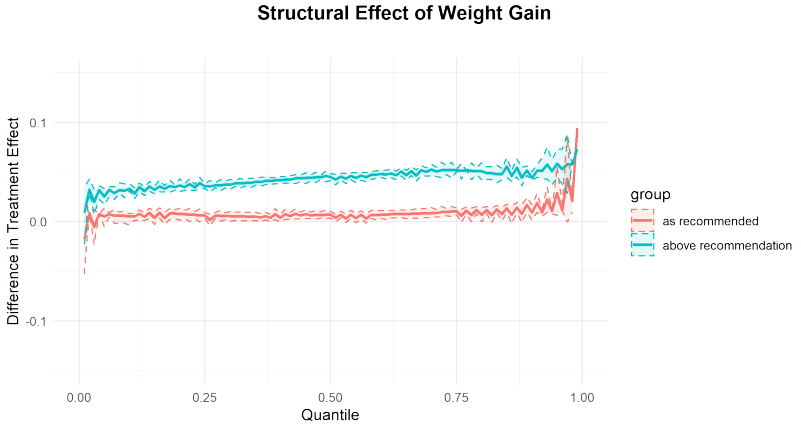


Figure A.45.: Propensity Trimming: Standardized Birth Weight - Structural Effect of Weight Gain (Recommendations)

*Note:* The figure shows the structural effect derived by decomposing the effect of maternal smoking during pregnancy on standardized birth weight. We only consider observations with a propensity of smoking falling between the 2.5-th percentile for treated observations and the 97.5-th percentile for untreated observations. 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.

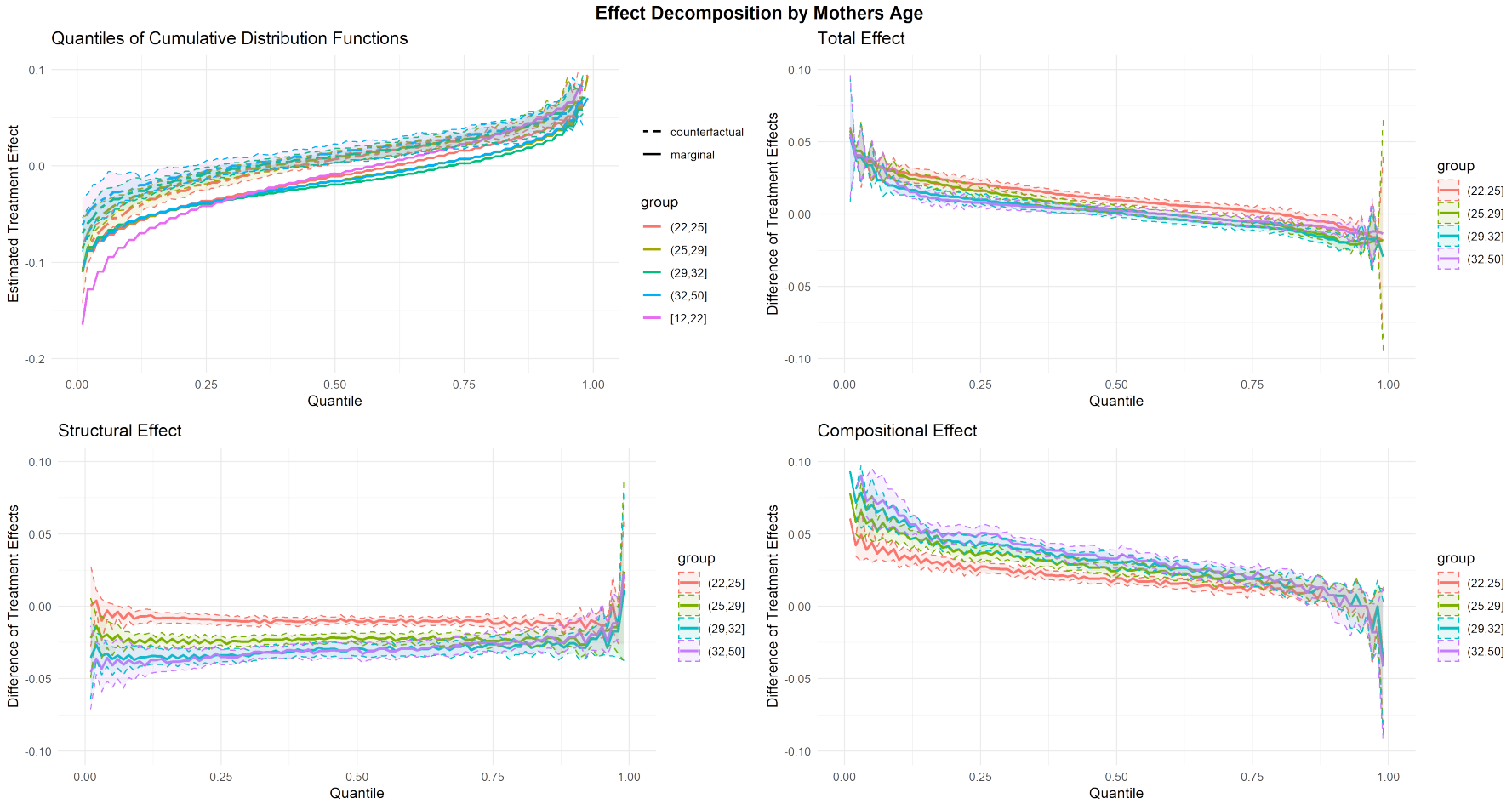


Figure A.46.: Propensity Trimming: Apgar Score - Effect Decomposition by Mother's Age

*Note:* The figure displays the decomposition of the effect of smoking during pregnancy on the Apgar score by mother's age. We only consider observations with a propensity of smoking falling between the 2.5-th percentile for treated observations and the 97.5-th percentile for untreated observations. The figure 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 (2.12). 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.

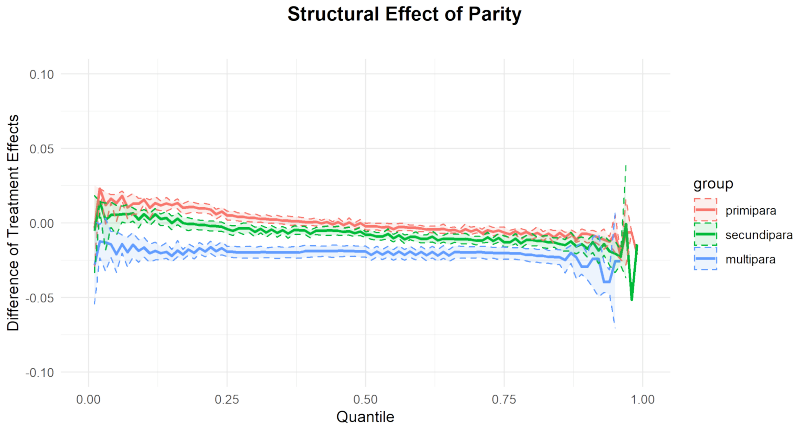


Figure A.47.: Propensity Trimming: Apgar Score - Structural Effect of Parity

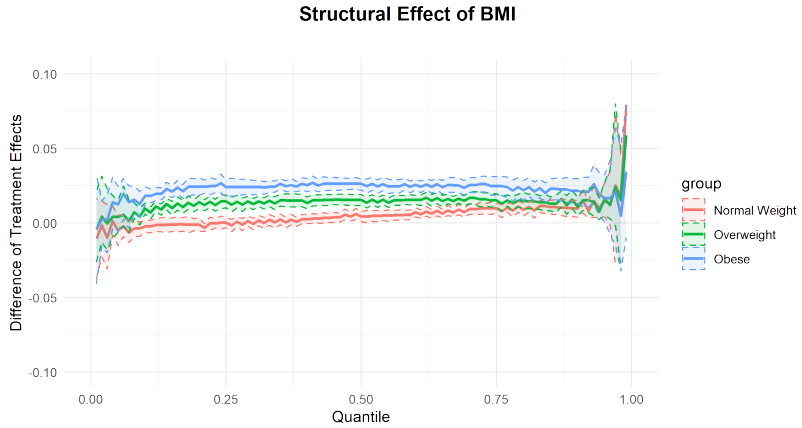


Figure A.48.: Propensity Trimming: Apgar Score - Structural Effect of prepregnancy BMI

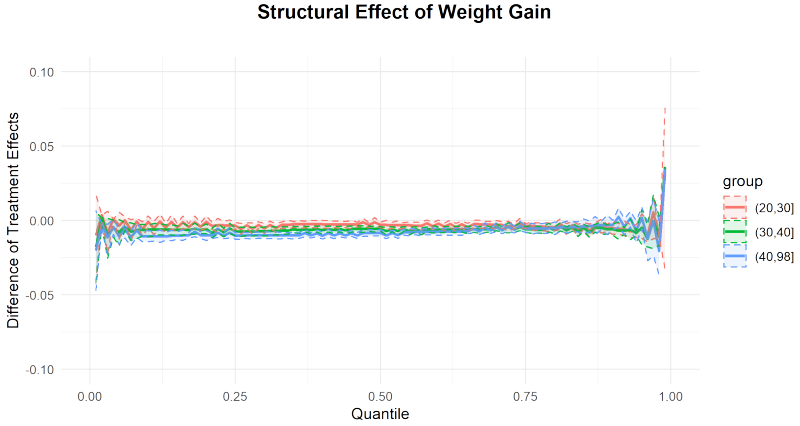


Figure A.49.: Propensity Trimming: Apgar Score - Structural Effect of Weight Gain

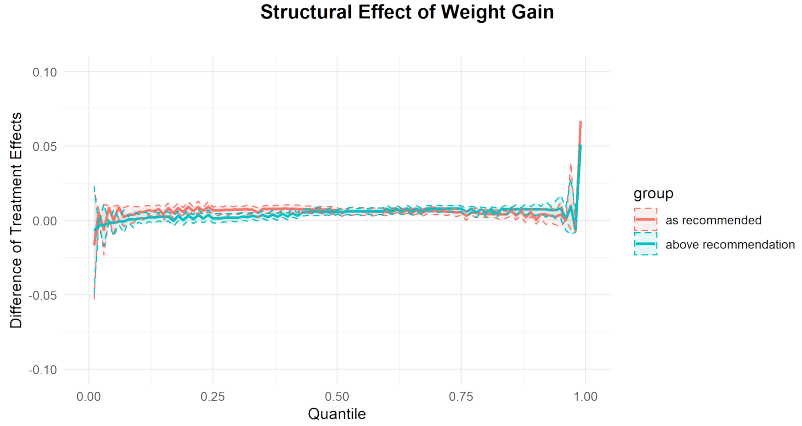


Figure A.50.: Propensity Trimming: Apgar Score - Structural Effect of Weight Gain (Recommendations)

*Note:* The figure shows the structural effect derived by decomposing the effect of maternal smoking during pregnancy on the 5-minute Apgar score. We only consider observations with a propensity of smoking falling between the 2.5-th percentile for treated observations and the 97.5-th percentile for untreated observations. 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.

# A.8. Decomposition: Full Decomposition Figures

## A.8.1. Standardized Birth Weight

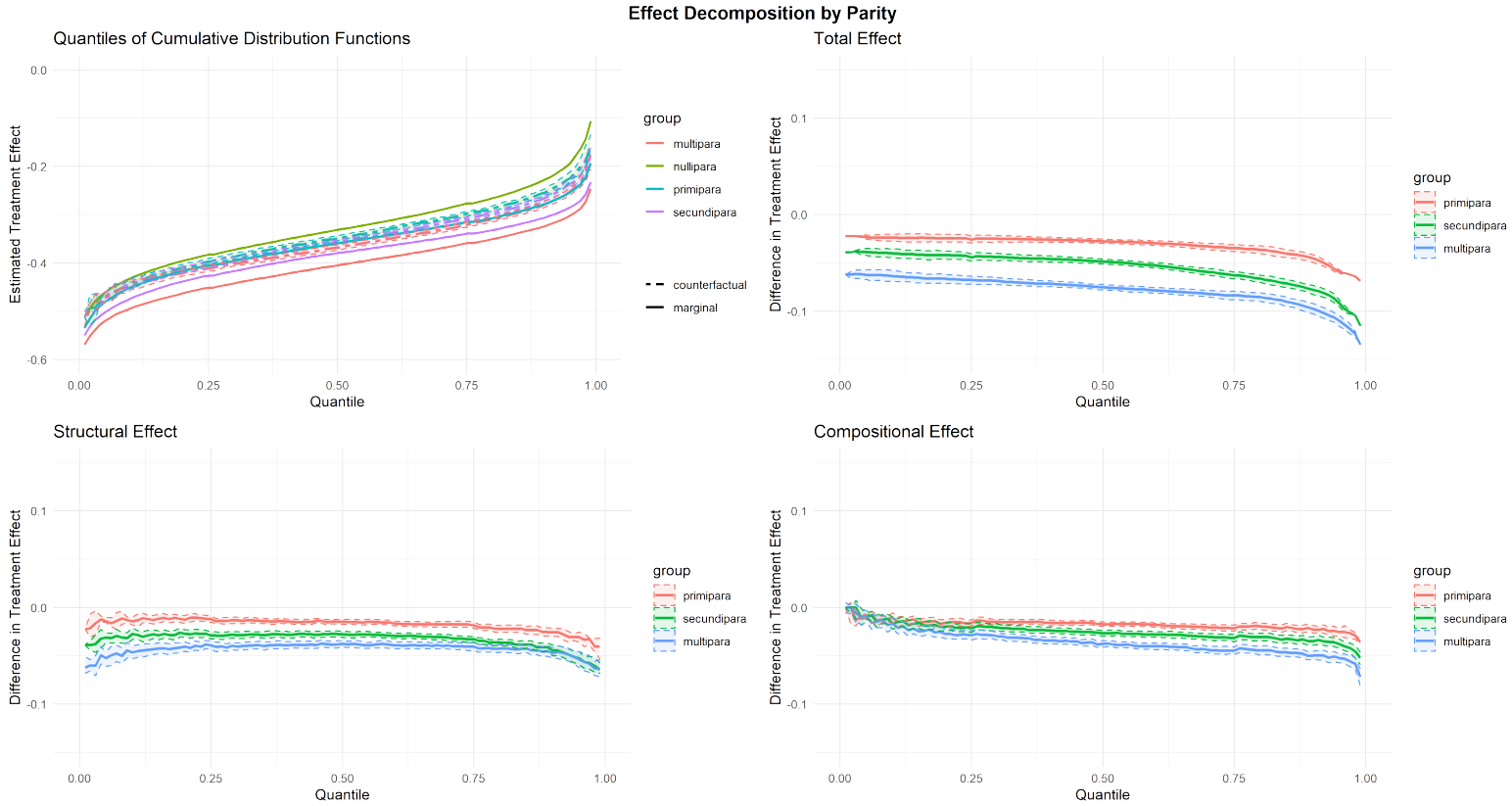


Figure A.51.: Effect Decomposition by Parity

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on standardized birth weight by parity. 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 reference group which are nulliparous mothers. Decomposition follows the procedure described in 1. Positive difference corresponds to effect mitigation with increasing parity, whereas negative difference corresponds to effect amplification with increasing parity respectively. Shaded areas show 95% confidence intervals.

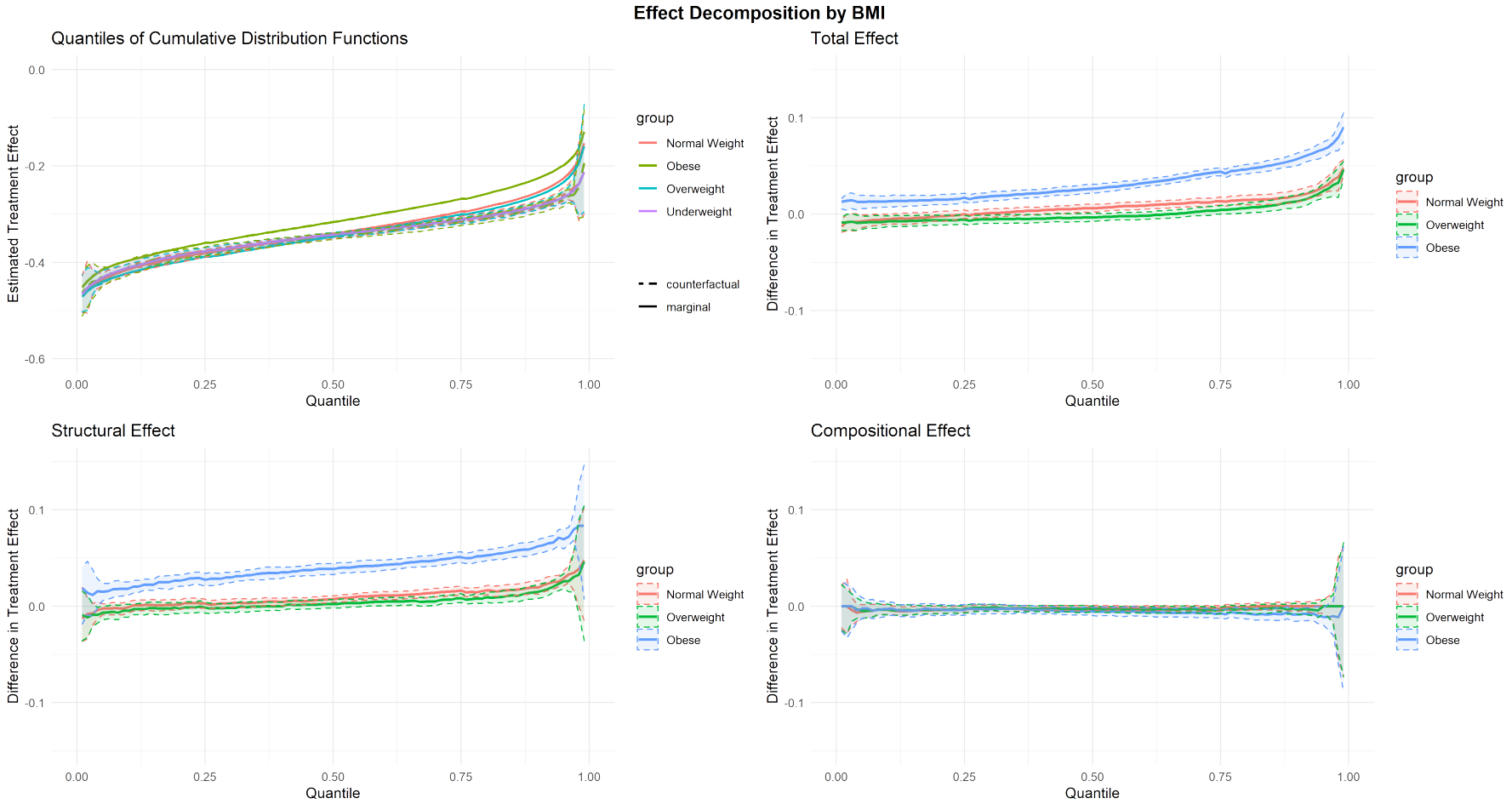


Figure A.52.: Effect Decomposition by prepregnancy BMI

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on standardized birth weight by prepregnancy BMI. 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 underweight mothers. Decomposition follows the procedure described in 1. Total, structural and compositional effect are defined as in equation (2.12). Positive difference corresponds to effect mitigation with increasing BMI, whereas negative difference corresponds to effect amplification with increasing BMI respectively. Shaded areas show 95% confidence intervals.

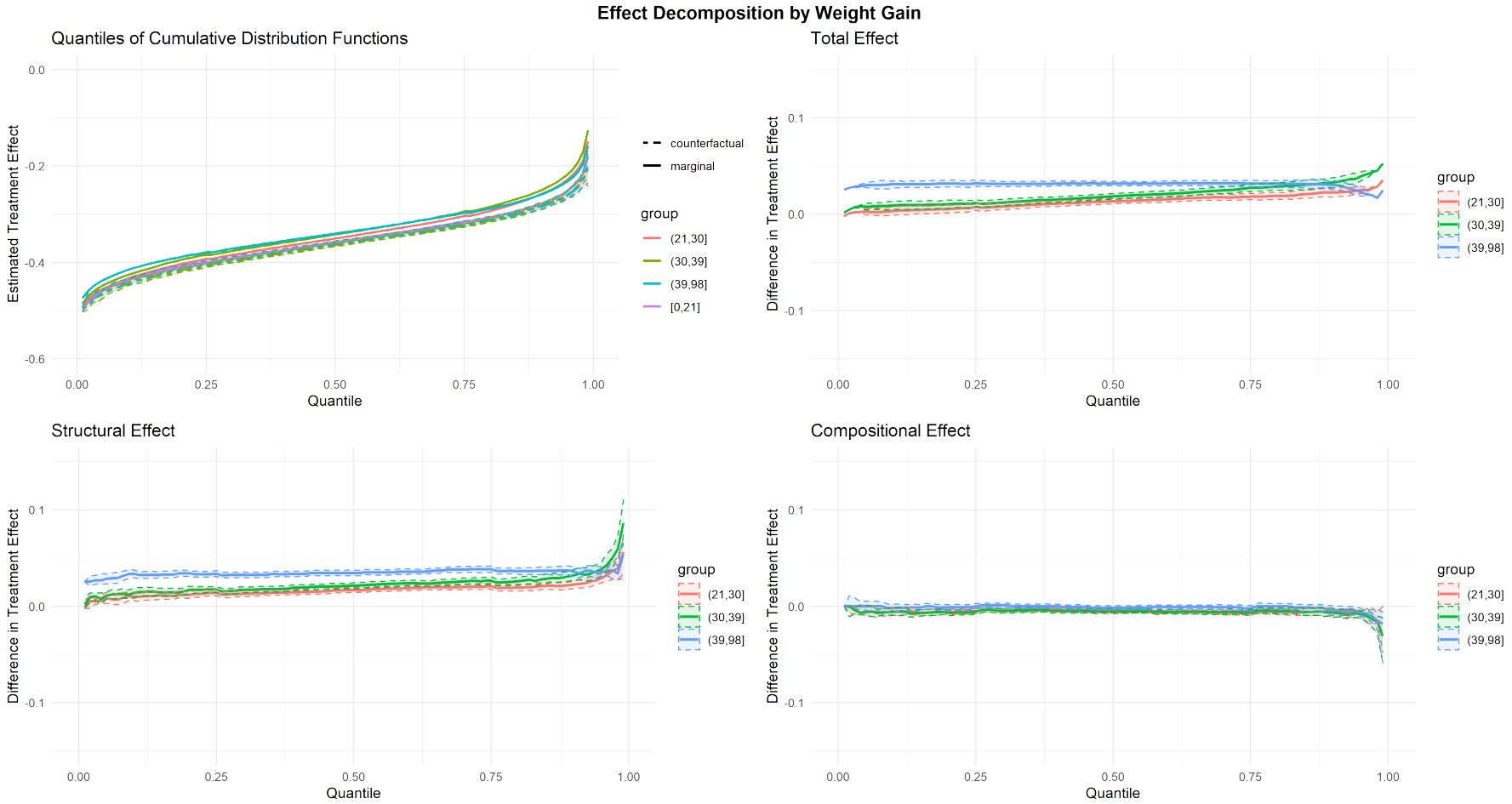


Figure A.53.: Effect Decomposition by pregnancy Weight Gain

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on standardized birth weight by weight gain in pounds. 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 gaining less than 22 pounds (lowest quartile). Decomposition follows the procedure described in 1. Total, structural and compositional effect are defined as in equation (2.12). Positive difference corresponds to effect mitigation with increasing weight gain, whereas negative difference corresponds to effect amplification with increasing weight gain respectively. Shaded areas show 95% confidence intervals.

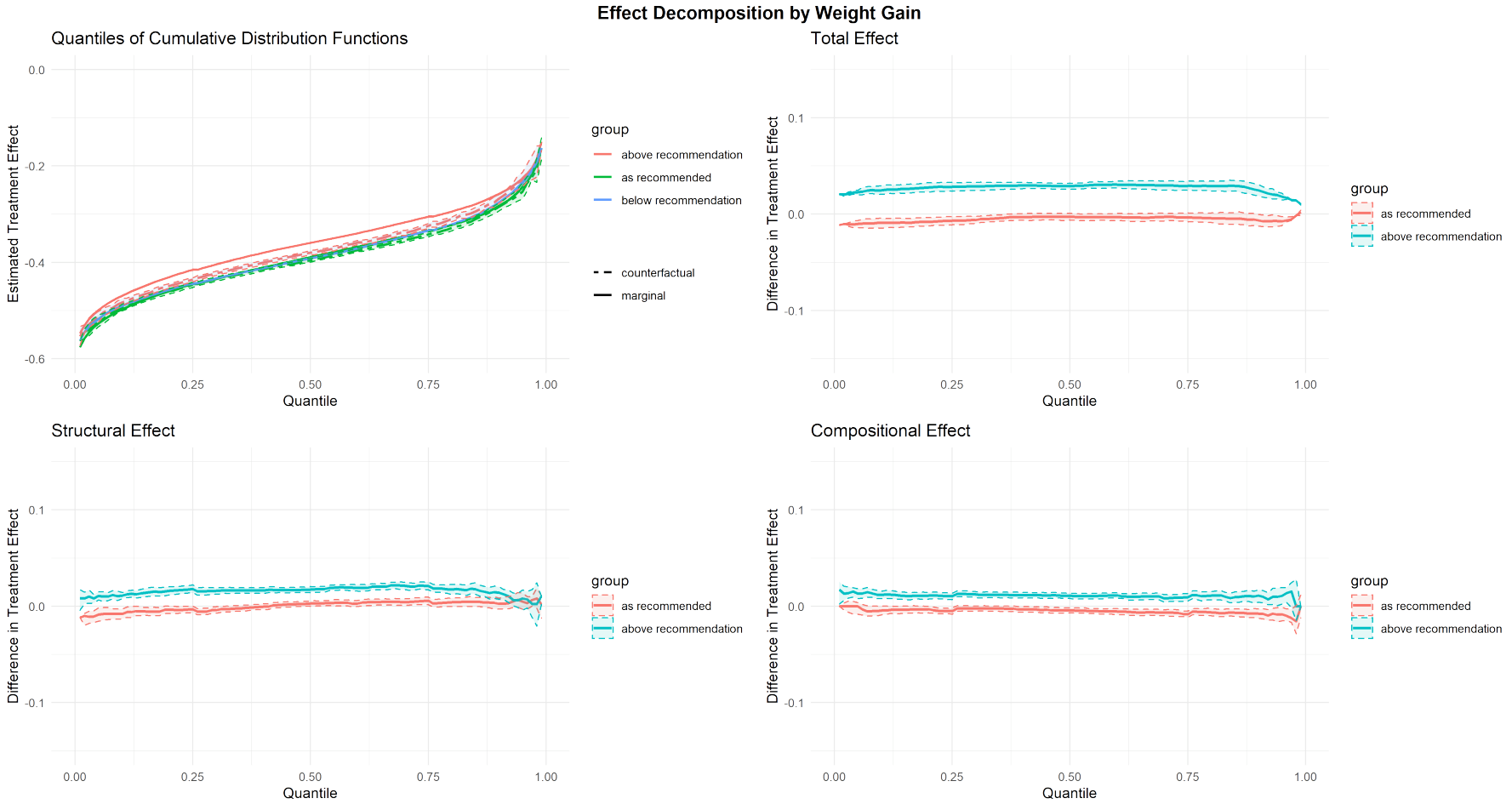


Figure A.54.: Effect Decomposition by Weight Gain Recommendations

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on standardized birth weight by weight gain recommendations. 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 gaining less than recommended based on guidelines published by CDC – National Center for Health Statistics (2021). Decomposition follows the procedure described in 1. Total, structural and compositional effect are defined as in equation (2.12). Shaded areas show 95% confidence intervals.

## A.8.2. Apgar Score

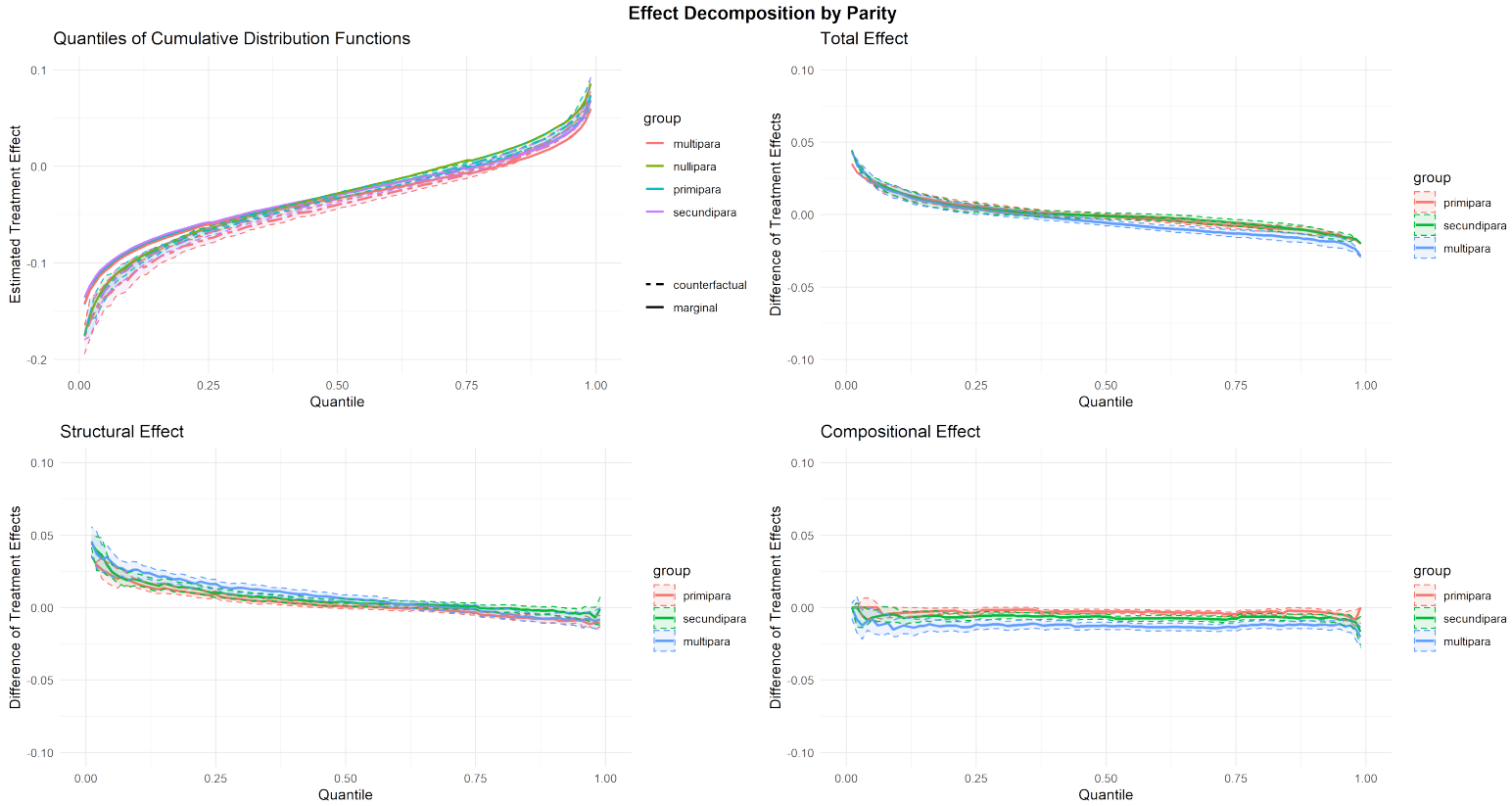


Figure A.55.: Apgar Score - Effect Decomposition by Parity

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on 5-minute Apgar score by parity. 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 reference group are nulliparous mothers. Decomposition follows the procedure described in 1. Positive difference corresponds to effect mitigation with increasing parity, whereas negative difference corresponds to effect amplification with increasing parity respectively. Shaded areas show 95% confidence intervals.

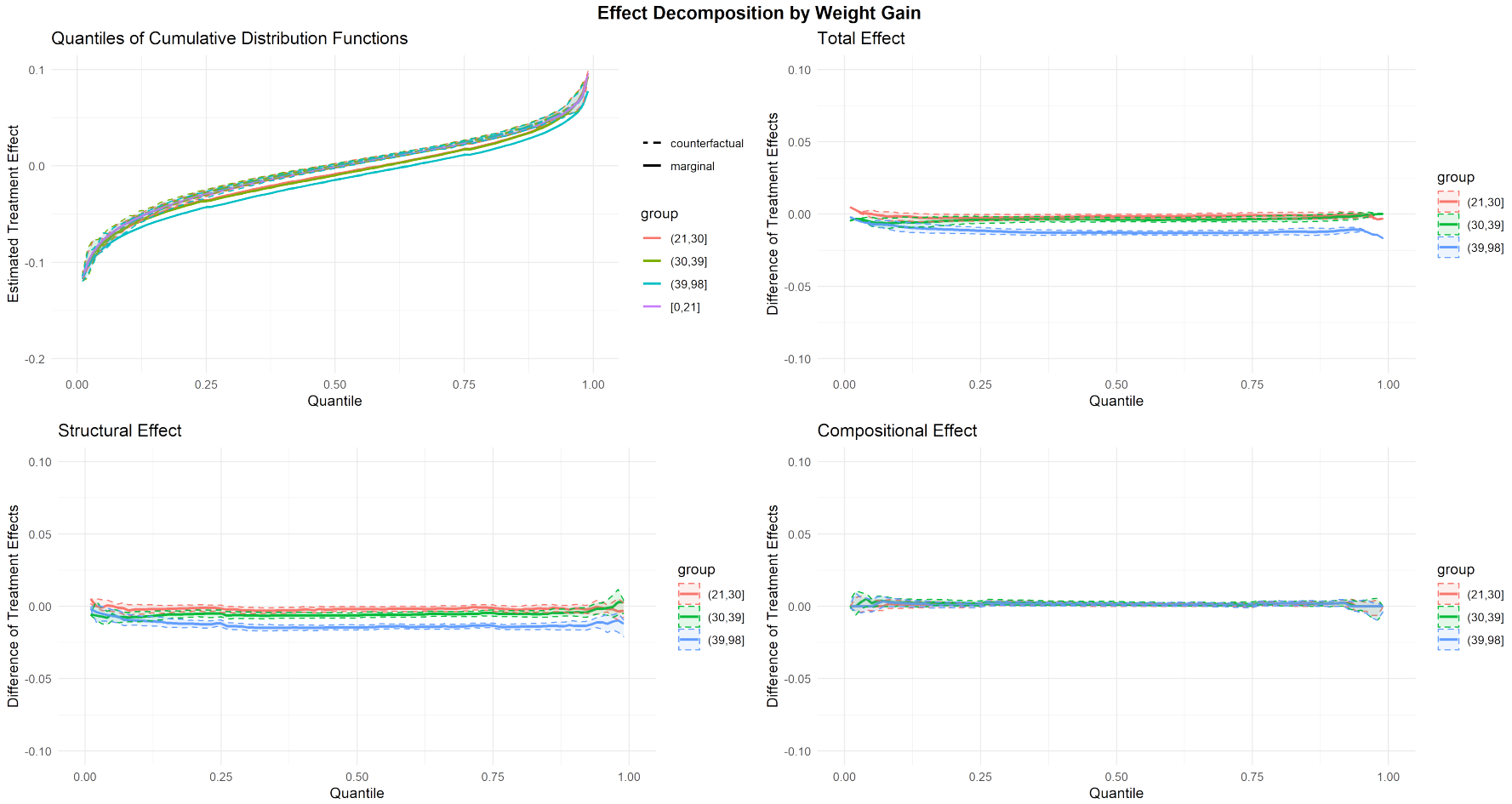


Figure A.56.: Apgar Score - Effect Decomposition by Weight Gain

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on 5-minute Apgar score by weight gain in pounds. 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 gaining less than 22 pounds (lowest quartile). Decomposition follows the procedure described in 1. Total, structural and compositional effect are defined as in equation (2.12). Positive difference corresponds to effect mitigation with increasing weight gain, whereas negative difference corresponds to effect amplification with increasing weight gain respectively. Shaded areas show 95% confidence intervals.

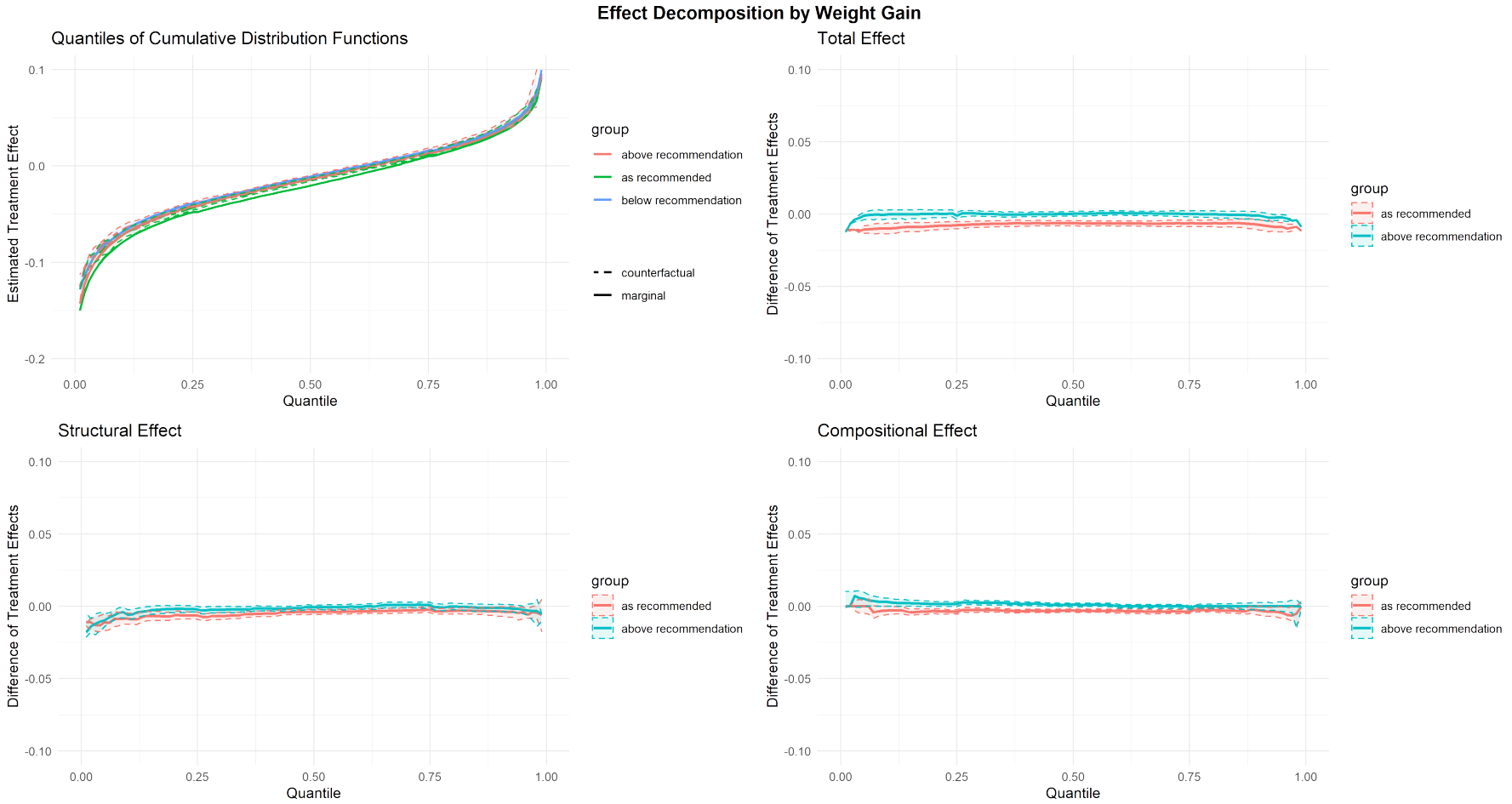


Figure A.57.: Apgar Score - Effect Decomposition by Weight Gain Recommendations

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on 5-minute Apgar score by weight gain recommendations. 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 gaining less than recommended based on guidelines published by CDC – National Center for Health Statistics (2021). Decomposition follows the procedure described in 1. Total, structural and compositional effect are defined as in equation (2.12). Shaded areas show 95% confidence intervals.

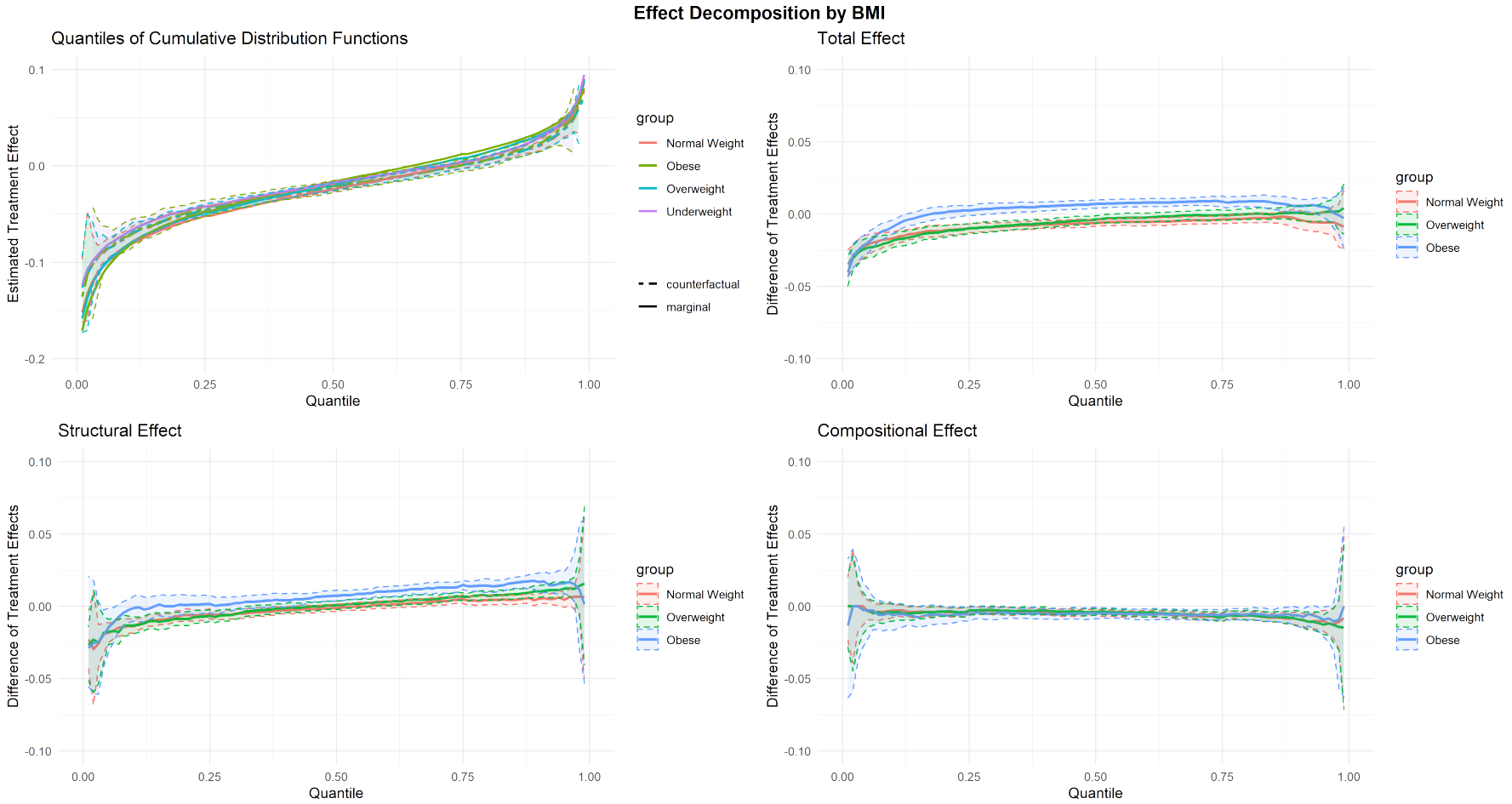


Figure A.58.: Apgar Score - Effect Decomposition by prepregnancy BMI

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on 5-minute Apgar score by prepregnancy BMI. 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 underweight mothers. Decomposition follows the procedure described in 1. Total, structural and compositional effect are defined as in equation (2.12). Positive difference corresponds to effect mitigation with increasing BMI, whereas negative difference corresponds to effect amplification with increasing BMI respectively. Shaded areas show 95% confidence intervals.

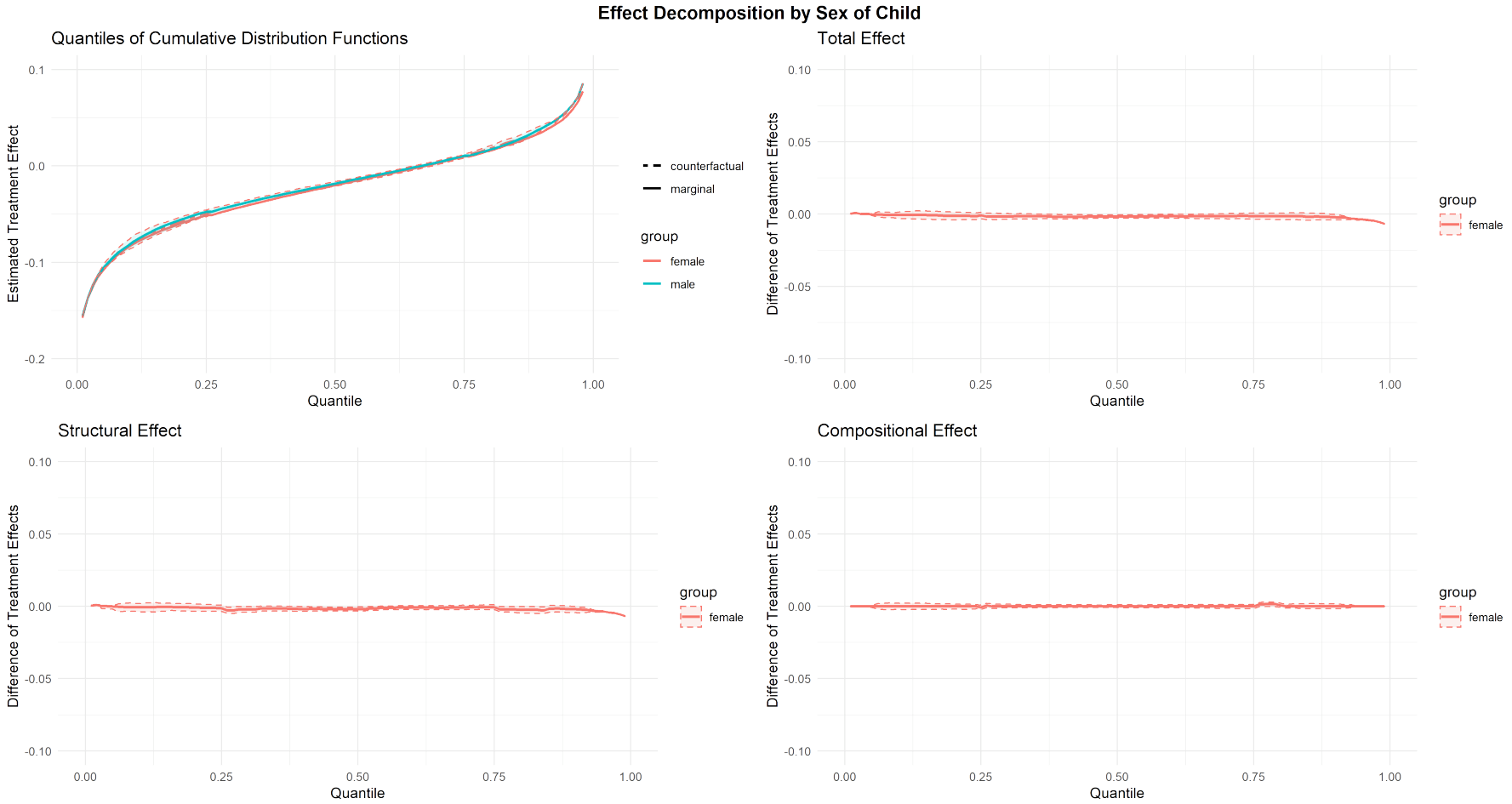


Figure A.59.: Apgar Score - Effect Decomposition by Sex of Newborn

*Note:* The figure displays the decomposition of the effect of maternal smoking during pregnancy on 5-minute Apgar score by sex of newborn. 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 underweight mothers. Decomposition follows the procedure described in 1. Total, structural and compositional effect are defined as in equation (2.12). Positive difference corresponds to effect mitigation with by sex. Shaded areas show 95% confidence intervals.

## A.9. Overview: Literature on Heterogeneity

Table A.3.: Overview Medical Literature on Heterogeneity

| | Outcome | Findings | N Obs | Method | Controls |
|---|---|---|---|---|---|
| Cnattingius et al. (1985): "Smoking, Maternal Age, and Fetal Growth" | birth weight for gestational age (standardized birth weight) | • smoking is the most important risk factor for fetal growth retardation<br>• among smokers reduction in birth weight is stronger with increasing maternal age<br>• maternal age does not influence standardized birth weight on its own | 3,022 | regression analysis | parity, maternal age, prepregnancy weight, maternal height, smoking, renal disease, previous live birth $< 2500$g, previous stillbirth, alcohol addiction, recurrent spontaneous abortions, hypertension, elevated AFP, uterine anomaly, urinary tract infection, vaginal bleeding, general disease |

Table A.3 : Continued from previous page

| | Outcome | Findings | Sample Size | Method | Controls |
|---|---|---|---|---|---|
| Cnattingius (1989): "Does age potentiate the smoking-related risk of fetal growth retardation? " | small-for-gestational-age (SGA) | • significant interaction between maternal age and moderate or heavy smoking<br>• relative risk for SGA for heavy smokers vs non-smokers is 1.9 in women aged 15-19<br>• relative risk for SGA for heavy smokers vs non-smokers is 3.4 in women aged 40-44 | 280,809 | logistic regression | maternal age, parity, relationship with father, smoking habits, type of birth (singleton, multiples) |
| Cnattingius et al. (1993): "Effect of age, parity, and smoking on pregnancy outcome: A population-based study" | low birth weight and preterm delivery, fetal death | • among multiparas, the smoking-related effect on the increase in the odds ratio of low birth weight and preterm delivery is greater than among nulliparas<br>• there is a smoking-related relative increase in the odds ratios for SGA births with advancing maternal age | 538,829 | logistic regression | maternal age, parity, smoking habits |

Table A.3 : Continued from previous page

| | Outcome | Findings | Sample Size | Method | Controls |
|---|---|---|---|---|---|
| Cnattingius (1997): "Maternal Age Modifies the Effect of Maternal Smoking on Intrauterine Growth Retardation but Not on Late Fetal Death and Placental Abruption" | late fetal death, placental abruption, and SGA birth | • effect modification of smoking by maternal age is found only regarding fetal growth<br>• for smoking women aged 40-44 years, the risk increase of SGA births was 4.5<br>• risk increase of SGA birth among smoking teenagers is only 2.0 | 1,057,711 | logistic regression | maternal age, cigarette smoking, parity, cohabitation with father |
| D'Souza et al. (1981): "Smoking in pregnancy: associations with skinfoldthickness, maternal weight gain, and fetal size at birth" | skinfold thickness, birth weight, head circumference | • heavy smokers gain significantly less weight than non-smokers<br>• no significant difference in skinfold thickness between smoker and non-smokers<br>• fetal growth retardation is not caused by nutritional deficiencies | 452 | mean comparison using $t$-test | smoking, maternal age, height, parity, duration of pregnancy |

Table A.3 : Continued from previous page

| | Outcome | Findings | Sample Size | Method | Controls |
|---|---|---|---|---|---|
| Haworth et al. (1980b): "Fetal growth retardation in cigarette-smoking mothers is not due to decreased maternal food intake" | birth weight, length, head circumference, APGAR Score | • smokers have significantly smaller infants, while having comparable pregnancy weight gain<br>• dietary intake of smokers is not less than that of the nonsmokers<br>• fetal growth retardation due to smoking is not caused by the mother's diminished intake of food | 536 | mean comparison, t-tests, regression analysis | dietary intake, smoking habits, maternal age, height, pregravid weight, pregnancy weight gain, parity, ethnic origin, educational level, family income, insurance status |
| Haworth et al. (1980a): "Relation of maternal cigarette smoking, obesity, and energy consumption to infant size" | birth weight and crown-heel length | • birth weight and length increases significantly with increasing maternal weight<br>• maternal obesity and cigarette smoking act independently of each other<br>• maternal overweight does not protect the fetus against growth-retardation by smoking | 536 | mean comparisons, testing for significance of mean differences | dietary intake, smoking habits, maternal age, height, prepregnancy weight, pregnancy weight gain, parity, ethnic origin, educational level, family income, insurance status, marital status |

Table A.3 : Continued from previous page

| | Outcome | Findings | Sample Size | Method | Controls |
|---|---|---|---|---|---|
| La Merrill et al. (2011): "Prepregnancy body mass index, smoking during pregnancy, and infant birth weight" | SGA, birth weight | • increasing prepregnancy BMI reduces the risk of SGA and increases birth weight<br>• effect of smoking during pregnancy on SGA and birth weight is noticeably reduced among obese and overweight women | 34,928 | GLM | maternal race, ethnicity, maternal birthplace, education, parity, delivery year, sex of newborn |
| Misra et al. (2005): " Maternal smoking and birth weight: Interaction with parity and mother's own in utero exposure to smoking" | birth weight | • maternal smoking reduces birth weight<br>• effect size on birth weight moderated by parity and the mother's own in utero exposure to smoking | 989 (+500) | OLS, Generalized Estimating Equations (GEE) | mother: race, parity, height, age, education, smoking, birth weight, IUGR, welfare program participation, hospitalization as child; grandmother: adult height, BMI, education, poverty ratio, smoking, sexually transmitted disease |

Table A.3 : Continued from previous page

| | Outcome | Findings | Sample Size | Method | Controls |
|---|---|---|---|---|---|
| Muscati et al. (1996): "Increased Energy Intake in Pregnant Smokers Does Not Prevent Human Fetal Growth Retardation" | SGA | • smoking is independently associated with higher energy intake, but lower maternal weight gain and birth weight<br>• energy intake is positively associated with a small increment in birth weight<br>• negative effect of smoking fetal growth retardation cannot be mitigated by increasing energy intake | 1,339 | logistic regression | maternal pregravid weight, height, pregnancy weight gain, smoking status, physical activity, energy intake throughout the duration of pregnancy, birth weight |
| Nabet et al. (2007): "Smoking during pregnancy according to obstetric complications and parity: results of the europop study" | preterm delivery | • smoking increases the risk of preterm delivery<br>• the risk of preterm delivery associated with smoking is higher for multiparae than primiparae | 9,389 | logistic regression | smoking during pregnancy, maternal age, maternal weight, height, marital status, educational level, working during pregnancy, obstetric history, complications during pregnancy, gestational age at birth, birth weight and the clinical condition of the newborn |

Table A.3 : Continued from previous page

| | Outcome | Findings | Sample Size | Method | Controls |
|---|---|---|---|---|---|
| Raymond et al. (1994): "Effects of maternal age, parity, and smoking on the risk of stillbirth " | risk of stillbirth; growth retardation | • women 35 years or older, smokers, and nulliparas have higher risks of stillbirth<br>• the association between smoking and stillbirth is explained by the combination of intrauterine growth retardation and placental complications | 638,242 | association tested $\chi^2$-test, logistic regression | maternal age, parity, smoking, pregnancy complications (hypertension, diabetes, placental complications), gestational age |
| Spinillo et al. (1994b): "Factors potentiating the smoking-related risk of fetal growth retardation" | fetal growth retardation | • factors independently increasing the smoking-related risk of fetal growth retardation: male fetus, nulliparity, maternal age 20 years or less, history of first trimester haemorrhage, low (less than 50 kg) prepregnancy weight | 1,041 | logistic regression | maternal smoking, maternal age, education, marital status, parity, prepregnancy weight, BMI, weight gain, previous LBW birth, sex, haemorrhage, hypertension, alcohol, coffee |

Table A.3 : Continued from previous page

| | Outcome | Findings | Sample Size | Method | Controls |
|---|---|---|---|---|---|
| Spinillo et al. (1994a): "Interaction between fetal gender and risk factors for fetal growth retardation" | fetal growth retardation | • fetal growth retardation is more common in female than male fetuses<br>• maternal smoking in pregnancy is a significant risk factor for growth retardation in both male and female fetuses<br>• effect of maternal smoking is significantly stronger in male fetuses | 1,312 | logistic regression | social class, education, prepregnancy weight, BMI, previous LBW infant, smoking, maternal age, hypertension, anemia, placenta previa, low weight gain, sex of newborn |
| Suzuki et al. (2011): "Gender differences in the association between maternal smoking during pregnancy and childhood growth trajectories: multilevel analysis" | BMI of child | • mean birth weight of both male and female children born to smoking mothers is significantly lower than for non-smokers<br>• smoking during pregnancy decreases infant birth weight regardless of gender | 1,619 | OLS | birth order, birth weight, gestational week of delivery and maternal BMI at the first pregnant checkup between the smoking and non-smoking mothers by the children's gender |

Table A.3 : Continued from previous page

| | Outcome | Findings | Sample Size | Method | Controls |
|---|---|---|---|---|---|
| Varvarigou et al. (2009): "Impact of Maternal Smoking on Birth Size: Effect of Parity and Sex Dimorphism" | birth weight, length and head circumference | • maternal smoking during pregnancy causes a delay in fetal growth, which is greater in male offspring<br>• the effect is enhanced with increasing parity but independent of maternal age | 2,108 | t-test, one-way ANOVA, Mann-Whitney U test, Spearman rank correlation | maternal smoking status and number of cigarettes smoked per day, age, parity |
| Wu Wen et al. (1990):"Smoking, maternal age, fetal growth, and gestational age at delivery" | birth weight, intrauterine growth retardation, and preterm delivery | • effect of smoking on fetal growth and gestational age is significantly greater with advancing maternal age<br>• smoking reduces birth weight by 134 g in young women, and 301 g in women older than 35 | 15,539 | logistic regression | race, parity, marital status, maternal weight, weight gain, and alcohol use |

Table A.3 : Continued from previous page

| | Outcome | Findings | Sample Size | Method | Controls |
|---|---|---|---|---|---|
| Zarén et al. (2000): "Maternal smoking affects fetal growth more in the male fetus" | birth weight, head circumference, ultrasound measurements | • negative effect of smoking on fetal growth is stronger for male fetuses<br>• birth weight reduction for male fetus: 8.2%; only 4.8% for female fetus | 856 | ANOVA $t$-test, OLS | age, parity, smoking, alcohol, height, weight gain, prepregnancy weight, BMI, gestation age |
| Zhang and Yang (2019): "Maternal smoking and infant low birth weight: Exploring the biological mechanism through the mother's prepregnancy weight status" | LBW | • increased maternal BMI reduces the odds of delivering a LBW infant<br>• BMI explains about 10.2% of the adverse impact of maternal smoking on having a LBW child | 6,550 | regression models, $\chi^2$-tests | race, marital status, poverty status, employment status, mothers age, education, weight gain, sex, parity |

# Appendix B.

# Appendix of Chapter 3

## B.1. Group Fixed Effects

To minimize 3.5, we make use of an iterative algorithm proposed by Bonhomme and Manresa (2015). Given a fixed number of groups, chosen by the researcher, we use an iterative algorithm consisting of an assignment and an update step, which are repeated until numerical convergence. The algorithm for GFE assignment is very similar to the well-known clustering algorithm kmeans.

---

**Procedure 2:** GFE estimator - Iterative:

---

For a given number of groups $g \in \{1, \ldots, G\}$:

1. Set random starting value $(\theta^{(0)}, \alpha^{(0)})$, $i = 0$.
2. Compute for all $s \in \{1, \ldots, N\}$:

$$g_s^{(i+1)} = \underset{g}{\operatorname{argmin}} \sum_{t=1}^{T} (\text{Smoking}_{st} - x'_{st}\theta - \alpha_{g_s t}^{(i)})^2$$

3. Compute

$$(\theta^{(i+1)}, \alpha^{(i+1)}) = \underset{\theta,\alpha}{\operatorname{argmin}} \sum_{s=1}^{N} \sum_{t=1}^{T} (\text{Smoking}_{st} - x'_{st}\theta - \alpha_{g_s^{(i+1)} t})^2$$

4. Set $i = i + 1$ and repeat until convergence.

---

# Appendix C.

# Appendix of Chapter 4

## Appendix

## C.1. Weather Data Overview

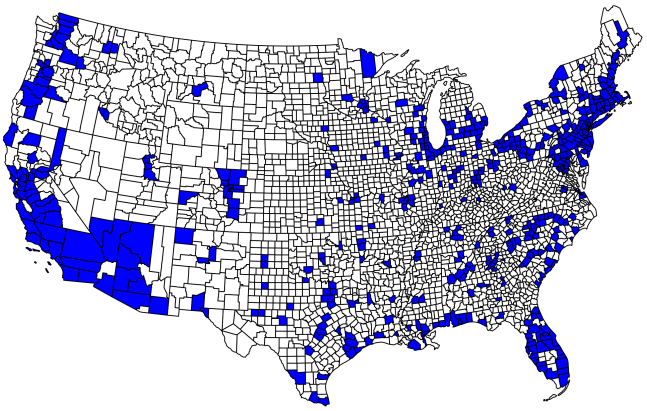

Figure C.1.: Counties with a population larger than 100,000 included in the analysis

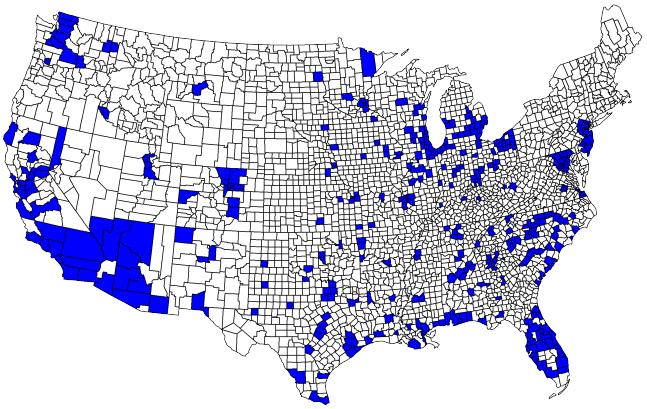

Figure C.2.: Counties with sufficient overlap in propensity score estimates

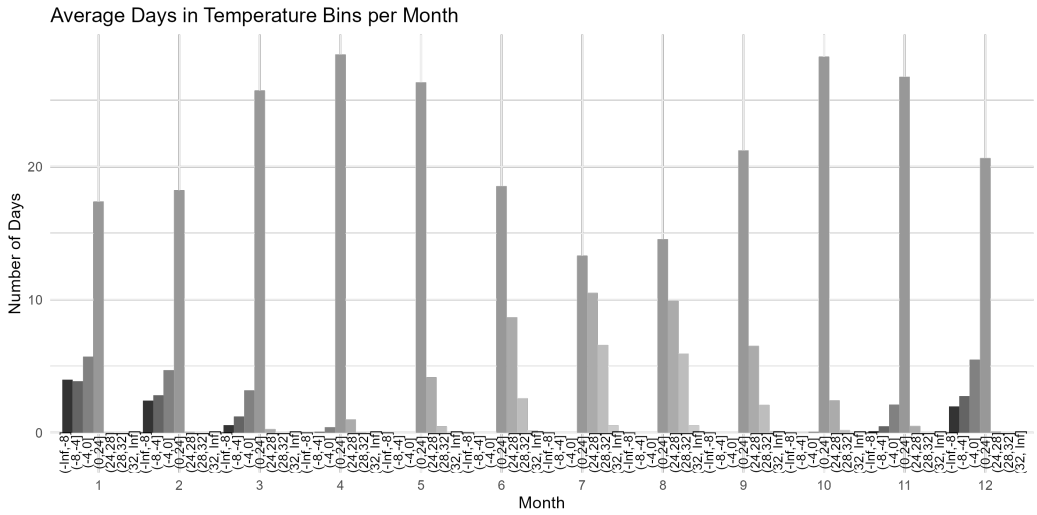


Figure C.3.: Average count of days in temperature bins per month

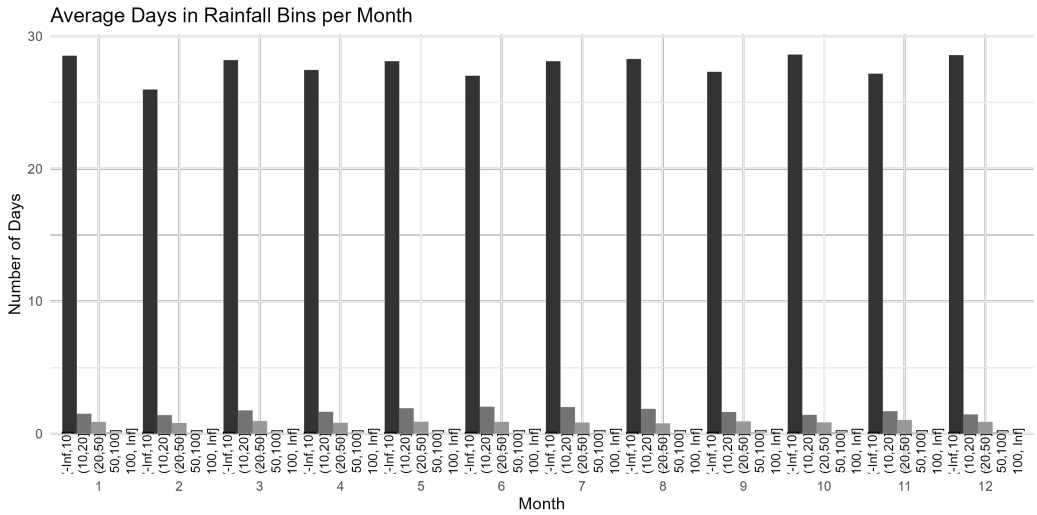


Figure C.4.: Average count of days in rainfall bins per month

## C.2. Mechanical Correlation of Gestation and Temperature Exposure

Table C.1 turns to an illustration of the mechanical correlation between gestation length and temperature exposure and the resulting problems with estimated effects. The estimates show effect patterns for the first and second trimesters, which are in line with what we found for the standardized outcomes. However, turning to the effects in the third trimester, we observe very large positive effects. The large positive effects arise from the mechanical correlation between gestation length and exposure to temperature. These mainly measure the effect of an additional day of the pregnancy rather than the actual effect of temperature exposure. This is because not all pregnancies observed last 9 months. We might therefore compare very lightweight preterm babies to full-term babies, whose count in each temperature bin is most likely to be larger, given a longer gestation length. Therefore, we see large positive effects of the exposure, rather than the expected smaller effects.

Table C.1.: Trimester Temperature Effects: Mechanical Correlation

| Dependent Variables: | Birth Weight | | | Gestation Length | | |
|---|---|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) | (5) | (6) |
| *Variables* | | | | | | |
| TR1: Avg. Temp. $< -8°C$ | 0.0170 | | | -0.0001 | | |
| | (0.0450) | | | (0.0003) | | |
| TR1: Avg. Temp. -8 - -4 °C | 0.0774 | | | 0.0000 | | |
| | (0.0513) | | | (0.0003) | | |
| TR1: Avg. Temp. -4 - 0 °C | 0.0308 | | | -0.0005** | | |
| | (0.0354) | | | (0.0002) | | |
| TR1: Avg. Temp. 24 - 28 °C | -0.0045 | | | -0.0004** | | |
| | (0.0310) | | | (0.0002) | | |
| TR1: Avg. Temp. 28 - 32 °C | -0.0075 | | | -0.0008* | | |
| | (0.0451) | | | (0.0004) | | |
| TR1: Avg. Temp. >32 °C | -0.1618 | | | -0.0001 | | |
| | (0.1158) | | | (0.0008) | | |
| TR2: Avg. Temp. $< -8°C$ | | -0.0164 | | | 0.0004 | |
| | | (0.0520) | | | (0.0002) | |
| TR2: Avg. Temp. -8 - -4 °C | | 0.0151 | | | -0.0004 | |
| | | (0.0561) | | | (0.0003) | |
| TR2: Avg. Temp. -4 - 0 °C | | -0.1307*** | | | -0.0004* | |
| | | (0.0418) | | | (0.0002) | |
| TR2: Avg. Temp. 24 - 28 °C | | -0.1174*** | | | -0.0006*** | |
| | | (0.0259) | | | (0.0002) | |
| TR2: Avg. Temp. 28 - 32 °C | | -0.1304*** | | | -0.0003 | |
| | | (0.0415) | | | (0.0003) | |
| TR2: Avg. Temp. > 32 °C | | -0.4329*** | | | -0.0010 | |
| | | (0.0633) | | | (0.0008) | |
| TR3: Avg. Temp. $< -8°C$ | | | 7.9138*** | | | 0.0770*** |
| | | | (0.3602) | | | (0.0037) |
| TR3: Avg. Temp. -8 - -4 °C | | | 5.2416*** | | | 0.0483*** |
| | | | (0.2737) | | | (0.0026) |
| TR3: Avg. Temp. -4 - 0 °C | | | 10.9569*** | | | 0.1043*** |
| | | | (0.3304) | | | (0.0035) |
| TR3: Avg. Temp. 24 - 28 °C | | | 7.9763*** | | | 0.0817*** |
| | | | (0.1633) | | | (0.0017) |
| TR3: Avg. Temp. 28 - 32 °C | | | 6.1534*** | | | 0.0646*** |
| | | | (0.2224) | | | (0.0025) |
| TR3: Avg. Temp. > 32 °C | | | 6.7083*** | | | 0.0743*** |
| | | | (0.2554) | | | (0.0028) |
| Weather Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Mother's Characteristics | Yes | Yes | Yes | Yes | Yes | Yes |
| *Fixed-effects* | | | | | | |
| Year of Birth | Yes | Yes | Yes | Yes | Yes | Yes |
| Month of Birth | Yes | Yes | Yes | Yes | Yes | Yes |
| County | Yes | Yes | Yes | Yes | Yes | Yes |
| State | Yes | Yes | Yes | Yes | Yes | Yes |
| Month of Birth-County | Yes | Yes | Yes | Yes | Yes | Yes |
| *Fit statistics* | | | | | | |
| Observations | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 |
| $R^2$ | 0.1105 | 0.1105 | 0.1295 | 0.0402 | 0.0402 | 0.1841 |
| Within $R^2$ | 0.1002 | 0.1002 | 0.1195 | 0.0295 | 0.0295 | 0.1750 |

*Notes:* Entries show the coefficient on the relevant temperature exposure measure. Samples uses all birth in counties with more than 100,000 inhabitants and no missing information in the variables of interest. Weather controls include average rainfall, average sunlight, and average snowfall in the 9 month after pregnancy start. Mother's controls include mother's age, mother's race, mother's Hispanic origin, mother's education, marital status, sex of child, month prenatal care began, number of prenatal visits, birth order, smoking during pregnancy, diabetes, and hypertension. Standard errors clustered by County of residence in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

## C.3. Decomposition

**Procedure 3:** Effect Decomposition

Repeat multiple times using different random splits into auxiliary and main sample:

- Split the sample into auxiliary (A) and main sample (M)
- On Auxiliary sample: train the causal forest
- 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 $\hat{F}_{X_k}(x) = \frac{1}{n_k} \sum_{i=1}^{n_k} 1_{\{X_{ki} \leq x\}}, \quad k \in \mathcal{K}$
    - Obtain estimates $\hat{F}_{\tau(x)^{(0)}|X_0}$ of the conditional distribution via $F_{Y|X}(y|x) = \Lambda(P(x)'\beta(y)), \quad \forall y \in \mathcal{Y}$
    - Obtain estimates of the counterfactual distributions via $F_{\tau\langle j|k\rangle}(y) := \int_{\mathcal{X}_k} F_{\tau_j|X_j}(y|x) dF_{X_k}(x)$
    - Obtain decomposition via

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

*Note:* This effect decomposition algorithm can also be used for quantiles by using following transformation $Q_{\tau\langle j|k\rangle}(y) := F^{-1}_{\tau\langle j|k\rangle}(p), \;\; p \in (0,1)$.

## C.4. Temperature Effects - Trimester

We further analyze sensitivity to the timing of the temperature exposure. Table C.2 shows the effects separately for each trimester, following equation (4.12). Results are not sensitive to including additional controls such as weather and mother's fixed effects. For standardized birth weight (columns (1), (2), (3)), we observe an increase in effect size with increasing pregnancy length. The effects of exposure to hot temperatures are most harmful in the second and third trimester. For the effects of cold exposure, we do only observe an increase in the third trimester. Regarding SGA, there is no strong effect in either the first or the second trimester. Significant effects are concentrated in the third trimester. Effects of temperature exposure on standardized Apgar Score are mostly not significant, exceptions are an increase in Apgar score found for extreme heat in the first and second trimester and a decrease for mild warm temperatures between 24-28°C in the third trimester.

Table C.2.: Trimester Temperature Effects

| Dependent Variables: | Standardized Birth Weight | | | SGA | | | Standardized Apgar Score | | |
|---|---|---|---|---|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| *Variables* | | | | | | | | | |
| TR1: Avg. Temp. <-8°C | 0.0694 | | | 0.0404 | | | -0.3315 | | |
| | (0.0931) | | | (0.0275) | | | (0.2334) | | |
| TR1: Avg. Temp. -8 - -4 °C | 0.1546 | | | -0.0261 | | | 0.3827 | | |
| | (0.1038) | | | (0.0266) | | | (0.2754) | | |
| TR1: Avg. Temp. -4 - 0 °C | 0.1807** | | | -0.0078 | | | -0.0037 | | |
| | (0.0725) | | | (0.0189) | | | (0.1486) | | |
| TR1: Avg. Temp. 24 - 28 °C | 0.0887* | | | -0.0204 | | | 0.0388 | | |
| | (0.0538) | | | (0.0140) | | | (0.0708) | | |
| TR1: Avg. Temp. 28 - 32 °C | 0.0902 | | | -0.0346** | | | -0.0054 | | |
| | (0.0605) | | | (0.0171) | | | (0.0939) | | |
| TR1: Avg. Temp. > 32 °C | -0.2429* | | | 0.0637*** | | | 0.4349*** | | |
| | (0.1288) | | | (0.0231) | | | (0.1571) | | |
| TR2: Avg. Temp. < −8°C | | -0.0979 | | | 0.0330 | | | -0.0808 | |
| | | (0.1148) | | | (0.0302) | | | (0.2343) | |
| TR2: Avg. Temp. -8 - -4 °C | | 0.0858 | | | -0.0834** | | | 0.2481 | |
| | | (0.1219) | | | (0.0335) | | | (0.2775) | |
| TR2: Avg. Temp. -4 - 0 °C | | -0.1895** | | | 0.0302 | | | 0.0661 | |
| | | (0.0860) | | | (0.0222) | | | (0.1641) | |
| TR2: Avg. Temp. 24 - 28 °C | | -0.0815 | | | -0.0021 | | | -0.0992 | |
| | | (0.0530) | | | (0.0139) | | | (0.0733) | |
| TR2: Avg. Temp. 28 - 32 °C | | -0.1102* | | | 0.0183 | | | -0.0633 | |
| | | (0.0609) | | | (0.0200) | | | (0.0967) | |
| TR2: Avg. Temp. > 32 °C | | -0.5523*** | | | 0.0612*** | | | 0.5212*** | |
| | | (0.0880) | | | (0.0176) | | | (0.1589) | |
| TR3: Avg. Temp. < −8°C | | | 0.3122 | | | -0.0691 | | | -0.2026 |
| | | | (0.2270) | | | (0.0481) | | | (0.2616) |
| TR3: Avg. Temp. -8 - -4 °C | | | 0.6775*** | | | -0.1600*** | | | 0.3490 |
| | | | (0.1902) | | | (0.0435) | | | (0.2903) |
| TR3: Avg. Temp. -4 - 0 °C | | | 0.8087*** | | | -0.1697*** | | | 0.2743 |
| | | | (0.2186) | | | (0.0413) | | | (0.1766) |
| TR3: Avg. Temp. 24 - 28 °C | | | -0.8682*** | | | 0.1283*** | | | -0.2701*** |
| | | | (0.1379) | | | (0.0258) | | | (0.0947) |
| TR3: Avg. Temp. 28 - 32 °C | | | -1.0127*** | | | 0.1313*** | | | 0.0743 |
| | | | (0.1579) | | | (0.0267) | | | (0.1979) |
| TR3: Avg. Temp. > 32 °C | | | -1.8491*** | | | 0.2694*** | | | -0.1215 |
| | | | (0.0852) | | | (0.0533) | | | (0.2480) |
| Weather Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Mother's Characteristics | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| *Fixed-effects* | | | | | | | | | |
| Year of Birth | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Month of Birth | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| County | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| State | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Month of Birth-County | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| *Fit statistics* | | | | | | | | | |
| Observations | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 |
| $R^2$ | 0.0900 | 0.0900 | 0.0901 | 0.0376 | 0.0376 | 0.0376 | 0.0283 | 0.0283 | 0.0283 |
| Within $R^2$ | 0.0815 | 0.0815 | 0.0816 | 0.0338 | 0.0338 | 0.0338 | 0.0025 | 0.0025 | 0.0025 |

*Notes:* Entries show the coefficient on the relevant temperature exposure measure, scaled by 1000 for ease of reading. This means that a one-unit change in *Avg. Temp. <-8° C* is associated with a 0.0004 unit change in standardized birth weight in column (1), holding all other variables constant. Samples uses all birth in counties with more than 100,000 inhabitants and no missing information in the variables of interest. Weather controls include average rainfall, average sunlight, and average snowfall in the 9 month after pregnancy start. Mother's controls include mother's age, mother's race, mother's Hispanic origin, mother's education, marital status, sex of child, month prenatal care began, number of prenatal visits, birth order, smoking during pregnancy, diabetes, and hypertension.
Standard errors clustered by County of residence in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

## C.5. Rainfall Effects

We additionally analyze effects of rainfall exposure. Table C.3 shows the results of equation (4.11) for exposure to rainfall using our preferred specification including mother's characteristics and weather controls. Milder rainfall of 10 to 50 mm per day shows slightly positive effects compared to exposure to rainfall between 0 and 10mm. Exposure to very extreme rainfall shows a weakly significant reduction of 0.0034 which corresponds to a reduction of 1.63 grams on average. Including additional controls (column (2)), none of the estimated effects remains significant. Similarly, we cannot observe any significant effect of rainfall on SGA birth (column (4)), indicating that most vulnerable infants are not affected by rainfall. For the standardized Apgar score, we see a significant reduction in exposure to extreme rainfall. Similarly, we cannot find significant effects of any rainfall bin when analyzing trimester-specific settings as specified in equation (4.12). See table C.4 for effect estimates for rainfall exposure. We therefore cannot find significant effects of rainfall exposure on any of the outcomes of health at birth. No effects of rainfall on health at birth is partly in line with previous literature. Andalón et al. (2016) find no significant effect of either drought or heavy rainfall on birth weight ad gestation length. In contrast, Rocha and Soares (2015) find positive effects of positive rainfall shocks during pregnancy on birth weight. They however study birth in rural Brazil, where no rainfall easily leads to water scarcity.

Table C.3.: Rainfall Effects

| Dependent Variables: | Standardized Birth Weight | | SGA | | Standardized Apgar Score | |
|---|---|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) | (5) | (6) |
| *Variables* | | | | | | |
| Rain 10-20mm | 0.6208*** | 0.1173 | -0.1530*** | -0.0297 | 0.0259 | 0.0183 |
| | (0.1122) | (0.1382) | (0.0230) | (0.0292) | (0.1942) | (0.1956) |
| Rain 20-50mm | 0.3631*** | -0.1776 | -0.1140*** | 0.0033 | -1.0747*** | -1.0087*** |
| | (0.1347) | (0.1677) | (0.0311) | (0.0405) | (0.3631) | (0.3666) |
| Rain 50-1000mm | 0.3772 | 0.1585 | -0.1329 | -0.1018 | 2.0544 | 2.4154* |
| | (0.4733) | (0.4463) | (0.1186) | (0.1057) | (1.3634) | (1.4536) |
| Rain +100mm | -3.3764* | -2.0670 | 0.4401 | 0.0979 | -12.3021*** | -12.0226** |
| | (1.9932) | (2.2034) | (0.5869) | (0.6275) | (4.2158) | (4.8551) |
| Weather Controls | | Yes | | Yes | | Yes |
| Mother's Characteristics | | Yes | | Yes | | Yes |
| *Fixed-effects* | | | | | | |
| Year of Birth | Yes | Yes | Yes | Yes | Yes | Yes |
| Month of Birth | Yes | Yes | Yes | Yes | Yes | Yes |
| County | Yes | Yes | Yes | Yes | Yes | Yes |
| State | Yes | Yes | Yes | Yes | Yes | Yes |
| Month of Birth-County | Yes | Yes | Yes | Yes | Yes | Yes |
| *Fit statistics* | | | | | | |
| Observations | 19,865,677 | 18,595,704 | 19,865,677 | 18,595,704 | 19,865,677 | 18,595,704 |
| $R^2$ | 0.0091 | 0.0900 | 0.0039 | 0.0376 | 0.0265 | 0.0283 |
| Within $R^2$ | 0.0000 | 0.0815 | 0.0000 | 0.0338 | 0.0000 | 0.0025 |

*Notes:* Entries show the coefficient on the relevant rainfall exposure measure, scaled by 1000 for ease of reading. This means that a one-unit change in *Rain 10-20mm* is associated with a 0.0006 unit change in standardized birth weight in column (1), holding all other variables constant. Samples uses all birth in counties with more than 100,000 inhabitants and no missing information in the variables of interest. Weather controls include average rainfall, average sunlight, and average snowfall in the 9 month after pregnancy start. Mother's controls include mother's age, mother's race, mother's Hispanic origin, mother's education, marital status, sex of child, month prenatal care began, number of prenatal visits, birth order, smoking during pregnancy, diabetes, and hypertension.
Standard errors clustered by County of residence in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table C.4.: Trimester Rainfall Effects

| Dependent Variables: | Standardized Birth Weight | | | SGA | | | Standardized Apgar Score | | |
|---|---|---|---|---|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| *Variables* | | | | | | | | | |
| 1st Trimester: Rain 10-20mm | 0.1488 | | | -0.0309 | | | 0.0202 | | |
| | (0.1178) | | | (0.0316) | | | (0.2056) | | |
| 1st Trimester: Rain 20-50mm | -0.0833 | | | 0.0496 | | | -0.9311*** | | |
| | (0.1713) | | | (0.0459) | | | (0.3505) | | |
| 1st Trimester: Rain 50-1000mm | -0.1022 | | | 0.0017 | | | 2.1038 | | |
| | (0.5296) | | | (0.1514) | | | (1.5005) | | |
| 1st Trimester: Rain +100mm | 0.6805 | | | -0.0834 | | | -16.5432*** | | |
| | (2.8214) | | | (0.7782) | | | (5.2608) | | |
| 2nd Trimester: Rain 10-20mm | | 0.0077 | | | 0.0217 | | | -0.0046 | |
| | | (0.1212) | | | (0.0359) | | | (0.2304) | |
| 2nd Trimester: Rain 20-50mm | | -0.2106 | | | -0.0070 | | | -0.6940* | |
| | | (0.1840) | | | (0.0469) | | | (0.3841) | |
| 2nd Trimester: Rain 50-1000mm | | 0.2515 | | | -0.0812 | | | 2.5004* | |
| | | (0.6458) | | | (0.1660) | | | (1.4930) | |
| 2nd Trimester: Rain +100mm | | -2.8722 | | | -0.4605 | | | -9.2522* | |
| | | (2.4008) | | | (0.7331) | | | (4.9711) | |
| 3rd Trimester: Rain 10-20mm | | | 0.1990 | | | -0.0765 | | | 0.1055 |
| | | | (0.3400) | | | (0.0662) | | | (0.2659) |
| 3rd Trimester: Rain 20-50mm | | | -0.3037 | | | -0.0542 | | | -1.4165*** |
| | | | (0.3140) | | | (0.0799) | | | (0.4768) |
| 3rd Trimester: Rain 50-1000mm | | | 0.4325 | | | -0.3017 | | | 2.6743 |
| | | | (0.8256) | | | (0.1933) | | | (1.6302) |
| 3rd Trimester: Rain +100mm | | | -5.3092 | | | 1.1469 | | | -5.6922 |
| | | | (3.4064) | | | (1.1729) | | | (5.2821) |
| Weather Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Mother's Characteristics | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| *Fixed-effects* | | | | | | | | | |
| Year of Birth | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Month of Birth | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| County | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| State | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Month of Birth-County | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| *Fit statistics* | | | | | | | | | |
| Observations | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 | 18,595,704 |
| $R^2$ | 0.0900 | 0.0900 | 0.0900 | 0.0376 | 0.0376 | 0.0376 | 0.0283 | 0.0283 | 0.0283 |
| Within $R^2$ | 0.0815 | 0.0815 | 0.0815 | 0.0338 | 0.0338 | 0.0338 | 0.0025 | 0.0025 | 0.0025 |

*Notes:* Entries show the coefficient on the relevant rainfall exposure measure, scaled by 1000 for ease of reading. This means that a one-unit change in *Rain 10-20mm* is associated with a 0.0001 unit change in standardized birth weight in column (1), holding all other variables constant. Samples uses all birth in counties with more than 100,000 inhabitants and no missing information in the variables of interest. Weather controls include average rainfall, average sunlight, and average snowfall in the 9 month after pregnancy start. Mother's controls include mother's age, mother's race, mother's Hispanic origin, mother's education, marital status, sex of child, month prenatal care began, number of prenatal visits, birth order, smoking during pregnancy, diabetes, and hypertension.
Standard errors clustered by County of residence in parentheses.
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

# Bibliography


Abell, T., Baker, L., and Ramsey, C. (1991). The effects of maternal smoking on infant birth weight. *Family medicine*, 23(2):103—107.

Adams, E. K., Miller, V. P., Ernst, C., Nishimura, B. K., Melvin, C., and Merritt, R. (2002). Neonatal health care costs related to smoking during pregnancy. *Health economics*, 11:193–206.

Adda, J. and Cornaglia, F. (2010). The Effect of Bans and Taxes on Passive Smoking. *American Economic Journal: Applied Economics*, 2(1):1–32.

Agency for Healthcare Research and Quality (2019). Research Protocol: Smoking Cessation Interventions During Pregnancy and the Postpartum Period. https://effectivehealthcare.ahrq.gov/products/smoking-pregnancy-infants/research-protocol.

Almond, D., Chay, K. Y., and Lee, D. S. (2005). The Costs of Low Birth Weight. *The Quarterly Journal of Economics*, 120(3):1031–1083.

Almond, D. and Currie, J. (2011). Killing Me Softly: The Fetal Origins Hypothesis. *Journal of Economic Perspectives*, 25(3):153–72.

American Academy of Pediatrics (2015). The Apgar Score. *Pediatrics*, 136(4):819–822.

American College of Obstetricians and Gynecologists. (2015). Committee Opinion No. 644: The Apgar Score. *Obstetrics and gynecology*, 126:e52–e55.

Andalón, M., Azevedo, J. P., Rodríguez-Castelán, C., Sanfelice, V., and Valderrama-González, D. (2016). Weather shocks and health at birth in Colombia. *World Development*, 82:69–82.

Anger, S., Kvasnicka, M., and Siedler, T. (2011). One last puff? Public smoking bans and smoking behavior. *Journal of Health Economics*, 30(3):591–601.

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

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

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

Athey, S. and Wager, S. (2019). Estimating Treatment Effects with Causal Forests: An Application. *arXiv preprint arXiv:1902.07409*.

Baba, S., Wikström, A.-K., Stephansson, O., and Cnattingius, S. (2013a). Changes in snuff and smoking habits in Swedish pregnant women and risk for small for gestational age births. *BJOG: An International Journal of Obstetrics & Gynaecology*, 120(4):456–462.

Baba, S., Wikström, A.-K., Stephansson, O., and Cnattingius, S. (2013b). Influence of Snuff and Smoking Habits in Early Pregnancy on Risks for Stillbirth and Early Neonatal Mortality. *Nicotine & Tobacco Research*, 16(1):78–83.

Belloni, A., Chernozhukov, V., and Hansen, C. (2013). Inference on Treatment Effects after Selection among High-Dimensional Controls†. *The Review of Economic Studies*, 81(2):608–650.

Bergmann, R., Bergmann, K., Schumann, S., Richter, R., and Dudenhausen, J. (2008). Rauchen in der Schwangerschaft: Verbreitung, Trend, Risikofaktoren. *Zeitschrift für Geburtshilfe und Neonatologie*, 212(03):80–86.

Bharadwaj, P., Johnsen, J. V., and Løken, K. V. (2014). Smoking bans, maternal smoking and birth outcomes. *Journal of Public Economics*, 115:72–93.

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2007). From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes*. *The Quarterly Journal of Economics*, 122(1):409–439.

Blinder, A. S. (1973). Wage Discrimination: Reduced Form and Structural Estimates. *The Journal of Human Resources*, 8(4):436–455.

Bonhomme, S. and Manresa, E. (2015). Grouped Patterns of Heterogeneity in Panel Data. *Econometrica*, 83(3):1147–1184.

Bratti, M., Frimpong, P. B., and Russo, S. (2021). Prenatal Exposure to Heat Waves and Child Health in Sub-saharan Africa. IZA Discussion Papers 14424.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Cardenas, A., Lutz, S. M., Everson, T. M., Perron, P., Bouchard, L., and Hivert, M.-F. (2019). Mediation by Placental DNA Methylation of the Association of Prenatal Maternal Smoking and Birth Weight. *American Journal of Epidemiology*, 188(11):1878–1886.

Carpenter, C., Postolek, S., and Warman, C. (2011). Public-Place Smoking Laws and Exposure to Environmental Tobacco Smoke (ETS). *American Economic Journal: Economic Policy*, 3(3):35–61.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2):138–154.

CDC – National Center for Health Statistics (2021). Weight Gain During Pregnancy. `https://www.cdc.gov/reproductivehealth/maternalinfanthealth/pregnancy-weight-gain.htm#weight`. Accessed: 2021-08-13.

Chang, G., Favara, M., and Novella, R. (2022). The origins of cognitive skills and non-cognitive skills: The long-term effect of in-utero rainfall shocks in India. *Economics & Human Biology*, 44:101089.

Chen, M., Chernozhukov, V., Fernandez-Val, I., and Melly, B. (2020a). *Counterfactual: Estimation and Inference Methods for Counterfactual Analysis*. R package version 1.2.

Chen, X., Tan, C. M., Zhang, X., and Zhang, X. (2020b). The effects of prenatal exposure to temperature extremes on birth outcomes: the case of China. *Journal of Population Economics*, 33:1263–1302.

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

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2018b). 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.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on Counterfactual Distributions. *Econometrica*, 81(6):2205–2268.

Cnattingius, S. (1989). Does age potentiate the smoking-related risk of fetal growth retardation? *Early Human Development*, 20(3):203–211.

Cnattingius, S. (1997). Maternal Age Modifies the Effect of Maternal Smoking on Intrauterine Growth Retardation but Not on Late Fetal Death and Placental Abruption. *American Journal of Epidemiology*, 145(4):319–323.

Cnattingius, S. (2004). The epidemiology of smoking during pregnancy: Smoking prevalence, maternal characteristics, and pregnancy outcomes. *Nicotine & Tobacco Research*, 6(Suppl_2):S125–S140.

Cnattingius, S., Axelsson, O., Eklund, G., and Lindmark, G. (1985). Smoking, Maternal Age, and Fetal Growth. *Obstetrics & Gynecology*, 66(4).

Cnattingius, S., Forman, M. R., Berendes, H. W., Graubard, B. I., and Isotalo, L. (1993). Effect of age, parity, and smoking on pregnancy outcome: A population-based study. *American Journal of Obstetrics and Gynecology*, 168(1, Part 1):16–21.

Cohen, G., Roux, J.-C., Grailhe, R., Malcolm, G., Changeux, J.-P., and Lagercrantz, H. (2005). Perinatal exposure to nicotine causes deficits associated with a loss of nicotinic receptor function. *Proceedings of the National Academy of Sciences*, 102(10):3817–3821.

Conley, D. and Bennett, N. G. (2000). Is Biology Destiny? Birth Weight and Life Chances. *American Sociological Review*, 65(3):458–467.

Conti, G., Hanson, M., Inskip, H., Crozier, S., Cooper, C., and Godfrey, K. M. (2020). Beyond birthweight: The origins of human capital. *IZA Discussion Papers, No. 13296, Institute of Labor Economics (IZA), Bonn.*

Creamer, M. R., Wang, T. W., Babb, S., Cullen, K. A., Day, H., Willis, G., Jamal, A., and Neff, L. (2019). Tobacco product use and cessation indicators among adults—United States, 2018. *Morbidity and mortality weekly report*, 68(45):1013.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Currie, J. and Almond, D. (2011). Chapter 15 - Human capital development before age five. In Card, D. and Ashenfelter, O., editors, *Handbook of Labor Economics*, volume 4 of *Handbook of Labor Economics*, pages 1315–1486. Elsevier.

Currie, J., Neidell, M., and Schmieder, J. F. (2009). Air pollution and infant health: Lessons from New Jersey. *Journal of Health Economics*, 28(3):688–703.

Currie, J. and Rossin-Slater, M. (2013). Weathering the storm: Hurricanes and birth outcomes. *Journal of Health Economics*, 32(3):487–503.

Currie, J. and Schwandt, H. (2013). Within-mother analysis of seasonal patterns in health at birth. *Proceedings of the National Academy of Sciences*, 110(30):12265–12270.

Currie, J., Zivin, J. G., Meckel, K., Neidell, M., and Schlenker, W. (2013). Something in the water: contaminated drinking water and infant health. *The Canadian journal of economics. Revue canadienne d'economique*, 46(3):791–810. 27134285[pmid].

Curtin, S. C. and Matthews, T. J. (2016). Smoking Prevalence and Cessation Before and During Pregnancy: Data From the Birth Certificate, 2014. *National vital statistics reports : from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 65:1–14.

Davis, J. M. and Heller, S. B. (2017). Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs. *American Economic Review*, 107(5):546–50.

DEHOGA (2008). *Nichtraucherschutzgesetze in den Bundesländern. Synopse zu den Landesgesetzen.* Deutscher Hotel- und Gaststättenverband (DEHOGA), Berlin, Germany.

Deryugina, T., Heutel, G., Miller, N. H., Molitor, D., and Reif, J. (2019). The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction. *American Economic Review*, 109(12):4178–4219.

Deschênes, O. and Greenstone, M. (2011). Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the US. *American Economic Journal: Applied Economics*, 3(4):152–185.

Deschênes, O., Greenstone, M., and Guryan, J. (2009). Climate change and birth weight. *American Economic Review*, 99(2):211–217.

Drouin, O., Sato, R., Drehmer, J. E., Nabi-Burza, E., Hipple Walters, B., Winickoff, J. P., and Levy, D. E. (2021). Cost-effectiveness of a Smoking Cessation Intervention for Parents in Pediatric Primary Care. *JAMA Network Open*, 4(4):e213927–e213927.

D'Souza, S. W., Black, P., and Richards, B. (1981). Smoking in pregnancy: associations with skinfold thickness, maternal weight gain, and fetal size at birth. *BMJ*, 282(6277):1661–1663.

Ekblad, M., Gissler, M., Korkeila, J., and Lehtonen, L. (2013). Trends and risk groups for smoking during pregnancy in Finland and other Nordic countries. *European Journal of Public Health*, 24(4):544–551.

Elek, P. and Bíró, A. (2021). Regional differences in diabetes across Europe – regression and causal forest analyses. *Economics & Human Biology*, 40:100948.

England, L. J., Grauman, A., Qian, C., Wilkins, D. G., Schisterman, E. F., Yu, K. F., and Levine, R. J. (2007). Misclassification of maternal smoking status and its effects on an epidemiologic study of pregnancy outcomes. *Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco*, 9:1005–13.

England, L. J., Kendrick, J. S., Wilson, H. G., Merritt, R. K., Gargiullo, P. M., and Zahniser, S. C. (2001). Effects of Smoking Reduction during Pregnancy on the Birth Weight of Term Infants. *American Journal of Epidemiology*, 154(8):694–701.

Evans, W. N. and Ringel, J. S. (1999). Can higher cigarette taxes improve birth outcomes? *Journal of Public Economics*, 72(1):135–154.

Fleitmann, S., Dohnke, B., Balke, K., Rustler, C., and Sonntag, U. (2010). Frauen und Rauchen. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 53(2):117–124.

Garn, S. M., Johnston, M., Ridella, S. A., and Petzold, A. S. (1981). Effect of Maternal Cigarette Smoking on Apgar Scores. *American Journal of Diseases of Children*, 135(6):503–506.

Gilman, S. E., Gardener, H., and Buka, S. L. (2008). Maternal Smoking during Pregnancy and Children's Cognitive and Physical Development: A Causal Risk Factor? *American Journal of Epidemiology*, 168(5):522–531.

Gulen, H., Jens, C., and Page, T. B. (2020). An application of causal forest in corporate finance: How does financing affect investment? *Available at SSRN*.

Hamilton, B. H. (2001). Estimating treatment effects in randomized clinical trials with noncompliance: the impact of maternal smoking on birthweight. *Health Economics*, 10(5):399–410.

Hankins, S. and Tarasenko, Y. (2016). Do Smoking Bans Improve Neonatal Health? *Health Services Research*, 51(5):1858–1878.

Harrod, C. S., Reynolds, R. M., Chasan-Taber, L., Fingerlin, T. E., Glueck, D. H., Brinton, J. T., and Dabelea, D. (2014). Quantity and Timing of Maternal Prenatal Smoking on Neonatal Body Composition: The Healthy Start Study. *The Journal of Pediatrics*, 165(4):707–712.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer.

Haworth, J., Ellestad-Sayed, J., King, J., and Dilling, L. A. (1980a). Fetal growth retardation in cigarette-smoking mothers is not due to decreased maternal food intake. *American Journal of Obstetrics and Gynecology*, 137(6):719–723.

Haworth, J., Ellestad-Sayed, J. J., King, J., and Dilling, L. A. (1980b). Relation of maternal cigarette smoking, obesity, and energy consumption to infant size. *American Journal of Obstetrics and Gynecology*, 138(8):1185–1189.

Hebel, J., Fox, N. L., and Sexton, M. (1988). Dose-response of birth weight to various measures of maternal smoking during pregnancy. *Journal of Clinical Epidemiology*, 41(5):483–489.

Hegyi, T., Carbone, T., Anwar, M., Ostfeld, B., Hiatt, M., Koons, A., Pinto-Martin, J., and Paneth, N. (1998). The Apgar Score and Its Components in the Preterm Infant. *Pediatrics*, 101(1):77–81.

Hernández-Díaz, S., Schisterman, E. F., and Hernán, M. A. (2006). The Birth Weight "Paradox" Uncovered? *American Journal of Epidemiology*, 164(11):1115–1120.

Holbrook, B. D. (2016). The effects of nicotine on human fetal development. *Birth Defects Research Part C: Embryo Today: Reviews*, 108(2):181–192.

Howland, R. E., Mulready-Ward, C., Madsen, A. M., Sackoff, J., Nyland-Funke, M., Bombard, J. M., and Tong, V. T. (2015). 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*, 19:1916–24.

Imbens, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature*, 58(4):1129–79.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

IQTIG, Institut für Qualitätssicherung und Transparenz im Gesundheitswesen (2016). Qualitätsreport 2016. `https://iqtig.org/downloads/berichte/2016/IQTIG_Qualitaetsreport-2016.pdf`.

Isen, A., Rossin-Slater, M., and Walker, R. (2017). Relationship between season of birth, temperature exposure, and later life wellbeing. *Proceedings of the National Academy of Sciences*, 114(51):13447–13452.

Jacob, J., Lehne, M., Mischker, A., Klinger, N., Zickermann, C., and Walker, J. (2017). Cost effects of preterm birth: a comparison of health care costs associated with early preterm, late preterm, and full-term birth in the first 3 years after birth. *The European journal of health economics : HEPAC : health economics in prevention and care*, 18:1041–1046.

Jamal, A., King, B. A., Neff, L. J., Whitmill, J., Babb, S. D., and Graffunder, C. M. (2016). Current cigarette smoking among adults—United States, 2005–2015. *Morbidity and Mortality Weekly Report*, 65(44):1205–1211.

Jones, A. M., Laporte, A., Rice, N., and Zucchelli, E. (2015). Do Public Smoking Bans have an Impact on Active Smoking? Evidence from the UK. *Health Economics*, 24(2):175–192.

Ju, H., Chadha, Y., Donovan, T., and O'Rourke, P. (2009). Fetal macrosomia and pregnancy outcomes. *Australian and New Zealand Journal of Obstetrics and Gynaecology*, 49(5):504–509.

Karlsson, M. and Ziebarth, N. R. (2018). Population health effects and health-related costs of extreme temperatures: Comprehensive evidence from Germany. *Journal of Environmental Economics and Management*, 91:93–117.

Kathleen Adams, E., Miller, V. P., Ernst, C., Nishimura, B. K., Melvin, C., and Merritt, R. (2002). Neonatal health care costs related to smoking during pregnancy. *Health Economics*, 11(3):193–206.

Knopik, V. S., Marceau, K., Palmer, R. H. C., Smith, T. F., and Heath, A. C. (2016). Maternal Smoking During Pregnancy and Offspring Birth Weight: A Genetically-Informed Approach Comparing Multiple Raters. *Behavior Genetics*, 46(3):353–364.

Kogan, M. D., Alexander, G. R., Kotelchuck, M., MacDorman, M. F., Buekens, P., Martin, J. A., and Papiernik, E. (2000). Trends in Twin Birth Outcomes and Prenatal Care Utilization in the United States, 1981-1997. *JAMA*, 284(3):335–341.

Kramer, M. S. (1987). Intrauterine Growth and Gestational Duration Determinants. *Pediatrics*, 80(4):502–511.

Kreif, N., DiazOrdaz, K., Moreno-Serra, R., Mirelman, A., Hidayat, T., and Suhrcke, M. (2022). Estimating heterogeneous policy impacts using causal machine learning: a case study of health insurance reform in Indonesia. *Health Services and Outcomes Research Methodology*, 22(2):192–227.

Kuntz, B. and Lampert, T. (2016). Social Disparities in Maternal Smoking during Pregnancy. *Geburtshilfe Frauenheilkd*, 76(03):239–247. 239.

Kuntz, B., Zeiher, J., Starker, A., Prütz, F., and Lampert, T. (2018). Smoking during pregnancy. Results of the cross-sectional KiGGS Wave 2 study and trends. *Journal of Health Monitoring*, 3(1):45–51.

Kvasnicka, M., Siedler, T., and Ziebarth, N. R. (2018). The health effects of smoking bans: Evidence from German hospitalization data. *Health Economics*, 27(11):1738–1753.

La Merrill, M., Stein, C. R., Landrigan, P., Engel, S. M., and Savitz, D. A. (2011). Prepregnancy Body Mass Index, Smoking During Pregnancy, and Infant Birth Weight. *Annals of Epidemiology*, 21(6):413–420.

Lampert, T., Von Der Lippe, E., and Müters, S. (2013). Verbreitung des Rauchens in der Erwachsenenbevölkerung in Deutschland. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, 56(5-6):802–808.

Le, K. and Nguyen, M. (2021). The impacts of temperature shocks on birth weight in Vietnam. *Population and Development Review*, 47(4):1025–1047.

Levy, D. E., Regan, S., Perez, G. K., Muzikansky, A., Friedman, E. R., Rabin, J., Rigotti, N. A., Ostroff, J. S., and Park, E. R. (2022). Cost-effectiveness of Implementing Smoking Cessation Interventions for Patients With Cancer. *JAMA Network Open*, 5(6):e2216362–e2216362.

Lien, D. S. and Evans, W. N. (2005). Estimating the Impact of Large Cigarette Tax Hikes: The Case of Maternal Smoking and Infant Birth Weight. *Journal of Human Resources*, XL(2):373–392.

Lightwood, J. M., Phibbs, C. S., and Glantz, S. A. (1999). Short-term health and economic benefits of smoking cessation: low birth weight. *Pediatrics*, 104:1312–20.

Luck, W., Nau, H., Hansen, R., and Steldinger, R. (1985). Extent of nicotine and cotinine transfer to the human fetus, placenta and amniotic fluid of smoking mothers. *Developmental pharmacology and therapeutics*, 8:384–95.

Ludwig, D. S. and Currie, J. (2010). The association between pregnancy weight gain and birthweight: a within-family comparison. *The Lancet*, 376(9745):984–990.

Miller, S. (2020). Causal forest estimation of heterogeneous and time-varying environmental policy effects. *Journal of Environmental Economics and Management*, 103:102337.

Misra, D. P., Astone, N., and Lynch, C. D. (2005). Maternal Smoking and Birth Weight: Interaction with Parity and Mother's Own in Utero Exposure to Smoking. *Epidemiology*, 16(3):288–293.

Miyake, Y., Tanaka, K., and Arakawa, M. (2013). Active and passive maternal smoking during pregnancy and birth outcomes: the Kyushu Okinawa Maternal and Child Health Study. *BMC Pregnancy and Childbirth*, 13(1):157.

Molina, O. and Saldarriaga, V. (2017). The perils of climate change: In utero exposure to temperature variability and birth outcomes in the Andean region. *Economics & Human Biology*, 24:111–124.

Muscati, S. K., Koski, K. G., and Gray-Donald, K. (1996). Increased Energy Intake in Pregnant Smokers Does Not Prevent Human Fetal Growth Retardation. *The Journal of Nutrition*, 126(12):2984–2989.

Nabet, C., Lelong, N., Ancel, P.-Y., Saurel-Cubizolles, M.-J., and Kaminski, M. (2007). Smoking during pregnancy according to obstetric complications and parity: results of the EUROPOP study. *European Journal of Epidemiology*, 22(10):715–721.

National Center for Health Statistics (2011-2018). Vital Statistics Natality Birth Data, 2011-2018. Public-use data file and documentation.

Nordström, M.-L. and Cnattingius, S. (1994). Smoking habits and birthweights in two successive births in Sweden. *Early Human Development*, 37(3):195–204.

North America Land Data Assimilation System (1979-2011). Daily Air Temperatures and Heat Index, years 1979-2011. CDC WONDER Online Database, released 2012.

Oaxaca, R. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, 14(3):693–709.

Pierce, J. P., Messer, K., White, M. M., Cowling, D. W., and Thomas, D. P. (2011). Prevalence of Heavy Smoking in California and the United States, 1965-2007. *JAMA*, 305(11):1106–1112.

Pollack, H. A. (2001). Sudden infant death syndrome, maternal smoking during pregnancy, and the cost-effectiveness of smoking cessation intervention. *American journal of public health*, 91:432–6.

Raymond, E. G., Cnattingius, S., and Kiely, J. L. (1994). Effects of maternal age, parity, and smoking on the risk of stillbirth. *BJOG: An International Journal of Obstetrics & Gynaecology*, 101(4):301–306.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, 56(4):931–954.

Rocha, R. and Soares, R. R. (2015). Water scarcity and birth outcomes in the Brazilian semiarid. *Journal of Development Economics*, 112:72–91.

Samuels, L., Nakstad, B., Roos, N., Bonell, A., Chersich, M., Havenith, G., Luchters, S., Day, L.-T., Hirst, J. E., Singh, T., Elliott-Sale, K., Hetem, R., Part, C., Sawry, S., Le Roux, J., and Kovats, S. (2022). Physiological mechanisms of the impact of heat during pregnancy and the clinical implications: review of the evidence from an expert group meeting. *International Journal of Biometeorology*, 66(8):1505–1513.

Schneider, S., Maul, H., Freerksen, N., and Pötschke-Langer, M. (2008). Who smokes during pregnancy? An analysis of the German Perinatal Quality Survey 2005. *Public Health*, 122(11):1210–1216.

Scholz, R., Voigt, M., Schneider, K. T. M., Rochow, N., Hagenah, H.-P., Hesse, V., and Straube, S. (2013). Analysis of the German Perinatal Survey of the Years 2007-2011 and Comparison with Data From 1995-1997: Maternal Characteristics. *Geburtshilfe und Frauenheilkunde*, 73:1247–1251.

Schwartz, R. M. (1989). What Price Prematurity? *Family Planning Perspectives*, 21(4):170–174.

Searles Nielsen, S., Dills, R. L., Glass, M., and Mueller, B. A. (2014). Accuracy of prenatal smoking data from Washington State birth certificates in a population-based sample with cotinine measurements. *Annals of epidemiology*, 24:236–9.

Sexton, M. and Hebel, J. R. (1984). A Clinical Trial of Change in Maternal Smoking and Its Effect on Birth Weight. *JAMA*, 251(7):911–915.

Smedberg, J., Lupattelli, A., Mårdby, A.-C., and Nordeng, H. (2014). 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*, 14(1):213.

Spinillo, A., Capuzzo, E., Nicola, S., Colonna, L., Iasci, A., and Zara, C. (1994a). Interaction between fetal gender and risk factors for fetal growth retardation. *American Journal of Obstetrics and Gynecology*, 171(5):1273–1277.

Spinillo, A., Capuzzo, E., Nicola, S. E., Colonna, L., Egbe, T. O., and Zara, C. (1994b). Factors potentiating the smoking-related risk of fetal growth retardation. *BJOG: An International Journal of Obstetrics & Gynaecology*, 101(11):954–958.

Straube, S., Voigt, M., Jorch, G., Hallier, E., Briese, V., and Borchardt, U. (2010). Investigation of the association of Apgar score with maternal socio-economic and biological factors: an analysis of German perinatal statistics. *Archives of Gynecology and Obstetrics*, 282(2):135–141.

Stürmer, T., Rothman, K. J., Avorn, J., and Glynn, R. J. (2010). 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*, 172:843–54.

Suzuki, K., Kondo, N., Sato, M., Tanaka, T., Ando, D., and Yamagata, Z. (2011). Gender differences in the association between maternal smoking during pregnancy and childhood growth trajectories: multilevel analysis. *International Journal of Obesity*, 35(1):53–59.

Suzuki, K., Shinohara, R., Sato, M., Otawa, S., and Yamagata, Z. (2016). Association Between Maternal Smoking During Pregnancy and Birth Weight: An Appropriately Adjusted Model From the Japan Environment and Children's Study. *Journal of Epidemiology*, 26(7):371–377.

Thorngren-Jerneck, K. and Herbst, A. (2001). Low 5-minute Apgar score: a population-based register study of 1 million term births. *Obstetrics & Gynecology*, 98(1):65–70.

Tibshirani, J., Athey, S., and Wager, S. (2020). *grf: Generalized Random Forests*. R package version 1.1.0.

Tong, V. T., Dietz, P. M., Morrow, B., D'Angelo, D. V., Farr, S. L., Rockhill, K. M., and England, L. J. (2013). Trends in Smoking Before, During, and After Pregnancy — Pregnancy Risk Assessment Monitoring System, United States, 40 Sites, 2000–2010. *Morbidity and Mortality Weekly Report: Surveillance Summaries*, 62(6):1–19.

U.S. Department of Health and Human Services (2010). *A Report of the Surgeon General: How Tobacco Smoke Causes Disease: What It Means to You*. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

Varvarigou, A. A., Asimakopoulou, A., and Beratis, N. G. (2009). Impact of Maternal Smoking on Birth Size: Effect of Parity and Sex Dimorphism. *Neonatology*, 95(1):61–67.

Vogler, G. P. and Kozlowski, L. T. (2002). Differential Influence of Maternal Smoking on Infant Birth WeightGene-Environment Interaction and Targeted Intervention. *JAMA*, 287(2):241–242.

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

Wilcox, A. J. (1993). Birth Weight and Perinatal Mortality: The Effect of Maternal Smoking. *American Journal of Epidemiology*, 137(10):1098–1104.

Wilde, J., Apouey, B. H., and Jung, T. (2017). The effect of ambient temperature shocks during conception and early pregnancy on later life outcomes. *European Economic Review*, 97:87–107.

Wilson, D., Wakefield, M., Owen, N., and Roberts, L. (1992). Characteristics of heavy smokers. *Preventive medicine*, 21:311–9.

World Health Organization (2015). *WHO global report on trends in prevalence of tobacco smoking 2015*. World Health Organization.

World Health Organization (2018). *COP24 special report: health and climate change*. World Health Organization.

Wu Wen, S., Goldenberg, R. L., Cutter, G. R., Hoffman, H. J., Cliver, S. P., Davis, R. O., and DuBard, M. B. (1990). Smoking, maternal age, fetal growth, and gestational age at delivery. *American Journal of Obstetrics and Gynecology*, 162(1):53–58.

Zarén, B., Lindmark, G., and Bakketeig, L. (2000). Maternal smoking affects fetal growth more in the male fetus. *Paediatric and Perinatal Epidemiology*, 14(2):118–126.

Zhang, W. and Yang, T.-C. (2019). Maternal Smoking and Infant Low Birth Weight: Exploring the Biological Mechanism Through the Mother's Pre-pregnancy Weight Status. *Population Research and Policy Review*, pages 1–19.

Zhang, X., Wang, Y., Chen, X., and Zhang, X. (2020). Associations between prenatal sunshine exposure and birth outcomes in China. *Science of The Total Environment*, 713:136472.