UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Targeted Maximum Likelihood Estimation for Evaluation of the Health Impacts of Air Pollution

Permalink

https://escholarship.org/uc/item/2670q5f4

Author

Sarovar, Varada

Publication Date

2017

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Targeted Maximum Likelihood Estimation for Evaluation of the Health Impacts of Air Pollution

by

Varada Sarovar

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Maya Petersen, Chair
Professor John Balmes
Professor Alan Hubbard
Professor Mark van der Laan

Spring 2017

Targeted Maximum Likelihood Estimation
for Evaluation of the Health Impacts of Air Pollution

Copyright 2017
by
Varada Sarovar

Abstract

Targeted Maximum Likelihood Estimation
for Evaluation of the Health Impacts of Air Pollution

by

Varada Sarovar

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Maya Petersen, Chair

The adverse effects of air pollution on human life is of serious concern for today's society. Two population groups that are especially vulnerable to air pollution are pregnant women and their growing fetuses, and the focus of this thesis is to study the effects of air pollution on these populations. In order to address the methodological limitations in prior research, we quantify the impact of air pollution on various adverse pregnancy outcomes, utilizing machine learning and novel causal inference methods. Specifically, we utilize two semi-parametric, double robust, asymptotically efficient substitution estimators to estimate the causal attributable risk of various pregnancy outcomes of interest. Model fitting via machine learning algorithms helps to avoid reliance on misspecified parametric models and thereby improve both the robustness and precision of our estimates, ensuring meaningful statistical inference. Under assumptions, the causal attributable risk that we estimate translates to the absolute change in adverse pregnancy outcome risk that would be observed under a hypothetical intervention to change pollution levels, relative to currently observed levels. The estimated causal attributable risk provides a quantitative estimate of a quantity with more immediate public health and policy relevance.

To my parents, my husband and my daughter -
for love, laughter and a beautiful life.

Contents

Contents **ii**

1 Introduction **1**
1.1 Overview . . . 1
1.2 Air pollution and adverse pregnancy outcomes . . . 2
1.3 Background and motivation . . . 2
1.4 A summary of prior research . . . 3
1.5 Limitations of current literature . . . 3
1.6 Research goals and significance . . . 4
1.7 Thesis layout . . . 5

2 A roadmap for causal inference **10**
2.1 Overview . . . 10
2.2 Define research question . . . 11
2.3 Estimate the statistical target parameter . . . 14
2.4 Interpret results under a causal framework . . . 18
2.5 Estimation challenges associated with big data . . . 19
2.6 Simulation . . . 22

3 Assessing the causal impact of prenatal exposure to nitrogen dioxide and ozone on stillbirth **25**
3.1 Introduction . . . 25
3.2 Methods . . . 26
3.3 Results . . . 31
3.4 Discussion . . . 33

4 Assessing the causal impact of prenatal traffic exposure on preterm birth **56**
4.1 Introduction . . . 56
4.2 Methods . . . 57
4.3 Results . . . 61
4.4 Discussion . . . 63

5 Conclusions **74**

Bibliography **75**

Acknowledgments

A very special thanks to my advisor, Maya Petersen, for her support, guidance and patience. Her high standards and occasional pushes helped me to complete the research reported in this thesis. In addition to Maya, I would like to thank Mark van der Laan, the wizard of Biostatistics, for his insightful comments and guidance during my Ph.D. I appreciate all the time he took to explain methodology to me. Without Maya's and Mark's supervision and support, this thesis would not have been possible. I am thankful to John Balmes, whose environmental health science class helped me to identify and shape my research interests, and also for introducing me to CHAPS-SJV collaborators. Alan Hubbard deserves a special thanks for his comments and questions during my qualifying exam, which motivated me to widen my research directions. It is an honor to have Maya, Mark, John and Alan on my quals and thesis committee.

A special thanks to research scientists from CalEPA/OEHHA - Brian Malig, Rupa Basu, Shelley Green, Rachel Broadwin, Keita Ebisu, Dharshani Perason - for being the best collaborators one could ask for. More than collaborators, they were my work family and I cannot thank enough for all the mentorship, kindness and support they provided. I am also grateful to researchers from CHAPS-SJV collaboration - Amy Padula, Wei Yang, Jennifer Mann, Gary Shaw, Katherine Hammond, Betsy Noth, Liza Lutzker, John Balmes - for providing me with data and support for one of my project. Special thanks to Sharon Norris who always saved me from not messing up all the required paper work and for calming me down during my quals prep. Finally, thanks to all my friends - Namita, Vinod, Charissa, Tina, Marie, Marla, Boriska, Michelle C, Pete, Michelle Y, Kevin, Marco, Carolyn - for adding fun during this journey.

I am happy that I can lean on my sister, Saritha, and her family - Shyam, Veda and Theja, always. Thanks to my uncles - Vishnu, Parameswaran, Vasudevan, Sanakaran, and aunts - Girija, Latha, Sudha, and cousins - Manu, Midhun, Jishnu, Yadu, Vidhu for your love throughout the years. Thanks to my maternal grandmom, who was always worried about me missing my meals, and my maternal uncles and aunts for caring about me always. I am thankful to my father-in-law, Sashidharan, and Mohan's uncle and aunt - Rajan and Jyothi for accepting me into your family wholeheartedly. I wish my grandparents, Narayanan & Leela and Kesavan, were alive to see me getting a Ph.D., they were always the first to encourage me in my studies.

I am indebted to my parents - Narayanan and Valsala - for all their love and care, and all the opportunities they provided me. They have been the soothing presence throughout my life. My husband - Mohan - has been and will be my best friend, source of confidence and my love. Mohan has been there for me, rain or shine, and I owe you everything. My daughter, Amudha Usha Sarovar (Ammu) is the most precious and the best thing that happen to me in my life. Thank you for making our life more delightful with your giggles, hugs and kisses. I love you and I dedicate my thesis to you guys :)

Chapter 1

Introduction

1.1 Overview

Ambient air pollution is identified as an important health problem in the United States as well as around the world. The World Health Organization (WHO) recognizes air pollution as a 'public health emergency', leading to one in every nine deaths in 2012 [1] and according to the *State of the Air 2016* report by the American Lung Association, around 52.1% of the population in the United States live in counties where air pollution levels, dominated by ozone or particulate matter, are considered dangerous to health [2]. Many of the air pollutants of concern are generated by human activities that include fuel combustion, mobile sources and industrial processes [3, 4]. In the United States, the Environmental Protection Agency (EPA), under the *Clean Air Act* [5], set permissible standards for six common ambient air pollutants, known as criteria pollutants: which are nitrogen dioxide (NO_2), ozone (O_3), particulate matter (PM), carbon monoxide (CO), sulfur dioxide (SO_2), and lead. In the United States, the EPA identifies motor vehicles as a major source of air pollution [6]. Although there are strict rules and regulations, such as the Clean Air Act, trends like the rise in metropolitan populations and subsequent rise in the number and use of vehicles has resulted in more traffic-related pollution [7].

Previous studies have linked air pollution with many adverse health outcomes [8], including cardiovascular diseases [9, 10], adverse pregnancy outcomes [11, 12, 13] and subsequent health problems [14, 15], respiratory diseases [16, 17], cancer [18, 19, 20], and death [13, 18]. Based on these studies, it is clear that increasing ambient air pollution threatens our health and wellbeing. However, air pollution will not affect everyone equally. Prior research has identified various air pollution vulnerability factors as well as population groups susceptible to adverse health effects [11, 14, 21, 22]. A population group that has been identified as particularly susceptible to air pollution is pregnant women and their growing fetuses [23, 24, 25]. The focus of this thesis is to analyze the causal effects of air pollution and its primary source, traffic density, on this vulnerable population.

1.2 Air pollution and adverse pregnancy outcomes

Prior research on adverse health outcomes in pregnant women and fetuses has linked air pollution to preterm birth (PTB) [12, 26, 27, 28], low birth weight (LBW) [12, 29, 30, 31], small for gestational age (SGA) [32, 28, 33], birth defects [34], subsequent health effects at later stages of life [35, 36, 37] and stillbirth (fetal death) [38, 39, 40, 41, 42].

Many biological pathways have been proposed to explain the adverse effect of air pollution on a growing fetus [43, 44, 45]. First, research suggests that air pollutants, such as NO_2, CO, SO_2 and PM_{10}, can modify blood coagulability and plasma viscosity [46, 47] and these can adversely affect the umbilical blood flow, leading to inadequate placental transfer of oxygen, which consequently effects fetal growth. Second, air pollution can trigger preterm delivery by increasing vulnerability to infection [44, 48, 49]. Third, O_3 and traffic-related air pollutants, such as nitrogen oxides (NO_x), PM, are capable of generating reactive oxygen species that lead to oxidative stress, and it has been hypothesized that oxidative stress can result in DNA damage and premature placental aging, resulting in fetal vulnerability [50, 51, 52, 53]. Fourth, air pollution can adversely effect the general health of a pregnant woman (e.g. gestational hypertensive disorders, increased stress levels) and hence the health of her growing fetus [54, 55, 56].

The burden of adverse pregnancy outcomes can lead to long-lasting consequences. For most women and their families, pregnancy is an important and happy event of their life; hence any adverse pregnancy outcome can pose a long-lasting emotional and psychological effect. Prior studies have found an association between adverse pregnancy outcomes and post traumatic stress disorder [57, 58]. Adverse birth events can also affect quality of life, relationships between couples and subsequent pregnancies [59, 60]. In addition to various psychosocial challenges, adverse pregnancy events can pose financial burden to the affected families and society [61, 62, 63, 64, 65].

1.3 Background and motivation

This research described in this dissertation is based on a collaboration with research scientists at California EPA/Office of Environmental Health Hazard Assessment (OEHHA) as well as with researchers from UC Berkeley and Stanford who are involved in the Children's Health & Air Pollution Study - San Joaquin Valley (CHAPS-SJV).

As part of the CalEPA/OEHHA collaboration, I was involved in a project focused on assessing the relationship between prenatal ambient air pollution exposure and stillbirth using a California cohort data from 1999 to 2009. I was interested in this project for various reasons. First, research assessing the relationship between air pollution and stillbirth is a relatively new topic, with most of the studies conducted outside the United States. Extrapolating the results of the two existing studies conducted within the United States, in New Jersey, to California is difficult because of differences in the sources and the characteristics of air pollution mixtures. Second, the large and diverse population in the dataset, as well as

availability of exposure data during the entire study period, were ideal for studying a rare outcome like stillbirth. Initially, I was involved in a project that established a strong and stable association between prenatal exposure to ambient air pollutants (specifically NO_2 and O_3) and stillbirth [39]. Subsequently, I extended this study as part of my thesis research to address some of the limitations in the previous analysis.

As part of the CHAPS-SJV collaboration, I was involved in a project focused on assessing the relationship between prenatal traffic exposure and preterm birth in the four most populated counties within the San Joaquin Valley air basin in California between 2000 and 2006. Researchers involved in this project had already established an association between prenatal traffic exposure and preterm birth using a parametric logistic regression model [27]. As part of my thesis research, I am addressing some of the limitations of prior research related to this topic.

In both cases, my thesis research advances the current state of understanding by going beyond establishment of association between air pollutant exposure and the relevant adverse birth outcomes, to assessing the *causal* influence of the air pollutants on the birth outcomes.

1.4 A summary of prior research

Ambient air pollution and stillbirth

Stillbirth (fetal death) is defined as the death of a fetus before or during delivery at or after 20 or 28 weeks of gestational age [66, 67]. The rate of stillbirth in the United States was around 1% as of 2013, affecting around $24,000$ pregnancies each year [66]. Assessing the relationship between ambient air pollution and stillbirth is a relatively new topic and a summary of the prior research is provided in Tables 1.1 and 1.2.

Traffic pollution and preterm birth

Preterm birth (PTB) is a major perinatology issue and is defined as birth that occurs before 37 weeks of gestational age. The prevalence rate of PTB in the United States was around 10% in 2016 [68]. A summary of prior research that assessed the relationship between prenatal traffic pollution or traffic related air toxic exposure and PTB is given in Tables 1.3 and 1.4.

1.5 Limitations of current literature

There are two major statistical methodology limitations to all previous studies on these topics. These are:

1. Parametric regression models were used to assess the relationship between exposure and outcome of interest, as well as to adjust for measured confounders. Utilizing parametric models, that are not supported by a priori knowledge, to adjust for measured confounders can result in biased estimates and incomplete control for confounding due to plausible model misspecification.

2. In prior studies, conditional odds ratios or regression based estimates were reported to evaluate the relationship between exposure and outcome. This parameters fail to quantify the risk of poor birth outcomes attributable to the exposure in the study population of interest, and also does not inform us about how much the outcome risk burden might be expected to change if an intervention (e.g. a new policy on ambient or traffic related air pollution level standards) modified the exposure level.

1.6 Research goals and significance

To address the above listed limitations in previous studies, we incorporate new machine learning and causal inference methods to analyze the impact of air pollution on various adverse birth outcomes. In the following, we quantify the impact of air pollution using the causal attributable risk (CAR) of various outcomes of interest, which is estimated using both targeted maximum likelihood estimation (TMLE) [69] and inverse probability of censoring weighted targeted maximum likelihood estimation (IPCW-TMLE) [70]. Both TMLE and IPCW-TMLE are semi-parametric, double robust, asymptotically efficient substitution estimators [71]. In addition, we utilize a machine learning algorithm, SuperLearner [72], for model fitting in order to minimize the reliance on parametric models. IPCW-TMLE also helps to overcome computational challenges that are posed by large data sets by utilizing a sub-sample for analysis.

Under assumptions to be detailed in chapter 2, the CAR of a particular outcome in a population, compares the absolute change in the outcome risk that would have been experienced under a hypothetical intervention to change pollution exposure levels relative to their currently observed levels. Evaluating CAR of various health outcomes, from different air pollution exposures, is relevant for evaluating the impact of new policies related to the corresponding air pollution level. Quantifying the health effect of potential intervention levels and establishing a causal relationship between air pollution and health outcomes are more convincing than establishing mere associations for policy relevant issues [73, 74].

Specific aims of the two studies presented in this thesis (chapters 3 and 4) are:

1. Quantify the causal attributable risk of stillbirth from prenatal ambient air pollution, focusing on NO_2 and O_3 exposures, in a California cohort dataset from 1999 to 2009.

2. Quantify the causal attributable risk of preterm birth from prenatal traffic exposure in four counties in the San Joaquin Valley air basin between 2000 to 2006

1.7 Thesis layout

This thesis consists of four chapters, including this introductory chapter. In this first chapter, a summary of the prior literature related to two adverse pregnancy outcomes and air pollution, as well as the limitations related to this prior research is provided. We also detail the aims of the studies that will be presented in detail in this thesis, as well as our motivations.

In the second chapter we present a summary of the well-established causal road map that will be followed in subsequent chapters. This includes a brief description of a semi-parametric efficient estimation approach [69] that we will apply to estimate the target parameter of interest that quantifies the impact of air pollution on the adverse pregnancy outcomes considered.

In chapter 3, we estimate the causal attributable risk of two criteria pollutants on stillbirth using a California cohort data from 1999 to 2009. Here we also address the exposure assignment challenge that could arise when a temporal trend exists both in exposure and in conception. Under assumptions, we found a weak causal link between exposure to the two pollutants considered and stillbirth. Moreover, we show that the significance of the causal effect depends on the exposure assignment that is used.

In chapter 4, we quantify the CAR of preterm birth from prenatal traffic exposure in the four most pollutant counties in San Joaquin Valley air basin from 2000 to 2006. Under assumptions, we were able to establish a causal link between prenatal traffic exposure and preterm birth.

Table 1.1: Summary of prior research related to ambient air pollution and stillbirth (Part 1)

Location & study period	Study design & analysis method	Pollutants & assigned exposure periods	Control of confounding	Main Results
England & 1962 - 1992 [75]	Cohort & conditional logistic regression	Black smoke (PM_4) & Trimesters, whole pregnancy period	Adjusted for infant sex, maternal age, parity & Townsend deprivation score	No significant association between black smoke exposure & stillbirth in any period considered
Czech Republic & 1986 - 1988 [76]	Ecologic & logistic regression	SO_2, NO_x, PM_{10} & Annual mean	Adjusted for district level socioeconomic characteristics	None of the pollutants showed significant association with the prevalence of stillbirth
Brazil & 1991 - 1992 [38]	Time series & Poisson regression	NO_2, SO_2, CO, O_3, PM_{10} & 2 to 14 day lagged moving average exposure. Lag period different for each pollutant	Adjusted for season & weather	Strong association for NO_2; but lesser for SO_2, CO
Taiwan & 2001 - 2007 [40]	Case-control study & logistic regression	NO_2, SO_2, CO, O_3, PM_{10} & Monthly, trimester and entire pregnancy exposure means	Adjusted for sex, maternal age, gestational age, municipal-level SES, season of conception, year of birth	Preterm : 1st trimester & 1st, 2nd & 3rd months of SO_2 exposure Term: 2nd trimester of PM_{10} & O_3 (both protective effect) All births: No effect
South Korea & 2001 - 2004 [77]	Prospective cohort & Logistic, linear regression	PM_{10} & Trimester means	Adjusted for infant sex, infant order, maternal age, BMI, season of birth, parents' education, alcohol drinking	Unadjusted: 1st trimester (protective effect) & 3rd trimester of PM_{10} Adjusted: 3rd trimester of PM10

Table 1.2: Summary of prior research related to ambient air pollution and stillbirth (Part 2)

Location & study period	Study design & analysis method	Pollutants & assigned exposure periods	Control of confounding	Main Results
Ohio & 2006 - 2010 [78]	Retrospective cohort & GEE model with logit link	$PM_{2.5}$ & Trimester and entire pregnancy exposure means	Adjusted for maternal age, race, education, quantity of prenatal care & cigarette smoking	Unadjusted & Adjusted: 3rd trimester of $PM_{2.5}$
New Jersey & 1998 - 2004 ($PM_{2.5}$ exposure available for 1999 - 2004) [42]	Time-stratified case-crossover & conditional logistic regression	NO_2, SO_2, CO, $PM_{2.5}$ & Lag day 2 and Lag day 2-6 before delivery	mean ambient temeprature	Unadjusted & adjusted: Lag 2 day exposure of SO_2 and CO
New Jersey & 1998 - 2004 ($PM_{2.5}$ exposure available for 1999 - 2004) [41]	Retrospective cohort & logistic regression	NO_2, SO_2, CO, $PM_{2.5}$ & Trimester and entire pregnancy exposure means	Adjusted for maternal age, race, education, prenatal care, smoking, neighborhood SES, year & month of conception & mean temperature	Unadjusted: 1st trimester of NO_2, 3rd trimester & entire pregnancy of SO_2 Adjusted: 1st trimester & entire pregnancy of NO_2 & 1st & 3rd trimesters of SO_2 & 2nd & 3rd trimesters of CO
California & 1999 - 2009 [39]	Retrospective Cohort & logistic regression	NO_2, SO_2, CO, $PM_{2.5}$, O_3 & Trimester and entire pregnancy exposure means	Adjusted for maternal age, race, education, infant sex, season of LMP, air basin name, conception year, apparent temperature	One pollutant model: NO_2 and $PM_{2.5}$ during the entire pregnancy, O_3 during 3rd trimester Two pollutant model: NO_2 during the entire pregnancy, O_3 during 3rd trimester
California & 1999 - 2009 [79]	Time-stratified case-crossover & meta-analysis	NO_2, SO_2, CO, $PM_{2.5}$, $PM_{2.5-10}$, O_3 Single day (2 to 6) and cumulative day lags	Adjusted for apparant temperature	Single day lag 4 for SO_2, CO & lag 2 for ,$PM_{2.5}$

Table 1.3: Summary of prior research related to prenatal traffic or traffic related air toxic exposure and PTB (Part 1)

Location & study period	Analysis method	Exposure of interest	Control of confounding	Main Results
East Kaohsiung, Taiwan & 1992 - 1997 [80]	Unconditional multiple logistic regression	Traffic exposure	Adjusted for maternal age, education, season, marital status, infant gender	Positive association between PTB and distance from freeway traffic
Los Angeles, California & 1994-1996 [81]	Conditional logistic regression	Traffic density	Adjusted for infant's sex, previous low birth weight or preterm infant, parity, interval since previous live birth, year of birth.	Significant association between traffic density & PTB; especially in the winter months among low SES.
Montreal, Canada & 1997-2001 [82]	Multilevel logistic regression	Proximity to highways	Adjusted for maternal age, civil status, country of birth, history of prior stillbirth, birth order, newborn sex & year of birth	Significant association between proximity to highway & PTB
San Joaquin Valley, California & 2000 - 2006 [27]	Logistic regression	Traffic density & entire pregnancy; Ambient air pollutants & 1st, 2nd, 3rd trimester, last month, last 6 weeks of pregnancy	Adjusted for birth weight, maternal age, race/ethnicity, education, prenatal care in the first trimester, Medi-Cal payment of birth costs	Significant associations in the unadjusted models for traffic density exposure. Significant associations in the 2nd trimester exposure of CO, NO_2, $PM_{2.5}$, PM_{10}
Valencia, Spain & 2004 - 2005 [26]	Multivariate logistic regression & Generalized Additive Models	NO_2, Benzene & 1st, 2nd, 3rd trimester & entire pregnancy exposure	Factors adjusted in both NO_2 & Benzene analysis: sex, maternal age, season of conception, working status at 3rd trimester, alcohol consumption during pregnancy, caffeine consumption, smoking, & rural zone of residence	NO_2: Significant association during 2nd, 3rd trimesters & entire pregnancy Benzene: Significant association during entire pregnancy

Table 1.4: Summary of prior research related to prenatal traffic or traffic related air toxic exposure and PTB (Part 2)

Location & study period	Analysis method	Exposure of interest	Control of confounding	Main Results
Los Angeles, California & 6/1/2004 - 3/30/2006 [83]	Conditional logistic regression	Various air toxics, criteria pollutants & 1st, 2nd, 3rd trimesters, last 30 days of pregnancy, entire pregnancy	Adjusted for maternal age, race/ethnicity, education, parity	Multi pollutant models: found strong association for entire pregnancy exposures to organic and elemental carbon, benzene, diesel, biomass burning & ammonium nitrate $PM_{2.5}$, & PAHs
North Carolina & 2004 - 2008 [84]	Generalized linear mixed models	Road proximity	Adjusted for maternal race, age, education, marital status, nativity, tobacco use during pregnancy, parity, season of birth, infant sex, tract-level urbanization, & tract-level median income	Road proximity associated with PTB
Stockholm, Sweden & 1998 - 2006 [85]	Mixed model logistic regression	NO_x & first trimester, the last 6 weeks of gestation, last 6 weeks at risk of PTB, full period of gestation.	Adjusted for maternal asthma, education, area of origin, age, smoking, body mass index, family situation, 1st trimester temperature & ozone, day & year of conception, & parity	No significant association observed between NO_x exposure and PTB
Netherlands & 1996 - 1997 [86]	Linear & logistic regression	NO_2, $PM_{2.5}$, and soot & entire pregnancy, 1st trimester and last month	Adjusted for maternal nationality, study region	No statistically significant association between pollutants considered and PTB
Amsterdam & 01/2003 - 03/2004 [87]	Linear & logistic regression	NO_2 & trimester, entire pregnancy exposure	Adjusted for maternal education	No statistically significant association between NO_2 exposure and PTB

Chapter 2

A roadmap for causal inference

2.1 Overview

When studying the health effects of air pollution, causal questions or inference are more informative and relevant than associations. To appreciate the difference between standard statistical inference and causal inference note that the former uses observed sample data from a data generating distribution to infer association between two variables, as well as extend analysis to unobserved data assuming that the experimental setup that generated the data remains the same. In contrast, causal inference is based on unobserved counterfactual data that would have been generated under a *different* hypothetical experimental setup [88, 89]. When addressing causal inference research questions it is useful to follow the well-established *causal roadmap* [90]. In this chapter, we present a summary of this roadmap, which we followed to address the research questions addressed in the thesis.

The causal roadmap is specified by the following steps:

1. Define research question
    a) Specify data and causal model
    b) Specify the causal target parameter of interest
    c) Specify the observed data and their link to causal model or counterfactual outcomes
    d) Identify the causal target parameter as a parameter of observed data distribution
2. Estimate the statistical target parameter
3. Interpret results under a causal framework

As we examine each one of these steps in more detail in this chapter, we will use a running example to illustrate concepts. This example is based on the traffic exposure and preterm birth cohort dataset studied in Chapter 4.

2.2 Define research question

The running example is based on the CHAPS-SJV cohort data (traffic exposure-PTB). The specific aim in that study is to quantify the causal attributable risk of preterm birth from prenatal traffic exposure in four most populated counties in the San Joaquin Valley air basin in California between 2000 to 2006

Defining the research question incorporates our knowledge about the system to be studied; it can be broken up using the steps mentioned above, and we will explain each step in detail using the running example.

Specify data and causal model

The first sub-step in defining the research question is to specify the data and causal model; this step helps to express our research question in mathematical terms. Structural causal models (SCM) [88, 91] were utilized to represent our causal model; the SCM framework integrates the approaches of structural equation modeling [92, 93], causal diagrams [91, 94] and potential outcomes or counterfactuals [95, 96],

We represent our data as $X = \{W, A, Y\}$. Here W denotes the baseline covariates, A is the exposure of interest and Y denotes the outcome of interest. X represents the endogenous variables that are relevant to our research question; it includes both measured or unmeasured variables that are affected by the other variables. In our running example: W includes various individual (e.g. maternal age, sex of infant), temporal (e.g. season of conception) and community or neighborhood factors (e.g. census tract variables), A is traffic density, and Y is preterm birth.

There will be some unmeasured background factors or error terms, also known as exogenous variables, that will effect the endogenous variables and they are represented by $U = \{U_W, U_A, U_Y\} \sim P_U$ with P_U denoting the distribution of U. In our example, U_W and U_Y could include some maternal or infant genetic characteristics, U_A could indicate the housing rental prices. $F = \{f_W, f_A, f_Y\}$ denotes a set of structural equations that represent each endogenous variable as a function of other variables, namely parent variables, that could include both exogenous and other endogenous variables that affect them. In our data, we did not assume any functional form for the structural equations; hence the non-parametric structural equations representing our data are given as:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(W, A, U_Y) \end{aligned} \tag{2.1}$$

Structural equations represent our background knowledge about the system to be studied and this is reflected in terms of the exclusion criteria (by restricting parent variables for each endogenous variables) and the independence assumptions (by assuming no common parent

for some endogenous variables) applied. $\mathcal{M}^{\mathcal{F}}$ denotes the structural causal model that include all the possible probability distributions for both endogenous and exogenous variables.

Specify the causal target parameter of interest

The second step in defining the research question is to specify the causal target parameters of interest. This step requires us to be specific about the type of intervention that we wish to apply to the current system as well as the resulting counterfactual outcomes. In the running example the aim of the study is to better understand changes in the preterm birth distribution in our study population from a plausible proposed intervention (e.g. a policy change) that could modify traffic pollution exposure level. Hence we decided to estimate the causal attributable risk (CAR), which is a population intervention parameter [97].

Structural causal models can represent the system to be studied under various desired experimental setups. For example under a new static intervention of $A = a$, on the exposure variable, the system to be studied represented in Eq. 2.1 will modify to

$$\begin{aligned} W &= f_W(U_W) \\ A &= a \\ Y_a &= f_Y(W, a, U_Y) \end{aligned} \tag{2.2}$$

In Eq. 2.2 the resulting outcome, under the desired intervention of $A = a$, is called the potential or the counterfactual outcome corresponding to the new intervention applied. The counterfactual outcome distribution is represented by $P_{U,X}$ which is the joint distribution of P_U and F. In our running example, counterfactual outcome Y_a for a mother is the birth outcome she would have had if she had been exposed to exposure $A = a$.

In addition to all the possible probability distributions for both endogenous and exogenous variables, $\mathcal{M}^{\mathcal{F}}$ also include the distributions of counterfactual outcomes under various intervention scenarios.

Based on the distribution of a counterfactual outcome Y_a, from a statistical intervention on the exposure variable $A = a$, the target parameter of interest is defined as

$$\Psi^{\mathcal{F}}_{P_{U,X}} = E_{P_{U,X}}(Y) - E_{P_{U,X}}(Y_a); a \in \mathcal{A} \tag{2.3}$$

The parameter represented in Eq. 2.3 compares the expected outcome that is currently observed with the expected counterfactual outcome under various hypothetical intervention scenarios applied to the exposure $A = a; a \in \mathcal{A}$. In the running example, $\mathcal{A} = \{1, 2, 3, 4\}$ represents a particular level/quartile of traffic density exposure, with 1 representing the lowest quartile of exposure and 4 representing the highest, and $\Psi^{\mathcal{F}}_{P_{U,X}}$ represents the impact of a particular level of traffic exposure on preterm birth relative to the current probability of preterm birth.

Specify the observed data and their link to the causal model or counterfactual outcomes

We assume that our observed data were created by sampling n times from a data generating system that is compatible with the causal model specified in the earlier step. Thus for each of the n subjects, the observed data structure can be represented as $O = \{W, A, Y\}$ and the observed data can be viewed as n independent and identical copies of $O \sim P_0$, where P_0 is the true, but unknown, distribution. Our statistical model $\mathcal{M}$ comprises of all the possible distributions, including the true, for the observed data; i.e. $P_0 \in \mathcal{M}$. Since $O \subset X$, the causal model implies the statistical model. Statistical models can be parametric and/or non-parametric.

To represent the observed data based on Neyman-Rubin counterfactual framework [95, 96], we assume the *stable unit treatment value assumption* (SUTVA) [98]. The two key points in this assumption are: firstly, the potential outcome for one subject is not effected by the exposure assigned to another subject; secondly there is only a single level for each exposure and hence the potential outcome under each exposure is well defined. Under SUTVA, we can represent the observed outcome for the i^{th} subject in terms of their observed exposure and potential counterfactual outcomes as: $Y_i = A_i Y_{i,1} + (1 - A_i)Y_{i,0}$. In reality, we will be observing only one of two potential outcomes for each subject and hence causal inference is a missing data problem.

Identify the causal target parameter as a parameter of observed data distribution

The final step in defining the research question is to identify the target causal parameter of interest from the observed data that we have. To identify the target causal parameter $\Psi^{\mathcal{F}}_{P_{U,X}}$ from the available observed data, we need to express it as a statistical parameter $\Psi(P_0)$ which is a function of observed data. To do this we need to make the following assumptions [71]:

1. Consistency assumption: This is a key assumption in causal inference. Under this assumption we assume that the the observed outcome (Y) under the observed exposure level (A) is equal to the counterfactual outcome (Y_A) under the observed exposure level. This is an untestable assumption. In the SCM framework, consistency assumption is implied by the definition of counterfactuals.

2. Randomization assumption: This assumption is also known as the "no unmeasured confounding" assumption because under it we assume that the measured covariates W are sufficient to control for confounding of the effect of A on Y. Mathematically, this assumption is expressed as $Y_a \perp\!\!\!\perp A|W, \forall\ a \in \mathcal{A}$. This is an untestable assumption.

3. Positivity assumption: This is also known as the experimental treatment assumption (ETA). Under ETA, we assume that in our population there is a positive probability

for receiving each level of exposure A within every combination of baseline covariate values that occurs with positive probability under P_0. Mathematically this is expressed as: $\min_{a \in \mathcal{A}} P_0(A = a|W) > 0$.

In our running example, we believe that most of the important confounders related to traffic exposure and preterm birth are included in the available observed data, however it is possible that some maternal or neighborhood/community characteristics that we do not track could potentially determine exposure and birth outcomes and thus act as confounders. Hence in this example it should be kept in mind that the untestable randomization assumption may not hold. Here, we also assume that there is enough variability within the traffic exposure quartiles, regardless of various covariate strata, and hence positivity assumption is reasonable.

Under these three assumptions, we can express the target parameter of interest in terms of the observed data as:

$$\Psi(P_0) = E_0(Y) - E_{W,0}[E_0(Y|A = a, W)]; a \in \mathcal{A} \tag{2.4}$$

This is the statistical estimand.

Identifying the target causal parameter in terms of the observed data structure completes the first step of the causal roadmap.

2.3 Estimate the statistical target parameter

The second step of causal roadmap is the estimation of the target parameter; and since we expressed our target parameter as a statistical estimand in the first step, this estimation step is essentially a statistical problem. There are different methods to estimate the identified statistical estimand. The G-computation method (or simple substitution method) makes use of an estimate of outcome regression $E_0(Y|A, W)$ [99, 100] and the inverse probability of treatment weighted (IPTW) method utilizes an estimate of treatment regression (also known as the propensity score) $P_0(A = a|W)$ [101]. The G-computation and IPTW estimators are consistent if the outcome regression or the treatment regression, respectively, are consistently estimated. In this thesis we make use of another class of estimators, which combines both outcome and treatment regression estimates, namely the *targeted maximum likelihood estimators* (TMLE) [69, 71, 102].

A generic description of TMLE involves the following steps [69, 71], and in what follows, we will explain each step in detail, using the running example.

1. Define the target parameter, which is a mapping from the statistical model to the parameter space $\Psi : \mathcal{M} \to \mathbb{R}$.

2. Compute the efficient influence curve (IC), $D^*_\Psi(P_0)(O)$, of the target parameter.

3. Define a loss function $L(P)$ such that expected value of that loss function, $E_0 L(P)$, is minimized at P_0; P_0 is the true distribution.

4. Perform TMLE updating steps:

   a) Form an estimate P_n^0 of P_0.

   b) Define least-favorable parametric working model through the initial estimate P_n^0, denoted as $P_n^0(\epsilon)$, such that $P_n^0(\epsilon = 0) = P_n^0$ and $\frac{d}{d\epsilon} L(P_n^0(\epsilon))|_{\epsilon=0} = D_\Psi^*(P_n^0)(O)$.

   c) Find $\epsilon_n^0 = \arg\min_\epsilon \sum_{i=1}^n L(P_n^0(\epsilon)(O_i))$.

   d) Update the initial estimate: $P_n^1 = P_n^0(\epsilon_n^0)$.

   e) Repeat the last three steps (b,c & d) until convergence at the K^{th} step, i.e. until $\epsilon_n^K = 0$. The final update at the K^{th} step is denoted as P_n^* and this is the TMLE of P_0.

5. Form the TMLE of the target parameter obtained by simply substituting the TMLE of P_0 as per the parameter mapping; i.e. $\psi_n^* = \Psi(P_n^*)$.

Define the target parameter

We have already provided our target parameter of interest, in terms of the observed data, in Eq. 2.4 and there are two points to note regarding this parameter. First, instead of directly targeting the difference $E_0(Y) - E_{W,0}[E_0(Y|A = a, W)]$, we will apply the targeted updating steps only to the second term of our target parameter $E_{W,0}[E_0(Y|A = a, W)]$. The first term of our target parameter, $E_0(Y)$, is estimated using the empirical mean of Y which is already an unbiased estimate, and hence this part does not require any further update. However, in an alternative approach, not implemented here, one could directly target the difference, which can be done in such a way that the estimate of $E_0(Y)$ still reduces to the empirical mean of Y. Second, note that $E_{W,0}[E_0(Y|A = a, W)]$, which we will refer as the *counterfactual term of the target parameter* in the subsequent steps, depends on P_0 through the conditional expectation of the outcome given the exposure and base line covariates, denoted as $\bar{Q}_0 = E_0(Y|A, W)$, and the marginal distribution of baseline covariates, denoted as $\bar{Q}_{W,0}$. Hence the second term of the target parameter, i.e. the counterfactual term, can be rewritten as a function of Q_0, where $Q_0 = (\bar{Q}_0, \bar{Q}_{W,0})$.

Compute the efficient influence curve of the target parameter

The next step is to compute the efficient influence curve (IC) of the target parameter. Fortunately, the efficient influence curve for $E_{W,0}[E_0(Y|A = a, W)]$, the counterfactual term of the target parameter, is already provided in the literature [71, 69]. If we denote conditional probability of exposure given the base line covariates as $g_0(A|W) = P_0(A|W)$ then the

efficient influence curve for the counterfactual term of the target parameter is given as:

$$D^*_{E_{W,0}[E_0(Y|A=a,W)]}(P_0)(O) = \frac{\mathbb{1}(A=a)}{g_0(a|W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(a,W) - E_{W,0}[E_0(Y|A=a,W)] \tag{2.5}$$

The efficient IC, given in Eq. 2.5, can be decomposed into two parts, as $D^*(P_0)(O) = D^*_Y(P_0)(O) + D^*_W(P_0)(O)$ based on their relation to the distribution of the outcome Y. In the upcoming steps we will be updating only the relevant part of the efficient IC, which is $D^*_Y(P_0)(O) = \frac{\mathbb{1}(A=a)}{g_0(a|W)}(Y - \bar{Q}_0(A,W))$ [71].

Also, using a simple calculation based on the definition of influence function [103], the efficient influence curve for the first part of the target parameter, $E_0(Y)$, can be written as $Y - E_0(Y)$. Applying the Delta method [71], the efficient influence curve for the target parameter, specified in Eq. 2.4, can be calculated to yield:

$$D^*_\Psi(P_0)(O) = Y - \frac{\mathbb{1}(A=a)}{g_0(a|W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(a,W) - \Psi(P_0) \tag{2.6}$$

Define a loss function

The next step is to define a loss function such that the risk, which is the expected value of the loss function, is minimized at the true distribution. We chose a loss function that depends on P_0, through the relevant parts of Q_0, specifically $\bar{Q}_0$. For example, for a binary outcome Y, the negative log loss function is represented as [71]: $L(O, \bar{Q}_0) = -log(\bar{Q}_0(A,W)^Y(1 - \bar{Q}_0(A,W))^{(1-Y)})$.

For our running example, we choose a negative log loss function.

Perform TMLE updating steps

Form an initial estimate P^0_n of P_0

Since we can write the target parameter of interest as a function of $Q_0 = (\bar{Q}_0, \bar{Q}_{W,0})$, in this step, we are forming an initial estimate of Q_0. The empirical probability distribution of W, denoted as $\bar{Q}_{W,n}$, was used to estimate $\bar{Q}_{W,0}$; this is a non-parametric maximum likelihood estimator and it will not add any bias to the target parameter. Hence in the subsequent steps the updating will be focused only on $\bar{Q}^0_n$, which is the initial estimate of $\bar{Q}_0 = E_0(Y|A,W)$. $\bar{Q}^0_n$ is based on the loss function and it can be obtained using a machine learning algorithm called SuperLearner [72]. Utilizing cross-validation, SuperLearner creates an optimal combination of fits obtained from user supplied individual prediction algorithms.

In our running example, to get an initial estimate of both $\bar{Q}_0$ and $g_0(A=a|W)$ we used SuperLearner, with a non-negative least squares loss function, a 10-fold cross validation and our library of candidate algorithms included main term logistic regression, logistic regression with all possible pairwise interactions, simple mean and the stepwise logistic regression. $\bar{Q}_{W,n}$ is estimated using the empirical distribution.

Define least-favorable parametric working model through the initial estimate

Once we have the initial estimate $\bar{Q}_n^0$ for $\bar{Q}_0$, the next step is to create a parametric working model, denoted as $\bar{Q}_n^0(\epsilon)$, through this initial estimate with the following properties [71]:

1. $\bar{Q}_n^0(\epsilon = 0) = \bar{Q}_n^0$. i.e. when $\epsilon = 0$, we are able to recover the initial estimate $\bar{Q}_n^0$ of $\bar{Q}_0$.
2. $\frac{d}{d\epsilon} L(\bar{Q}_n^0(\epsilon))|_{\epsilon=0} = D_Y^*(Q_n^0, g_n)(O)$. i.e. taking a derivative of the loss of the parametric working model (i.e. the score of the working model) with respect to ϵ and evaluating it at $\epsilon = 0$ will give the appropriate component of the efficient IC.

In our analysis, we used the following least favorable parametric working model:

$$\bar{Q}_n^0(\epsilon)(A, W) = expit\big(log\frac{\bar{Q}_n^0}{(1 - \bar{Q}_n^0)}(A, W) + \epsilon H_n^*(A, W)\big) \tag{2.7}$$

where $H_n^*(A, W) = \frac{\mathbb{1}(A=a)}{g_n(A=a|W)}$ and $g_n(A = a|W)$ is an estimate of $g_0(A = a|W)$ that can be obtained using SuperLearner.

Updating the initial estimate

Once we have the parametric working model through the initial estimate of $\bar{Q}_0$, the next step is to identify the ϵ value that minimize the risk within this parametric working model family. i.e. $\epsilon_n^0 = \arg\min_\epsilon \sum_{i=1}^n L(\bar{Q}_n^0(\epsilon)(O_i))$. The TMLE update is then defined as the parametric working model evaluated at the optimal ϵ_n^0. i.e. $\bar{Q}_n^* = \bar{Q}_n^0(\epsilon_n^0)$ and the resulting TMLE of Q_0 is denoted as $Q_n^* = (\bar{Q}_n^*, \bar{Q}_{W,n})$.

In our running example, in the updating step, the optimal ϵ_n^0 is obtained by performing a logistic regression of Y on $H_n^*(A, W)$, with initial estimate of $\bar{Q}_0$ as an offset and by suppressing the intercept. The resulting maximum likelihood estimate of the coefficient of $H_n^*(A, W)$ is the optimal ϵ_n^0. The TMLE update of $\bar{Q}_n^0$ is obtained as $logit(\bar{Q}_n^*) = logit(\bar{Q}_n^0) + \epsilon_n^0 H_n^*(A, W)$.

Estimating the target parameter

The target parameter is then estimated by applying the same mapping Ψ, that defines the parameter as a function of the true distribution, to an estimate of the true distribution obtained using the observed data. Thus, along with the empirical means of both Y and W, by plugging in the updated estimate $\bar{Q}_n^*$ into the parameter mapping, we can estimate the the statistical estimand of the target parameter of interest given in Eq. 2.4 as

$$\Psi(P_n^*) = \frac{1}{n}\sum_{i=1}^n Y_i - \frac{1}{n}\sum_{i=1}^n \bar{Q}_n^*(a, W_i); a \in \mathcal{A} \tag{2.8}$$

TMLE is a substitution estimator [71].

TMLE solves efficient IC equation

When ϵ converges to zero, we obtain the minimum risk model within the chosen parametric working model family. i.e.

$$\begin{aligned} 0 &= \frac{d}{d\epsilon} EL(P_n^*(\epsilon)(O))|_{\epsilon=0} = E\frac{d}{d\epsilon} L(P_n^*(\epsilon)(O))|_{\epsilon=0} \\ &= ED_\Psi^*(P_n^*)(O) \end{aligned} \tag{2.9}$$

In Eq. 2.9 the last equality is based on the property of the chosen parametric working model family; $\frac{d}{d\epsilon} L(P_n^*(\epsilon)(O))|_{\epsilon=0} = D_\Psi^*(P_n^*)(O)$. Eq. 2.9 implies that the TMLE of P_0, P_n^*, solves the efficient influence curve equation $0 = \sum_{i=1}^n D_\Psi^*(P_n^*)(O_i)$ and hence it inherits many attractive asymptotical properties [71, 102], such as:

1. TMLE is double robust. i.e. TMLE will be consistent if either $g_n(A = a|W)$, the initial estimator of the treatment regression $g_0(A = a|W)$, or $\bar{Q}_n^0$, the initial estimator of the outcome regression $\bar{Q}_0$, is consistent.

2. TMLE is asymptotically linear as well as asymptotically efficient under the following conditions:

   a) Both $g_n(A = a|W)$ and $\bar{Q}_n^0$ are consistent.

   b) $D_\Psi^*(P_n^*)$ belongs to the Donsker class with probability that approaches 1.

   c) The second order reminder term, $\int_w \left(\frac{(g_n - g_0)(a|W)}{g_n(a|W)}\right)(\bar{Q}_n^0 - \bar{Q}_0)(a, W)dP_0(w) = o_p(\frac{1}{\sqrt{n}})$.

   If an estimator is asymptotically efficient, it will have the smallest possible variance (i.e. it will attain the Cramer-Rao lower bound) among regular estimators as sample size n goes to infinity. If an estimator is asymptotically linear, it implies that this estimator will be asymptotically normal with mean zero and its variance well approximated by the sample variance of influence curve divided by the sample size n. This provides a basis for central limit theorem based statistical inference for TMLE.

2.4 Interpret results under a causal framework

Interpretation of the results is the last step of the causal roadmap. Under the above assumptions, TMLE is asymptotically normal, and we can conservatively estimate its variance using the sample variance of the estimated efficient influence curve [71]. We could construct a 95% confidence interval for the target parameter as: $\Psi(P_n^*) \pm 1.96\ \hat{\sigma}$; where $\hat{\sigma}^2 = \frac{\sum_{i=1}^n IC_{n,\Psi(P_n^*)}^2(O_i)}{n}$. As an estimated working influence curve, for the target parameter, we can use:

$$IC_{n,\Psi(P_n^*)} = Y - \frac{\mathbb{1}(A = a)}{g_n(A = a|W)}(Y - \bar{Q}_n^*(A, W)) - \bar{Q}_n^*(A = a, W) - \Psi(P_n^*); a \in \mathcal{A} \tag{2.10}$$

In our running example, the causal attributable risk of preterm birth, under assumptions, compares the absolute change in preterm birth risk that would have been experienced by our study population under a hypothetical intervention to change traffic related pollution levels comparative to their currently observed levels.

2.5 Estimation challenges associated with big data

Applying machine learning algorithms, such as SuperLearner, on large data sets can be computationally intensive. The birth cohort data sets collected over several years are examples of such large data sets. For example the birth cohort that we analyzed in chapter 3 of this thesis, includes eleven years of birth information from all of California and it consists of around 3.5 million observations. The running example data set, that we treat throughout this chapter, includes six years of data from four counties in California and it consists of around 300,000 observations. Hence in our analysis, to cope with the computational challenges, we also incorporate a TMLE for two-stage designs namely the inverse-probability-of censoring-weighted targeted maximum likelihood estimator (IPCW-TMLE) [70]. IPCW-TMLE inherits the desired properties of original TMLE that we discussed above, but utilizing a sub-sample of the original cohort to estimate our target parameter of interest.

Revisiting the causal roadmap

Before explaining the details of target parameter estimation part, using two-stage TMLE, we will go through the relevant steps of the causal roadmap that will be different, when utilizing a sub-sample of the whole cohort data, from the one provided earlier, where we utilized the whole cohort data collected from the target population in our analysis. We will use the same running example as before to illustrate the relevant parts of the causal roadmap, even though this cohort was comparatively smaller and does not present any computational issues.

Define research question

In this step of the causal roadmap, if one uses a sub-sample of the cohort collected from the target population, the main differences occur in the observed data and the assessment of identifiability of the target parameter based on the observed data.

Suppose we have a data structure $X = \{W, A, Y\}$, representing a full-data as mentioned earlier, with distribution $P_{X,0}$. In the two-stage sampling designs, that are relevant to this thesis, the second stage data comprises of a sub-sample of this full-data structure, and a particular subject is sub-sampled to the second stage, with a known probability, based on their observed outcome, Y. The reduced sub-sample data structure can be viewed as a missing data structure relative to the full-data structure. The observed data structure can be represented as $O^R = (Y, \Delta, \Delta X) \sim P_0; P_0 \in \mathcal{M}^R$, where Δ is a indicator of inclusion to the second stage sub-sample and we observe $X = \{W, A, Y\}$ when $\Delta = 1$.

From this reduced data structure, in order to identify our target causal parameter of interest specified earlier we need an additional assumption, namely the missing at random (MAR) assumption, which is mathematically stated as $P_{X,0}(\Delta|X) = P_{X,0}(\Delta|Y)$; i.e. the missing mechanism, $\Pi_0(Y) \equiv P_{X,0}(\Delta = 1|Y)$, that constructed the sub-sample from the full-data is determined only by the value of Y. Under MAR and the positivity assumption, the reduced data distribution of P_0 is implied in terms of the full-data distribution, $P_{O,0}$, as well as the missing mechanism distribution, Π_0, as:

$$\begin{aligned} P_{X,0}(Y = y, A = a, W = w) &= \frac{P_0(Y = y, A = a, W = w, \Delta = 1)}{P(\Delta = 1|Y = y)} \\ &= \frac{P_0(Y = y, A = a, W = w, \Delta = 1)}{\Pi_0(Y = y)} \end{aligned}$$

The missing mechanism or the sampling probability, if unknown, can be estimated using the data. Here, however, the missing mechanism is known. Under assumptions we can express the target parameter of interest in terms of the observed data, consisting of n independent and identical copies of O^R, as in Eq. 2.4.

IPCW-TMLE estimation of the counterfactual term of the target parameter

Our target parameter $\Psi^R : \mathcal{M}^R \to \mathbb{R}$ remains the same as Ψ, but we define our loss function differently. The new loss function, for the reduced observed data structure, is defined by giving appropriate weights to the full-data loss function as $L(O^R, \bar{Q}_0) = \frac{\Delta}{\Pi_0(Y)} L(O, \bar{Q}_0)$. As before, this loss function depends on P_0 through the relevant parts of Q_0 and the updating steps of IPCW-TMLE will focus only on $\bar{Q}_0$ of Q_0. However, finding the initial estimates of both $\bar{Q}_0$ and $g_0(A|W)$ using SuperLearner as well as the updating steps are based on the new loss function. To find these initial estimates, we apply SuerLearner on the full-data with appropriate observational weights, denoted as $\frac{\Delta}{\Pi_0(Y)}$, associated to each observation. Here, Π_0 is known, based on a known sampling procedure.

Once we have the initial estimate of $\bar{Q}_0$, the next step is to define the parametric working model through it and we use the same parametric working model as before, specified in Eq. 2.7. However, the minimum risk model within this parametric model working family should be found using the new loss function for the reduced observed data structure; i.e. $\epsilon_n^0 = \arg\min_\epsilon EL(\bar{Q}_n^0(\epsilon)(O_i^R)) = \arg\min_\epsilon \sum_{i=1}^n \frac{\Delta_i}{\Pi_0(Y_i)} L(\bar{Q}_n^0(\epsilon)(O_i))$. Then as before, the IPCW-TMLE update is defined as the parametric working model evaluated at the optimal ϵ_n^0 obtained and the final IPCW-TMLE of Q_0 is represented as $Q_n^* = (\bar{Q}_n^*, \bar{Q}_{W,n})$.

Using the empirical means of both Y and W as well as the IPCW-TMLE updated estimate $\bar{Q}_n^*$, we can estimate the the statistical estimand of the target parameter of interest given in Eq. 2.4 as

$$\Psi(P_n^*) = \frac{1}{n}\sum_{i=1}^n Y_i - \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i}{\Pi_0(Y_i)} \bar{Q}_n^*(a, W_i); a \in \mathcal{A} \tag{2.11}$$

The efficient influence curve of the target parameter at $P_0 = P_{P_{X,0},\Pi_0(Y)}$ can be represented in terms of the efficient IC of the target parameter at $P_{X,0}$ as [70]

$$D_{\Psi}^{R^*}(P_{X,0}, \Pi_0(Y))(O) = \frac{\Delta}{\Pi_0(Y)} D_{\Psi}^*(P_{X,0}) - \left(\frac{\Delta}{\Pi_0(Y)} - 1\right) E_0(D_{\Psi}^*(P_{X,0})|\Delta = 1, Y) \quad (2.12)$$

Here $D_{\Psi}^*(P_{X,0})$ is the efficient IC for the full-data parameter that is specified in Eq. 2.6. If the missing mechanism is estimated non-parametrically, it follows that in Eq. 2.12 the expectation of second term under estimated distribution is zero and the IPCW-TMLE solves the efficient influence curve [70]. Here, we sacrifice full efficiency by using the known Π_0 and as a result, the IPCW-TMLE solves the weighted full-data efficient influence curve equation.

Estimation of the target parameter in the running example

In our running example, the sub-sample data includes all the preterm births with $Y = 1$ and five randomly chosen term births with $Y = 0$ per preterm birth; thus the sub-sample consists of $29,359$ preterm births and randomly chosen $146,795$ out of $226,689$ term births. Here, conditioning on the outcome Y, each observation included in the sub-sample had a known sampling probability and they are $\Pi_0(Y = 1) = 1$ and $\Pi_0(Y = 0)$ equal to the proportion of term births included in the subsample in the second stage. $\bar{Q}_{W,n}$, is a discrete distribution and here it adds an observational weight $\frac{\Delta}{n \times \Pi_0(Y)}$ for each observations. When obtaining the initial estimates of both $\bar{Q}_0$ and $g_0(A = a|W)$, we utilized SuperLearner with a non-negative least squares loss function, a 10-fold cross validation as before, but our candidate algorithms in the library included weighted versions of the individual algorithms mentioned earlier; i.e. we used the full-data with appropriate observational weights applied to each observation.

In the first updating step, the optimal ϵ_n^0 is obtained by performing a weighted logistic regression of Y on $H_n^*(A, W)$, with initial estimate of $\bar{Q}_0$ as an offset and by suppressing the intercept. The resulting maximum likelihood estimate of the coefficient of $H_n^*(A, W)$ is the optimal ϵ_n^0. The first IPCW-TMLE update of $\bar{Q}_n^0$ is obtained as $logit(\bar{Q}_n^*) = logit(\bar{Q}_n^0) + \epsilon_n^0 H_n^*(A, W)$. Based on the final updated fit, the target parameter can be estimated as shown in Eq. 2.11.

Interpret results under a causal framework

Since IPCW-TMLE is asymptotically normal, we can construct a 95% confidence interval for the target parameter as: $\Psi(P_n^*) \pm 1.96\ \hat{\sigma}$; where $\hat{\sigma}^2 = \frac{\sum_{i=1}^n IC_{n,\Psi(P_n^*)}^2(O^R)}{n}$. The influence curve for the target parameter can be estimated by putting observational weights to the estimated full-data IC [71, 70], given in Eq. 2.10. i.e.

$$IC_{n,\Psi(P_n^*)} = \frac{\Delta}{\Pi_0(Y)}\Big(Y - \frac{\mathbb{1}(A = a)}{g_n(A = a|W)}(Y - \bar{Q}_n^*(A, W)) - \bar{Q}_n^*(A = a, W) - \Psi(P_n^*)\Big); a \in \mathcal{A} \quad (2.13)$$

In our running example, the variance of the IPCW-TMLE estimated target parameter is approximated using the weighted variance of the estimated influence curve of the full-data target parameter.

2.6 Simulation

We develop a simulation to (1) estimate causal attributable risk by implementing IPTW, TMLE and IPCW-TMLE, (2) compare bias and confidence interval coverage of these three estimators, (3) illustrate the finite sample performance of these three estimators, and (4) illustrate the double robustness property of both TMLE and IPCW-TMLE. Simulations are done in R version 3.2.4.

Data generation and estimation processes

The first step of the simulation was data generation and estimation of the true value of the target parameter of interest. Our target parameter of interest was the causal attributable risk as specified in Eq. 2.3. For a sample size of 10 million observations, we generated various baseline covariates $W = \{W_1, W_2, W_3, W_4\}$, binary exposure A and the binary outcome variable Y as follows:

$W_1 \sim Bernoulli(0.5)$ $\quad$ $W_2 \sim Bernoulli(0.25)$

$W_3 \sim Uniform(0, 1)$ $\quad$ $W_4 = expit(0.5 - W_2 + W_3)$

$P(A = 1|W) = expit(-1 \ - 2 * W1 \ + 1.75 * W3 \ + .2 * W4)$

$sinW3 = sin(\pi \times W_3)$

$P(Y = 1|A, W) = expit(-13.5 + 1.3 \times W_1 \times W_4 + 7 \times A \times sinW3 + 6.5 \times (1 - A) \times W_2 + 1.2 \times W_4)$

The prevalence probability of the outcome was around 12.9% in the resulting population and the true value of the target parameter of interest was -12.2, and here the negative value of the target parameter indicates that the counterfactual mean of outcome under exposure $A = 1$ is higher than that of the observed mean of outcome.

From this population, we sampled n observations ($n = \{1000, 2000\}$) to form a dataset, and estimated the target parameter of interest, the causal attributable risk as specified in Eq. 2.4, for $A = 1$, using the IPTW, TMLE and IPCW-TMLE methods. This was repeated $m = 1000$ times to collect statistics on the performance of the estimators. When estimating the target parameter using IPCW-TMLE, the sub-sample included all cases, with $Y = 1$, and five randomly sampled controls, with $Y = 0$, per case. Both the outcome and treatment regressions were estimated with the correctly specified logistic model with appropriate main

and interaction terms as per the data generating process, and misspecified regression models that only included an intercept term. The performance of the estimators were compared based on the estimated bias and the confidence interval coverage values.

In Fig. 2.1 the top panels show the bias of the estimates under various models. As expected, IPTW exhibited higher bias when the treatment regression model was misspecified. However because of the "double-robustness" property of both TMLE and IPCW-TMLE, lower bias values were obtained when either the outcome regression or the treatment regression was misspecified. Similar behavior can be seen in terms of the confidence interval coverage plots, show in the bottom panels of Fig. 2.1. In addition, the performance of TMLE and IPCW-TMLE were comparable in terms of bias and confidence interval coverage.

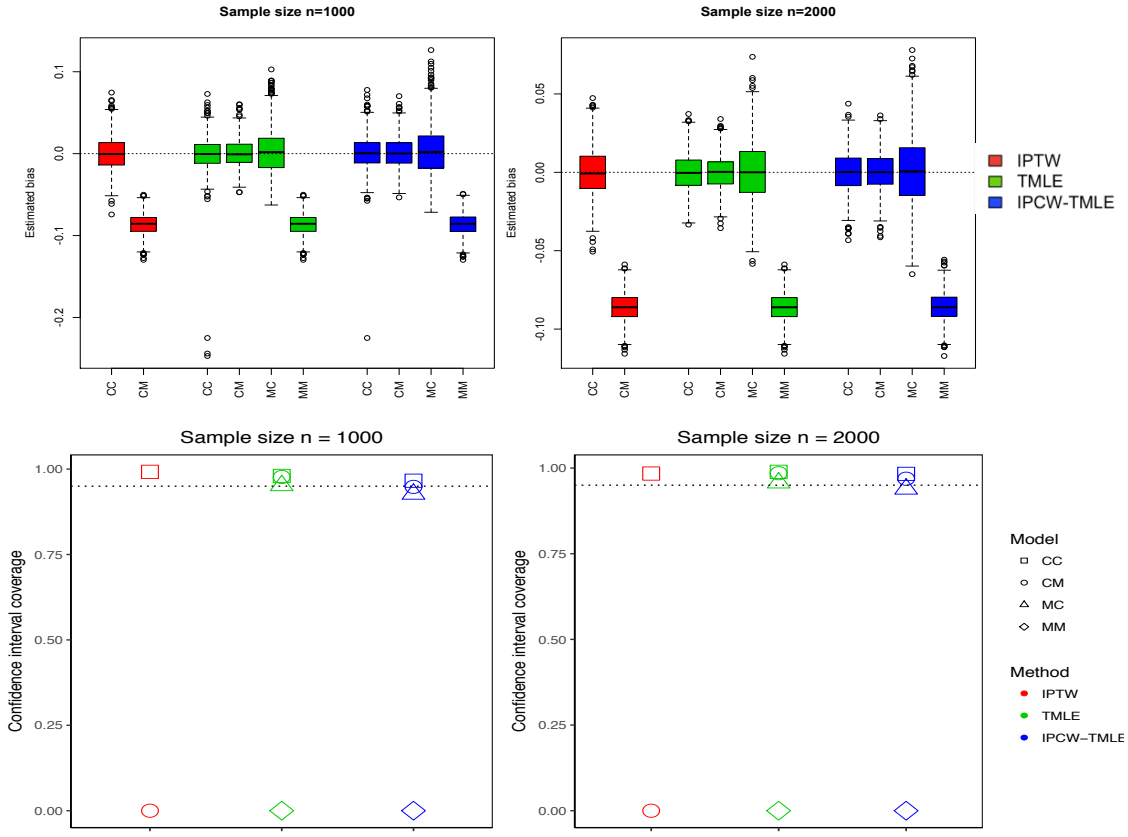

Figure 2.1: Estimated Bias (top panels) and confidence interval coverage (bottom panels) obtained while estimating the target parameter of interest using IPTW (red), TMLE (green) and IPCW-TMLE (blue), for sample sizes n = 1000 and n = 2500. CC indicates both the outcome regression and the treatment regression are correctly specified. CM indicates the outcome regression is correctly specified, but the treatment regression misspecified. MC indicates the outcome regression is misspecified, but the the treatment regression correctly specified. MM indicates both the outcome regression and the treatment regression are misspecified. In the confidence interval coverage plots (bottom panels), the dashed line indicates the 95% confidence interval and CC is represented using 'square' symbol, CM with 'circle', MC with 'triangle' and MM with 'diamond'. The double robustness property of both TMLE and IPCW-TMLE are clear from the estimated bias and confidence interval coverage values. Also, the properties of TMLE and IPCW-TMLE are similar.

Chapter 3

Assessing the causal impact of prenatal exposure to nitrogen dioxide and ozone on stillbirth

3.1 Introduction

The adverse effect of air pollution on human life is a serious concern for today's society. Many studies have been conducted to understand the relationship between air pollution and various human health problems [9, 104, 105, 106]. Children and human fetuses are particularly vulnerable populations, and hence many studies have concentrated on the adverse effects of air pollution on these populations. Prenatal exposure to environmental pollution affects newborns' health, both at birth and at later stages of development [11, 38, 107, 108]; ambient air pollution may also increase the risk of miscarriage, stillbirth, neonatal and postnatal mortality [12, 39, 109, 110, 111].

The Environmental Protection Agency (EPA) identifies motor vehicles as a major source of air pollution in the Unites States [6]. Rise in metropolitan populations and subsequent rise in the number and use of vehicles has resulted in more traffic-related pollution [7]. Traffic emission is the origin for both nitrogen dioxide (NO_2) and ozone (O_3), and well-studied adverse health effects of NO_2 and O_3 exposures include respiratory problems and increased incidence of respiratory-related emergency room visits [16, 112, 113, 114]. Research investigating the relationship between ambient air pollution and stillbirth is relatively a new topic; in particular studies focussing on both NO_2, O_3 or both, is limited [38, 39, 40, 41, 42] and their results have been mixed. For example, no association between NO_2 and stillbirth was found in a study conducted in the Czech Republic [76]. In a Brazilian study [38], an association was found between NO_2 and increased risk of still birth, but no association was reported between O_3 and stillbirth. No statistically significant associations were observed between either pollutants and stillbirth in a study conducted in Taiwan [40]. In a study conducted in New Jersey, NO_2 exposure during the third trimester was associated with

stillbirth [41], but no association was found for a short term NO_2 exposure [42]. In our previous study [39], using the same California cohort data from 1999 to 2009 that we used in the current analysis, both significant and borderline significant associations were found with both NO_2 and O_3 on stillbirth during various pregnancy periods and the relationship was stable even after adjusting for other pollutants included in the study.

In prior studies, the effect of prenatal exposures to NO_2, O_3 or both on stillbirth risk was estimated by applying a logistic regression [39, 40, 42], a generalized estimating equation model [41] or a Poisson regression model [38]. Even though these studies inform us about the association between stillbirth and a pollutant exposure level, there are some limitations. First, parametric regression models were used to assess the relationship between prenatal pollutant exposures and stillbirth and to adjust for measured confounders. Reliance on parametric models to adjust for measured confounders can result in biased estimates and incomplete control for confounding due to model misspecification. Second, prior studies evaluated the association between stillbirth and pollutant exposure based on a conditional odds ratio or a regression based estimate. However, these parameters fail to quantify the stillbirth risk burden attributable from the pollution exposure in the study population of interest, and also it does not inform us about how much the stillbirth risk burden will change if an intervention (e.g. a new policy on air pollution level standards) modifies the exposure level. To address these two limitations, in this analysis, we build on prior research by applying a semi-parametric efficient estimation approach, targeted maximum likelihood estimation (TMLE) [69], to estimate the casual attributable risk of air pollution exposure on stillbirth in a population that has been studied previously using parametric models [39]. We use SuperLearning, a flexible machine learning method [72], to estimate outcome regression and treatment mechanism when implementing TMLE, in order to avoid reliance on misspecified parametric models, and thereby improve both the robustness and precision of our estimates and ensure meaningful statistical inference. The causal attributable risk that we estimate (under an assumption of no unmeasured confounding) corresponds to the absolute change in stillbirth risk that would be observed under a hypothetical intervention to change ambient pollution levels relative to currently observed levels.

3.2 Methods

Study population and outcome of interest

Our study population included all live births and stillbirths that occurred in California between January 1, 1999 and December 31, 2009; births were identified by the Office of Health Information Research of the California Department of Public Health (California Office of Vital Statistics). Birth and fetal death certificates were used to extract information on date of live birth or stillbirth and gestational age of the infant or fetus. Our study cohort included only those mothers who resided at a non-missing California zip code and had singleton births. No information was available about pregnancies that were terminated before 20 weeks of

gestational age and though our study cohort included pregnancies between 20 to 44 weeks of gestational age, in order to avoid selection bias [115] we limited our analysis to those pregnancies that survived until the third trimester (i.e. till 27th week). Our outcome of interest was stillbirth (fetal death), defined in the State of California as the death of a fetus who is at least 20 weeks of gestational age prior to complete expulsion or extraction from the mother (California Code of Regulations, Title 17, Section 916).

Exposure of interest

The exposure of interest was prenatal exposure to NO_2 and O_3 exposures during the third trimesters of pregnancy and we defined the third trimester as the time between 27th week of gestation to delivery. California Air Resources Board provided hourly measures of levels of both pollutants and we used the daily 1-hour maximum values of each pollutant to calculate the corresponding exposure. We used the population-weighted centroid of the 2000 US Census zip code tabulation area, associated with the maternal residential zip code, to assign exposures and ArcGIS software (Esri, Redlands, California) to calculate the distances from this centroid to nearby monitors. To improve accuracy of exposure assessment, we assigned O_3 exposures to those mothers living within 10 km of a corresponding O_3 monitors and because of the known spatial heterogeneity of NO_2 [116, 117], we decided to limit our analysis to only those mothers whose maternal residence is within 5 km of the closest monitor.

Exposure period definitions

We calculated the third trimester exposure to both NO_2 and O_3 in two ways. In the first option, based on the actual gestational age, we calculated the third trimester exposure to both NO_2 and O_3, as the average of weekly mean exposures; provided that at least 5 days of monitored data were available within each week and exposure data were available for at least 75% of the weeks included until the event date (stillbirth or live birth date) in the third trimester. The drawback of defining the third trimester exposure period in this way was that the event date determines the length of exposure period. For example, if the event date occurred in the third trimester, the exposure period could include a minimum of 1 week (for an event date in the 28th week) to a maximum of 17 weeks (for an event date in the 44th week). This approach, while accurately capturing cumulative trimester exposures pre-dating the birth outcome, is subject to bias due to temporal trends in exposure and conception. We therefore also employed a second approach in which, irrespective of actual gestational age at event date, we calculated the third trimester exposures as the average of weekly exposure from 27 weeks until the expected date of delivery i.e. until 40 weeks. Here, each subject will have a fixed 14 weeks of exposure period irrespective of their event date; hence before calculating the cumulative trimester exposure, we made sure that weekly exposures were available for at least 11 out of the 14 weeks in the third trimester. This approach removes the impact of the outcome on the period during which the exposure is assessed, but at the cost that the exposure period may now extend beyond the outcome date.

In both exposure period assigning options, exposures of interest were first calculated as continuous values and then categorized based on quartiles.

Potential confounders

From the available population information provided in the birth or fetal death certificates, we identified three categories of potential confounders: individual factors, temporal factors and neighborhood or community level factors. The first category of the individual factors included maternal age, education, race/ethnicity, air basin of maternal residence as well as sex of the fetus or infant. There are 15 air basins in California to manage and monitor air pollution, and within each of the air basins geographical and meteorological features are comparable; air basin of maternal residence were determined based on the zip codes provided. Second, temporal confounders included season of last menstrual period (LMP) and year of conception. Season of LMP was calculated using the information of LMP and year of conception that divides deliveries into year-long groups was calculated based on the conception date (i.e. LMP date + 2 weeks). Third, neighborhood or community level factors included the following variables (given in percentage) from Census 2000: high school graduates, house ownership, non-Hispanic White, employed, under poverty and under 2 times poverty. Individuals with missing confounder values were excluded from the analysis.

Statistical method

Parameters of interest

Our goal was to estimate the impact of being exposed to a particular level of NO_2 or O_3 on stillbirth during the third trimester of pregnancy. In order to quantify the burden of stillbirth in our population of interest, we estimated the causal attributable risk (CAR), which compares the outcome distribution under a hypothetical intervention on a population with observed outcome distribution [97]. Based on the counterfactual framework (Neyman-Rubin framework) [95, 96], our causal parameter of interest is defined as

$$E(Y) - E(Y_a); a \in \mathcal{A} = \{1, 2, 3, 4\} \tag{3.1}$$

Here Y represents the observed birth outcome, which is equal to one if a stillbirth occurred, and Y_a represents the counterfactual birth outcome that a mother would have had if she had been exposed to exposure quartile $A = a \in \mathcal{A}$. Here entries in $\mathcal{A}$ represent a particular level/quartile of NO_2 or O_3 exposure, with 1 representing the lowest quartile of exposure and 4 representing the highest.

In order to identify the causal parameter of interest from the observed data distribution, we make the following assumptions [71]. First, we assume consistency $Y = Y_A$; i.e. we assume that the the observed outcome (Y) a mother experienced under the observed exposure level is equal to the counterfactual outcome (Y_A) under the observed exposure level. Second, we make the randomization assumption (no unmeasured confounding), which assumes that

given our set of baseline covariates (which we denote as W), the potential outcomes Y_a is independent of the exposure A; i.e. $Y_a \perp\!\!\!\perp A|W$, $\forall\ a \in \mathcal{A}$. Under this assumption, we assume that measured covariates W are sufficient to control for confounding of the effect of A on Y. Finally, the positivity assumption assumes that there is a positive probability for receiving each quartile of exposure A within every combination of baseline covariate values among the population; i.e. $\min_{a \in \mathcal{A}} P(A = a|W) > 0$. Under these assumptions, the causal parameter of interest can be rewritten as [97, 99]

$$E(Y) - E_W[E(Y|A = a, W)]; a \in \mathcal{A} = \{1, 2, 3, 4\} \tag{3.2}$$

This is a function of the observed data distribution alone and it represents the statistical estimand.

Estimation of parameters of interest

The first term of the above statistical estimand, E(Y), can be estimated as a simple empirical mean. There are various methods to estimate the second component of the above statistical estimand. The G-computation (or simple substitution method) utilizes an estimate of the outcome regression, $E(Y|A, W)$ [99, 100]. The inverse probability of treatment weighted (IPTW) method utilizes an estimate of the treatment regression (also known as the propensity score), $P(A|W)$ [101]. The G-computation and IPTW estimators are consistent if the outcome regression or the treatment regression, respectively, are consistently estimated. Here, we utilize another class of estimation methods called targeted maximal likelihood estimators (TMLE) [69] that combine initial estimators of both outcome and treatment regressions. TMLE for the second term of our target parameter, $E_W[E(Y|A = a, W)]$, is a two step procedure where in the first step, initial estimates of both outcome and treatment regressions are obtained and in the second bias-reduction step, the initial estimate of the outcome regression is updated using the estimated treatment regression, targeted towards the parameter of interest. TMLE has the property that it will be consistent, if either one of the two initial estimators is consistent (the so-called "double robustness" property), and it will be efficient (have the lowest asymptotic variance among reasonable estimators) if both are estimated consistently. In our analysis, both the outcome and treatment regression estimates were obtained using SuperLearner [72]. SuperLearner is a data-adaptive algorithm that utilizes cross validation to build an optimal combination of fits that are obtained from user supplied individual prediction algorithms.

Incorporating machine-learning algorithms for analysis on a large data set, like the birth cohort that we are analyzing, can be computationally intensive. Hence, we used a TMLE for a two stage design known as an inverse probability of censoring weighted targeted maximum likelihood estimator (IPCW-TMLE) [70]. The original data set was comprised of the baseline covariates, exposure and outcome on all subjects. We then sampled from this cohort conditional on outcome status resulting in an artificially reduced data structure given as $O^R = (Y, \Delta, \Delta X)$, where X represents the original observed data in the first

stage $X = (W, A, Y)$ and Δ is an indicator about the second stage inclusion, i.e. $\Delta = 1$ if an observation is included in the second stage and $\Delta = 0$ otherwise. Given an outcome Y, each observation thus had a known sampling probability, which we denote by $\Pi_0(Y)$, equal to one if the outcome was a stillbirth and otherwise equal to the proportion of live births sampled. Each step in the estimation procedure (estimation of the initial outcome and treatment regressions using SuperLearner, followed by the targeted updating step, and plugging in the final estimate to the G-computation formula) was implemented using the inverse of these sampling weights; specifically, the sampling weight is equal to $\frac{\Delta}{\Pi_0(Y)}$.

Since under assumptions the IPCW-TMLE is asymptotically linear, its variance can be approximated by the sample variance of its estimated influence curve divided by sample size. Using the Delta method [71], the influence curve for the standard TMLE of the causal attributable risk can be derived and it is explicitly

$$Y - \frac{I(A = a)}{P(A = a|W)}\big(Y - Q_0(A, W)\big) + Q_0(A = a, W) - \Psi(Q_0) \tag{3.3}$$

where $Q_0(A, W)$ denotes the true conditional expectation of the outcome (i.e. the outcome regression $E(Y|A, W)$). If $\Pi_0(Y)$ is estimated using a correctly specified parametric model then the influence curve of the IPCW-TMLE is equal to this full-data influence curve multiplied by the sampling weights $\frac{\Delta}{\Pi_0(Y)}$ minus its projection onto the tangent space of model for $\Pi_0(Y)$ [70]. Thus the variance of the IPCW-TMLE estimator can be conservatively estimated by treating the sampling weights as known, plugging in estimates of the outcome regression and propensity score, and taking the sample variance of the weighted influence curve estimate and dividing by sample size.

The 95% confidence interval for our target parameter is thus estimated as

$$\hat{\psi}_n \pm 1.96\frac{\sqrt{\hat{\sigma}^2}}{\sqrt{N}} \tag{3.4}$$

where $\hat{\psi}_n$ is the estimated causal attributable risk, $\hat{\sigma}^2$ is the sample variance of the weighted influence curve estimate and N is the total number of observations in our final study population.

To implement IPCW-TMLE, first we sampled from our original data. We included all the stillbirths and randomly chose 25 live births per stillbirth. All the stillbirths were given observational weight equal to one and those live births included in the sample were assigned a weight of (proportion of live births included in the subsample in the second stage)$^{-1}$. To implement SuperLearner in our analysis, we used a non-negative least squares loss function, 10-fold cross validation and our library of candidate algorithms included weighted main term logistic regression, simple weighted mean and the stepwise weighted logistic regression. In primary analysis, we bounded our estimated treatment regression from below at 0.01 [118]. We also applied other lower bound values (no lower bound, 0.025 and 0.05) to the estimated treatment regression values to see the influence of these bounds on estimated causal parameter value.

In order to compare results qualitatively with that of prior research that utilized the same data [39], we also conducted a traditional main-term logistic regression analysis including all the confounders, that we adjusted for when estimating CAR, as well as quartiles of NO_2 or O_3 exposure as a four level categorical variable.

Analyses were conducted using SAS version 9.4 and R version 3.1.2 and the random seed was set to 1 while randomly sampling the live births.

3.3 Results

Map of California showing air basins, along with both NO_2 and O_3 monitoring stations, is provided in Figure 3.1. Our study cohort included a total number of 3,574,788 eligible singleton pregnancies with gestational age between 20 to 44 weeks. We were able to assign exposure using option 1 (where we defined the exposure period based on the event date) for all 3,574,788 pregnancies; however only 3,441,667 pregnancies had exposure assigned using option 2 (where the exposure period had a fixed length, irrespective of the event date). We included only those subjects with a non-missing exposure and baseline covariates as well as we limited our analysis to those pregnancies that survived until the third trimester. Details on exclusion criteria applied to our study cohort, for the NO_2 and O_3 analyses, are provided in Figures 3.2 and 3.3 resepectively. To assess the impact of NO_2 exposure on stillbirth, our final study population included 1,113,651 subjects with exposure assigned using option 1 exposure period definition and 1,112,797 subjects with exposure assigned using option 2 exposure period definition; and these numbers for O_3 analyses were 2,668,464 and 2,668,100 respectively. The distributions of various demographic and clinical covariates in our final study populations for NO_2 analysis are given in Tables 3.1 through 3.3 and that for O_3 analysis are given in Tables 3.4 through 3.6. The total number of stillbirths and live births in our final study populations was different for NO_2 and O_3 analyses; however, the characteristics of the populations were similar. In our final study populations for both analyses, majority of mothers were Hispanic, aged 25 to 34 years, and not educated beyond high school and most of the fetuses were male. The quartile specific exposure matrices for both NO_2 and O_3 are provided in Tables 3.7 and 3.8 respectively and outcome specific exposure matrices are given in Table 3.9. The exposure distributions were very similar irrespective of the exposure period definitions that we applied.

We observed practical positivity violations, i.e. the predicted probability of treatment mechanism was close to zero and in order to mitigate the issues related to it, a lower bound to the predicted treatment mechanism was applied while estimating the target parameter. The summary statistic of predicted probabilities of the treatment mechanism, before applying a lower bound, is given in Tables 3.10 and 3.11 respectively. First focusing on propensity scores obtained in the NO_2 exposure analysis as shown in Tables 3.10; regardless of the exposure period definitions that we used, around 25% of study subjects had propensity score smaller than 0.01 in fourth exposure quartile during the third trimester. Around 2% and 2.5% of subjects had a propensity score below 0.01, in the first and third exposure

quartiles respectively. For the O_3 exposure analysis, as given in Table 3.11, irrespective of the exposure period definition, no subject had a propensity score value below 0.01 in the first three exposure quartiles, but around 11% subjects had a propensity score smaller than 0.01 in the fourth exposure quartile for both definitions of the exposure period.

Within each exposure quartile, an estimate of $E(Y|A = a), a \in \mathcal{A}$, obtained using the empirical mean of stillbirth and an estimate of the counterfactual outcome, $E(Y_a)$, obtained using IPCW-TMLE from both NO_2 or O_3 exposures are given in Table 3.12. We denote the estimate of $E(Y|A = a)$ as an *unadjusted* estimator and that of $E(Y_a)$ as an *adjusted* estimator. If exposure indeed increases risk of stillbirth in the third trimester, we would expect their values in exposure quartile 1 will be lowest and that in exposure quartile 4 will be highest. First focusing on NO_2 analysis results. In trimester 3, when the exposure period was defined using option 1, we observed an increasing trend in the unadjusted stillbirth risk with respect to exposure except in the fourth exposure quartile, and though we did not see any increasing trend in adjusted stillbirth risk with respect to exposure, the adjusted counterfactual risk in the fourth exposure quartile was highest compared with other three quartiles. When the exposure period was defined using option 2, we observed an increasing trend in both unadjusted and IPCW-TMLE adjusted stillbirth risk estimators with respect to exposure with highest stillbirth risk in the fourth exposure quartile. Now focusing on ozone exposure results: when exposure was assigned using option 1, an increasing trend was found in adjusted stillbirth risk estimates with respect to exposure; but when the exposure was assigned using option 2, an increasing trend among the same was present only in the first three exposure quartiles.

Figure 3.4 shows the estimated causal attributable risk (CAR) of stillbirth, our target parameter of interest, from both NO_2 and O_3 exposures in the third trimester. Here the estimated CAR is the difference between the unadjusted estimate of the empirical stillbirth risk in the whole data and the adjusted estimate of the counterfactual stillbirth risk within each exposure quartile. In these plots a negative parameter value indicates that the estimated adjusted stillbirth risk, under a hypothetical intervention to set a pollutant exposure equal to a particular exposure quartile, is higher than the unadjusted stillbirth risk; i.e. the more negative the estimated CAR value, greater the harmful effect of that pollutant exposure quartile on a growing fetus, relative to current levels of pollutant. First focus on the estimated CAR of stillbirth from NO_2 exposure in the third trimester. When the exposure was assigned using option 1 exposure period, we found that exposing all pregnant women to the fourth exposure quartile would increase the stillbirth risk by 0.291% (95% CI: 0.236%, 0.346%). However, when the exposure was assigned using option 2 exposure period, none of the CAR estimates were significantly different from zero. Now focusing on the estimated CAR of stillbirth from O_3 exposure in the third trimester. When the exposure was assigned using option 1 exposure period, we found a trend in estimated CAR of stillbirth, with a protective effect in the first exposure quartile (estimated CAR = 0.019%, 95% CI: 0.007%, 0.031%) and an increasing risk in the fourth exposure quartile (estimated CAR = -0.056%, 95% CI: -0.086%, -0.027%). When the O_3 exposure was assigned using option 2 the same trend was observed for the first three quartiles; however, some of the effect estimates were not

significant and for the fourth exposure quartile, the estimated CAR was significantly positive, suggesting a protective effect. We believe that the practical positivity violations that we observed, especially in the 4th exposure quartile of O_3 exposure might have produced the counterintuitive results that we observed as well as the large confidence intervals in some of the exposure quartiles.

In order to investigate the potential impact of positivity violations, we conducted sensitivity analyses in which we estimated the casual attributable risk under various values (0.05, 0.025, 0.01 and no cut off) to the lower bound of propensity score. Results are shown in Figures 3.5 and 3.6. In Figure 3.5, we observed some of the lower bounds produced a significant change in the CAR estimates from NO_2 exposure, especially for exposure quartile 3 and 4, regardless of the exposure assigning options; but no patterns were observed. In Figure 3.6, the effect of various lower bound to the propensity score was more visible in exposure quartile 4, especially when the O_3 exposure was assigned using option 1. In both Figures 3.5 and 3.6, CAR estimates calculated using option 2 exposure definition had larger confidence interval compared with that calculated using option 1 exposure; this was true with all the lower bound values that we applied to the propensity scores.

Table 3.13 shows the results obtained from the traditional logistic regression analysis with all the main-term baseline covariates that we adjusted for when estimating CAR and the quartiles of third trimester NO_2 or O_3 exposure under two exposure scenarios. However, the reported conditional odds ratio of stillbirth is a different statistical parameter than the IPCW-TMLE estimated CAR. Even if we correctly specify the parametric regression model and suppose the randomization assumption holds, the exponentiated regression coefficients of exposure quartiles estimate the conditional causal odds ratio, which is different from the estimated CAR under the equivalent causal assumptions. However, we can utilize Table 3.13 results for a qualitative comparison. For exposure assigned using option 2, the parametric logistic regression results suggests a detrimental effect of exposure to the third quartile of O_3 (adjusted odds ratio = 1.078, 95% CI: 1.008, 1.153), while IPCW-TMLE found no significant effect. This may be attributable to more complete control for measured confounders when utilizing IPCW-TMLE. When comparing the adjusted odds ratio of stillbirth under exposure assignment option 1 for both NO_2 and O_3, provided in Table 3.13, with that of our prior analysis [39] we can see that the results are slightly different. Our current analysis is different from that of our prior [39] in the following aspects: a) here we limited our analysis to include only those fetuses that survived until the third trimester, b) we included census variables to adjust for neighborhood or community level confounders along with other listed baseline covariates, and c) we performed the logistic regression with quartile of NO_2 or O_3 exposure.

3.4 Discussion

We estimated the CAR of stillbirth from two criteria pollutants during the third trimester of pregnancy using a semi-parametric targeted approach. While estimating the CAR of stillbirth, we examined sensitivity to different definitions of exposure period, which is relevant

especially when there are temporal trends in both conception and exposure of interest. When assigning the exposure using option 1, the event date determined the exposure period length. However, since both conception and pollutant levels can vary seasonally, this definition may lead us to conclude an effect when there is only simple temporal co-variation of exposure and stillbirth [119]. Therefore, we defined exposure assignment option 2 in which a fixed exposure period was assigned based on the conception date. The disadvantage with option 2 is that the assigned exposure level may not be consistent with the pregnancy period and hence it could weaken the relationship between exposure and stillbirth.

Based on the estimated CARs obtained when using the exposure period definition that depends on the event date (option 1), our results suggest that prenatal exposure to highest levels (i.e. exposure quartile 4) of both NO_2 and O_3 during the third trimester is harmful. However we may be seeing this effect because of the temporal trends in pollution and conception, for we did not see a similar trend when we used the fixed exposure period irrespective of the actual gestational age experienced by a mother (option 2). Thus our results imply the causal estimates are affected by the exposure period definition used.

As mentioned in the introduction, previous studies have examined the impact of prenatal nitrogen dioxide and ozone exposure on stillbirth based on parametric models and results to date have been mixed. A direct comparison between our current analysis results and that of previous studies is not possible because of the difference in estimated parameters reported. In the prior studies a difference in stillbirth associated with exposure during various gestational periods, while holding the baseline covariates as constant, were given in terms of odds ratio. But in our analysis, using counterfactual scenarios that arise by applying a hypothetical intervention exposure level, we compared the counterfactual risk of stillbirth with the observed risk of stillbirth in our study population. Although a direct comparison is not possible, our results strengthen these previous studies that established an association between NO_2 and O_3 exposure and stillbirth by establishing a weak causal link, under assumptions, especially when we assigned the exposure with exposure period depending on the event date (i.e. option 1); however the significance of the causal effect estimates depended on the exposure period definition.

Exact biological mechanisms, explaining the effect of air pollution on the growing fetus that may lead to stillbirth, are yet to be established, even though many biologic pathways have been proposed [43, 45]. Studies have suggested that air pollution (especially NO_2, CO, SO_2 and PM_{10}) can alter the blood coagulability and plasma viscosity [46, 47]. This may adversely affect the umbilical blood flow, leading to inadequate placental transfer of oxygen and thus consequently affecting the fetal growth. Air pollutants such as nitrogen oxides, O_3, particulate matter are capable of generating reactive oxygen species and thus leading to oxidative stress [50] and it has been hypothesized that oxidative stress can result in DNA damage and in premature placental aging resulting in fetal vulnerability [51, 52].

There are some limitations to our study. We estimated our parameter of interest under two major assumptions - randomization assumption and positivity assumption. In this analysis we might have excluded some important maternal (e.g. smoking status, individual SES, perinatal risk factors) and community factors (e.g. safety features, access to amenities

and resources that support a healthy living) because of their unavailability in our data set during the entire study period. However, prior research analyzing infant mortality has not observed a change in relationship with air pollution exposure after considering maternal smoking status [120]. Also, we assume individual maternal education status as well as the neighborhood level confounders considered in our analysis could act as a proxy for individual SES information as well as the care a mother received during her pregnancy. In addition to unmeasured individual level confounders, unmeasured confounding at the community and geographic level remains a concern [121, 122, 123]. While we adjusted for air basin as well as a number of factors at the level of census tract, unmeasured factors that vary spatially and correlate with pollution exposure may in part explain our results. We assumed that there will be enough variability within NO_2 or O_3 exposure quartiles, irrespective of various baseline covariate strata, and hence it is reasonable to assume a positivity assumption; however, in our analysis, we observed practical positivity violations. Even though we tried to minimize impact by setting a lower bound on the estimated propensity score, we observed some counterintuitive results as well as large variance in those exposure quartiles where we had large positivity violations. An additional limitation is the potential misclassification in the assigned exposure that may arise due to the fact that we used ambient air pollution exposure obtained from the closest monitor based on maternal residence zip code at the delivery time. Also, actual exposure experienced by each mother depends on other factors, like time spend outdoor and at the assigned zip code, the distance of the maternal residence from the assigned air pollution monitor and all these factors might have also contributed to exposure misclassification. We obtained the date of stillbirth occurrence from the provided fetal death certificate data, but prior studies show that a lag between the actual stillbirth and complete expulsion or extraction of fetus from it's mother is possible and this may cause inaccuracy in the reported stillbirth date. In our analysis, we considered only nitrogen dioxide and ozone at a time, however each mother might have experienced a pollution mixture that contains many other pollutants as well. Future studies addressing these limitations would be beneficial.

Our analysis has several strengths. In our knowledge this is the first study that calculated the causal attributable risk of stillbirth from two criteria air pollutants, using a semi-parametric targeted estimation technique incorporating machine learning. Estimating a causal attributable risk, that compares the health effect of a potential exposure level with that of observed, is relevant for assessing the impact of new policies related to air pollution and for designing appropriate strategies to help those population who will be affected by the proposed air pollution standards. Our analysis could easily extend to any policy relevant hypothetical level of both nitrogen dioxide and ozone as well as any other pollutant of interest and, under the key assumption that relevant confounders are measured, could quantify it's causal impact on stillbirth. In a study cohort, that consists all the important individual and neighborhood level factors related to both air pollution exposure and birth outcome, quantifying the health effect of potential intervention and establishing a causal relationship between the exposure and outcome would be more convincing than a mere association in policy relevant issues [73, 74]. In our analysis we used IPCW-TMLE, to estimate the causal

attributable risk of stillbirth, which allowed us to work with a sub-sample of the final study population and thus helped to overcome the computational challenges posed by the large amount of data in our study cohort. Similar to the standard TMLE, IPCW-TMLE is also double robust [70], which means the estimator of our parameter of interest obtained using IPCW-TMLE will be consistent if either the outcome regression or the propensity score are estimated consistently. In our analysis, we incorporated a machine learning algorithm, SuperLearner, to estimate both the outcome and treatment regressions, thus avoid bias from parametric misspecification and helps minimize variance and ensure accurate statistical inference for IPCW-TMLE [71, 72].

In our future research, we are interested in exploring the effects of other criteria pollutants on stillbirth. To address the challenges in defining the exposure period based on gestational length, we are also interested in exploring other methodologies such as time-to-event or survival approach to estimate the risk of stillbirth. Survival analysis will be useful to explore week specific probability of stillbirth risk and will have more flexibility for comparing stillbirth and live birth on a smaller time scale; and this is important as the fetal development and hence the stillbirth risk probability varies on a smaller time period than trimesters.

In conclusion, we were able to quantify the causal attributable risk from the third trimester prenatal exposure of nitrogen dioxide or ozone on stillbirth using California cohort data. In our knowledge, this is the first epidemiological study that quantified the causal impact from ambient air pollution on stillbirth using a semi-parametric targeted estimation approach. Since studies assessing the causal impact of air pollution exposure on stillbirth is a new research area, further studies in other locations as well as data base with more stillbirth risk factor information are recommended.

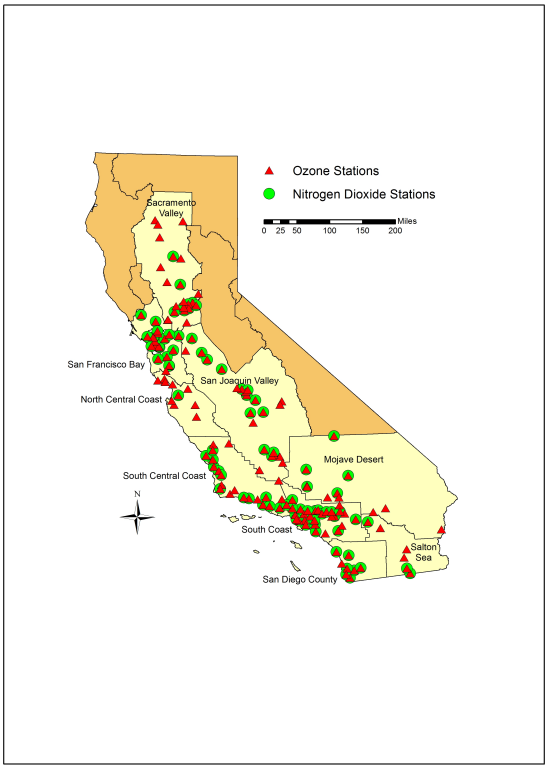

Figure 3.1: Map of California showing air basins with nitrogen dioxide and ozone monitors

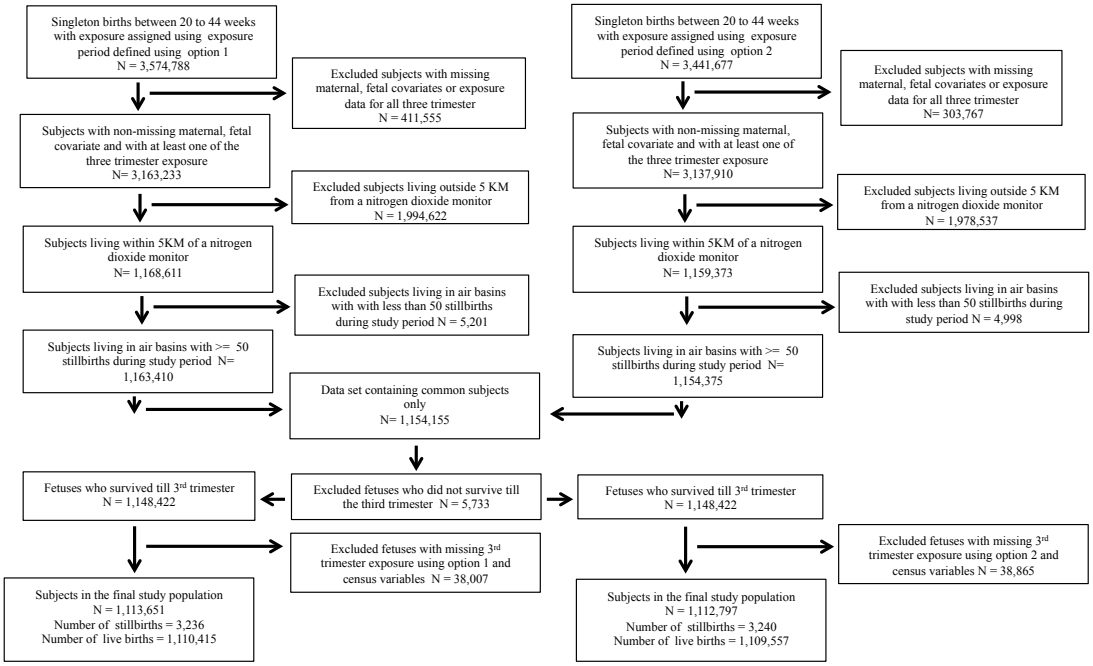

Figure 3.2: Flow chart showing the study population, based on exclusion criteria for nitrogen dioxide analysis, in California 1999 to 2009

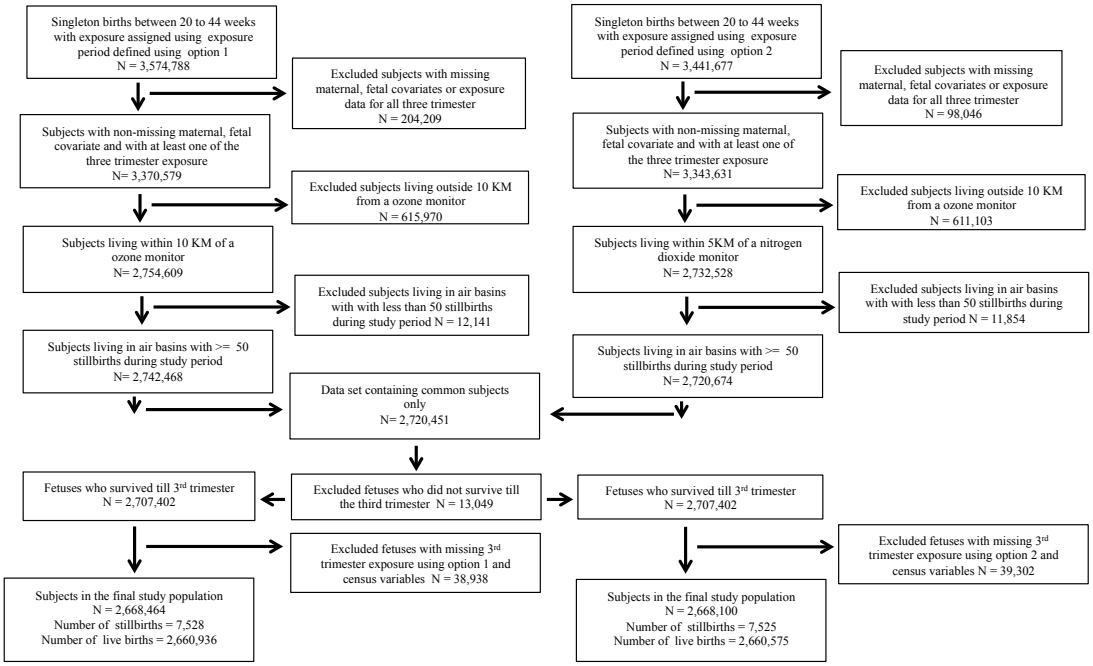

Figure 3.3: Flow chart showing the study population, based on exclusion criteria for ozone analysis, in California 1999 to 2009

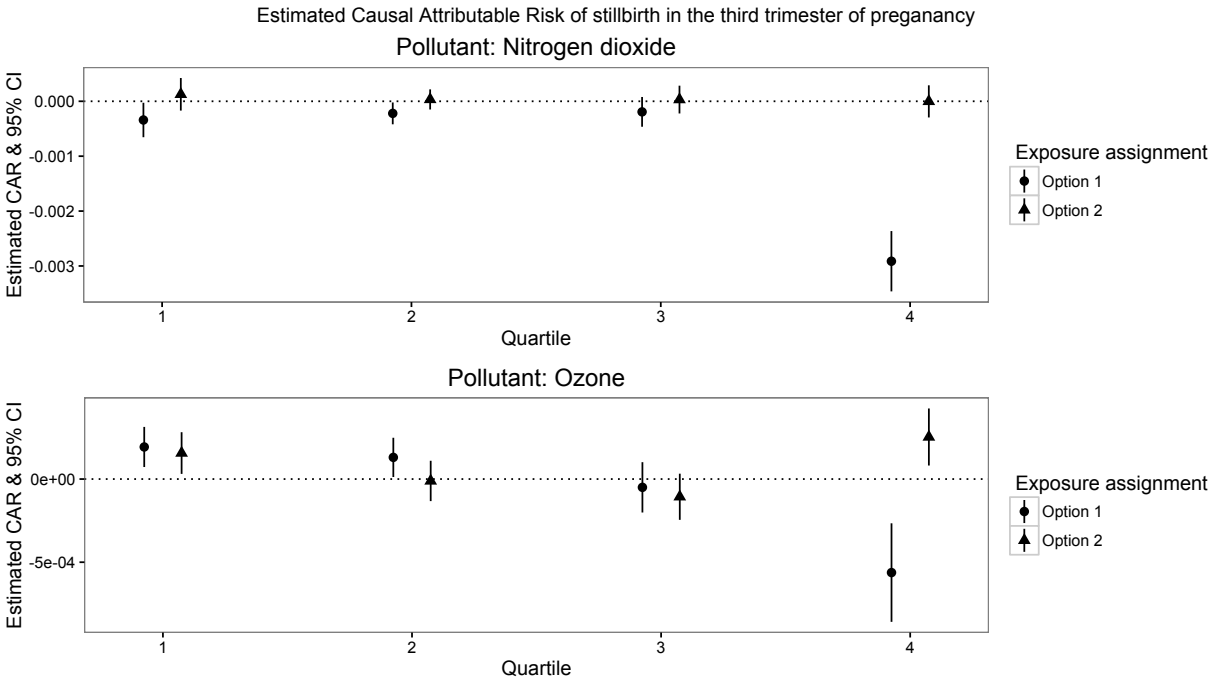

Figure 3.4: Estimated causal attributable risk of stillbirth from the nitrogen dioxide and ozone exposures in the third trimester of pregnancy. Fetuses that survived until the 3rd trimester were included in the analysis and propensity score were given a lower bound at 0.01. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

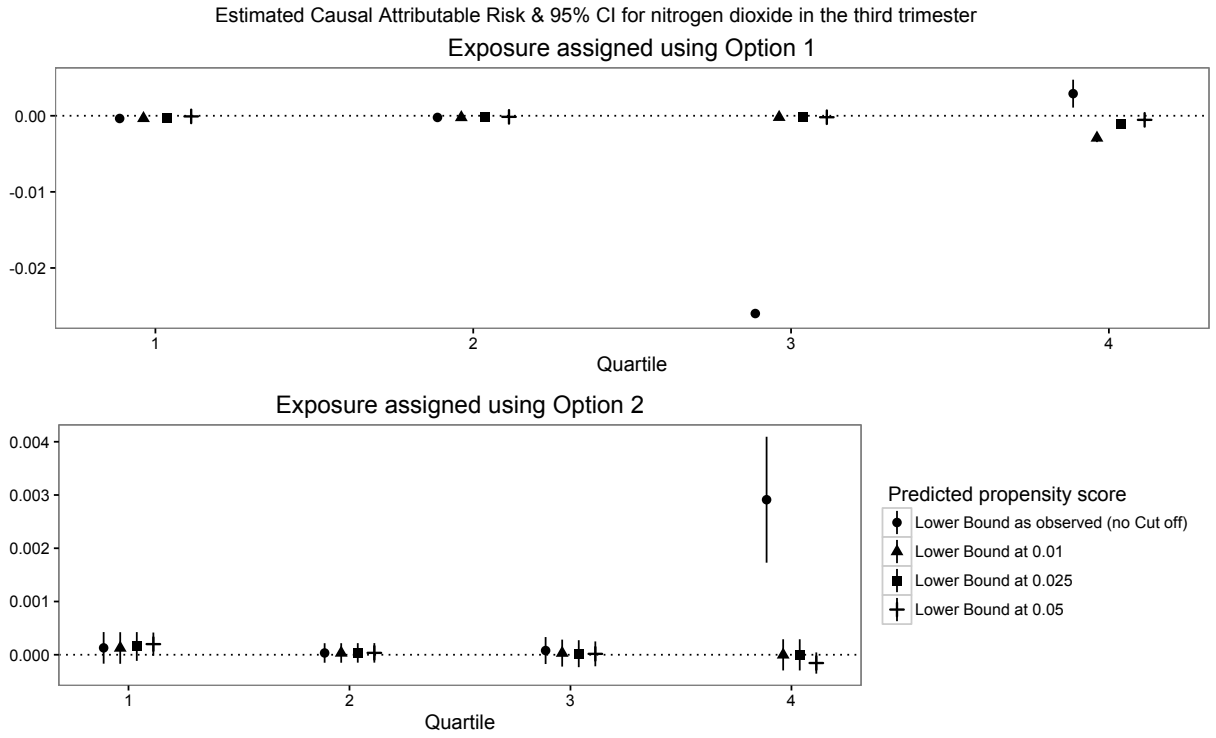

Figure 3.5: Estimated causal attributable risk of stillbirth from the nitrogen dioxide exposure in the third trimester of pregnancy. Fetuses that survived until the 3rd trimester were included in the analysis and propensity score were given various lower bounds. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

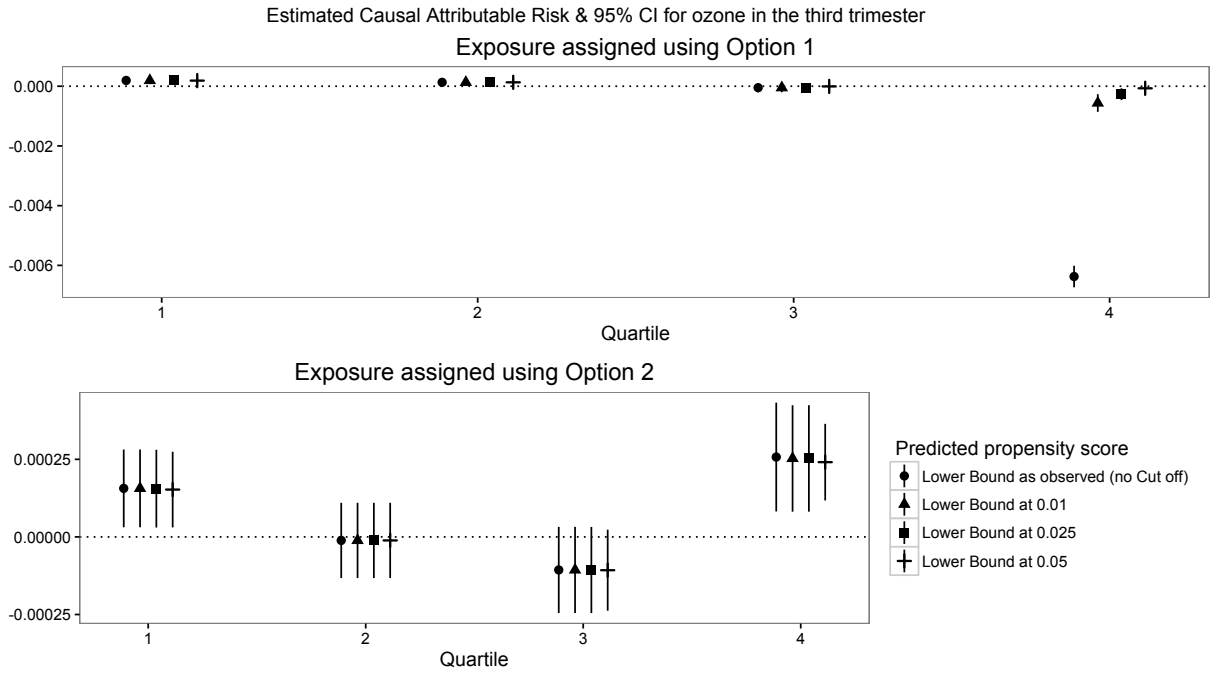

Figure 3.6: Estimated causal attributable risk of stillbirth from the ozone exposure in the third trimester of pregnancy. Fetuses that survived until the 3rd trimester were included in the analysis and propensity score were given various lower bounds. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

Table 3.1: Characteristics of stillbirths and live births in our final study population that we used to estimate CAR of stillbirth from third trimester nitrogen dioxide exposure (Part 1). This includes only those fetuses who survived until the third trimester. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

Characteristics	Pollutant: Nitrogen dioxide							
	Exposure period definition: Option 1				Exposure period definition: Option 2			
	Stillbirth		Live birth		Stillbirth		Live birth	
	N	%	N	%	N	%	N	%
Total	3,236	100	1,110,415	100	3,240	100	1,109,557	100
Maternal age (Years)								
< 25	1,169	36.1	401,343	36.1	1,169	36.1	400,969	36.1
$25 - 34$	1,484	45.9	547,086	49.3	1,489	46.0	546,681	49.3
> 34	583	18.0	161,986	14.6	582	18.0	161,907	14.6
Maternal education								
High school or less	2,181	67.4	678,597	61.1	2184	67.4	678,070	61.1
Some college	6,53	20.2	221,922	20.0	653	20.2	221,731	20.0
College or beyond	402	12.4	209,896	18.9	403	12.4	209,756	18.9
Maternal race/ethnicity								
Non-Hispanic white	770	23.8	301,422	27.1	772	23.8	301,197	27.1
Non-Hispanic black	388	12.0	68,737	6.2	387	11.9	68,686	6.2
Hispanic	1,774	54.8	632,319	56.9	1779	54.9	631,769	56.9
Non-Hispanic asian	251	7.8	93,201	8.4	251	7.7	93,178	8.4
Other non-Hispanic	53	1.6	14,736	1.3	51	1.6	14,727	1.3

Table 3.2: Characteristics of stillbirths and live births in our final study population that we used to estimate CAR of stillbirth from third trimester nitrogen dioxide exposure (Part 2). This includes only those fetuses who survived until the third trimester. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

Characteristics	Pollutant: Nitrogen dioxide							
	Exposure period definition: Option 1				Exposure period definition: Option 2			
	Stillbirth		Live birth		Stillbirth		Live birth	
	N	%	N	%	N	%	N	%
Total	3,236	100	1,110,415	100	3,240	100	1,109,557	100
Season of LMP								
Cool (November - April)	1,599	49.4	562,509	50.7	1,604	49.5	562,152	50.7
Warm (May - October)	1,637	50.6	547,906	49.3	1,636	50.5	547,405	49.3
California airbasin								
Mojave Desert	76	2.3	23,268	2.1	77	2.4	23,281	2.1
North Central Coast	73	2.3	27,532	2.5	73	2.3	27,560	2.5
Sacramento Valley	252	7.8	95,280	8.6	252	7.8	95,134	8.6
Salton Sea	33	1.0	15,213	1.4	34	1.0	15,138	1.4
San Diego County	252	7.8	90,218	8.1	253	7.8	90,069	8.1
San Francisco Bay	284	8.8	96,190	8.7	284	8.8	96,205	8.7
San Joaquin Valley	674	20.8	192,878	17.4	674	20.8	192,652	17.4
South Central Coast	113	3.5	46,873	4.2	114	3.5	46,672	4.2
South Coast	1,477	45.6	522,963	47.1	1,479	45.6	522,846	47.1

Table 3.3: Characteristics of stillbirths and live births in our final study population that we used to estimate CAR of stillbirth from third trimester nitrogen dioxide exposure (Part 3). This includes only those fetuses who survived until the third trimester. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

Characteristics	Pollutant: Nitrogen dioxide							
	Exposure period definition: Option 1				Exposure period definition: Option 2			
	Stillbirth		Live birth		Stillbirth		Live birth	
	N	%	N	%	N	%	N	%
Total	3,236	100	1,110,415	100	3,240	100	1,109,557	100
Fetal sex								
Male	1,661	51.3	568,025	51.2	1,659	51.2	567,608	51.2
Female	1,575	48.7	542,390	48.8	1,581	48.8	541,949	48.8
Gestational length (weeks)								
27 - 36	1,990	61.5	100,668	9.1	1,998	61.7	100,692	9.1
37 - 44	1,246	38.5	1,009,747	90.9	1,242	38.3	1,008,865	90.9
Gestation length (weeks)	Mean (SD)		Mean (SD)		Mean (SD)		Mean (SD)	
	34.52 (4.43)		38.91 (2.03)		34.50 (4.42)		38.91 (2.03)	

Table 3.4: Characteristics of stillbirths and live births in our final study population that we used to estimate CAR of stillbirth from third trimester ozone exposure (Part 1). This includes only those fetuses who survived until the third trimester. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

Characteristics	Pollutant: Ozone							
	Exposure period definition: Option 1				Exposure period definition: Option 2			
	Stillbirth		Live birth		Stillbirth		Live birth	
	N	%	N	%	N	%	N	%
Total	7,528	100	2,660,936	100	7,525	100	2,660,575	100
Maternal age (Years)								
< 25	2,545	33.8	909,615	34.2	2,537	33.7	909,522	34.2
$25 - 34$	3,460	46.0	1,336,153	50.2	3,463	46.0	1,335,908	50.2
> 34	1,523	20.2	415,168	15.6	1,525	20.3	415,145	15.6
Maternal education								
High school or less	4,966	66.0	1,575,240	59.2	4,969	66.0	1,575,066	59.2
Some college	1,581	21.0	535,667	20.1	1,584	21.0	535,591	20.1
College or beyond	981	13.0	550,029	20.7	972	12.9	549,918	20.7
Maternal race/ethnicity								
Non-Hispanic white	1,758	23.4	716,405	26.9	1,756	23.3	716,190	26.9
Non-Hispanic black	824	10.9	156,549	5.9	820	10.9	156,493	5.9
Hispanic	4,222	56.1	1,484,112	55.8	4,231	56.2	1,484,019	55.8
Non-Hispanic asian	622	8.3	271,939	10.2	618	8.2	271,942	10.2
Other non-Hispanic	102	1.4	31,931	1.2	100	1.3	31,931	1.2

Table 3.5: Characteristics of stillbirths and live births in our final study population that we used to estimate CAR of stillbirth from third trimester ozone exposure (Part 2). This includes only those fetuses who survived until the third trimester. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

Characteristics	Pollutant: Ozone							
	Exposure period definition: Option 1				Exposure period definition: Option 2			
	Stillbirth		Live birth		Stillbirth		Live birth	
	N	%	N	%	N	%	N	%
Total	7,528	100	2,660,936	100	7,525	100	2,660,575	100
Season of LMP								
Cool (November - April)	3,823	50.8	1,352,666	50.8	3,816	50.7	1,352,218	50.8
Warm (May - October)	3,705	49.2	1,308,270	49.2	3,709	49.3	1,308,357	49.2
California air basin								
Mojave Desert	198	2.6	59,721	2.2	199	2.6	59,734	2.2
North Central Coast	190	2.5	73,422	2.8	190	2.5	73,429	2.8
Sacramento Valley	561	7.5	210,103	7.9	557	7.4	209,873	7.9
Salton Sea	119	1.6	47,439	1.8	118	1.6	47,420	1.8
San Diego County	551	7.3	229,781	8.6	550	7.3	229,545	8.6
San Francisco Bay	630	8.4	231,439	8.7	624	8.3	231,448	8.7
San Joaquin Valley	1,082	14.4	319,432	12.0	1084	14.4	319,282	12.0
South Central Coast	369	4.9	139,588	5.2	369	4.9	139,602	5.2
South Coast	3,828	50.9	1,350,011	50.7	3,834	51.0	1,350,242	50.8

Table 3.6: Characteristics of stillbirths and live births in our final study population that we used to estimate CAR of stillbirth from third trimester ozone exposure (Part 3). This includes only those fetuses who survived until the third trimester. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

Characteristics	Pollutant: Ozone							
	Exposure period definition: Option 1				Exposure period definition: Option 2			
	Stillbirth		Live birth		Stillbirth		Live birth	
	N	%	N	%	N	%	N	%
Total	7,528	100	2,660,936	100	7,525	100	2,660,575	100
Fetal sex								
Male	3883	51.6	1,362,207	51.2	3883	51.6	1,362,020	51.2
Female	3645	48.4	1,298,729	48.8	3642	48.4	1,298,555	48.8
Gestational length (weeks)								
27 - 36	4,665	62.0	239,203	9.0	4,667	62.0	239,378	9.0
37 - 44	2,863	38.0	2,421,733	91.0	2,858	38.0	2,421,197	91.0
Gestation length (weeks)	Mean (SD)		Mean (SD)		Mean (SD)		Mean (SD)	
	34.54 (4.36)		38.90 (2.02)		34.53 (4.36)		38.90 (2.02)	

Table 3.7: Quartile specific NO_2 exposure summary in the final study population. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date. Our final study population included only those fetuses who survived until the third trimester.

NO_2 Exposure period definition using option 1 (exposure period determined by the event date)							
Trimester	Exposure quartile	Minimum (ppb)	First Quartile (ppb)	Median (ppb)	Third Quartile (ppb)	Maximum (ppb)	Number of subjects
3	1	3.34	17.10	20.84	23.50	25.87	278,394
	2	25.88	28.26	30.58	32.81	35.12	278,430
	3	35.12	37.41	39.76	42.32	45.04	278,414
	4	45.04	48.12	51.71	57.53	108.00	278,413
NO_2 Exposure period definition using option 2 (fixed exposure period, irrespective of the event date)							
Trimester	Exposure quartile	Minimum (ppb)	First Quartile (ppb)	Median (ppb)	Third Quartile (ppb)	Maximum (ppb)	Number of subjects
3	1	4.122	17.15	20.90	23.54	25.90	278,180
	2	25.90	28.30	30.59	32.85	35.14	278,188
	3	35.14	37.41	39.76	42.33	45.03	278,091
	4	45.03	48.10	51.67	57.52	94.92	278,338

Table 3.8: Quartile specific O_3 exposure summary in the final study population. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date. Our final study population included only those fetuses who survived until the third trimester.

O_3 Exposure period definition using option 1 (exposure period determined by the event date)							
Trimester	Exposure quartile	Minimum (ppb)	First Quartile (ppb)	Median (ppb)	Third Quartile (ppb)	Maximum (ppb)	Number of subjects
3	1	5.17	27.86	32.06	35.36	38.00	666,566
	2	38.00	40.53	43.12	45.75	48.48	667,628
	3	48.48	50.95	53.41	56.07	59.06	667,143
	4	59.06	62.95	68.58	77.27	119.80	667,127
O_3 Exposure period definition using option 2 (fixed exposure period, irrespective of the event date)							
Trimester	Exposure quartile	Minimum (ppb)	First Quartile (ppb)	Median (ppb)	Third Quartile (ppb)	Maximum (ppb)	Number of subjects
3	1	6.92	28.04	32.23	35.46	38.05	666,498
	2	38.05	40.57	43.13	45.73	48.49	667,041
	3	48.49	50.93	53.38	56.02	58.98	667,343
	4	58.98	62.89	68.46	77.05	111.00	667,218

Table 3.9: Outcome specific exposure summary in the final study population. Fetuses that survived until the 3rd trimester were included in the analysis and propensity score were given various lower bounds. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

Trimester	Assigned exposure	Pollutant: Nitrogen dioxide		Pollutant: Ozone	
		Stillbirth	Live birth	Stillbirth	Live birth
		Median (IQR) (ppb)	Median (IQR) (ppb)	Median (IQR) (ppb)	Median (IQR) (ppb)
3	Option 1: exposure period length was determined by the event date	35.6 (19.2)	35.1 (19.2)	48.6 (22.6)	48.5 (21.1)
	Option 2: fixed exposure period, irrespective of the event date	35.7 (19.3)	35.1 (19.1)	48.7 (21.3)	48.5 (20.9)

Table 3.10: Propensity score before setting a lower bound to mitigate practical positivity violation that occurred while estimating the causal attributable risk from NO_2 exposure on stillbirth. Fetuses that survived until the 3rd trimester were included in the analysis and propensity score were given various lower bounds. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

NO_2 Exposure period definition using option 1 (exposure period was determined by the event date)							
Trimester & sub-sample size	Exposure quartile	Minimum	First Quartile	Median	Third Quartile	Maximum	No. of subjects with propensity score < 0.01
Trimester 3 &	1	0.004915	0.038538	0.138013	0.416321	0.991608	1,707
no. of subjects	2	0.0104	0.1255	0.2136	0.3669	0.6387	0
in sub-sample	3	0.0000	0.1619	0.2690	0.3384	0.6404	2,168
= 84,136	4	0.001118	0.009122	0.099295	0.457283	0.975077	21,724
NO_2 Exposure period definition using option 2 (fixed exposure period, irrespective of the event date)							
Trimester & sub-sample size	Exposure quartile	Minimum	First Quartile	Median	Third Quartile	Maximum	No. of subjects with propensity score < 0.01
Trimester 3 &	1	0.004827	0.038726	0.136817	0.408772	0.992231	1,916
no. of subjects	2	0.01094	0.12570	0.21779	0.37145	0.64842	0
in sub-sample	3	0.0000	0.1616	0.2708	0.3419	0.6506	2,168
= 84,240	4	0.001361	0.008275	0.093302	0.453999	0.976648	22,729

Table 3.11: Propensity score before setting a lower bound to mitigate practical positivity violation that occurred while estimating the causal attributable risk from O_3 exposure on stillbirth. Fetuses that survived until the 3rd trimester were included in the analysis and propensity score were given various lower bounds. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

O_3 Exposure period definition using option 1 (exposure period was determined by the event date)							
Trimester	Exposure quartile	Minimum	First Quartile	Median	Third Quartile	Maximum	No. of subjects with propensity score < 0.01
Trimester 3 &	1	0.0129	0.1476	0.2331	0.3352	0.8198	0
number of subjects	2	0.07636	0.18397	0.22347	0.27995	0.70253	0
in sub-sample	3	0.01147	0.18134	0.23466	0.27974	0.69408	0
= 195,728	4	0.000685	0.100636	0.222579	0.381306	0.948545	20,785
O_3 Exposure period definition using option 2 (fixed exposure period, irrespective of the event date)							
Trimester	Exposure quartile	Minimum	First Quartile	Median	Third Quartile	Maximum	No. of subjects with propensity score < 0.01
Trimester 3 &	1	0.01148	0.15304	0.23457	0.33024	0.81258	0
number of subjects	2	0.08748	0.18454	0.22272	0.27988	0.70349	0
in sub-sample	3	0.01078	0.18200	0.23661	0.28167	0.68744	0
= 195,650	4	0.000123	0.102001	0.224878	0.375222	0.880484	21,822

Table 3.12: Observed and predicted counterfactual probability of stillbirth (given in %) from nitrogen dioxide or ozone exposures. Fetuses that survived until the third trimester (till the 27th week) were included in the analysis and propensity score were given a lower bound at 0.01. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

Trimester	Exposure quartile	Pollutant: Nitrogen dioxide			
		Exposure period definition: Option 1		Exposure period definition: Option 2	
		Probability of stillbirth (%)		Probability of stillbirth (%)	
		Unadjusted (95% CI)	Adjusted (95% CI)	Unadjusted (95% CI)	Adjusted (95% CI)
3	1	0.276 (0.256, 0.295)	0.325 (0.292, 0.357)	0.273 (0.254, 0.293)	0.278 (0.248, 0.309)
	2	0.287 (0.267, 0.307)	0.313 (0.291, 0.335)	0.289 (0.269, 0.309)	0.288 (0.267, 0.308)
	3	0.302 (0.282, 0.322)	0.310 (0.281, 0.339)	0.299 (0.279, 0.319)	0.288 (0.261, 0.315)
	4	0.297 (0.277, 0.318)	0.582 (0.526, 0.638)	0.303 (0.283, 0.324)	0.291 (0.261, 0.321)
Trimester	Exposure quartile	Pollutant: Ozone			
		Exposure period definition: Option 1		Exposure period definition: Option 2	
		Probability of stillbirth (%)		Probability of stillbirth (%)	
		Unadjusted (95% CI)	Adjusted (95% CI)	Unadjusted (95% CI)	Adjusted (95% CI)
3	1	0.297 (0.284, 0.310)	0.263 (0.249, 0.276)	0.282 (0.270, 0.295)	0.263 (0.249, 0.276)
	2	0.265 (0.253, 0.278)	0.269 (0.256, 0.282)	0.277 (0.264, 0.290)	0.283 (0.269, 0.297)
	3	0.265 (0.253, 0.277)	0.287 (0.271, 0.304)	0.283 (0.270, 0.296)	0.293 (0.277, 0.308)
	4	0.301 (0.288, 0.315)	0.338 (0.308, 0.369)	0.286 (0.273, 0.299)	0.257 (0.239, 0.275)

Table 3.13: Odds Ratios for stillbirth from nitrogen dioxide or ozone exposures, obtained using a traditional logistic regression analysis. Fetuses that survived until the third trimester (till the 27th week) were included in the analysis. Here option 1 indicates that the event date (date of stillbirth or live birth) determines the exposure period and option 2 indicates that exposure was assessed over a fixed gestational period, irrespective of a mother's observed gestational age based on her event date.

Trimester	Exposure quartile	Pollutant: Nitrogen dioxide	
		Exposure period definition: Option 1	Exposure period definition: Option 2
		Odds Ratio (95% CI)	Odds Ratio (95% CI)
3	1	1.000 (Reference)	1.000 (Reference)
	2	1.001 (0.900, 1.114)	1.006 (0.904, 1.120)
	3	1.063 (0.941, 1.200)	1.051 (0.930, 1.187)
	4	0.998 (0.862, 1.156)	1.053 (0.909, 1.221)
Trimester	Exposure quartile	Pollutant: Ozone	
		Exposure period definition: Option 1	Exposure period definition: Option 2
		Odds Ratio (95% CI)	Odds Ratio (95% CI)
3	1	1.000 (Reference)	1.000 (Reference)
	2	0.950 (0.890, 1.014)	1.046 (0.980, 1.117)
	3	0.960 (0.897, 1.027)	1.078 (1.008, 1.153)
	4	1.047 (0.978, 1.121)	1.035 (0.965, 1.109)

Chapter 4

Assessing the causal impact of prenatal traffic exposure on preterm birth

4.1 Introduction

Increasing air pollution is a threat to public health worldwide. According to the State of the Air 2016 report by the American lung association, around 52.1% of the population in the US live in counties where air pollution levels, dominated by ozone or particle, are considered dangerous to health [2]. Traffic is a major source of air pollution and traffic related health issues are a major concern in today's society [7]. Many epidemiological studies have analyzed the adverse effects of traffic pollution on various health issues [124, 36, 37, 125] that includes many adverse pregnancy outcomes [126, 127, 128, 110]. Preterm birth (PTB) is a major issue in perinatology; for it poses subsequent health and developmental challenges on the fetus as well as can cause economic burden to the family [61, 62, 63]. In 2016, preterm birth rate in United States was around 10% [68].

Prior studies reported a positive [83, 26, 27, 80, 84, 81, 82] or a null association [85, 86, 87] between traffic-related air pollution and preterm birth; however, there are some limitation to these studies. First, parametric regression models were used to assess the relationship between traffic-related air pollution and measured confounders. Utilizing misspecified regression models to adjust for measured confounders can result in biased estimates and in incomplete control for confounding. Second, in prior studies, the relationship between traffic-related air pollution and preterm birth was assessed based on conditional odds; this parameter fails to quantify the preterm risk burden from traffic-related air pollution in the study population of interest, and it does not inform us about the preterm risk burden change associated with a modified exposure level (for example; a new policy or intervention that could modify the traffic pollution standards). In order to address these limitations, we decided to estimate a causal attributable risk of preterm birth from prenatal traffic pollution exposure

in a population that has been previously studied using parametric regression model [27]. Specifically, we decided to apply targeted maximum likelihood estimation (TMLE) [69], a semi-parametric efficient estimation approach, to estimate the causal attributable risk of preterm birth. When implementing TMLE we incorporated SuperLearner, a flexible data adaptive algorithm [72], which can improve both the robustness and precision of our estimates and ensure meaningful statistical inference. Applying machine learning algorithms on big data sets, that are similar to our birth cohort data, can be computationally challenging. Hence we also illustrate the use of another class of TMLE, known as the TMLE for two stage design [70], which allows us to work with a subsample of the cohort to estimate the causal attributable risk of preterm birth. The causal attributable risk of preterm birth, under assumptions, compares the absolute change in preterm birth risk that would have been experienced by our study population under a hypothetical intervention to change traffic related pollution levels comparative to their currently observed levels. To our knowledge, this is the first epidemiological study that estimated the causal attributable risk of preterm birth from prenatal traffic pollution exposure using a semi-parametric targeted method incorporating machine learning.

4.2 Methods

Study population and outcome of interest

Our study population included all the live births that occurred in the four most populated counties in the San Joaquin Valley of California (Fresno, Kern, Stanislaus, and San Joaquin) between 2000 to 2006. They were identified from the birth certificate data provided by the California Department of Public Health. Preterm birth was our outcome of interest, and it was defined as birth that occurred before 37 weeks of gestational age. Our data included only those pregnancies, that resulted in live births, between 20 to 42 week of gestational age; information about births that occurred before 20 weeks were not available to us.

Exposure of interest

Traffic density was our exposure of interest; it is a dimensionless quantity that summarize the traffic activity in proximity of a maternal residence location. As a first step to calculate the exposure of interest, using ArcGIS software (ESRI, Redlands, California) maternal residence locations (obtained from the birth certificate) were geocoded and ZP4 software (Semaphore Corporation, Aptos, California) was used to correct the residential addresses. Using the traffic count data received from Tele Atlas/Geographic Data Technology (GDT) in 2005, we applied a distance-decayed annual average daily traffic (AADT) volume to estimate the required traffic density. Detailed exposure assignment description was provided in a prior study that utilized the same exposure data [27]. Prenatal traffic density exposure during the entire pregnancy (i.e. from conception to birth) was first calculated as a continuous value

and then categorized based on quartiles. No temporal trend was associated with the traffic density exposure.

Potential confounders

From the available population information obtained from the California Department of Public Health and Census 2000, three categories of potential confounders were identified: individual factors, temporal factors and community or neighborhood factors. Individual factors included various maternal factors (county of residence, age, education and race/ethnicity), info on whether prenatal care started in first trimester, method of payment for delivery, parity indicator and sex of the infant. The method of payment for delivery information represents whether Medi-Cal (Medicaid) or other government program paid the birth costs. Temporal factors included season of conception and year of birth. Using the extracted Census variables, we created an indicator with the following characteristics was created: unemployment rate $> 10\%$, income from public assistance $> 15\%$ and families below the federal poverty level $> 20\%$ [27]; this represented the neighborhood or community factors.

Our study was approved by the Office for Protection of Human Subjects, University of California Berkeley and the California State Committee for the Protection of Human Subjects.

Statistical method

Parameters of interest

The aim of our study was to quantify the impact of traffic exposure on preterm birth during the entire pregnancy and for this we estimate the causal attributable risk (CAR), which compares the outcome distribution under a hypothetical intervention on a population with the outcome distribution observed [97]. Our causal parameter of interest was based on the Neyman-Rubin counterfactual framework [95, 96] and it is defined as

$$E(Y) - E(Y_a); a \in \mathcal{A} = \{1, 2, 3, 4\} \tag{4.1}$$

Here the observed birth outcome, which is equal to 1 if a preterm birth occurred, and the counterfactual birth outcome a mother would have had if she had been exposed to exposure quartile $A = a \in \mathcal{A}$ are represented as Y and Y_a respectively. The entries in $\mathcal{A}$ represent a particular level/quartile of traffic density exposure, with 1 representing the lowest level of exposure and 4 representing the highest.

Following assumptions are required to identify the causal parameter of interest from the observed data distribution [71]. First, under consistency assumption we assume that the the observed outcome (Y) a mother experienced under the observed exposure level is equal to the counterfactual outcome (Y_A) under the observed exposure level. Second, under the randomization assumption, we assume that given our set of baseline covariates (which we denote as W), the potential outcomes Y_a is independent of the exposure A; i.e. $Y_a \perp\!\!\!\perp A|W, \forall$

$a \in \mathcal{A}$. Randomization assumption is also known as no unmeasured confounding assumption, for under this assumption, we assume that measured covariates W are sufficient to control for confounding of the effect of A on Y. Finally, under the positivity assumption, we assume that there is a positive probability for receiving each quartile of exposure A within every combination of baseline covariate values among the population; i.e. $\min_{a \in \mathcal{A}} P(A = a|W) > 0$. Under these three assumptions, the causal parameter of interest can be rewritten as a function of observed data distribution alone [97, 99]

$$E(Y) - E_W[E(Y|A = a, W)]; a \in \mathcal{A} = \{1, 2, 3, 4\} \tag{4.2}$$

This represents the statistical estimand. Here, note that the first term in the above statistical estimand, $E(Y)$, represents the empirical mean of the outcome Y and in the subsequent steps, we will refer the second term of the above estimand, $E_W[E(Y|A = a, W)]$, as the *counterfactual part of the target parameter*.

Estimation of parameters of interest

There are various methods to estimate the counterfactual part of our target parameter. The G-computation method (or simple substitution method) make use of the outcome regression estimate, $E(Y|A, W)$ [99, 100], and the inverse probability of treatment weighted (IPTW) method uses a treatment regression estimate (also known as the propensity score), $P(A|W)$ [101]. The G-computation and IPTW estimators are consistent if the outcome regression or the treatment regression, respectively, are consistently estimated. In our analysis, we utilize another class of estimation method called targeted maximal likelihood estimators (TMLE) [69], that combine estimators of both outcome and treatment regressions and is double robust.

Estimation of the parameter using the entire final study population

TMLE for the counterfactual term of our target parameter, $E_W[E(Y|A = a, W)]$, is a two step procedure; the initial estimates of both outcome and treatment regressions are obtained in the first step and in the second bias-reduction step the initial estimate of the outcome regression is updated towards the parameter of interest utilizing the initial estimate of treatment regression. TMLE is robust to misspecification of either the outcome regression or the treatment regression (the so-called "double robustness" property) and it will be efficient, with lowest asymptotic variance among reasonable estimators, if both regression models are correctly identified. We estimated both the outcome and treatment regressions using Super-Learner [72], a data-adaptive algorithm, that based on cross-validation creates an optimal combination of fits obtained from user supplied individual prediction algorithms.

Under assumptions, TMLE is asymptotically linear, with its variance explained by the sample variance of estimated influence curve of the target parameter divided by sample size. The influence curve for the causal attributable risk can be derived using the Delta method

[71] and it is explicitly

$$Y - \frac{I(A=a)}{P(A=a|W)}\big(Y - Q_0(A,W)\big) + Q_0(A=a,W) - \Psi(Q_0) \tag{4.3}$$

where $Q_0(A,W)$ denotes the true conditional expectation of the outcome.

The 95% confidence interval for our target parameter is thus estimated as

$$\hat{\psi}_n \pm 1.96\frac{\sqrt{\hat{\sigma}^2}}{\sqrt{N}} \tag{4.4}$$

where $\hat{\psi}_n$ is the estimated causal attributable risk, $\hat{\sigma}^2$ is the sample variance of the influence curve estimate and N is the total number of observations in our final study population.

In our analysis to estimate the outcome and treatment regressions using SuperLearner, we used a non-negative least squares loss function, a 10-fold cross validation and our library of candidate algorithms included main term logistic regression, logistic regression with all possible pairwise interactions, simple mean and the stepwise logistic regression.

Estimation of the parameter using a sub-sample of the final study population

Incorporating machine-learning algorithms, like SuperLearner, on large birth cohort data sets can be computationally intensive. Hence, we also illustrate the use of another class of TMLE, known as an inverse probability of censoring weighted targeted maximum likelihood estimator (IPCW-TMLE) [70] that utilizes a sub-sample of the final study population to estimate the target parameter. In IPCW-TMLE, a sub-sample of the final study cohort was created conditional on outcome status. If Δ is an indicator about the sub-sample inclusion, with $\Delta = 1$ if an observation is included in the sub-sample, we can represent the sub-sample data structure as $O^R = (Y, \Delta, \Delta X)$, where X represents the original data set $X = (W, A, Y)$. Conditional on an outcome Y, each observation had a known sampling probability, denoted by $\Pi_0(Y)$; $\Pi_0(Y)$ is equal to one if the outcome was a preterm birth and otherwise equal to the proportion of term births sampled. IPCW-TMLE is also a two-step procedure which involve initial estimation of the outcome and treatment regressions using SuperLearner as well as the targeted updating step, both implemented using the inverse of the sampling weights; specifically, the sampling weight is equal to $\frac{\Delta}{\Pi_0(Y)}$. Since, IPCW-TMLE is also asymptotically linear, under assumptions, the standard error of the IPCW-TMLE estimator can be conservatively estimated by taking the sample variance of the weighted influence curve estimate, provided in Eq. 4.3, and dividing by sample size.

To implement IPCW-TMLE, first we created a sub-sample of our original data, that included all the preterm births and randomly chose 5 term births per preterm birth. The sampling probability for preterm births and term births were equal to one and proportion of term births included in the subsample in the second stage respectively. The loss function, number of folds and the candidate algorithms included in SuperLearner library were same as before; but appropriate sampling weight was assigned to each observations.

In order to compare results qualitatively with that of prior research, we also conducted a traditional main-term logistic regression analysis including all the confounders, that we adjusted for when estimating CAR, as well as traffic exposure quartiles as a four level categorical variable. Our analyses were conducted using R version 3.1.2 and the random seed was set to 1 while randomly sampling the term births.

4.3 Results

All live births that occurred between 2000 to 2006, in four most populated counties within the San Joaquin Valley air basin, were included in our original cohort and it included 329,650 live births. We excluded mothers with a missing file number (N=248) and who had multiple births (N=8,373). We also excluded those fetuses whose gestational age was either missing or not between 20 to 42 weeks (N=44,713) as well as those fetuses with birth weight missing or < 500 or > 5000 grams (n=764). 1,025 live births were removed because of the lack of a last menstrual period (LMP) date. We excluded those subjects with a missing exposure value (N=12,345) or a confounder value (N=6,134). Exclusion criteria applied to our study cohort can be found in Figure 4.1. Our final study population included 256,048 live births, whose descriptive statistics by gestational age can be found in Table 4.1 through Table 4.4.

Our data set included 11.5% preterm births and 88.5% term births; with 8.5% preterm births occurring between 34 to 36 weeks of gestational age. Majority of our study population included male fetuses (51.3%) and around 35% of the fetuses were the first born child. Higher percentage of conceptions occurred in winter (25.7%) and fall (25.6%). More than half of the mothers were Hispanic and had a high school degree. Most of the mothers received prenatal care starting in the first trimester (81.3%), with majority of them receiving government aid to pay the delivery related costs (53.7%) and about 25% mothers had C-section. The mean, 5th and 95th percentiles of our exposure of interest were also provided in Table 4.4; and from this it is clear that the provided summary statistics were higher for the preterm births in comparison with term births. In Table 4.5, we have also provided quartile specific traffic density values.

The propensity score, while estimating the causal attributable risk of preterm birth using both TMLE and IPCWP-TMLE are given in Table 4.6. We can see that the predicted propensity score are well bounded and away from zero in each exposure quartiles, which implies that there is no practical positivity violation observed in our analysis [118]. Within each quartile, the estimate of $E(Y|A = a)$ is obtained using the quartile specific empirical mean of preterm birth; we refer this an *unadjusted* estimator. Similarly, for each traffic exposure quartile $A = a$, the counterfactual estimate $E(Y_a)$ is obtained using both TMLE and IPCW-TMLE and we refer this an *adjusted* estimator. Both unadjusted and adjusted estimators are provided in Table 4.7. There is an increasing trend in the unadjusted probability of preterm birth, with the lowest probability of preterm birth in quartile 1. When we adjusted for confounders the resulted counterfactual probabilities in the first two exposure quartiles were higher and that in the 3^{rd} and 4^{th} quartiles were lower in comparison

with their counterpart unadjusted preterm birth probabilities; however, except within the third exposure quartile, the increasing trend in the preterm probability remains as before. The adjusted counterfactual probabilities of preterm birth were similar when we estimated them using either the whole study population applying TMLE or a sub-sample of the study population using IPCW-TMLE.

Estimated causal attributable risk using both TMLE and IPCW-TMLe are given in Figure 4.2. Here the estimated CAR is the difference between the unadjusted estimate of the empirical probability of preterm birth in the whole data and the adjusted estimate of the counterfactual probability of preterm birth within each exposure quartile. If an adjusted counterfactual probability estimate is higher than that of the unadjusted empirical probability, it will result in a negative causal attributable risk. A positive CAR value indicates a protective effect and the more negative the estimated CAR value, the higher harmful effect on mothers exposed to that particular exposure quartile. Except within exposure quartile 3, we observed a trend in the estimated CAR values with an increasing preterm risk as we move from lower to higher exposure quartiles, but the significant CAR estimates were observed only in first and fourth exposure quartiles and the effect estimates using both TMLE and IPCW-TMLE were similar. Our CAR estimates using TMLE suggest that shifting all pregnant women's exposure to the first quartile would decrease preterm birth by 0.295% (95% CI: 0.057%, 0.532%) and increasing it to the fourth quartile would increase preterm birth by 0.276% (95% CI: 0.051%, 0.500%). When utilizing IPCW-TMLE, the CAR estimates in first and fourth exposure quartiles were 0.292% (95% CI: 0.055%, 0.528%) and -0.246% (95% CI: -0.469%, -0.022%) respectively.

Conditional odds ratio of preterm birth, obtained from a traditional main-term logistic regression analysis, with the same confounders that we adjusted for when estimating CAR and traffic exposure quartiles, is given in Table 4.8. This statistical parameter is different from CAR, even when the parametric model is correctly specified and randomization assumptions holds. However, Table 4.8 results can be utilized for a qualitative comparison with prior research. Parametric logistic regression results suggest a harmful effect of exposure to both the 2nd and 4th traffic exposure quartiles; however the TMLE or IPCW-TMLE results found harmful effect in the 4th exposure quartile only. Even though we utilized the same data, the results shown in Table 4.8 are slightly different from that of prior research [27] and this could be because of the following reasons: a) in our current analysis we included a slightly different baseline covariates (we included season of conception and an indicator representing community factors, but excluded birth weight), and b) we report the conditional odds ratio of preterm birth in the higher exposure quartiles relative to that of lowest exposure quartile. In our current analysis, even though both conditional odds ratio and CAR estimates led to a conclusion of a detrimental effect from exposure to the highest quartile, note that CAR provides a quantitative estimate of a quantity with more immediate public health and policy relevance.

4.4 Discussion

Incorporating data-adaptive algorithm, we estimated the causal attributable risk of preterm birth from prenatal traffic exposure in four most populated counties in San Joaquin Valley air basin in California in a semi-parametric targeted way and our results suggests that exposing pregnant woman to the highest traffic exposure is harmful. We observed preterm birth risk can increase relative to the current observed level by around 0.25% if all the pregnant women are exposed to the highest exposure quartile.

As mentioned in the introduction, prior studies, based on parametric models, have analyzed the impact of prenatal traffic exposure on preterm birth and results were mixed. Because of the difference in the estimated parameters, it is not possible to make a direct comparison between our current study results with that of prior studies. In previous studies, preterm risk difference associated with different traffic exposure levels while holding the baseline covariates as constant, was given in terms of odds ratio. In our analysis, by applying a hypothetical intervention traffic exposure level, a counterfactual probability of preterm birth risk was compared with that was currently observed in our study population. The confounders that we adjusted in our analysis were also different from the prior studies and this might have reflected in the parameter estimates that we reported. Because of these dissimilarities, even though a direct comparison is not possible, our analysis where we were able to establish a causal link between prenatal traffic density exposure and preterm birth supports the prior studies that established an association between them.

Several plausible biologic pathways have been proposed through which ozone and traffic related air pollutants, such as nitrogen oxides (NO_X), particulate matter (PM), carbon monoxide (CO) or carbon dioxide (CO_2), could adversely affect various birth outcomes [43, 44, 45]. Air pollution can increase vulnerability to infection, which can trigger preterm delivery [44, 48, 49]. Traffic related air pollutants were associated with oxidative stress and inflammation [43, 44, 53] and both of these can influence the pregnancy duration. Air pollutants can also change the blood coagulability and plasma viscosity; this influences the umbilical blood flow and placental oxygen transfer adversely affecting the fetal growth [46, 47].

There are some limitations to our study. Randomization assumption was one of the major assumption that we made while estimating the causal attributable risk of preterm birth. Despite making this assumption, we acknowledge that our data does not include information on various maternal factors (e.g. smoking, individual SES, individual perinatal risk factors) and neighborhood characteristics (e.g. access to nutritious food and other amenities to support a healthy living, safety features), that could potentially determine both exposure and birth outcome [129, 121, 122, 123]; hence in our analysis this untestable assumption maybe unlikely to hold. Second limitation is the possibility of exposure misclassification in the assigned exposure for each mothers due to the lack of information about their time spend outdoors or at the assigned zip code etc and these factors could alter the actual exposure experienced. However, geocoding the maternal residential street addresses while assigning

exposure might have helped to overcome this issue to some extend.

Our analysis has several strengths. We used a diverse population with large sample size. Since maternal addresses were geocoded our assigned exposure may have been more accurate than the exposure estimate obtained from simply assigning a mother to a nearest monitoring station based on her residential zip code. Addition to various individual and temporal factors, we included few neighborhood level confounders; this is important as a neighborhood environment can influence birth outcome [121, 122, 123]. Evaluating CAR of various health issues from air pollution exposure is relevant for evaluating the impact of new policies related to air pollution levels; this can help authorities for outlining proper prevention strategies to help those population who will be affected by the proposed new standards. Also, quantifying the health effect of potential intervention levels and establishing a causal relationship between air pollution and health outcomes are more convincing than a mere association in policy relevant issues [73, 74]. In our analysis, to estimate our target parameter, we utilized both TMLE and IPCW-TMLE that used either the entire study population or a sub-sample of it respectively. As mentioned earlier, both methods are doubly robust to model misspecification. Also, by using only a sub-sample of data for analysis, IPCW-TMLE helps to overcome the computational challenges of big data analysis that can be common when analyzing birth registry data. Also, incorporating a data-adaptive algorithm in our our analysis, to estimate both the outcome and treatment regressions, helps to remove parametric misspecification bias and to ensure accurate statistical inference for TMLE methods [72, 71].

In our future research, we are interested in estimating the causal attributable risk of various subsets of preterm birth, specifically extremely PTB with gestational age between 20-27 weeks, very PTB (VPTB) with gestational age between 28-31 weeks, moderate PTB (MPTB) with gestational age between 32-34 weeks and late PTB (LPTB) with gestational age of 35-36 weeks. Assessing the risk within various subsets of preterm birth is important as many changes happen during human fetal development in short period of time.

To conclude, we calculated the causal attributable risk of preterm birth from prenatal traffic exposure using cohort data from San Joaquin Valley air basin in California. Our study is relevant for it helps to quantify the burden of preterm birth from traffic pollution and it can easily extend to any relevant hypothetical pollution exposure level and thus could inform the authorities to make necessary prevention strategies to alleviate the risks associated with a specific exposure pollution level. This is important as air pollution from traffic is a serious global health concern and based on our results, we recommend further studies in other locations.

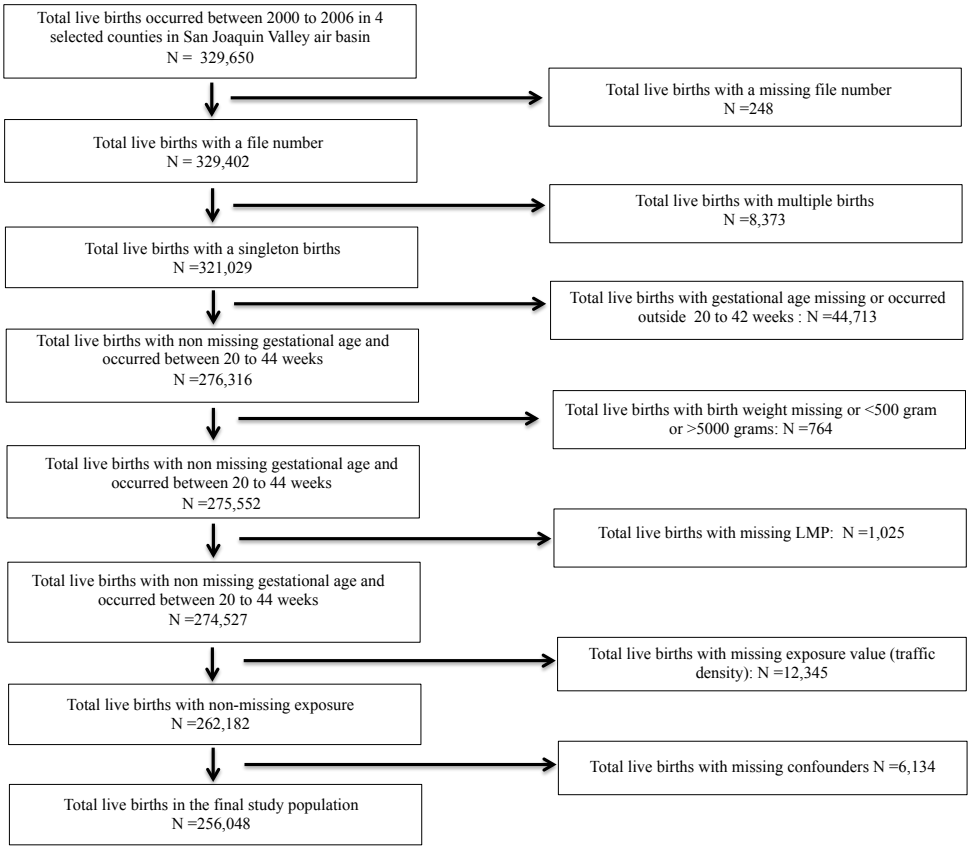

Figure 4.1: Flow chart showing the exclusion criteria applied to create our final study population

Table 4.1: Characteristics of our study population included in the analysis (Part 1). Our final study population included 256,048 live births that occurred in four most populated counties within San Joaquin Valley air basin in California between 2000 to 2006.

	Gestational age (weeks)						
	37 - 42	34 - 36	32 - 33	28 - 31	24 - 27	20 - 23	Total
N (%)	226,689 (88.5)	21,580 (8.5)	3,794 (1.5)	2,628 (1.0)	1,037 (0.4)	320 (0.1)	256,048 (100)
Characteristics	%						
First born	35.3	32.9	34.3	35.8	41.9	38.1	35.1
Male	50.9	54.0	54.6	55.2	56.3	59.4	51.3
C-Section	24.8	28.4	34.8	42.4	42.3	30.3	25.5
Prenatal care in 1st trimester	82.0	78.2	73.9	67.8	69.9	59.7	81.3
Low SES	17.3	20.2	23.3	22.6	21.7	20.6	17.7
Medi-Cal payment costs	52.8	59.7	63.7	63.4	58.8	62.5	53.7
Birth year							
2000	13.2	12.5	12.6	13.1	15.6	15.6	13.1
2001	13.4	12.6	11.8	13.2	12.5	11.6	13.3
2002	13.7	13.0	13.3	12.3	12.1	13.1	13.6
2003	13.8	13.8	13.8	13.3	13.5	14.1	13.8
2004	14.4	14.6	14.7	14.2	14.7	16.6	14.4
2005	15.1	15.9	15.3	16.7	15.0	14.4	15.2
2006	16.4	17.5	18.5	17.3	16.6	14.7	16.6

Table 4.2: Characteristics of our study population included in the analysis (Part 2). Our final study population included 256,048 live births that occurred in four most populated counties within San Joaquin Valley air basin in California between 2000 to 2006.

	Gestational age (weeks)						Total
	37 - 42	34 - 36	32 - 33	28 - 31	24 - 27	20 - 23	
N	226,689	21,580	3,794	2,628	1,037	320	256,048
(%)	(88.5)	(8.5)	(1.5)	(1.0)	(0.4)	(0.1)	(100)
Characteristics	%						
Maternal age (years)							
< 20	13.3	15.2	16.9	20.2	19.7	23.4	13.6
20-24	28.9	28.7	28.6	28.2	26.7	26.2	28.9
25-29	27.7	25.5	23.3	22.8	21.6	21.6	27.3
30-34	19.3	18.2	18.1	15.4	19.0	17.2	19.2
>34	10.8	12.4	13.2	13.3	13.0	12.5	11.0
Maternal race/ethnicity							
Non-Hispanic white	30.7	26.0	23.0	24.2	23.7	21.9	30.1
African-American	4.9	6.7	8.0	9.1	9.5	8.1	5.1
Hispanic	55.7	57.3	59.2	56.1	57.5	61.2	55.9
Asian	7.4	8.7	8.4	9.3	8.3	5.9	7.6
Other	1.3	1.4	1.3	1.3	1.1	2.8	1.3
Maternal education							
< High school	12.1	12.5	13.8	11.6	11.3	11.2	12.1
High school	53.2	57.7	59.4	63.1	60.6	61.9	53.8
Some college	21.5	19.8	18.5	17.7	20.2	17.8	21.2
College degree	13.2	10.0	8.3	7.7	8.0	9.1	12.8

Table 4.3: Characteristics of our study population included in the analysis (Part 3). Our final study population included 256,048 live births that occurred in four most populated counties within San Joaquin Valley air basin in California between 2000 to 2006.

	Gestational age (weeks)						
	37 - 42	34 - 36	32 - 33	28 - 31	24 - 27	20 - 23	Total
N	226,689	21,580	3,794	2,628	1,037	320	256,048
(%)	(88.5)	(8.5)	(1.5)	(1.0)	(0.4)	(0.1)	(100)
Characteristics	%						
Maternal county of residence							
Fresno	32.7	35.2	34.9	33.1	37.7	35.0	33.0
Kern	23.2	24.8	26.8	25.4	23.3	26.9	23.5
San Joaquin	25.4	22.9	21.1	22.9	22.4	20.6	25.1
Stanislaus	18.7	17.0	17.1	18.5	16.6	17.5	18.5
Season of conception							
Winter (Dec. - Feb.)	25.9	24.1	23.5	22.9	22.4	20.9	25.7
Spring (Mar. - May)	24.8	25.3	25.7	25.2	26.0	27.5	24.9
Summer (Jun. - Aug.)	23.8	24.2	25.4	25.3	25.4	23.8	23.9
Fall (Sept. - Nov.)	25.4	26.5	25.4	26.6	26.2	27.8	25.6
Pregnancy Complications							
Diabetes	1.6	2.3	2.2	1.6	1.3	0.9	0.2
Hypertension	0.2	0.3	0.4	0.6	0.5	0.0	1.7

Table 4.4: Characteristics of our study population included in the analysis (Part 4). Our final study population included 256,048 live births that occurred in four most populated counties within San Joaquin Valley air basin in California between 2000 to 2006.

	Gestational age (weeks)						
	37 - 42	34 - 36	32 - 33	28 - 31	24 - 27	20 - 23	Total
N (%)	226,689 (88.5)	21,580 (8.5)	3,794 (1.5)	2,628 (1.0)	1,037 (0.4)	320 (0.1)	256,048 (100)
Characteristics	Mean (SD)						
Infant birth weight (gram)	3,426 (466)	3,000 (581)	2,649 (762)	2,207 (954)	1,612 (1,113)	1,779 (1,287)	3,357 (545)
Exposure of interest	Mean (5th percentile, 95th percentile)						
Traffic density	34.94 (0.0, 138.82)	37.35 (0.0, 146.59)	38.73 (0.0, 147.16)	38.14 (0.0, 150.43)	39.31 (0.0, 154.87)	35.38 (0.0, 123.25)	35.25 (0.0, 140.14)

Table 4.5: Quartile specific traffic density - a dimensionless quantity- in the final study population. Our final study population included 256,048 live births that occurred in four most populated counties within San Joaquin Valley air basin in California between 2000 to 2006.

Exposure quartile	Number of subjects	Minimum	Median	Inter-quartile range (IQR)	Maximum
1	64,011	0	0	0	0.62
2	64,012	0.62	6.88	8.21	16.49
3	64,012	16.49	28.18	14.22	45.83
4	64,013	45.83	81.69	64.59	554.47

Table 4.6: Propensity score while estimating the counterfactual probability of preterm birth from prenatal traffic exposure. We estimated the causal attributable risk of preterm birth using both TMLE and IPCW-TMLE techniques. Our final study population included 256,048 live births that occurred in four most populated counties within San Joaquin Valley air basin in California between 2000 to 2006. TMLE analysis included all the subjects in this population (N = 256,048); however in the IPCW-TMLE analysis, we used a sub-sample of this cohort that consists of all the preterm births (N = 29,359) and five term births per preterm birth.

Propensity score, when estimating the causal attributable risk of preterm birth using TMLE						
No. of subjects included in analysis	Exposure quartile	Minimum	First Quartile	Median	Third Quartile	Maximum
256,048	1	0.06497	0.21759	0.25017	0.28649	0.57556
	2	0.15150	0.20830	0.24540	0.28060	0.42290
	3	0.17030	0.22570	0.24370	0.27280	0.35940
	4	0.07968	0.19714	0.23741	0.27727	0.59966
Propensity score, when estimating the causal attributable risk of preterm birth using IPCW-TMLE						
No. of subjects included in analysis	Exposure quartile	Minimum	First Quartile	Median	Third Quartile	Maximum
176,154	1	0.06865	0.21692	0.24889	0.28497	0.57363
	2	0.14790	0.20810	0.24410	0.28040	0.4162
	3	0.17350	0.22620	0.24430	0.27180	0.36620
	4	0.07721	0.19934	0.23854	0.27708	0.58896

Table 4.7: Unadjusted (empirical probability) and adjusted (counterfactual probability) of preterm birth (given in %) within each prenatal traffic exposure quartile during the entire pregnancy period. We estimated the adjusted counterfactual probability of preterm birth using both TMLE and IPCW-TMLE techniques. Our final study population included 256,048 live births that occurred in four most populated counties within San Joaquin Valley air basin in California between 2000 to 2006. TMLE analysis included all the subjects in this population (N = 256,048); however in the IPCW-TMLE analysis, we used a sub-sample of this cohort that consists of all the preterm births (N = 29,359) and five term births per preterm birth.

Exposure quartile	Probability of preterm birth (%)		
	Unadjusted (95% CI)	Adjusted (95% CI) using TMLE	Adjusted (95% CI) using IPCW-TMLE
1	10.73 (10.61, 10.85)	11.17 (10.91, 11.44)	11.17 (10.91, 11.44)
2	11.45 (11.33, 11.57)	11.56 (11.31, 11.81)	11.57 (11.32, 11.83)
3	11.47 (11.35, 11.60)	11.45 (11.20, 11.70)	11.45 (11.20 11.69)
4	12.21 (12.09, 12.34)	11.74 (11.49, 12.00)	11.71 (11.46, 11.97)

Table 4.8: Odds Ratios for preterm birth from prenatal traffic exposure during the entire pregnancy period, obtained using a traditional logistic regression analysis. Our final study population included 256,048 live births that occurred in four most populated counties within San Joaquin Valley air basin in California between 2000 to 2006.

Exposure quartile	Odds Ratio (95% confidence interval)
1	1.000 (Reference)
2	1.041 (1.004, 1.080)
3	1.034 (0.997, 1.073)
4	1.064 (1.026, 1.103)

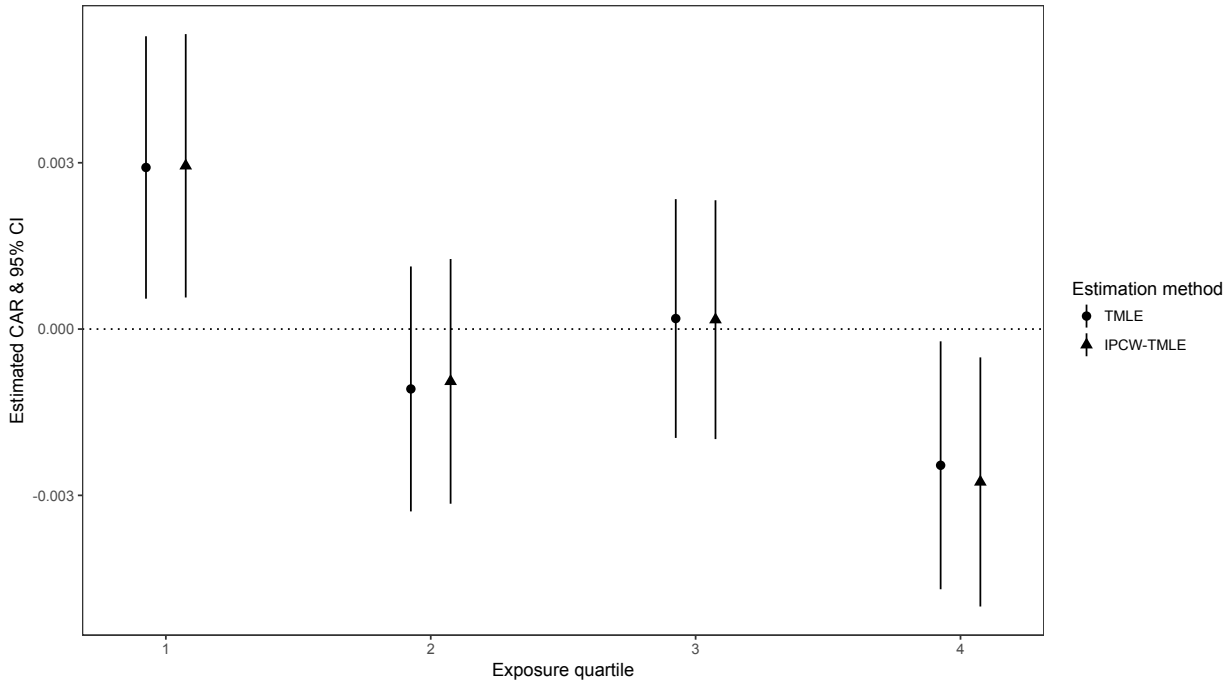

Figure 4.2: Causal attributable risk of preterm birth from prenatal traffic exposure during the entire pregnancy, estimated using both TMLE and IPCW-TMLE. Our final study population included 256,048 live births that occurred in four most populated counties within San Joaquin Valley air basin in California between 2000 to 2006. TMLE analysis included all the subjects in this population (N = 256,048); however in the IPCW-TMLE analysis, we used a sub-sample of this cohort that consists of all the preterm births (N = 29,359) and five term births per preterm birth.

Chapter 5

Conclusions

As a summary, we present the main conclusions drawn from the studies presented in this thesis.

1. We demonstrated the utility of a data-adaptive algorithm and various semi-parametric efficient approaches to estimate causal attributable risk (CAR) of adverse pregnancy outcomes.

2. Based on the CAR estimates of stillbirth, our study suggests that exposure to highest levels of both NO_2 and O_3 during the third trimester of pregnancy is harmful but only when exposure period is assigned based on the date of birth outcome, an approach subject to temporal confounding. Hence, in this analysis, we also address a method for assigning exposure to deal with the challenges of temporal trends in pollution as well as in conception and our results imply the causal estimates are affected by the exposure period definition applied.

3. In the study assessing the relationship between traffic pollution and preterm birth, our estimates suggest that exposing all pregnant women to the lowest exposure level would decrease the preterm birth risk by around 0.30% and exposing them to the highest level would increase the preterm birth risk by around 0.28%.

4. Both studies could be improved by utilizing information on additional neighborhood or community level characteristics as well as individual and temporal factors that could influence both exposure and outcomes.

5. The causal attributable risk (CAR) of adverse pregnancy outcomes has immediate public health impact and importance. We recommend researchers to follow the analysis presented in this thesis when analyzing air pollution related adverse pregnancy outcomes.

Bibliography

[1] World Health Organization. Ambient air pollution: A global assessment of exposure and burden of disease. http://apps.who.int/iris/bitstream/10665/250141/1/9789241511353-eng.pdf, 2016.

[2] State of the Air 2016 — American Lung Association. http://www.lung.org/assets/documents/healthy-air/state-of-the-air/sota-2016-full.pdf, 2016.

[3] Air Emissions Sources. US EPA. https://www.epa.gov/air-emissions-inventories/air-emissions-sources.

[4] Sources of Air Pollution. National park service. https://www.nature.nps.gov/air/aqbasics/sources.cfm.

[5] Clean Air Act. https://www.epa.gov/clean-air-act-overview, 1956.

[6] Air Contaminants: Traffic Pollutants. California Environmental Health Tracking Program. http://cehtp.org/faq/air/air_contaminants_traffic_pollutants.

[7] Panel on the Health Effects of Traffic-Related Air Pollution. Traffic-related air pollution: a critical review of the literature on emissions, exposure, and health effects. (17), 2010.

[8] Air pollution and health risk. US EPA. https://www3.epa.gov/airtoxics/3_90_022.html.

[9] Robert D. Brook, Barry Franklin, Wayne Cascio, Yuling Hong, George Howard, Michael Lipsett, Russell Luepker, Murray Mittleman, Jonathan Samet, Sidney C. Smith, and Ira Tager. Air pollution and cardiovascular disease: A statement for healthcare professionals from the expert panel on population and prevention science of the American Heart Association. *Circulation*, 109(21):2655–2671, 2004.

[10] Nicholas L Mills, Ken Donaldson, Paddy W Hadoke, Nicholas A Boon, William MacNee, Flemming R Cassee, T Sandstrom, Anders Blomberg, David E Newby, and Thomas Sandström. Adverse cardiovascular effects of air pollution. *Nat Clin Pract Cardiovasc Med*, 6(1):36–44, 2009.

[11] Radim J. Šrám, Blanka Binková, Jan Dejmek, and Martin Bobak. Ambient air pollution and pregnancy outcomes: A review of the literature. *Environmental Health Perspectives*, 113(4):375–382, 2005.

[12] D M Stieb, L Chen, M Eshoul, and S Judek. Ambient air pollution, birth weight and preterm birth: A systematic review and meta-analysis. *Environmental Research*, 117:100–111, 2012.

[13] Nazeeba Siddika, Hamudat A Balogun, Adeladza K Amegah, and Jouni J K Jaakkola. Prenatal ambient air pollution exposure and the risk of stillbirth: systematic review and meta-analysis of the empirical evidence. *Occupational and environmental medicine*, 73(9):573–581, 2016.

[14] World Health Organization. Regional Office for Europe. and European Centre for Environment and Health. Effects of air pollution on children's health and development : a review of the evidence. page 185 p., 2005.

[15] Lilian Calderón-Garcidueñas, Ricardo Torres-Jardón, Randy J Kulesza, Su-Bin Park, and Amedeo D'Angiulli. Air pollution and detrimental effects on children's brain. The need for a multidisciplinary approach to the issue complexity and challenges. *Frontiers in human neuroscience*, 8(AUG):613, 2014.

[16] J. Sunyer. Urban air pollution and chronic obstructive pulmonary disease: A review. *European Respiratory Journal*, 17(5):1024–1033, 2001.

[17] Guarnieri M and Balmes J R. Outdoor air pollution and asthma. *Lancet*, 383:318–319, 2014.

[18] C A III Pope, R T Burnett, M J Thun, Calle E E, D Krewski, K Ito, and G Thurston. Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *JAMA*, 287(9), 2002.

[19] Y Zhao, S Wang, K Aunan, H M Seip, and J Hao. Air pollution and lung cancer risks in China - a meta-analysis. *Science of the Total Environment*, 366(2-3):500–513, 2006.

[20] D Pyatt and S Hays. A review of the potential association between childhood leukemia and benzene. *Chem Biol Interact*, 184(1-2):151–164, 2010.

[21] Anna Makri and Nikolaos Stilianakis. Vulnerability to air pollution health effects. *International journal of hygiene and environmental health*, 211(3-4):326–336, 2008.

[22] M.L. Bell, A. Zanobetti, and F. Dominici. Evidence on vulnerability and susceptibility to health risks associated with short-term exposure to particulate matter: a systematic review and meta-analysis. *American journal of epidemiology*, 178(6), 2013.

[23] Tracey J. Woodruff, Jennifer D. Parker, Amy D. Kyle, and Kenneth C. Schoendorf. Disparities in exposure to air pollution during pregnancy. *Environmental Health Perspectives*, 111(7):942–946, 2003.

[24] Marie Lynn Miranda, Pamela Maxson, and Sharon Edwards. Environmental contributions to disparities in pregnancy outcomes. *Epidemiologic Reviews*, 31(1):67–83, 2009.

[25] K P Stillerman, D R Mattison, L C Giudice, and T J Woodruff. Environmental exposures and adverse pregnancy outcomes: a review of the science. *Reprod Sci*, 15(7):631–650, 2008.

[26] Sabrina Llop, Ferran Ballester, Marisa Estarlich, Ana Esplugues, Marisa Rebagliato, and Carmen Iñiguez. Preterm birth and exposure to air pollutants during pregnancy. *Environmental Research*, 110(8):778–785, 2010.

[27] A M Padula, K M Mortimer, I B Tager, S K Hammond, F W Lurmann, W Yang, D K Stevenson, and G M Shaw. Traffic-related air pollution and risk of preterm birth in the San Joaquin Valley of California. *Annals of Epidemiology*, 24(12):888–895.e4, 2014.

[28] Hien Q. Le, Stuart A. Batterman, Julia J. Wirth, Robert L. Wahl, Katherine J. Hoggatt, Alireza Sadeghnejad, Mary Lee Hultin, and Michael Depa. Air pollutant exposure and preterm and term small-for-gestational-age births in Detroit, Michigan: Long-term trends and associations. *Environment International*, 44(1):7–17, 2012.

[29] NL Fleischer, M Merialdi, and A van Donkelaar. Outdoor Air Pollution, Preterm Birth, and Low Birth Weight: Analysis of the World Health Organization Global Survey on Maternal and Perinatal Health. *Environmental Health*, 2014.

[30] M et al. Pedersen. Ambient air pollution and low birthweight: a European cohort study (ESCAPE). *Lancet Respir Med.*, 1(9):695–704, 2013.

[31] E H Ha, Y C Hong, B E Lee, B H Woo, J Schwartz, and D C Christiani. Is air pollution a risk factor for low birth weight in Seoul? *Epidemiology*, 12(6):643–648, 2001.

[32] Prakesh S Shah, Taiba Balkhair, and Knowledge Synth Grp Determinants P. Air pollution and birth outcomes: A systematic review. *Environment International*, 37(2):498–516, 2011.

[33] Pei Chen Lee, James M. Roberts, Janet M. Catov, Evelyn O. Talbott, and Beate Ritz. First trimester exposure to ambient air pollution, pregnancy complications and adverse birth outcomes in Allegheny County, PA. *Maternal and Child Health Journal*, 17(3):545–555, 2013.

[34] Martine Vrijheid, David Martinez, Sandra Manzanares, Payam Dadvand, Anna Schembari, Judith Rankin, and Mark Nieuwenhuijsen. Ambient air pollution and risk of congenital anomalies: A systematic review and meta-analysis. *Environmental Health Perspectives*, 119(5):598–606, 2011.

[35] Bradley S Peterson, Virginia A Rauh, Ravi Bansal, Xuejun Hao, Zachary Toth, Giancarlo Nati, Kirwan Walsh, Rachel L Miller, Franchesca Arias, David Semanek, and Frederica Perera. Effects of prenatal exposure to air pollutants (polycyclic aromatic hydrocarbons) on the development of brain white matter, cognition, and behavior in later childhood. *JAMA psychiatry*, 72(6):531–40, 2015.

[36] N Künzli, P-O Bridevaux, L-J S Liu, R Garcia-Esteban, C Schindler, M W Gerbase, J Sunyer, D Keidel, and T Rochat. Traffic-related air pollution correlates with adult-onset asthma among never-smokers. *Thorax*, 64(8):664–670, 2009.

[37] Michael Jerrett, Rob McConnell, Jennifer Wolch, Roger Chang, Claudia Lam, Genevieve Dunton, Frank Gilliland, Fred Lurmann, Talat Islam, and Kiros Berhane. Traffic-related air pollution and obesity formation in children: a longitudinal, multilevel analysis. *Environmental health : a global access science source*, 13(1):49, 2014.

[38] Luiz A A Pereira, Dana Loomis, Gleice M S Conceição, Alfésio L F Braga, Rosângela M. Arcas, Humberto S. Kishi, Júlio M. Singer, György M. Böhm, and Paulo H N Saldiva. Association between air pollution and intrauterine mortality in Sao Paulo, Brazil. *Environmental Health Perspectives*, 106(6):325–329, 1998.

[39] Rochelle Green, Varada Sarovar, Brian Malig, and Rupa Basu. Association of stillbirth with ambient air pollution in a California cohort study. *American journal of epidemiology*, 181(11):874–882, 2015.

[40] Bing-Fang Hwang, Yungling Leo Lee, and Jouni J K Jaakkola. Air pollution and stillbirth: a population-based case-control study in Taiwan. *Environmental health perspectives*, 119(9):1345–9, 2011.

[41] Ambarina S. Faiz, George G. Rhoads, Kitaw Demissie, Lakota Kruse, Yong Lin, and David Q. Rich. Ambient air pollution and the risk of stillbirth. *American Journal of Epidemiology*, 176(4):308–316, 2012.

[42] Ambarina S Faiz, George G Rhoads, Kitaw Demissie, Yong Lin, Lakota Kruse, and David Q Rich. Does ambient air pollution trigger stillbirth? *Epidemiology (Cambridge, Mass.)*, 24(4):538–44, 2013.

[43] S Kannan, D P Misra, J T Dvonch, and A Krishnakumar. Exposures to airborne particulate matter and adverse perinatal outcomes: a biologically plausible mechanistic framework for exploring potential effect modification by nutrition. *Environ Health Perspect*, 114:1636–1642, 2006.

[44] Felipe Vadillo-Ortega, Alvaro Osornio-Vargas, Miatta A. Buxton, Brisa N. Sánchez, Leonora Rojas-Bracho, Martin Viveros-Alcaráz, Marisol Castillo-Castrejón, Jorge Beltrán-Montoya, Daniel G. Brown, and Marie S. O'Neill. Air pollution, inflammation and preterm birth: A potential mechanistic link. *Medical Hypotheses*, 82(2):219–224, 2014.

[45] Mariana Matera Veras, Nilmara de Oliveira Alves, Lais Fajersztajn, and Paulo Saldiva. Before the first breath: prenatal exposures to air pollution and lung development. *Cell and Tissue Research*, pages 1–11, 2016.

[46] S Panasevich, K Leander, M Rosenlund, P Ljungman, T Bellander, U de Faire, G Pershagen, and F Nyberg. Associations of long- and short-term air pollution exposure with markers of inflammation and coagulation in a population sample. *Occupational and environmental medicine*, 66(11):747–753, 2009.

[47] Annette Peters, Angela Döring, H. Erich Wichmann, and Wolfgang Koenig. Increased plasma viscosity during an air pollution episode: A link to mortality? *Lancet*, 349(9065):1582–1587, 1997.

[48] R. S. Gibbs, R. Romero, S. L. Hillier, D. A. Eschenbach, and R. L. Sweet. A review of premature birth and subclinical infection. *American Journal of Obstetrics and Gynecology*, 166(5):1515–1528, 1992.

[49] Remy Slama, Lyndsey Darrow, Jennifer Parker, Tracey J Woodruff, Matthew Strickland, Mark Nieuwenhuijsen, Svetlana Glinianaia, Katherine J Hoggatt, Srimathi Kannan, Fintan Hurley, Jaroslaw Kalinka, Radim Sram, Michael Brauer, Michelle Wilhelm, Joachim Heinrich, and Beate Ritz. Meeting report: Atmospheric pollution and human reproduction. *Environmental Health Perspectives*, 116(6):791–798, 2008.

[50] Maura Lodovici and Elisabetta Bigagli. Oxidative stress and air pollution exposure. *Journal of Toxicology*, 2011, 2011.

[51] Brent Altemose, Mark G Robson, Howard M Kipen, Pamela Ohman Strickland, Qingyu Meng, Jicheng Gong, Wei Huang, Guangfa Wang, David Q Rich, Tong Zhu, and Junfeng Zhang. Association of air pollution sources and aldehydes with biomarkers of blood coagulation, pulmonary inflammation, and systemic oxidative stress. *Journal of Exposure Science and Environmental Epidemiology*, (May):1–7, 2016.

[52] R. Smith, K. Maiti, and R. J. Aitken. Unexplained antepartum stillbirth: A consequence of placental aging? *Placenta*, 34(4):310–313, 2013.

[53] F J Kelly. Oxidative stress: its role in air pollution and adverse health effects. *Occupational Environmental MEdicine*, (60):612–616, 2003.

[54] Edith H van den Hooven, Yvonne de Kluizenaar, Frank H Pierik, Albert Hofman, Sjoerd W van Ratingen, Peter Y J Zandveld, Johan P Mackenbach, Eric A P Steegers, Henk M E Miedema, and Vincent W V Jaddoe. Air pollution, blood pressure, and the risk of hypertensive complications during pregnancy: the generation R study. *Hypertension (Dallas, Tex. : 1979)*, 57(3):406–412, 2011.

[55] Barbara Hackley, M S N CNM, Abigail Feinstein, and Jane Dixon. Air Pollution: Impact on Maternal and Perinatal Health. 52(5):435–443, 2007.

[56] Yanfen Lin, Leilei Zhou, Jian Xu, Zhongcheng Luo, Haidong Kan, Jinsong Zhang, Chonghuai Yan, and Jun Zhang. The impacts of air pollution on maternal stress during pregnancy. *Scientific Reports*, 7:40956, 2017.

[57] Louise B. Andersen, Lisa B. Melvaer, Poul Videbech, Ronald F. Lamont, and Jan S. Joergensen. Risk factors for developing post-traumatic stress disorder following childbirth: A systematic review. *Acta Obstetricia et Gynecologica Scandinavica*, 91(11):1261–1272, 2012.

[58] P. Turton, P. Hughes, C. D H Evans, and D. Fainman. Incidence, correlates and predictors of post traumatic stress disorder in the pregnancy after stillbirth. *British Journal of Psychiatry*, 178(JUNE):556–560, 2001.

[59] Evelyn Regina Egle Couto, Bruna Vian, Zoraide Gregório, Marcelo Luis Nomura, Renata Zaccaria, and Renato Passini Jr. Quality of life, depression and anxiety among pregnant women with previous adverse pregnancy outcomes. *São Paulo Medical Journal = Revista Paulista De Medicina*, 127(4):185–189, 2009.

[60] Katherine J Gold, Ananda Sen, and Rodney A Hayward. Marriage and cohabitation outcomes after pregnancy loss. *Pediatrics*, 125(5):e1202–7, 2010.

[61] Dag Moster, Rolv Terje Lie, and Trond Markestad. Long-term medical and social consequences of preterm birth. *The New England journal of medicine*, 359(3):262–73, 2008.

[62] Richard E Behrman and Adrienne Stith Butler. *Preterm Birth: Causes, Consequences, and Prevention*. National Academies Press, 2007.

[63] S Petrou. Economic consequences of preterm birth and low birth weight. *Br J Obstet Gyn*, 110 (suppl:17–23, 2003.

[64] Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *Quarterly Journal of Economics*, 2005.

[65] Alexander E P Heazell, Dimitrios Siassakos, Hannah Blencowe, Christy Burden, Zulfiqar A. Bhutta, Joanne Cacciatore, Nghia Dang, Jai Das, Vicki Flenady, Katherine J. Gold, Olivia K. Mensah, Joseph Millum, Daniel Nuzum, Keelin O'Donoghue, Maggie

Redshaw, Arjumand Rizvi, Tracy Roberts, H. E Toyin Saraki, Claire Storey, Aleena M. Wojcieszek, and Soo Downe. Stillbirths: Economic and psychosocial consequences. *The Lancet*, 387(10018):604–616, 2016.

[66] M F MacDorman and E C Gregory. Fetal and Perinatal Mortality: United States, 2013. *Natl Vital Stat Rep*, 64(8):1–24, 2015.

[67] Simon Cousens, Hannah Blencowe, Cynthia Stanton, Doris Chou, Saifuddin Ahmed, Laura Steinhardt, Andreea A. Creanga, Özge Tunalp, Zohra Patel Balsara, Shivam Gupta, Lale Say, and Joy E. Lawn. National, regional, and worldwide estimates of stillbirth rates in 2009 with trends since 1995: A systematic analysis. *The Lancet*, 377(9774):1319–1330, 2011.

[68] March of dimes. http://www.marchofdimes.org/materials/premature-birth-report-card-united-states.pdf.

[69] Mark J. van der Laan and Daniel Rubin. Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, 2(1), 2006.

[70] Sherri Rose and Mark J. van der Laan. A Targeted Maximum Likelihood Estimator for Two-Stage Designs. *The International Journal of Biostatistics*, 7(1):1–21, 2011.

[71] Mark J van der Laan and Sherri Rose. *Targeted Learning Causal Inference for Observational and Experimental Data*. Springer Series in Statistics, 2011.

[72] Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article25, 2007.

[73] Corwin Matthew Zigler and Francesca Dominici. Point: Clarifying policy evidence with potential-outcomes thinking-beyond exposure-response estimation in air pollution epidemiology. *American Journal of Epidemiology*, 180(12):1133–1140, 2014.

[74] Daniel Westreich, Jessie K. Edwards, Elizabeth T. Rogawski, Michael G. Hudgens, Elizabeth A. Stuart, and Stephen R. Cole. Causal impact: Epidemiological approaches for a public health of consequence. *American Journal of Public Health*, 106(6):1011–1012, 2016.

[75] M. S. Pearce, S. V. Glinianaia, J. Rankin, S. Rushton, M. Charlton, L. Parker, and T. Pless-Mulloli. No association between ambient particulate matter exposure during pregnancy and stillbirth risk in the north of England, 1962-1992. *Environmental Research*, 110(1):118–122, 2010.

[76] M Bobak and D a Leon. Pregnancy outcomes and outdoor air pollution: an ecological study in districts of the Czech Republic 1986-8. *Occupational and environmental medicine*, 56(8):539–543, 1999.

[77] Ok-Jin Kim, Eun-Hee Ha, Byung-Mi Kim, Ju-Hee Seo, Hye-Sook Park, Woo-Jae Jung, Bo-Eun Lee, Young-Ju Suh, Young-Ju Kim, Jong-Tae Lee, Ho Kim, and Yun-Chul Hong. PM10 and pregnancy outcomes: a hospital-based cohort study of pregnant women in Seoul. *Journal of occupational and environmental medicine*, 49(12):1394–402, 2007.

[78] E Defranco, E Hall, M Hossain, A Chen, E N Haynes, D Jones, S Ren, L Lu, and L Muglia. Air pollution and stillbirth risk: exposure to airborne particulate matter during pregnancy is associated with fetal death. *PLoS ONE*, 10(3):e0120594, 2015.

[79] Basu R, V Sarovar, and Malig B. Short-term air pollution exposure and the risk of stillbirth in California. In *28th annual meeting of the International Society for Environmental Epidemiology (ISEE)*, 2016.

[80] Chun-Yuh Yang, Chih-Ching Chang, Hung-Yi Chuang, Chi-Kung Ho, Trong-Neng Wu, and Shang-Shyue Tsai. Evidence for increased risks of preterm delivery in a population residing near a freeway in Taiwan. *Archives of environmental health*, 58(10):649–54, 2003.

[81] Ninez A. Ponce, Katherine J. Hoggatt, Michelle Wilhelm, and Beate Ritz. Preterm birth: The interaction of traffic-related air pollution with economic hardship in Los Angeles neighborhoods. *American Journal of Epidemiology*, 162(2):140–148, 2005.

[82] M Généreux, N Auger, M Goneau, and M Daniel. Neighbourhood socioeconomic status, maternal education and adverse birth outcomes among mothers living near highways. *Journal of epidemiology and community health*, 62(8):695–700, 2008.

[83] M Wilhelm, J K Ghosh, J Su, M Cockburn, M Jerrett, and B Ritz. Traffic-related air toxics and preterm birth: a population-based case-control study in Los Angeles county, California. *Environmental Health: A Global Access Science Source*, 10:89, 2011.

[84] Marie Lynn Miranda, Sharon E Edwards, Howard H Chang, and Richard L Auten. Proximity to roadways and pregnancy outcomes. *Journal of Exposure Science and Environmental Epidemiology*, 23(1):32–38, 2012.

[85] David Olsson, Ingrid Mogren, Kristina Eneroth, and Bertil Forsberg. Traffic pollution at the home address and pregnancy outcomes in Stockholm, Sweden. *BMJ open*, 5(8):e007034, 2015.

[86] U Gehring, A H Wijga, P Fischer, J C de Jongste, M Kerkhof, G H Koppelman, H A Smit, and B Brunekreef. Traffic-related air pollution, preterm birth and term birth weight in the PIAMA birth cohort study. *Environmental Research*, 111(1):125–135, 2011.

[87] U Gehring, M van Eijsden, M B Dijkema, M F van der Wal, P Fischer, and B Brunekreef. Traffic-related air pollution and pregnancy outcomes in the Dutch ABCD birth cohort study. *Occup Environ Med*, 68(1):36–43, 2011.

[88] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(0):96–146, 2009.

[89] Miguel A. Hernán. A definition of causal effect for epidemiological research. *Journal of epidemiology and community health*, 58:265–271, 2004.

[90] M L Petersen and M J van der Laan. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology*, 25(3):418–426, 2014.

[91] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

[92] Arthur S Goldberger. Structural Equation Models in the Social Sciences. *Econometrica*, 40979-1001, 1972.

[93] O D Duncan. *Introduction to structural equation models*. Academic Press, 1975.

[94] Judea Pearl. Causal Diagrams for Empirical Research. *Biometrika Trust, Oxford University Press*, 82(4):669–688, 1995.

[95] J Neyman. On the application of probability theory to agricultural experiments: principles (in Polish with German summary). *Roczniki Nauk Rolniczch*, 10:21–51, 1923.

[96] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[97] Alan E. Hubbard and Mark J. Van Der Laan. Population intervention models in causal inference. *Biometrika*, 95(1):35–47, 2008.

[98] Jasjeet S. Sekhon. The Neyman-Rubin Model of Causal Inference and Estimation Via Matching Methods. *The Oxford Handbook of Political Methodology*, 2008.

[99] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[100] Jonathan M. Snowden, Sherri Rose, and Kathleen M. Mortimer. Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7):731–738, 2011.

[101] M A Hernan, B Brumback, and J M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of {HIV}-positive men. *Epidemiol*, 11(5):561–570, 2000.

[102] Mark J. van der Laan and Richard J. C. M. Starmans. Entering the Era of Data Science: Targeted Learning and the Integration of Statistics and Computational Data Analysis. *Advances in Statistics*, 2014:1–19, 2014.

[103] Frank R Hampel. The Influence Curve and Its Role in Robust Estimation. *Source Journal of the American Statistical Association*, 69(346):383–393, 1974.

[104] Francesca Dominici, Lianne Sheppard, and Merlise Clyde. Health Effects of Air Pollution: A Statistical Review. *International Statistical Review*, 71(2):243–276, 2003.

[105] Johnston SL Chauhan AJ. Air pollution and infection in respiratory illness. *British Medical Bulletin*, 68(1):95–112, 2003.

[106] S Genc, Z Zadeoglulari, S.H. Fuss, and K Genc. The Adverse Effects of Air Pollution on the Nervous System. *Journal of Toxicology*, 2012:1–23, 2012.

[107] Svetlana Glinianaia, Judith Rankin, R Bell, Tanja Pless-Mulloli, and D Howel. Particulate air pollution and fetal health: a systematic review of the epidemiologic evidence. *Epidemiology*, pages 36–45, 2004.

[108] Beate Ritz. The Effect of Ambient Carbon Monoxide on Low Birth Weight among Children Born in Southern California between 1989 and 1993. *Environmental Health Perspectives*, 107(1):17–25, 1999.

[109] Rochelle S. Green, Brian Malig, Gayle C. Windham, Laura Fenster, Bart Ostro, and Shanna Swan. Residential exposure to traffic and spontaneous abortion. *Environmental Health Perspectives*, 117(12):1939–1944, 2009.

[110] Amy M. Padula, Ira B. Tager, Suzan L. Carmichael, S. Katharine Hammond, Frederick Lurmann, and Gary M. Shaw. The association of ambient air pollution and traffic exposures with selected congenital anomalies in the San Joaquin Valley of California. *American Journal of Epidemiology*, 177(10):1074–1085, 2013.

[111] Tracey J. Woodruff, Jeanne Grillo, and Kenneth C. Schoendorf. The relationship between selected causes of postneonatal infant mortality and particulate air pollution in the United States. *Environmental Health Perspectives*, 105(6):608–612, 1997.

[112] Gudrun Weinmayr, Elisa Romeo, Manuela de Sario, Stephan K. Weiland, and Francesco Forastiere. Short-Term effects of PM10 and NO2 on respiratory health among children with asthma or asthma-like symptoms: A systematic review and Meta-Analysis. *Environmental Health Perspectives*, 118(4):449–457, 2010.

[113] J. R. Balmes. The role of ozone exposure in the epidemiology of asthma. *Environmental Health Perspectives*, 101(SUPPL. 4):219–224, 1993.

[114] Brian J. Malig, Dharshani L. Pearson, Yun Brenda Chang, Rachel Broadwin, Rupa Basu, Rochelle S. Green, and Bart Ostro. A time-stratified case-crossover study of ambient ozone exposure and emergency department visits for specific respiratory diagnoses in California (2005-2008). *Environmental Health Perspectives*, 124(6):745–753, 2016.

[115] Zeyan Liew, Jørn Olsen, Xin Cui, Beate Ritz, and Onyebuchi A. Arah. Bias from conditioning on live birth in pregnancy cohorts: An illustration based on neurodevelopment in children after prenatal exposure to organic pollutants. *International Journal of Epidemiology*, 44(1):345–354, 2015.

[116] Stefanie E Sarnat, Mitchel Klein, Jeremy A Sarnat, W Dana Flanders, Lance A Waller, James A Mulholland, Armistead G Russell, and Paige E Tolbert. An examination of exposure measurement error from air pollutant spatial variability in time-series studies. *Journal of exposure science & environmental epidemiology*, 20(2):135–46, 2010.

[117] G T Goldman, J A Mulholland, A G Russell, M J Strickland, M Klein, L A Waller, and P E Tolbert. Impact of exposure measurement error in air pollution epidemiology: effect of error type in time-series studies. *Environ Health*, 10:61, 2011.

[118] Maya L. Petersen, Kristin E. Porter, Susan Gruber, Yue Wang, and Mark J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

[119] L A Darrow, M J Strickland, M Klein, L A Waller, W D Flanders, A Correa, M Marcus, and P E Tolbert. Seasonality of birth and implications for temporal studies of preterm birth. *Epidemiology*, 20(5):699–706, 2009.

[120] Lyndsey A. Darrow, Tracey J. Woodruff, and Jennifer D. Parker. Maternal smoking as a confounder in studies of air pollution and infant mortality. *Epidemiology*, 17(5):592–593, 2006.

[121] Lynne C. Messer, Jay S. Kaufman, Nancy Dole, David A. Savitz, and Barbara A. Laraia. Neighborhood Crime, Deprivation, and Preterm Birth. *Annals of Epidemiology*, 16(6):455–462, 2006.

[122] Patricia O'Campo, Jessica G. Burke, Jennifer Culhane, Irma T. Elo, Janet Eyster, Claudia Holzman, Lynne C. Messer, Jay S. Kaufman, and Barbara A. Laraia. Neighborhood deprivation and preterm birth among non-Hispanic Black and white women in eight geographic areas in the United States. *American Journal of Epidemiology*, 167(2):155–163, 2008.

[123] A Metcalfe, P Lail, W A Ghali, and R S Sauve. The association between neighbourhoods and adverse birth outcomes: a systematic review and meta-analysis of multi-level studies. *Paediatr Perinat Epidemiol*, 25(3):236–245, 2011.

[124] Annette Peters, Stephanie von Klot, Margit Heier, Ines Trentinaglia, Allmut Hörmann, H. Erich Wichmann, and Hannelore Löwel. Exposure to Traffic and the Onset of Myocardial Infarction. *New England Journal of Medicine*, 351(17):1721–1730, 2004.

[125] Marie Pedersen, Thorhallur I. Halldorsson, Sjurdur F. Olsen, Dorrit Hjortebjerg, Matthias Ketzel, Charlotta Grandström, Ole Raaschou-Nielsen, and Mette Sørensen. Impact of road traffic pollution on pre-eclampsia and pregnancy-induced hypertensive disorders. *Epidemiology*, page 1, 2016.

[126] Michael Jerrett, Ketan Shankardass, Kiros Berhane, W. James Gauderman, Nino Künzli, Edward Avol, Frank Gilliland, Fred Lurmann, Jassy N. Molitor, John T. Molitor, Duncan C. Thomas, John Peters, and Rob McConnell. Traffic-related air pollution and asthma onset in children: A prospective cohort study with individual exposure measurement. *Environmental Health Perspectives*, 116(10):1433–1438, 2008.

[127] Ferran Ballester, Marisa Estarlich, Carmen Iñiguez, Sabrina Llop, Rosa Ramón, Ana Esplugues, Marina Lacasaña, and Marisa Rebagliato. Air pollution exposure during pregnancy and reduced birth size: a prospective birth cohort study in Valencia, Spain. *Environmental health : a global access science source*, 9:6, 2010.

[128] S Johnson, C Hollis, P Kochhar, E Hennessy, D Wolke, and N Marlow. Autism Spectrum Disorders in Extremely Preterm Children. *Journal of Pediatrics*, 156(4):525–531.e2, 2010.

[129] Nina B. Kyrklund-Blomberg, Fredrik Granath, and Sven Cnattingius. Maternal smoking and causes of very preterm birth. *Acta Obstetricia et Gynecologica Scandinavica*, 84(6):572–577, 2005.