Topics in Survival Analysis

by

Lucia Catherine Petito

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 Nicholas P. Jewell, Chair
Professor Barbara Abrams
Professor Alan Hubbard

Spring 2017

Topics in Survival Analysis

Abstract

Topics in Survival Analysis

by

Lucia Catherine Petito

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Nicholas P. Jewell, Chair

This dissertation covers three distinct topics in survival analysis: 1) current status data in the context of group testing subject to misclassification; 2) marginal structural modeling of a safety outcome from clinical trial data; and 3) the relationship between preterm birth and weight gain in pregnancy. Abstracts for each chapter separately are presented below.

Chapter 2. Group testing, introduced by Dorfman (1943), has been used to reduce costs when estimating the prevalence of a binary characteristic based on a screening test of k groups that include n independent individuals in total. If the unknown prevalence is low, and the screening test suffers from misclassification, it is also possible to obtain more precise prevalence estimates than those obtained from testing all n samples separately (Tu et al., 1994). In some applications, the individual binary response corresponds to whether an underlying time-to-event variable T is less than an observed screening time C, a data structure known as current status data. Given sufficient variation in the observed Cs, it is possible to estimate the distribution function, F, of T nonparametrically, at least at some points in its support, using the pool-adjacent-violators algorithm (Ayer et al., 1955). Here, we consider nonparametric estimation of F based on group tested current status data for groups of size k where the group tests positive if and only if any individual's unobserved T is less than its corresponding observed C. We investigate the performance of the group-based estimator as compared to the individual test nonparametric maximum likelihood estimator, and show that the former can be more precise in the presence of misclassification for low values of $F(t)$. Potential applications include testing for the presence of various diseases from pooled samples where interest focuses on the age at incidence distribution rather than overall prevalence. We apply this estimator to the age-at-incidence curve for hepatitis C infection in a sample of U.S. women who gave birth to a child in 2014, where group assignment is done at random and based on maternal age. We discuss the relationship to other work in the literature, and potential extensions.

Chapter 3. Marginal structural modeling was first developed to address time-dependent confounding in studies where the effect of a time-varying exposure on an outcome is of interest. This chapter begins by introducing the reader to the concept of time-dependent confounding, and describes inverse probability weighting estimators for parameters of marginal structural models. The second part of chapter 3 contains an application of marginal structural modeling in a drug safety study. Studies in pharmacoepidemiology are often conducted in rich data sources, such as clinical trials or administrative databases, where large quantities of information are collected repeatedly over time. These data sources can and should be exploited, but traditional methods often cannot incorporate all available data, and fail to take time-dependent confounding into account. Marginal structural modeling and weighted estimators, tools often used in observational studies, can help to alleviate these challenges.

Our objective in this study was to estimate the relation between rheumatoid arthritis (RA) disease activity, cholesterol levels, and major adverse cardiovascular events (MACE) in patients with moderate to severe rheumatoid arthritis who are currently prescribed tocilizumab, accounting for the presence of time-dependent confounding, such as other inflammatory markers, lipid levels, and rheumatoid arthritis disease measures. We studied 3,986 patients enrolled in one of five clinical trials used to study tocilizumab, who then joined one of three long-term extension studies. We used a weighted logistic regression model to explore associations between pre-treatment levels of RA disease activity and cholesterol on the 5-year risk of MACE. We then used a logistic marginal structural model to explore causal relations between pre- and post-treatment RA disease activity and cholesterol levels, and 5-year risk of MACE, adjusting for time-dependent confounders. We did not find evidence that pre- or post-treatment levels of RA disease activity, HDL cholesterol, and LDL cholesterol were associated with increased risk of MACE in patients with moderate to severe rheumatoid arthritis taking tocilizumab, once time-dependent confounding from inflammatory markers and other lipid levels was taken into account. After adjustment for time dependent confounding, traditional markers of disease activity and cholesterol were not associated with an increased risk of cardiac events among RA patients treated with tocilizumab.

Chapter 4. The relationship between weight gain in pregnancy and preterm birth is still contested due to their inherent dependence. In the first part of Chapter 4, we wanted to quantify the relationship between pregnancy weight gain with early and late preterm birth and evaluate whether associations differed between non-Hispanic (NH) black and NH white women. We analyzed a retrospective cohort of all live births to NH black and NH white women in the U.S. 2011-2015 (n = 10,714,983). We used weight gain z-scores in multiple logistic regression models, stratified by prepregnancy body mass index (BMI) and race, to calculate population attributable risks (PAR) and PAR percentages for early and late preterm birth. We found that both low and high pregnancy weight gain were related to preterm birth, but these associations varied by BMI and race, and differed from associations with late preterm birth. For high weight gain and early preterm birth, the PAR percentage ranged from 8-10% in NH black women and from 6-8% in NH white women. Racial differences were small or nonexistent for late preterm birth, with PAR percentages ranging from 2-7% in NH black women and from 3-7% in NH white women. We conclude that these findings add to evidence that moderate gestational weight gain could help prevent preterm birth, and suggest that the impact may be greatest for early preterm birth in NH black women.

The second part of Chapter 4 is a preliminary analysis assessing the variety of measures of weight gain in pregnancy and their relationship with preterm birth. Serial GWG measurements provide ideal data, but are rarely available in population health datasets. The electronic medical records from 160,635 women in Sweden have been compiled to be the largest dataset in the world that contains repeated weight gain measures through pregnancy. Here, we describe the pattern of weight gain in pregnancy in 103,661 Swedish pregnancies, and assess whether the observed pattern before 37 weeks' gestation differs between preterm and term pregnancies.

This dissertation is dedicated to my parents, Christopher and Sue Petito, my brothers Chris and Billy Petito, and my wonderful fiancé Mark Levin.

Without all of your love and support, none of this would be possible.

Contents

Contents **ii**

List of Figures **iv**

List of Tables **vii**

1 Introduction **1**
1.1 Terms and Quantities in Survival Analysis 1
1.2 Organization of Dissertation 4

2 Misclassified Group Tested Current Status Data **6**
2.1 Introduction 6
2.2 Notation and Likelihood Function 7
2.3 An expectation-maximization-pool- adjacent-violators algorithm 8
2.4 Elementary Example 11
2.5 Simulations 12
2.6 Application to Hepatitis C Data 16
2.7 Discussion 17

3 An Application of Marginal Structural Modeling **30**
3.1 Gentle Introduction to Marginal Structural Models 30
3.2 Using marginal structural modeling to evaluate the roles of lipid levels, inflammatory markers, and rheumatoid arthritis disease activity in assessing risk of cardiovascular events 37
3.3 Implementation of the Marginal Structural Model from Longitudinal Data . 55

4 Weight gain in pregnancy and preterm birth **59**
4.1 Introduction 59
4.2 Weight gain during pregnancy and the black-white disparity in preterm birth 62
4.3 Weight gain during pregnancy and preterm birth in a Swedish Cohort 78

Bibliography **93**
Chapter 1 93

Chapter 2 93
Chapter 3 96
Chapter 4 98

List of Figures

2.1 Elementary example of data configuration with two groups, each of size 2, where the first group has tested positive and the second group has tested negative. . . 19

2.2 Results from six simulations of estimation of F, with 1,000 runs each, showing results for varying sample sizes n and group sizes k. The black lines are the average estimates of $\hat{F}(t)$ over the 1,000 simulations with the solid line representing the true cumulative distribution function, Weibull(4, 25), and the dashed and dotted lines respectively representing the estimates from the pool-adjacent-violators algorithm and the expectation-maximization-pool-adjacent-violators algorithm. The grey lines are the 2.5^{th} and 97.5^{th} quantiles from the simulation runs for each estimator, using the same line types. 19

2.3 Graphical representation of the finite sample percent relative bias from twelve simulations replicated 1,000 times with 5,000 individuals each, based on varying group sizes k and misclassification rates (α, β), the latter denoted on the right hand column. The solid black line displays results for the expectation-maximization-pool-adjacent-violators algorithm for group tests, and the dashed black line represents the pool-adjacent-violators algorithm for misclassified individual test data. The dashed grey line represents the reference of zero percent bias. 20

2.4 Log of the pseudo-relative efficiency of the expectation-maximization-pool-adjacent-violators algorithm and the adjusted pool-adjacent-violators algorithm from twelve simulations with 1,000 runs with 5,000 individuals each, based on varying grouping sizes k and misclassification rates (α, β), the latter denoted on the right hand y-axis. The solid black line is a lowess curve showing the overall trend in pseudo-relative efficiency as t increases. The dashed black line represents equal width 95% Monte Carlo quantile intervals for reference; if the solid black line is below 0, the width of the expectation-maximization-pool-adjacent-violators 95% Monte Carlo quantile interval is smaller than that from the individual test pool-adjacent-violators algorithm. 21

2.5 Four estimates of the cumulative incidence of hepatitis C in non-Hispanic white, child-bearing women ages 13-40 living in the U.S. in 2014 when grouping is assigned within common values of age. Grouping sizes considered were $k = 1$, 2, 5, and 10 The solid black line is the estimate from the individual or group tested results, and the solid grey lines represent the upper and lower bounds for 95% confidence intervals. 22

2.6 Four estimates of the cumulative incidence of hepatitis C in non-Hispanic white, child-bearing women ages 13-40 living in the U.S. in 2014 when group testing assigned at random. The solid line is the pool-adjacent violators estimate from the individual test results, and the dotted, short dashed, and long dashed lines are the estimates from the expectation-maximization pool-adjacent-violators algorithm from artificially grouping the individual test results into groups of sizes 2, 5, and 10, respectively. 22

2.7 Results from simulations done on a grid of ten possible censoring times with $n = 1,000$ individuals observed at each censoring time. Grouping was done within the grid points. Groups of sizes 1, 2, 5, and 10, and misclassification rates of 0, 1%, 5%, 10%, and 20% were examined. Estimates and variances were calculated as described in Section 4 of the supplemental information. Points represent the estimates and confidence bars represent asymptotically normal 95% confidence intervals. 29

3.1 Directed acyclic graph representing the confounding process assumed in the prepregnancy hypertension - preeclampsia example. 31

3.2 Directed acyclic graph representing the process assumed (when ignoring repeated measures) for studying the risk of having a major adverse cardiovascular event before the end of the study. The principal exposures of interest are shown in circles. Confounders (demographics, health status, concomitant medication use, and pre-treatment initiation levels of all laboratory values) are indicated by a square. Solid thick arrows denote the causal paths of interest. Dashed thin arrows indicate potential pathways for confounding. 53

3.3 Directed acyclic graph representing the hypothesized longitudinal process for major adverse cardiovascular events. The principal exposures of interest are shown in circles. Time independent confounders (e.g. demographics, health status, and concomitant medication use) and time dependent confounders (e.g. laboratory values) are shown in squares. Solid thick arrows denote the causal paths of interest. Dashed thin arrows indicate potential pathways for confounding. All confounders used in analyses are listed in table 3.1. 54

4.1 An illustration of when term versus preterm births occur. The blue rectangles indicate weight gain that happens while women are still at risk for preterm birth, while the red rectangle indicates weight gain that happens after a woman is not at risk of preterm birth. 60

4.2 Predicted gestational weight gain at each week of gestation at z-scores of -1 standard deviation (SD), 0 SD, and +1 SD 77

4.3 Sample selection from all women who delivered singleton infants in the Stockholm and Gotland regions of Sweden, 2008-2014 86

4.4 Kaplan-Meier curve describing probability of women remaining pregnant at each week of gestation in 103,661 women who delivered live, singleton infants in the Stockholm and Gotland regions of Sweden, 2008-2014. Note confidence intervals are present, but so small they become negligible. 89

4.5 Average weight gain at each week of pregnancy in term and preterm pregnancies, stratified by early pregnancy BMI 90

4.6 Average weight gain at each week of pregnancy in each early pregnancy BMI group, stratified by gestational duration 91

4.7 The points represent differences in average weight gain between term and preterm pregnancies at each weeks of gestation in each early pregnancy BMI group. The black line at 0kg represents no difference in average weight gain between term and preterm pregnancies. The blue line is a lowess curve showing the smoothed trend in these residuals. Values below 0kg indicate higher average weight gain in preterm pregnancies. 92

List of Tables

2.1 Results from the expectation-maximization-pool-adjacent-violators algorithm in groups of size two and the pool-adjacent-violators algorithm for individuals in a setting with n =10,000 individuals on 10 fixed censoring times for several fixed, equal and constant specificity and sensitivity. Grouping was done across the Cs, although there were 1,000 individuals at each C. Here we present the bias of each estimator (compared to $F(T)$) multiplied by 1,000 (for example, a reported bias of 1.4 corresponds to a true bias of 0.0014) and the width of the 95% Monte Carlo quantile intervals, multiplied by 100 (e.g. a reported width of 1.53 corresponds to an actual width of 0.0153). 25

2.2 Results from the expectation-maximization-pool-adjacent-violators algorithm in groups of size five and the pool-adjacent-violators algorithm for individuals in a setting with n =10,000 individuals on 10 fixed censoring times for several fixed, equal and constant specificity and sensitivity. Grouping was done across the Cs, although there were 1,000 individuals at each C. Here we present the bias of each estimator (compared to $F(T)$) multiplied by 1,000 (for example, a reported bias of 1.4 corresponds to a true bias of 0.0014) and the width of the 95% Monte Carlo quantile intervals, multiplied by 100 (e.g. a reported width of 1.53 corresponds to an actual width of 0.0153). 26

2.3 Results from the expectation-maximization-pool-adjacent-violators algorithm in groups of size ten and the pool-adjacent-violators algorithm for individuals in a setting with n =10,000 individuals on 10 fixed censoring times for several fixed, equal and constant specificity and sensitivity. Grouping was done across the Cs, although there were 1,000 individuals at each C. Here we present the bias of each estimator (compared to $F(T)$) multiplied by 1,000 (for example, a reported bias of 1.4 corresponds to a true bias of 0.0014) and the width of the 95% Monte Carlo quantile intervals, multiplied by 100 (e.g. a reported width of 1.53 corresponds to an actual width of 0.0153). 27

3.1 Variables included in exposure models used to create weights for longitudinal marginal structural model 47

3.2 Description of imputed data at baseline 48

3.3 Description of pre-treatment patient characteristics 49

3.4 Description of time-varying patient characteristics 50
3.5 Description of censoring observed in dataset 50
3.6 Odds Ratios (95% Confidence Intervals) from weighted and unweighted logistic regression models for the pre-treatment effects of rheumatoid arthritis disease activity score, HDL-cholesterol, and LDL-cholesterol on major adverse cardiac events in patients with moderate to severe RA taking tocilizumab. 51
3.7 Odds Ratios (95% Confidence Intervals) from marginal structural model and unweighted regression model for the post-treatment effects of RA Disease Activity Score, HDL Cholesterol, and LDL Cholesterol on Major Adverse Cardiac Events in patients with moderate to severe RA taking tocilizumab. 52

4.1 Characteristics of the study sample and rates of preterm birth by maternal race, U.S. singleton pregnancies in non-Hispanic black and non-Hispanic white women (2011-2015) 69
4.2 Crude rates of early preterm birth by maternal race, prepregnancy BMI, and pregnancy weight gain z-score categories, U.S. singleton pregnancies in non-Hispanic black and non-Hispanic white women (2011-2015) 71
4.3 Crude rates of late preterm birth by maternal race, prepregnancy BMI, and pregnancy weight gain z-score categories, U.S. singleton pregnancies in non-Hispanic black and non-Hispanic white women (2011-2015) 72
4.4 Population attributable risks and population attributable risk percentages for early preterm birth ($<$ 32 wk) due to low and high pregnancy weight gain, stratified by race and prepregnancy BMI and adjusted for confounders, in U.S. singleton pregnancies in non-Hispanic black and non-Hispanic white women (2011-2015) . 73
4.5 Population attributable risks and population attributable risk percentages for late preterm birth ($\geq$ 32 wk and $<$ 37 wk) due to low and high pregnancy weight gain, stratified by race and prepregnancy BMI and adjusted for confounders, in U.S. singleton pregnancies in non-Hispanic black and non-Hispanic white women (2011-2015) 75
4.6 Comparison of characteristics of included and excluded women who delivered live, singleton infants in the Stockholm and Gotland regions of Sweden, 2008-2014 . . 87
4.7 Characteristics of 103,661 women who delivered live, singleton infants in the Stockholm and Gotland regions of Sweden, 2008-2014 88

Acknowledgments

Through my time at Berkeley, I have interacted with many people, all of whom have touched my life and enriched it. It is with great pleasure that I acknowledge them here, and thank them for their roles in my life.

My advisor, Dr. Nicholas P. Jewell, has been an unending source of support, mentorship, and encouragement through my time at Berkeley. I have enjoyed each and every interaction I have had with him, beginning in Spring 2012, when he told me that "graduate school is 5 years, and there is no reason you should be miserable the whole time." He taught me to critically think about statistical questions and concisely express myself (in writing), and provided so many opportunities to develop my skill as a collaborator and a teacher. I appreciate all of the advice, both career-oriented and personal, he has so willingly given over the past five years. His guidance allowed me to grow personally and professionally throughout graduate school.

Working with Dr. Barbara Abrams for the past three years has been one of my greatest joys in graduate school. I thank her for her support, academically, emotionally, and financially. Her expertise in perinatal, nutritional, and life course epidemiology have shaped my view of women's health issues. Her ability to critically think about public health issues and generate research questions is astounding and inspiring. I greatly appreciate the opportunity she gave me to be a part of her research team. She always encouraged me to pursue and develop my talent as a communicator as well as a statistician.

Dr. Alan Hubbard has also been a wonderful mentor of mine. He has provided many opportunities for me to develop as a teacher. His cheery attitude and practical approach to statistics always encourage me to do better work. The motto "Do good science" has been forever etched in my brain. I want to thank him for his guidance, patience, and for serving on my dissertation committee.

The other faculty members in the Division of Biostatistics, Dr. Sandrine Dudoit, Dr. Lexin Li, Dr. Maya Petersen, Dr. Steve Selvin, and Dr. Mark van der Laan, provided me with an excellent statistical education and grounded me in the way I frame statistical questions. I thank them for all of the advice they have provided over the last five years, both inside and outside the classroom. Dr. Mahasin Mujahid provided me with my first opportunity to do methodologic research in epidemiology, for which I will be forever grateful. Additionally, I thank Dr. Marco Carone for all of the knowledge he imparted to me. He has been an excellent mentor and I aspire to be as good a teacher as he is someday.

Dr. Maureen Lahiff, Sharon Norris and Burke Bundy have also been an invaluable part of my support network. Maureen provided so much guidance, helping me to shape how I approach students as a teacher. Sharon is so helpful - without her assistance navigating

administrative barriers my Ph.D. would not have been possible. Burke has provided useful instructions for parallelizing statistical computations, and interesting conversation in the lounge.

The students in both the Biostatistics and Epidemiology departments have been a great source of emotional and academic support over the past five years. Learning and laughing with you all has been my pleasure. I want to thank Stephanie Leonard especially for being a wonderful office-mate. I aspire to her very matter-of-fact approach to research, and her advice and friendship has been a great source of emotional support to me.

My participation in the UC Berkeley Symphony Orchestra has played an essential role in the maintenance of my mental health throughout graduate school. I have greatly appreciated the leadership and mentorship opportunities they provided. I also thank Christina Simpson, Madison Alan-Lee, and Mosa Tsay for two wonderful years of chamber music making. My violin teacher, Antoine van Dongen, has also been a pleasure to work with for the past 4 years. His unfailing positive attitude and musicianship are inspirational.

Thanks to *Biometrika* for accepting the material presented in Chapter 2 for publication, titled, “Misclassified Group Tested Current Status Data.” The National Vital Statistics System, maintained by the Center for Disease Control, is the source for the datasets used in Chapter 2 and Chapter 4.1.

Grants R01 HL122212 from the National Heart, Lung, and Blood Institute, R01 HD072008 from the National Institute of Child Health & Human Development, and R01 MD006014 from the National Center on Minority Health and Health Disparities and an internship in Real World Data Science at Genetech-Roche have provided support for the projects included in this dissertation.

Genentech-Roche graciously provided the dataset used in Chapter 3, a compilation of data collected in clinical trials and long-term-extension studies in rheumatoid arthritis patients taking tocilizumab (Actemra). Dr. Sara Gale, scientist in the Real World Data Science division of Genetech-Roche, mentored me through a summer internship and has been a great source of support in determining my next career step. I would also like to thank Dr. Khaled Sarsour, Dr. Micki Klearman, and Dr. Dave Musselman for their invaluable input in formulating the scientific question, and editing and shaping the final paper.

I would like to thank Stephanie Leonard, Dr. Barbara Abrams, Dr. Olof Stephansson, Dr. Jennifer Hutcheon, Dr. Lisa Bodnar, Dr. Mahasin Mujahid, and Dr. Yvonne Cheng for their contributions to Chapter 4.1. Dr. Olof Stephansson and Dr. Kari Johansson at Karolinska Institutet provided the dataset used in Chapter 4.2, weight measurements from all women in the Stockholm and Gotland regions of Sweden from 2008-2014. I want to thank Kari for preparing the dataset for me, and Olof for his help in formulating the research question.

None of my accomplishments would have been possible without the support of my family, Sue, Chris, Chris, and Billy Petito. My parents taught me to dream, and instilled a value of hard work in me. My brothers are my unquestioning supporters always. Their love (and interest in my work) from 3,000 miles away has meant the world to me.

Finally, my fiance, best friend, and partner for life, Mark Levin, has provided an unending supply of love, support, and fun distractions. Thank you for always supporting me no matter which direction my dreams go. Although I am not always the most eloquent, I love you and having you in my life means the world to me. Thank you for your patience and understanding, through the good times, and the more stressful times.

Chapter 1

Introduction

1.1 Terms and Quantities in Survival Analysis

Survival analysis is the branch of statistics concerned with making inference about "time-to-event" distributions. Specifically, the time until an event occurs from a pre-specified starting time is the quantity of interest. Some common examples of questions that survival analysis addresses that arise in epidemiology are: what is the median age of death in the general population? What is the median life expectancy after an individual has been infected with human immunodeficiency virus (HIV)? What is the median age at onset of Altzheimer's disease? What is the proportion of patients with breast cancer that will survive 5 years after treatment ends?

In order to define a random variable T that represents a time-to-event, three elements must be carefully and explicitly specified [1, 2]. The first is the terminating event. This event is the focus of the analysis, and is commonly death, disease occurrence, disease recovery, or the level of biological marker surpassing a threshold (e.g. cluster of differentiation 4 (CD4) count first dropping below 200 cells/mm^3). The next quantity to be defined is the initiating event. Initiating events are usually milestone events such as birth or initiation of treatment (e.g. anti-retro viral therapy initiation for HIV patients, date of hip replacement surgery). The last element is the scale used to measure 'time' elapsed between the initiating and terminating events. Absolute time is frequently used, but occasionally measuring time in terms of number of risky events can be more helpful in understanding the process. For example, the number of sexual contacts may be a better time scale to use when studying the time to contracting HIV than absolute time.

Broadly speaking, the distribution, F, of the random variable T, the time of event, is the quantity of interest [1, 2]. F is a cumulative distribution function, defined as $F(t) = Pr(T \leq t)$, which gives the probability that the event occurs at or before time t. The random variable T also has an associated survival function, which, not surprisingly, is most commonly used in survival analysis. This survival function represent the probability that the event occurs after time t, and is written as:

$$S(t) = Pr(T > t) = \int_t^\infty f(u)du = 1 - F(t).$$

The survival function S has several properties that are worth mentioning. Survival functions are always monotonically decreasing, meaning that at no point is $S(u) < S(u + \delta)$, for any $u \in [0, \infty)$ and $\delta > 0$. Also, it is assumed that at time 0, no events have occurred yet, meaning that $S(0) = 1$. The survival function S is strictly positive - $S(u) > 0 \forall u \in [0, \infty)$ - although $S(t)$ approaches 0 as t approaches ∞. Finally, if S is differentiable, then its derivative (the associated density function) can be defined as:

$$s(t) = S'(t) = \frac{d}{dt} S(t) = \frac{d}{dt} \int_t^\infty f(u)du = \frac{d}{dt}[1 - F(t)] = -f(t)$$

The hazard function of T, generally denoted $\lambda(t)$, is the event rate at time t conditional on survival until time t or later [1, 2]. We can define λ more precisely as

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}.$$

The hazard function is nonnegative ($\lambda(t) \geq 0 \forall t \in [0, \infty)$), and integrates to ∞. It is otherwise unconstrained - it can be increasing, decreasing, or discontinuous unlike the survival function. It has an associated cumulative hazard function, defined as:

$$\Lambda(t) = \int_0^t \lambda(u)du,$$

which represents the accumulation of hazard from the beginning until time t. Λ can also be written in terms of the survival function:

$$\Lambda(t) = -\log S(t) \implies S(t) = \exp(-\Lambda(t)).$$

This relation is the basis for many developments in survival analysis.

One of the main challenges that differentiates survival analysis from traditional statistical methods is incompleteness in the data [1, 2]. In a perfect world, T would be observable for all individuals. However, the observation of T is often obscured by the presence of censoring, C. There are three main types of censoring: interval censoring, left censoring, and right censoring. Interval censoring occurs when the event of interest occurs between two points of observation; for example, when cholesterol levels cross a threshold between primary care physician visits, but the precise moment in time when that threshold was achieved is unknown. Left censoring exists when the random variable T is known to have happened on the interval $[0, t]$. This case is rare; it exists only when subjects in a study have been enrolled but experienced the event in question at an unknown time before the study begins, e.g. a mild cardiovascular event.

If it is only known that the random variable T may have occurred after some time t, that observation is considered to be right-censored at time t [1, 2]. This type of censoring can occur for several reasons. First, the study design may contribute to the right censoring of participants. For example, a study end follow-up of individuals before all of the individuals have experienced the event of interest. Also, the study may have been funded to detect a certain number of events (e.g. cardiovascular events), and may stop follow-up of individuals once that threshold has been reached. Individuals may also be right-censored "randomly," in that they may voluntarily choose to leave the study early due to unrelated health complications, needing to move away from the study site, or other personal reasons. Typically, all types of censoring C are assumed to be independent of the event time T.

Another case where T is unobservable for certain individuals is called truncation [1, 2]. When individuals are censored, they are still available in the study design process to be recruited as a participant. When individuals are truncated, they are unable to be a part of the sampling pool used to recruit study participants. Individuals can be either left or right truncated. Left truncation occurs most often in cross-sectional study designs, where individuals who, for example, died before study recruitment were necessarily ineligible to be a part of the study. Right truncation exists in scenarios when the variable T has an upper bound, so individuals with large values of T cannot be recruited. For example, right truncation can occur in studies that were conducted over a fixed time interval. Consider a study of individuals who contracted HIV after January 1, 1980 and developed AIDS by December 31, 1989. If the variable T we are studying is time to onset of AIDS, individuals who contracted HIV after January 1, 1980 but who took longer than the study duration to develop AIDS were not eligible to participate in the study. Although statistically there is not much we can do about truncation, as researchers, we can be aware of the biases that it potentially introduces into our studies.

Statisticians have developed methods to estimate F, or its complement $S = 1 - F$, in the presence of censoring. The initial development in this field was the Kaplan-Meier estimator, which allows for the nonparametric estimation of the survival function S over the range of times when having the event is possible, even in the presence of censoring [3]. Kaplan-Meier works by dividing the range over which T is defined into intervals, and within each interval calculating the survival probability as the number of individuals surviving to the end of the interval divided by the number of individuals at risk for the event at the beginning of the interval. In a setting with no censoring or truncation, the Kaplan-Meier reduces to the complement of the empirical cumulative distribution function, $\hat{F}$. The Kaplan-Meier curve is the non-parametric maximum likelihood estimator for right-censored data, and can be computed in subgroups to assess possible differences in survival [3, 4].

The most cited statistical paper of all time, which as of May 1, 2017 has been cited 44,880 times according to Google Scholar, derives the Cox proportional hazards (PH) model [5]. Instead of directly modeling the survival function S, Cox suggests modeling the hazard function because it is related to the survival function as described above. Specifically, the Cox proportional hazards model allows for the adjustment of measured covariate information in a similar manner to linear regression. If each individual has k covariates measured, say $X = \{X_1, \ldots, X_k\}$, then a Cox PH model estimates $\beta = \{\beta_1, \ldots, \beta_k\}$ in the following equation:

$$\lambda(t \mid X) = \lambda_0(t) \exp(\beta_1 X_1 + \cdots + \beta_k X_k) = \lambda_0(t) \exp(X \cdot \beta).$$

This model assumes that the hazard changes multiplicatively with the covariates. Cox PH models do not require the specification of a functional form for the baseline hazard, making them attractive to researchers who do not want to impose extra assumptions. Extensions of the Cox PH models exist to accommodate many special cases, including time-varying covariates and multiple events per individual [6]. Other types of parametric models, such as accelerated failure time models [7] and additive hazards models [8] have been developed to estimate the hazard in the presence of covariate information.

1.2 Organization of Dissertation

This dissertation is organized into three chapters, each of which concerns an extension of typical survival analysis required when other properties of the data must also be addressed.

Chapter 2 concerns a special case of interval censored data, misclassified group tested current status data. Interval censored data exists when the only available information about the random variable T is that it exists on an interval. Current status data is the extreme form of this censoring, where the only information is about whether T occurred on the interval $[0, t]$, or whether it could occur after time t. Group tested current status data consists of the observation times $t_1, \ldots, t_k$ for each individual in a group of size k, but only contains one test result Y^* for each group, as opposed to the individual test results $y_1, \ldots, y_k$. Here, we propose a non-parametric maximum likelihood estimator for F, the cumulative distribution function of T, in group tested current status data, where each test result is subject to a fixed misclassification rate.

Another issue that arises in survival analysis, or generally any longitudinal analysis, is time-dependent confounding. In longitudinal studies, when the effect of a time-varying exposure on an outcome is the parameter of interest, time-dependent confounding exists when there is a time-varying covariate that influences the outcome, and both influences future levels of the exposure and is influenced by past levels of the exposure. Chapter 3 begins by presenting inverse probability weighting methods to estimate parameters of a marginal structural model. Then, we apply marginal structural modeling with inverse probability weights to the joint estimation of the effects of pre- and post-treatment initiation levels of rheumatoid arthritis disease activity, high-density lipoprotein cholesterol, and low-density lipoprotein cholesterol on the risk of major adverse cardiovascular events in a sample of patients with moderate to severe rheumatoid arthritis who are taking a regimen of tocilizumab, a drug used to treat rheumatoid arthritis.

Chapter 4 addresses the study of weight gain in pregnancy and preterm birth. In human pregnancy, gestational duration is generally 37 to 42 weeks. Preterm births are defined as pregnancies where gestational duration only lasts between viability (22 weeks) and 37 weeks. This can be considered a special case of survival analysis, where we are interested in survival at a particular point on the survival curve, $S(37)$. The relation between weight gain in pregnancy and preterm birth is difficult because a dependence on time is inherently built into the measurement of weight gain in pregnancy. In this chapter, we first review the epidemiology of and statistical issues arising when studying preterm birth and gestational weight gain. Then, we present a study of the relation between gestational weight gain and the black-white disparity in preterm birth rates in the U.S. birth certificate data from 2011-2015. Finally, we consider different ways of incorporating weight gain in pregnancy into statistical analyses of the overall relation between pregnancy weight gain and preterm birth in a cohort of Swedish births from 2008-2014.

Chapter 2

Misclassified Group Tested Current Status Data

2.1 Introduction

Group testing of a binary response has once again become a topic of great interest in the last decade [9–11]. The idea was first introduced in 1943 as a potential cost-saving measure for the detection of syphilis in U.S. army recruits [12]. Group testing reduces the number of tests by allocating, randomly or otherwise, n individuals into J groups of equal size k, and only testing each pooled group once, in order to provide an estimate of the prevalence of a binary characteristic in a population.

More recent work has considered potential issues with group testing, such as dilution effects, non-random group assignment, and misclassification [13–16]. Tu et al. (1994) [17] suggest that if the unknown prevalence of a binary characteristic is sufficiently low and the screening test suffers from misclassification, more precise estimates of the prevalence can be obtained from J group tests than from testing all n individuals separately. The intuition behind this finding is complex. When a test has a rate of misclassification independent of the number of individuals in the pooled sample, performing fewer tests could increase precision of the prevalence estimate due to fewer tests being performed, and thereby leading to less noise in the observations. This is particularly the case when the prevalence is sufficiently small, making it uncommon that two positives will occur in the same group.

The data structure where an individual's binary response corresponds to an underlying time-to-event variable T occurring before an observed screening time C is known as current status data, or interval censoring type I [18, 19]. The nonparametric maximum likelihood estimator of the distribution function, F, of T, for current status data is the pool-adjacent-violators algorithm, although it is only possible to use this estimator if there is sufficient variation in the observed Cs [20, 21].

In this paper, we develop a simple algorithm to compute a nonparametric maximum likelihood estimator of F for group tested current status data, and extend it to settings where the test is subject to misclassification. When misclassification is present, we hypothesize that there will sometimes be substantial gains in precision for values of T at which the prevalence is sufficiently small, as described in Tu et al. (1994) [17] in the case of estimating a single fixed prevalence.

2.2 Notation and Likelihood Function

We assume that the underlying data (prior to grouping) arise from n independent realizations of a bivariate random variable, $\Phi = \{1(T < C), C\}$, where the survival random variable T and screening random variable C follow distribution functions F and G respectively. Throughout, we assume that T and C are independent. The observed data are based on grouping these realizations at random into blocks of size k, where for convenience we assume that n/k is an integer. It is trivial to extend all the results below to situations where block sizes may vary. Thus each original unit corresponds to the j^{th} individual in the i^{th} group, where $i = 1, \ldots, n/k$ and $j = 1, \ldots, k$. The group tested result from the i^{th} group, Δ_i, is the only test result available, whereas individual screening times, C_{ij}, are observed for all participants. Specifically, $\Delta_i = 0$ if and only if $\Phi_{ij} = 0$ for all $j = 1, \ldots, k$, and $\Delta_i = 1$ otherwise. The group test detects the presence of one or more positives in the group, but cannot distinguish between a single, or several, positive Φ_{ij}s. The immediate goal is to estimate the distribution function F.

Due to the assumed independence of T and C we can focus on the conditional likelihood of the data given the observed screening times $\{C_{ij} : i = 1, \ldots, n/k; j = 1, \ldots, k\}$. Since $\text{pr}(\Delta_i = 0 \mid C_{ij} : j = 1, \ldots, k) = \prod_{j=1}^{k} \text{pr}(\Phi_{ij} = 0 \mid C_{ij})$, this conditional likelihood is given by

$$\mathcal{CL} \ = \textstyle\prod_{i=1}^{n/k} \{S(c_{i1}) \times \cdots \times S(c_{ik})\}^{1-\delta_i} \{1 - S(c_{i1}) \times \cdots \times S(c_{ik})\}^{\delta_i}, \tag{2.1}$$

where $S = 1 - F$ is the survival function of T. This conditional likelihood applies to differing methods for selecting the screening times C and assigning the observations to groups for testing. At one extreme, the C values in each group are selected completely at random; at the other end of the spectrum, individuals with common values of C are assigned to the same group. The latter sampling scheme is only fully feasible if the distribution function G is discrete. While the estimation strategy pursued here applies generally, estimation is much simpler with common C values in each group, and asymptotic properties of the estimator are more easily derived. For example, with common values of C in each grouping, with fixed group size k, the likelihood (1) simplifies to that for the standard current status data problem with underlying survival function $S_k(c_i) = S(c_i)^k$. Estimates, and inference, regarding

S_k can then be immediately translated to corresponding statements regarding S itself.

In practice, with a continuous G, it may be advantageous to group individuals with approximately the same value of C. This development assumes a perfect screening test of whether the true group test result was positive, $\Delta_i = 1$, or not. We can extend these ideas to permit misclassification of the test results, which we now denote by Y to distinguish the potentially misclassified test result from the true result Δ. Assume that the test has known sensitivity and specificity, independent of both the screening times, C, and group size, given by $\alpha = \mathrm{pr}(Y = 1 \mid \Delta = 1)$ and $\beta = \mathrm{pr}(Y = 0 \mid \Delta = 0)$ with the assumption that $\alpha + \beta > 1$. Then, the conditional likelihood of the potentially misclassified data, given the observed screening times $\{C_{ij} : i = 1, \ldots, n/k; j = 1, \ldots, k\}$, can be written as

$$\mathcal{CL}(\alpha, \beta) \;=\; \prod_{i=1}^{n/k} \{1 - \alpha + \gamma S(c_{i1}) \times \cdots \times S(c_{ik})\}^{1-y_i} \{\alpha - \gamma S(c_{i1}) \times \cdots \times S(c_{ik})\}^{y_i} \quad (2.2)$$

where $\gamma = \alpha + \beta - 1$.

2.3 An expectation-maximization-pool-adjacent-violators algorithm

Development of the Algorithm

Group tested current status data can be formulated as a missing data problem. First consider the setting without misclassification of test results. While the full set of screening times C_{ij} are observed, only group tested results, Δ_i, are available, whereas a complete data set would include all individual test results, Φ_{ij}. This missing information setting naturally allows use of the expectation-maximization algorithm (Dempster et al., 1977).

To implement the expectation-maximization algorithm, we calculate the expected value of the true individual test result, Φ_{ij}, given the observed value of the group tested result, Δ_i, based on a current estimate of F. These calculations are straightforward when there is no misclassification:

$$E(\Phi_{ij} \mid \Delta_i = 0, C_{i1} = c_{i1}, \ldots, C_{ik} = c_{ik}) \;=\; 0, \quad (2.3)$$

$$E(\Phi_{ij} \mid \Delta_i = 1, C_{i1} = c_{i1}, \ldots, C_{ik} = c_{ik}) \;=\; F(c_{ij})\{1 - S(c_{i1}) \times \cdots \times S(c_{ik})\}^{-1}. \quad (2.4)$$

For misclassified data with sensitivity α and specificity β, computing the expected value of an individual true disease status Φ_{ij} given the potentially misclassified observed group test result Y_i becomes slightly more complicated, see Supplementary Material. Letting $\gamma = \alpha + \beta - 1$ this step becomes:

$$E(\Phi_{ij} \mid Y_i = 1, C_{i1} = c_{i1}, \ldots, C_{ik} = c_{ik}) \;=\; \alpha F(c_{ij})\{\alpha - \gamma S(c_{i1}) \times \cdots \times S(c_{ik})\}^{-1}, (2.5)$$

$$E(\Phi_{ij} \mid Y_i = 0, C_{i1} = c_{i1}, \ldots, C_{ik} = c_{ik}) \;=\; \frac{(1-\alpha)F(c_{ij})}{(1-\alpha) + \gamma S(c_{i1}) \times \cdots \times S(c_{ik})}. \quad (2.6)$$

For the maximization step, we simply use a weighted version of the pool-adjacent-violators algorithm on the full data set ϕ_{ij}, where $\phi_{ij} = 0$ with weight 1 if $\delta_i = 0$, per (3). On the other hand, per (4), if $\delta_i = 1$, $\phi_{ij} = 1$ with weight given by the right-hand side of (4), together with additional observations $\phi_{ij} = 0$ with weight given by one minus the right-hand side of (4). The complete algorithm is thus described as follows:

Step 1. *Initialize values of $f_{ij}^{(0)} = \hat{F}^{(0)}(c_{ij})$ for each individual and set a threshold τ for convergence.*

Step 2. *Expectation. For each individual $j \in \{1, \ldots, k\}$ in group i, calculate the probability that each individual tested positive, f_{ij}^*, given their group's test result. For perfectly classified results, δ_i, use*

$$f_{ij}^* = \begin{cases} f_{ij}^{(0)} \left\{1 - \prod_{J=1}^{k}(1 - f_{iJ}^{(0)})\right\}^{-1}, & \delta_i = 1, \\ 0, & \delta_i = 0. \end{cases} \tag{2.7}$$

For group tested results subject to misclassification, y_i, with sensitivity α and specificity β such that $\gamma = \alpha + \beta - 1$, use

$$f_{ij}^* = \begin{cases} \alpha f_{ij}^{(0)} \left\{\alpha - \gamma \prod_{J=1}^{k}(1 - f_{iJ}^{(0)})\right\}^{-1}, & y_i = 1, \\ (1-\alpha) f_{ij}^{(0)} \left\{1 - \alpha + \gamma \prod_{J=1}^{k}(1 - f_{iJ}^{(0)})\right\}^{-1}, & y_i = 0. \end{cases} \tag{2.8}$$

Step 3. *Maximization. Use the group tested results, δ_i or y_i, as the observations for each individual, and the probabilities from step 2 as the weights in the weighted pool-adjacent-violators algorithm to calculate updated estimates of $f_{ij}^{(1)} = \hat{F}^{(1)}(c_{ij})$.*

Step 4. *Repeat steps 2 and 3 using the estimate of $\hat{F}$ from step 3 as initial values for step 2 until convergence, for example until*

$$\sum_{i=1}^{n/k} \sum_{j=1}^{k} \left\{\hat{F}^{(t+1)}(c_{ij}) - \hat{F}^{(t)}(c_{ij})\right\}^2 < \tau.$$

It is important to run the algorithm under several choices of starting values, not only to reduce the possibility of converging to a local extrema, but also to discover different non-unique versions of the nonparametric maximum likelihood estimator. We recommend choosing a large set of random starting values of F at the observed set of C_{ij}s by generating random Uniform(0, 1) values ordered so that the starting values are monotonically increasing with C_{ij}s.

Comments Regarding Asymptotics

Asymptotic results for standard current status data are non-standard. The nonparametric maximum likelihood estimator is known to be consistent, although converging only at the rate $n^{1/3}$, but with a non-Gaussian limiting distribution known as Chernoff's distribution [21] for situations where the monitoring time distribution, G, is continuous; Banerjee (2012) [22] provides a concise discussion of this result. Rather than using Wald-type pointwise confidence intervals derived from this limit, Banerjee & Wellner [23, 24] suggest the use of a likelihood ratio approach to construct confidence bands.

On the other hand, when G has finite support, the likelihood is parametric, since F can then only be estimated at this finite number of support points: the observed censoring times. As expected from this observation, the nonparametric maximum likelihood estimator now converges to a Gaussian limit at rate $n^{1/2}$ with the asymptotic variance at a specific monitoring time C_0 given simply by $F(C_0)\{1 - F(C_0)\}\{g(C_0)\}^{-1}$, which is straightforward to estimate using the obvious plug-in estimates [25, 26]. The hybrid problem where the number of support points grows with the sample size is discussed beautifully in Tang, Banerjee & Kosorok (2012) [27]. Sal y Rosas & Hughes (2011) [28] describe the inversion of a likelihood ratio test to obtain pointwise confidence intervals for F when the data are subject to misclassification.

These results can only be directly applied to the group testing scenario in the simplest situations. For the extreme situation with only one monitoring time, estimation of $F(C_0)$ reduces to the simple estimation of prevalence. This scenario has been studied extensively in the group testing literature with misclassification; for example, Tu et al. (1994) [17]provides asymptotically normal confidence intervals with convergence rate $n^{1/2}$. Generalizing slightly, the situation with a finite support for C, and with no misclassification, simplifies to the case considered by Yu et al. (1998) [25] if individuals within a group all share a common value of C. In this case, $\mathrm{pr}(\Delta = 1 \mid C) = 1 - S(C)^k$ so that asymptotic results for the nonparametric maximum likelihood estimator applied to the group tested data immediately apply to the plug-in estimator for S, or F, at the finite number of screening times C using the delta method. We anticipate that this will extend straightforwardly in the presence of misclassification and also suggest that the use of the bootstrap will be effective here.

Even with a finite number of monitoring times, the situation becomes more complex when screening times are randomly assigned to the groups. This is clear even with only two monitoring times and pair groupings at random. Further, there are as yet no known asymptotic results for the nonparametric maximum likelihood estimator of §2.3 with a continuous screening time distribution, although we anticipate that convergence will remain at rate $n^{1/3}$.

2.4 Elementary Example

An Analytic Solution

For illustration, consider a simple example in a setting without misclassified test results, where there are two groups each containing two individuals; that is, $n = 4$ and $k = 2$. There are twelve possible combinations of group assignments and test results, corresponding to three different possible sets of pair assignments with each pair having two possible test outcomes. Consideration of the conditional likelihood (1) reveals a simple solution in all but one of these cases; we focus on the remaining case with the grouping as shown in Figure 2.1 with $\Delta_1 = 1$ and $\Delta_2 = 0$.

The conditional likelihood (1) in this setting is:

$$\mathcal{CL}_4 = \{1 - S(c_1)S(c_3)\}\, S(c_2)S(c_4).$$

It is immediate that the nonparametric maximum likelihood estimator must have $\hat{S}(c_1) = \hat{S}(c_2)$ and $\hat{S}(c_3) = \hat{S}(c_4)$. Thus, the nonparametric maximum likelihood estimator is not unique but achieved by any set of $\{\hat{S}(c_1), \ldots, \hat{S}(c_4)\}$ with $\hat{S}(c_1) = \hat{S}(c_2)$, $\hat{S}(c_3) = \hat{S}(c_4)$, and $\hat{S}(c_2)\hat{S}(c_3) = 0.5$. We show how the expectation-maximization-pool-adjacent-violators algorithm converges to one such solution, the specific value depending directly on the starting values for $\hat{F}^{(0)}(c_i) = 1 - \hat{S}^{(0)}(c_i)$.

Given an initial set of probabilities: $\hat{F}^{(0)}(C_1) = f_1$, $\hat{F}^{(0)}(C_2) = f_2$, $\hat{F}^{(0)}(C_3) = f_3$, and $\hat{F}^{(0)}(C_4) = f_4$ such that $f_1 \leq f_2 \leq f_3 \leq f_4$, the first step of the algorithm calculates the expectation of each of the initial conditional probabilities, f_i^* for $i = 1, \ldots, 4$, as given in (7) and (8); that is the probability that an individual was positive given the known group tested result. For two of these probabilities, in a setting without misclassification, this calculation is trivial: the pair tested negative so neither of the individuals was positive. Thus we can set $f_2^* = f_4^* = 0$. For the pair that tested positive, this calculation follows directly from (7):

$$f_1^* = \text{pr}(T_1 \leq C_1 \mid \Delta_1 = 1) = f_1\{1 - (1 - f_1)(1 - f_3)\}^{-1} = f_1\{f_1 + f_3 - f_1 f_3\}^{-1},$$
$$f_3^* = \text{pr}(T_3 \leq C_3 \mid \Delta_1 = 1) = f_3\{1 - (1 - f_1)(1 - f_3)\}^{-1} = f_3\{f_1 + f_3 - f_1 f_3\}^{-1}.$$

The next step of the algorithm is to make these f_j^*s monotonic, recalling that $f_2^* = f_4^* = 0$, using the pool-adjacent-violators algorithm. This yields the updated estimates of F:

$$\hat{F}^{(1)}(C_1) = \hat{F}^{(1)}(C_2) = f_1^*/2 = f_1\{2(f_1 + f_3 - f_1 f_3)\}^{-1}, \quad (2.9)$$
$$\hat{F}^{(1)}(C_3) = \hat{F}^{(1)}(C_4) = f_3^*/2 = f_3\{2(f_1 + f_3 - f_1 f_3)\}^{-1}. \quad (2.10)$$

These steps are then iterated until a determination of convergence based on comparing, say, the sum of the squared differences between $\hat{F}^{(m)}$ and $\hat{F}^{(m+1)}$ at each observed C to a pre-specified threshold τ.

Multiple Convergence Values

As we demonstrated in §2.4, the initial values for the pair who tested negative, $f_2 = \hat{F}^{(0)}(C_2)$ and $f_4 = \hat{F}^{(0)}(C_4)$, are not relevant to the update step in our expectation-maximization-pool-adjacent-violators algorithm. Therefore, when discussing convergence of the algorithm we will only consider initial values for $\hat{F}^{(0)}(C_1) = f_1$ and $\hat{F}^{(0)}(C_3) = f_3$.

In all settings where $f_1 = f_3 = f$, the update step given by (9) and (10) becomes

$$\hat{F}^{(1)}(C_1) = \hat{F}^{(1)}(C_3) = \hat{f} = 1\{2(2-f)\}^{-1}.$$

Thus, at convergence, $f = \{2(2-f)\}^{-1}$, so that the algorithm converges to $f = 1 - 2^{-1/2}$, the only solution in [0,1]. This is, of course, also expressed as $\hat{S}(C_j) = 2^{-1/2}$ for $j = 1, \ldots, 4$.

For any other set of starting values, the ratio of f_1/f_2 remains unchanged by the iterations. We can thus write $f_1 = rf_3$ where $0 < r < 1$, and r remains fixed as determined by the starting values for f_1 and f_3 . At convergence, equation (9) then simplifies to

$$f_3 = f_3 \left\{2\left(rf_3 + f_3 - rf_3^2\right)\right\}^{-1}.$$

Convergence thus occurs when $rf_3 + f_3 - rf_3^2 = 1/2$. After an application of the quadratic formula, this simplifies to

$$f_3 = \left\{r + 1 - \left(r^2+1\right)^{1/2}\right\}(2r)^{-1}, \tag{2.11}$$

the only feasible solution. It immediately follows that at convergence, $f_1 = \left\{r + 1 - (r^2+1)^{1/2}\right\}/2$, to respect $(1-f_1)(1-f_3) = 0.5$, as noted in §2.4.

This simple case demonstrates the non-uniqueness of the nonparametric maximum likelihood estimator, with the algorithm converging to a specific solution for $\hat{F}$ determined by the ratio of the starting values of the F at C_1 and C_3. When using this algorithm in an applied setting, we suggest repeating it many times, using a different set of randomly drawn starting values each time, and then computing the likelihood function to identify as many different unique solutions to the optimization as possible.

2.5 Simulations

Design of Simulations

We carry out two series of simulations to examine the behavior of the expectation- maximization-pool-adjacent-violators algorithm for group tested data as it compares to the pool-adjacent-violators algorithm, the nonparametric maximum likelihood estimator for individual-level current status data [29]. We consider two scenarios, one where the tests are subject to no misclassification, and a second, where the test is subject to misclassification with known, constant error rates. In the latter case, the comparative estimator for misclassified individual-level current status data was derived by McKeown & Jewell (2010) [30]. We consider both continuous and discrete independent screening times. The former are described and discussed below whereas the latter are available in Supplementary Material.

Each simulation is characterized by a set of fixed parameters: n, the number of individuals; k, the group size; and α and β, the sensitivity and specificity of the screening test, respectively. We set $\alpha = \beta = 1$ in scenarios without misclassification. We first simulate traditional current status data for each individual from the distribution of the true event times, F, and the censoring distribution, G. Each run of the simulations begins by simulating data of sample size n at the individual level, and subsequently assigning individuals to groups randomly.

The distribution F of the event times T is Weibull with shape and scale parameters 4 and 25 respectively; here F has mean 22.7 and variance 40.4. For the perfectly classified test simulations, the screening distribution G for C is Uniform(0, 36), allowing almost all of the distribution F to be identified. The necessary binary datum Φ is then determined from the generated individual values of T and C. The values of Δ, the group tested results, follow immediately from the values of Φ from each individual in the group, as described in §2.2. Each simulation is performed 1,000 times in 6 different settings - $n = (1000, 5000)$ and groupings of sizes $k = (2, 5, 10)$.

For misclassified test results, we are most interested in examining performance of the expectation-maximization-pool-adjacent-violators estimator in the left tail of F, where false positive test results could have the largest effect on the estimate of F (Tu et al., 1994). Thus, while F remains the same Weibull distribution, we now take C to be Uniform (0, 14) to ensure that $F(t) \leq 10\%$. Here we select a single sample size $n =$ 5,000 in 12 different settings with group sizes k in $\{2, 5, 10\}$ and misclassification rates of $\alpha = \beta$ in $\{0.8, 0.9, 0.95, 0.99\}$. In these simulations, the observed misclassified data are obtained by subjecting first each individual test result Φ to misclassification under the specified test characteristics; second the group tested outcome Y is separately generated by misclassifying the corresponding group test result Δ. Here we have used the same test classification probabilities, assuming independence between the group size and the error rates of the testing procedure.

In each run of both simulations, perfectly classified versus misclassified data, we compute both the appropriate expectation-maximization-pool-adjacent-violators algorithm for the group tested data and the appropriate pool-adjacent-violators algorithm for individual data. To select initial values for the expectation-maximization-pool-adjacent-violators algorithm, we first draw n values uniformly on the range [0, 1] and sort them from smallest to largest. We then order the observations so the Cs are monotonically increasing, and match the ordered initial probabilities to the ordered data. Although we noted earlier that for a specific application we recommend examining multiple starting values, here we opt to only randomly select one set of initial values for each simulated dataset, thereby only achieving one of potentially many possible nonparametric maximum likelihood estimate.

The averages of the estimates of F from each algorithm over the 1,000 runs are calculated for each t in the support of G. To calculate the estimate of F at a value of C not observed in a specific simulation, we assume left-continuity of both estimators in situations where this is not imposed by monotonicity. To provide a sense of the variability of each estimator, we also calculate the 2.5^{th} and 97.5^{th} quantiles of the estimates over the 1,000 simulations. For the second set of simulations, we use these quantities to compute a measure of pseudo-relative efficiency, the ratio of the widths of these 95% Monte Carlo quantile intervals: $\{(q_{97.5} - q_{2.5})(\text{group})\ \} \ / \ \{(q_{97.5} - q_{2.5})(\text{individual})\}$. It is less relevant to focus on variances of the simulated estimates since we hypothesize this estimator does not converge to a Gaussian distribution, nor at rate $n^{1/2}$.

The Supplementary Material contains results from two simulations in samples of size $n = 10{,}000$, with 10 fixed, equal frequency screening times, C, and true event probabilities at each screening time fixed at $\{0.005, 0.01, \ldots, 0.05\}$. In the first simulation, we randomly group individuals within values of C to allow for the presentation of asymptotically normal confidence intervals, as described in §2.3, and in the second, we group across screening times and again present the widths of the 95% Monte Carlo quantile intervals.

Results: Perfectly Classified Data

Figure 2.2 displays the results from applying the expectation-maximization-pool-adjacent-violators algorithm and the pool-adjacent-violators algorithm to data generated in the six simulations where there is no misclassification of the test results. These simulations show that the finite sample bias is small except perhaps when the group size is large, e.g. $k = 10$, and $F(t)$ is small. Even then, this bias declines systematically as the sample size increases. As anticipated, in all situations, the bias is also smaller for the estimator based on individual test results. Similarly, and also to be expected, the latter is also more precise, though the gain decreases for larger sample sizes and smaller k. This being said, the group tested estimator stands up remarkably well given that the screening costs are reduced by 50%, 80% and 90% when $k = 2, 5, 10$, respectively, assuming that costs are proportional to the number of tests.

Because the asymptotic properties of this expectation-maximization-pool-adjacent-violators algorithm are currently unknown, to demonstrate variability in the estimates we present the 95% Monte Carlo quantile interval as dashed and dotted lines in Fig. 2.2. The width of this interval for the pool-adjacent-violators algorithm from individual data is always smaller than that from the expectation-maximization-pool-adjacent-violators algorithm applied to group tested data. This is to be expected, as there is no misclassification in these simulations. Smaller grouping sizes k in the expectation-maximization-pool-adjacent-violators algorithm provide 95% quantile intervals more similar to those estimated from individual data, and as n increases for fixed k, the width of the 95% quantile interval decreases. Overall, Fig. 2 demonstrates that the expectation-maximization-pool-adjacent-violators algorithm provides an unbiased estimate of the true underlying distribution, F.

Results: Misclassified Data

Figures 2.3 and 2.4 present results from the twelve simulations in settings with $n = 5{,}000$ individuals and varying group sizes and misclassification rates. Figure 2.3 demonstrates that the percent relative bias of both of these estimators in these finite samples is large, e.g. greater than 100%, for estimates of $F(t)$ that are very small, e.g. less than 0.002, and very close to 0 for estimates of $F(t)$ that are greater than 0.02, even at large group sizes with high misclassification rates. Although the individual-based estimator is less biased at small group sizes and low misclassification rates, we do see similar or lower amounts of bias from the group testing estimator at higher misclassification rates, e.g., $\alpha = \beta = 0.8, 0.9$, particularly with the larger grouping sizes $k = 5, 10$ and at lower values of T. Ultimately, the shapes of the finite sample relative bias curves for both of these estimators are very similar so, at the very least, grouping does not introduce substantial amounts of additional bias.

With regard to variability, the comparison of the the widths of the 95% Monte Carlo quantile intervals associated with both estimators, as shown in Fig. 2.4, demonstrates considerable advantage for our estimator from group tested data at low t and high levels of misclassification. For example, $T = 10$ corresponds to a true prevalence of 2.5%. If a test is subject to 10% misclassification, $\alpha = \beta = 0.9$, then test results from data grouped into pools of size 10 will provide a more or equally precise estimate of $F(t)$ for $T < 10$ than data from individual tests. This implies that if the cumulative failure rate in question is less than 2.5%, a testing procedure that involves groups of size 10 will cost 90% less than testing everyone individually, and will result in a less-biased and more precise estimate of $F(t)$ in this range. In general, the specific threshold t below which such precision gains can be expected depends on both the group size and missclassification rate as suggested by Tu et. al (1994) [17] for estimation of a single fixed prevalence.

Supplementary Material includes results from simulations of group tested current status on a grid, with grouping done solely within common observation times, which more easily ensures a sufficiently small maximum value of F. As seen in Tu et. al (1994) [17], we see a reduction in the size of 95% confidence intervals as the group size increases, and separately a reduction in the size of the 95% confidence intervals as the misclassification rates decrease. Additionally, there appears to be no substantial increase in bias as group size increases.

2.6 Application to Hepatitis C Data

To investigate the performance of our estimator in a practical setting, we use publicly available data from the 2014 U.S. Birth Data File, created by the National Center for Health Statistics, to investigate the age-at-incidence distribution for Hepatitis C in non-Hispanic white women of childbearing age. These data include all such women of ages 13-40 who gave birth in 2014. We are thus making the tacit assumption that women who gave birth are a representative sample of women of the same ages that could have given birth with regard to their risk of infection with hepatitic C. This is not exactly correct but seems to present a reasonable approximation, at least for sexually active women. Of the 1,981,521 eligible women, we randomly sampled 10%, creating our sample of $N = 197{,}840$ observations, for greater ease of illustration and computation. The data include the mother's age in years, and her Hepatitis C status at the birth of her child. Of the $N = 197{,}840$ women in our illustration, only 901 of them tested positive for Hepatitis C, a cumulative incidence of 0.46%. When accounting for potential misclassification of these test results, we used the sensitivity, $\alpha = 0.987$ and specificity, $\beta = 0.999$, associated with the most commonly used test for Hepatitis C: an enzyme immunoassay test. Although hepatitis C can be spread through sexual contact, it is primarily transferred through blood transmission, and an increase in the incidence of hepatitis C after age 25 would imply that people are beginning to or continuing to engage in risky drug behavior.

These data are based on individual blood testing of each mother separately. To illustrate the methods here, we consider group testing of pooled blood samples, representing a potentially enormous savings of test costs depending on the size of the grouping used. These savings persist even if specific infected individuals need to be identified. As discussed above, given the low misclassification results we anticipate some loss of accuracy in estimating the prevalence that may nonetheless be worth the considerable cost reduction. We created artificial group test results two ways: 1) by assigning the data into groups of sizes 2, 5, and 10 within values of mother's age, and 2) by randomly assigning the data into groups of sizes 2, 5, and 10. Then, each group test was assigned a positive result if at least one individual test was positive. For grided group assignments, we computed point estimates and 95% confidence intervals adjusted for misclassification using the method described in §2.3. For random group assignments, we computed the adjusted pool-adjacent violators algorithm on the individual test results, and, for comparison, the expectation-maximization pool-adjacent-violators algorithm on the group tested results.

Figure 2.5 displays the results from the estimates from individual and group tested results with groups of sizes 2, 5, and 10 in a setting where group assignment is done with women who are the same age. These results are satisfying, as they lead to the same public health implications. Although the estimates are slightly different, they increase with group size, the major jumps in the estimates occur at ages 19 and 21 in each of these group sizes. Here we can be fairly certain that any public health intervention to potentially reduce the public health burden due to Hepatitis C infection would best occur during adolescence, ideally before risky behaviors such as drug use and unprotected sexual activity begin. In this example, major cost reductions could occur, assuming costs proportional to number of tests, by reducing the number of tests performed, without changing the conclusions from the analysis.

Figure 2.6 displays the results of the estimates from individual and group tested results with groups of sizes 2, 5, and 10 in a setting where group assignment is done completely at random. Unlike the estimates in Fig. 2.5 from data grouped according to maternal age, here the estimates from data in each group size provide slightly different implications. The results from the individual tests suggest an essentially flat cumulative incidence of hepatitis C after age 21 having reached a cumulative incidence of approximately 0.38%. This has significant implications for a public health intervention, potentially indicating, for example, that any future hepatitis C vaccination would be most effective if implemented during even late adolescence; no vaccine currently exists although several candidates are under development. The group tested results from groups of size 2 also support this conclusion, although they suggest that the cumulative incidence does not increase after age 19. However, the results from groups of both sizes 5 and 10 tell a slightly different story. While these estimates increase to a cumulative incidence of roughly 0.4% before age 20, they then both continue to increase with age to somewhere between 0.45–0.55% by age 40, suggesting that a substantial fraction of hepatitis C infections occur post-adolescence.

Because these estimates seem to imply public health interventions at different times in life, it is important to consider which estimate is most reliable in this particular setting. As noted, there is very little misclassification in the testing procedure so that we would expect that the results from the adjusted pool-adjacent violators algorithm based on individual data would be more accurate, albeit obtained at significantly higher cost. However, the pool-adjacent-violators algorithm adjusted for misclassification has a limitation - it automatically estimates cumulative incidences that are less than $1 - \beta$ as 0. Because the cumulative incidences at the early ages are very low - less than 0.5%, if we had set $\beta \leq 0.995$ in this application, our estimate from the individual data adjusted for misclassification would have been 0 at all ages. This suggests a potential issue with individual test results that may not be as present in group tested results.

2.7 Discussion

In this paper we propose a modified expectation-maximization algorithm to estimate a distribution function from data obtained by group tested current status screening with test misclassification. Simulations show that the estimator based on group tested data adds relatively little additional small-sample bias compared to an estimator based on individual data, at far lower cost, although this conclusion necessarily requires a larger n as the grouping size

k increases. Additionally, when substantial misclassification is present, and $F(t)$ is low, estimates from the expectation-maximization-pool-adjacent-violators algorithm done in groups of size 5 or larger may be less biased with improved precision, although inferential properties for this procedure need further development. This raises the option that a significantly less expensive testing procedure might result in a less biased and more precise estimate for the left tail of F.

In the presence of misclassification, these observations suggest possible hybrid grouping strategies that may improve precision at low values of $F(t)$ and maintain perfomance at higher levels, all in comparison to individual tests whose costs are far greater. That is, if possible, if the screening times are known in advance of pooling, it will likely be advantageous to first group individuals according to the observed Cs, and then use larger group sizes at the smaller values of C and decrease the group size as C increases, even to individual tests. Simulations to examine variations of these possibilties are currently underway. As noted earlier, when individuals in a group have similar C values, it is possible to also use an approximate individual group tested current status estimator by treating all C values in the group as the same.

There are a number of important extensions to these results. As noted, the pool-adjacent-violators estimator for classic current status data converges at a rate of $n^{1/3}$ [21] with a non-standard asymptotic limit. We conjecture that the same asymptotics will hold for the group tested estimator, although this remains to be established. In practice, in a setting with misclassified individual current status data, the m-out-of-n bootstrap [30] has been shown to provide one method of obtaining valid inference procedures. We look forward to further theoretical progress in this area.

It is natural to anticipate that misclassification rates may depend on the group size. This may occur, for example, if the screening test is more sensitive to detecting a positive group when there are more individual positives in the pool, related to the so-called dilution effect [13, 31]. Second, covariate-adjusted regression analysis has been a primary focus of the statistical literature on group testing [32–35]. In addition, in many applications, interest focuses on regression effects or group comparisons of time-to-event properties rather than estimation of the underlying distribution function itself, often through use of standard multiplicative or additive regression models. Such regression models has been widely studied for individual current status data (see [19]). Future work will examine the use of additive hazard regression models for group tested current status data.

Figures

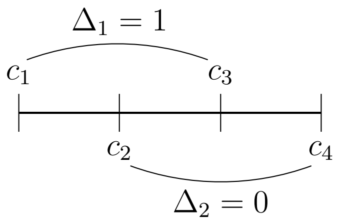


Figure 2.1: Elementary example of data configuration with two groups, each of size 2, where the first group has tested positive and the second group has tested negative.

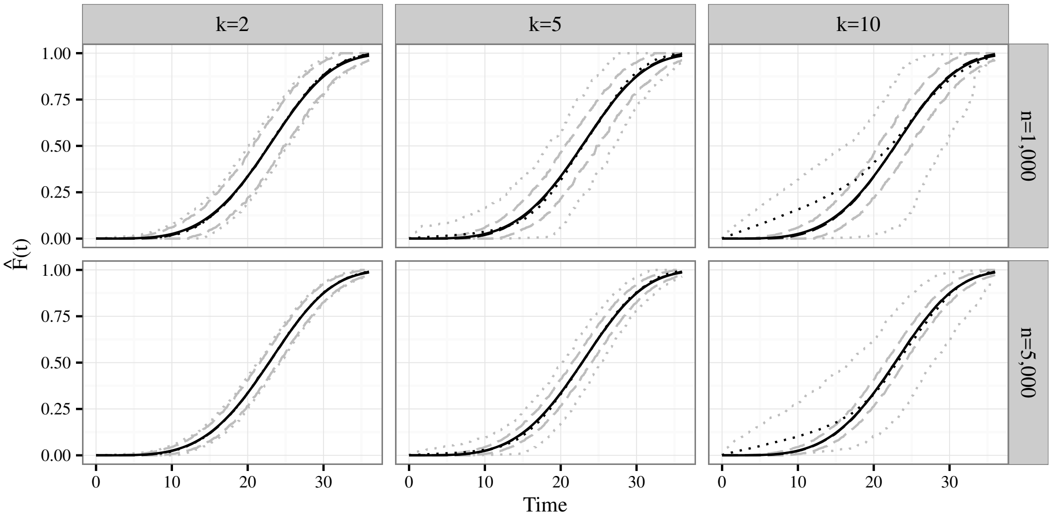


Figure 2.2: Results from six simulations of estimation of F, with 1,000 runs each, showing results for varying sample sizes n and group sizes k. The black lines are the average estimates of $\hat{F}(t)$ over the 1,000 simulations with the solid line representing the true cumulative distribution function, Weibull(4, 25), and the dashed and dotted lines respectively representing the estimates from the pool-adjacent-violators algorithm and the expectation-maximization-pool-adjacent-violators algorithm. The grey lines are the 2.5^{th} and 97.5^{th} quantiles from the simulation runs for each estimator, using the same line types.

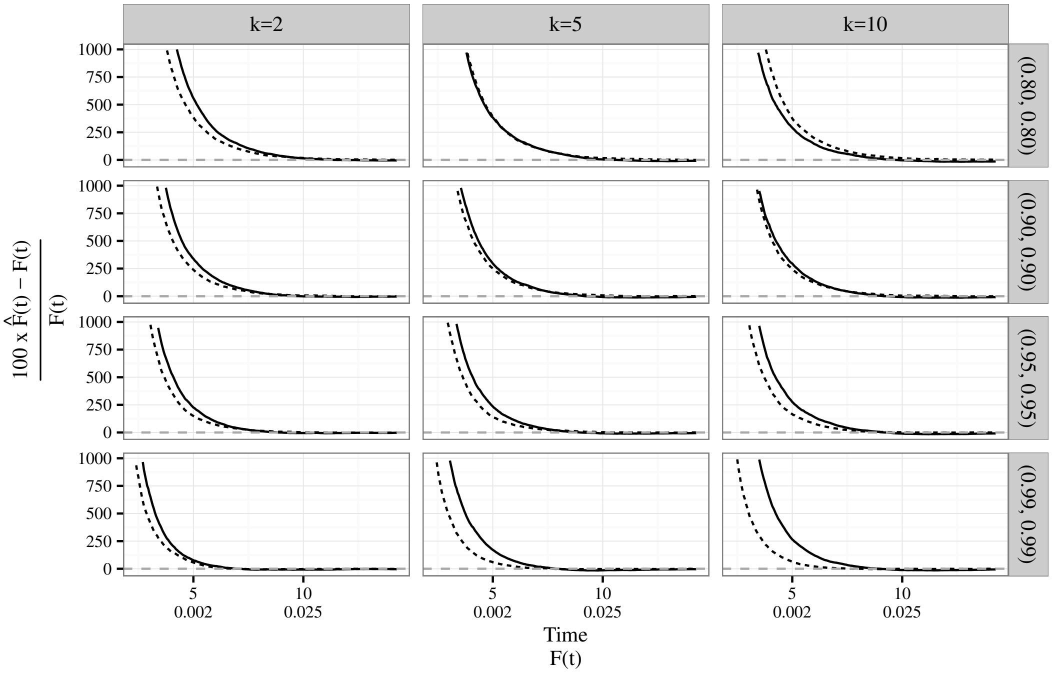


Figure 2.3: Graphical representation of the finite sample percent relative bias from twelve simulations replicated 1,000 times with 5,000 individuals each, based on varying group sizes k and misclassification rates (α, β), the latter denoted on the right hand column. The solid black line displays results for the expectation-maximization-pool-adjacent-violators algorithm for group tests, and the dashed black line represents the pool-adjacent-violators algorithm for misclassified individual test data. The dashed grey line represents the reference of zero percent bias.

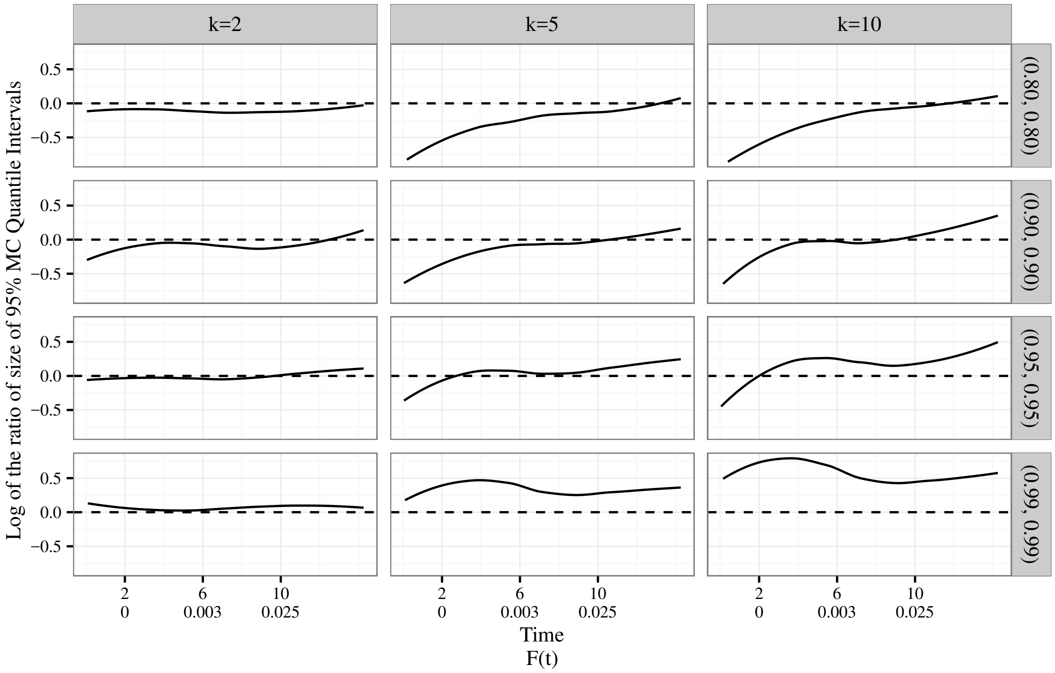


Figure 2.4: Log of the pseudo-relative efficiency of the expectation-maximization-pool-adjacent-violators algorithm and the adjusted pool-adjacent-violators algorithm from twelve simulations with 1,000 runs with 5,000 individuals each, based on varying grouping sizes k and misclassification rates (α, β), the latter denoted on the right hand y-axis. The solid black line is a lowess curve showing the overall trend in pseudo-relative efficiency as t increases. The dashed black line represents equal width 95% Monte Carlo quantile intervals for reference; if the solid black line is below 0, the width of the expectation-maximization-pool-adjacent-violators 95% Monte Carlo quantile interval is smaller than that from the individual test pool-adjacent-violators algorithm.

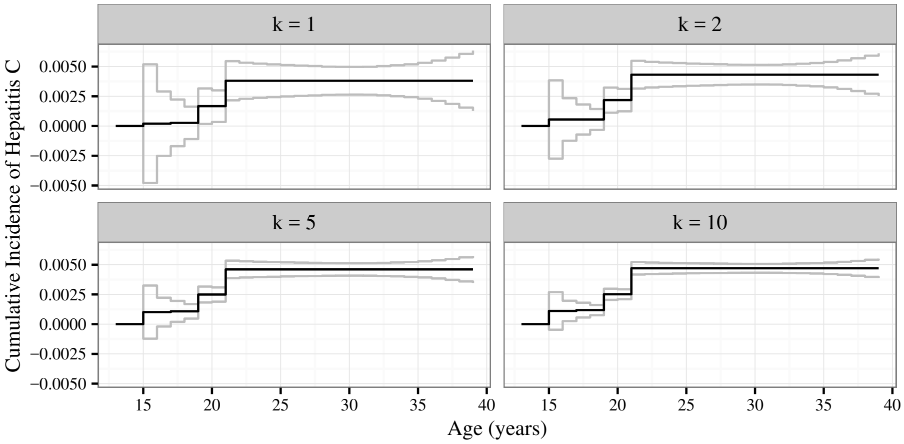


Figure 2.5: Four estimates of the cumulative incidence of hepatitis C in non-Hispanic white, child-bearing women ages 13-40 living in the U.S. in 2014 when grouping is assigned within common values of age. Grouping sizes considered were $k = 1$, 2, 5, and 10 The solid black line is the estimate from the individual or group tested results, and the solid grey lines represent the upper and lower bounds for 95% confidence intervals.

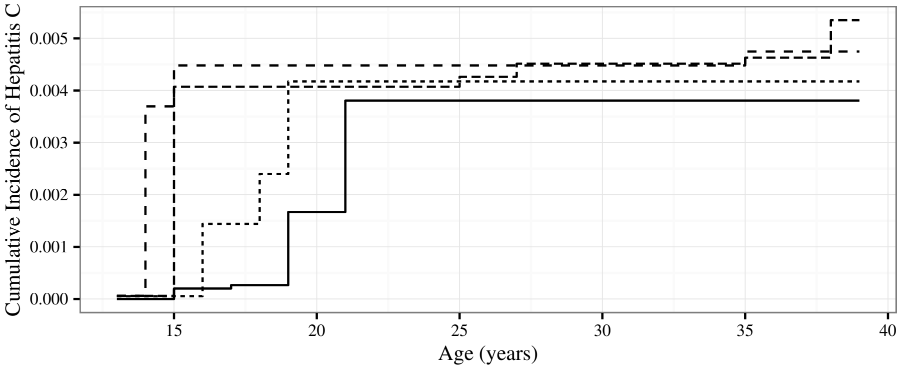


Figure 2.6: Four estimates of the cumulative incidence of hepatitis C in non-Hispanic white, child-bearing women ages 13-40 living in the U.S. in 2014 when group testing assigned at random. The solid line is the pool-adjacent violators estimate from the individual test results, and the dotted, short dashed, and long dashed lines are the estimates from the expectation-maximization pool-adjacent-violators algorithm from artificially grouping the individual test results into groups of sizes 2, 5, and 10, respectively.

Supplementary material

Derivation of Expectation Step under Misclassification

In this section, we derive the expectation step for the expectation-maximization-pool-adjacent-violators algorithm if the group testing procedure is subject to misclassification. Assume that the group assignment algorithm generates groups of size k, and that the sensitivity and specificity, α and β, of the test are known and constant, regardless of group size. The misclassified group tested result Y_i is known, and we want to know what the probability is of an individual in the group being a true positive, $\text{pr}[\Phi_{ij} = 1 \mid Y_i, c_{i1}, \ldots, c_{ik}]$.

Statement. Define $\gamma = \alpha + \beta - 1$ for $\alpha, \beta > 0.5$. The expected value of a true individual test result Φ_{ij} given the potentially misclassified group tested result Y_i and the individual censoring times $c_{i1}, \ldots, c_{ik}$ is:

$$\begin{aligned} E(\Phi_{ij} \mid Y_i = 1, c_{i1}, \ldots, c_{ik}) &= \frac{\alpha F(c_{ij})}{\alpha - \gamma S(c_{i1}) \times \cdots \times S(c_{ik})} \\ E(\Phi_{ij} \mid Y_i = 0, c_{i1}, \ldots, c_{ik}) &= \frac{(1-\alpha) F(c_{ij})}{(1-\alpha) + \gamma S(c_{i1}) \times \cdots \times S(c_{ik})}. \end{aligned}$$

Here, if we substitute $\alpha = \beta = 1$ into these equations, we get the "E" step from perfectly classified data, as seen as (3) and (4) in the main paper.

Derivation. Here we only consider one group, so we will omit the "i" subscripts everywhere.

We begin with the case where $Y_i = 1$:

$$\begin{aligned} E(\Phi_j \mid Y = 1, c_1, \ldots, c_k) &= \text{pr}(\Phi_j = 1 \mid Y = 1, c_1, \ldots, c_k) \\ &= \frac{\text{pr}(Y = 1 \mid \Phi_j = 1, c_1, \ldots, c_k)\text{pr}(\Phi_j = 1 \mid c_1, \ldots, c_k)}{\text{pr}(Y = 1 \mid c_1, \ldots, c_k)} \end{aligned}$$

Breaking this into smaller pieces, we first see that the conditional individual probability of a true positive, by definition is $\text{pr}(\Phi_j = 1 \mid c_1, \ldots, c_k) = F(c_j)$, and the probability of a positive misclassified group test result, given that one individual in the group is positive is simply the sensitivity, $\text{pr}(Y = 1 \mid \Phi_j = 1, c_1, \ldots, c_k) = \text{pr}(Y = 1 \mid \Delta = 1, c_1, \ldots, c_k) = \alpha$. Remember that Φ_j is the true individual test result, and Δ is the true grouped test result, or $\Delta = 1 - \mathbf{1}(\Phi_1 = \cdots = \Phi_k = 0)$.

The probability of a positive misclassified test becomes a bit trickier:

$$\begin{aligned} \text{pr}(Y = 1 \mid c_1, \ldots, c_k) &= \text{pr}(Y = 1 \mid \Delta = 1, c_1, \ldots, c_k)\text{pr}(\Delta = 1 \mid c_1, \ldots, c_k) \\ &\quad + \text{pr}(Y = 1 \mid \Delta = 0, c_1, \ldots, c_k) + \text{pr}(\Delta = 0 \mid c_1, \ldots, c_k) \\ &= \alpha\{1 - S(c_1) \times \cdots \times S(c_k)\} + (1-\beta)S(c_1) \times \cdots \times S(c_k) \end{aligned}$$

Putting it all together,

$$\begin{aligned} E(\Phi_{ij} \mid Y_i = 1, c_{i1}, \dots, c_{ik}) &= \frac{\alpha F(c_j)}{\alpha\{1 - S(c_1) \times \cdots \times S(c_k)\} + (1-\beta)S(c_1) \times \cdots \times S(c_k)} \\ &= \frac{\alpha F(c_j)}{\alpha - \gamma S(c_1) \times \cdots \times S(c_k)} \end{aligned}$$

where $\gamma = \alpha + \beta - 1$.

The case where $Y_i = 0$ is very similar. We again begin by cleverly applying Bayes rule:

$$\begin{aligned} E(\Phi_j \mid Y = 0, c_1, \dots, c_k) &= \mathrm{pr}(\Phi_j = 1 \mid Y = 0, c_1, \dots, c_k) \\ &= \frac{\mathrm{pr}(Y = 0 \mid \Phi_j = 1, c_1, \dots, c_k)\mathrm{pr}(\Phi_j = 1 \mid c_1, \dots, c_k)}{\mathrm{Pr}(Y = 0 \mid c_1, \dots, c_k)} \end{aligned}$$

and realize that all the quantities in the expression are identical to or clear complements of quantities calculated in the case where $Y_i = 1$.

- $\mathrm{pr}(\Phi_j = 1 \mid c_1, \dots, c_k) = F(c_j)$
- $\mathrm{pr}(Y = 0 \mid \Phi_j = 1, c_1, \dots, c_k) = 1 - \mathrm{pr}(Y = 1 \mid \Phi_j = 1, c_1, \dots, c_k) = 1 - \alpha$
- $\mathrm{pr}(Y = 0 \mid c_1, \dots, c_k) = 1 - \mathrm{pr}(Y = 1 \mid c_1, \dots, c_k) = (1 - \alpha) + \gamma S(c_1) \times \cdots \times S(c_k)$

Thus,

$$E(\Phi_j \mid Y_i = 0, c_1, \dots, c_k) = \frac{(1-\alpha)F(c_j)}{1 - \alpha + \gamma S(c_1) \times \cdots \times S(c_k)}.$$

Results from Simulations with Fixed Censoring Times and Random Grouping Structure

Tables 1-3 contain the results from the simulations with fixed censoring times described in §5.1 of the main paper. All results are directly comparable to conclusions drawn from the simulations with random censoring times.

T	1	2	3	4	5	6	7	8	9	10
F(T)	0.005	0.01	0.015	0.02	0.025	0.03	0.035	0.04	0.045	0.05
$(\alpha, \beta) = (0.8, 0.8)$										
Group Estimator										
bias	-1.4	-1.7	-1.5	-1.1	-0.6	-0.2	-0.2	0.3	-1.3	-6.1
width	2.11	2.59	2.97	3.18	3.37	3.47	3.70	4.06	4.73	6.29
Individual Estimator										
bias	0.1	0.2	0.1	0.5	0.4	0.1	0.2	-0.6	-1.9	-6.6
width	2.54	3.04	3.58	4.06	4.21	3.98	4.49	4.61	5.13	6.40
$(\alpha, \beta) = (0.9, 0.9)$										
Group Estimator										
bias	-0.3	-0.5	-0.2	-0.2	0.1	0.4	0.1	0.3	-0.5	-3.2
width	1.82	2.08	2.32	2.51	2.82	2.83	2.79	2.97	3.52	3.93
Individual Estimator										
bias	0.0	0.3	0.3	0.0	0.2	0.4	0.3	0.2	-1.5	-3.9
width	1.94	2.46	2.88	2.81	2.91	3.06	3.02	3.13	3.54	4.31
$(\alpha, \beta) = (0.95, 0.95)$										
Group Estimator										
bias	0.2	-0.0	0.2	0.6	0.4	0.4	-0.1	-0.4	-0.9	-3.3
width	1.49	1.82	2.14	2.19	2.25	2.40	2.39	2.62	2.99	3.54
Individual Estimator										
bias	0.2	0.1	0.0	-0.1	-0.1	0.2	-0.2	-0.3	-0.7	-2.7
width	1.53	2.11	2.15	2.31	2.31	2.44	2.52	2.56	2.89	3.44
$(\alpha, \beta) = (0.99, 0.99)$										
Group Estimator										
bias	0.5	0.5	0.4	0.4	0.1	-0.1	0.0	0.1	0.0	1.8
width	1.27	1.60	1.71	1.96	1.92	2.05	2.06	2.18	2.33	2.86
Individual Estimator										
bias	0.3	0.3	0.3	0.2	-0.1	-0.1	0.2	0.3	0.2	-1.2
width	1.16	1.38	1.53	1.68	1.80	1.81	1.80	1.99	2.15	2.55

Table 2.1: Results from the expectation-maximization-pool-adjacent-violators algorithm in groups of size two and the pool-adjacent-violators algorithm for individuals in a setting with n =10,000 individuals on 10 fixed censoring times for several fixed, equal and constant specificity and sensitivity. Grouping was done across the Cs, although there were 1,000 individuals at each C. Here we present the bias of each estimator (compared to $F(T)$) multiplied by 1,000 (for example, a reported bias of 1.4 corresponds to a true bias of 0.0014) and the width of the 95% Monte Carlo quantile intervals, multiplied by 100 (e.g. a reported width of 1.53 corresponds to an actual width of 0.0153).

T	1	2	3	4	5	6	7	8	9	10
F(T)	0.005	0.01	0.015	0.02	0.025	0.03	0.035	0.04	0.045	0.05
$(\alpha, \beta) = (0.8, 0.8)$										
Group Estimator										
bias	1.7	2.0	2.2	2.3	1.7	1.6	0.7	-1.1	-3.6	-9.9
width	1.90	2.33	2.82	3.23	3.32	3.67	3.87	4.30	4.85	6.45
Individual Estimator										
bias	-0.3	0.1	0.1	0.4	0.3	0.1	-0.5	-1.2	-2.6	-7.5
width	2.50	3.11	3.58	4.17	4.33	4.25	4.50	4.42	4.89	6.21
$(\alpha, \beta) = (0.9, 0.9)$										
Group Estimator										
bias	1.4	1.8	1.8	1.6	0.9	0.4	0.1	-0.8	-2.2	-6.6
width	1.79	2.25	2.56	2.78	2.94	2.87	3.05	3.36	3.62	5.01
Individual Estimator										
bias	-0.2	0.2	0.2	0.9	0.4	0.0	-0.2	-0.2	-1.1	-4.2
width	1.96	2.42	2.81	2.95	2.90	3.04	3.13	3.06	3.33	4.26
$(\alpha, \beta) = (0.95, 0.95)$										
Group Estimator										
bias	1.4	1.4	1.2	0.7	0.6	0.4	-0.3	-0.4	-1.0	-4.0
width	1.59	2.03	2.43	2.61	2.72	2.72	2.85	2.93	3.28	4.34
Individual Estimator										
bias	0.4	0.3	0.1	0.1	0.3	-0.1	-0.5	-0.3	-0.7	-2.5
width	1.56	2.17	2.33	2.22	2.35	2.42	2.38	2.44	2.61	3.33
$(\alpha, \beta) = (0.99, 0.99)$										
Group Estimator										
bias	1.0	1.1	1.1	0.7	0.4	0.3	-0.1	-0.3	-1.4	-3.7
width	1.60	1.96	2.32	2.38	2.47	2.42	2.52	2.63	2.91	3.69
Individual Estimator										
bias	0.0	0.1	0.3	0.1	0.1	0.2	0.1	-0.1	-0.4	-1.6
width	1.12	1.43	1.60	1.68	1.68	1.94	1.94	1.94	2.14	2.55

Table 2.2: Results from the expectation-maximization-pool-adjacent-violators algorithm in groups of size five and the pool-adjacent-violators algorithm for individuals in a setting with n =10,000 individuals on 10 fixed censoring times for several fixed, equal and constant specificity and sensitivity. Grouping was done across the Cs, although there were 1,000 individuals at each C. Here we present the bias of each estimator (compared to $F(T)$) multiplied by 1,000 (for example, a reported bias of 1.4 corresponds to a true bias of 0.0014) and the width of the 95% Monte Carlo quantile intervals, multiplied by 100 (e.g. a reported width of 1.53 corresponds to an actual width of 0.0153).

T	1	2	3	4	5	6	7	8	9	10
F(T)	0.005	0.01	0.015	0.02	0.025	0.03	0.035	0.04	0.045	0.05
$(\alpha, \beta) = (0.8, 0.8)$										
Group Estimator										
bias	2.4	3.5	3.8	3.7	3.3	2.2	0.4	-1.8	-5.7	-13.7
width	1.59	2.52	3.03	3.45	3.55	4.07	4.07	4.54	5.14	7.62
Individual Estimator										
bias	-0.0	0.9	0.9	0.9	0.8	0.2	0.1	-0.4	-1.6	-6.4
width	2.38	2.98	3.50	3.97	4.28	4.15	4.39	4.64	5.11	6.64
$(\alpha, \beta) = (0.9, 0.9)$										
Group Estimator										
bias	1.8	2.5	2.6	2.6	2.1	1.3	0.3	-1.0	-3.2	-9.5
width	1.96	2.50	2.77	3.00	3.22	3.36	3.49	3.68	4.40	5.99
Individual Estimator										
bias	0.4	0.7	0.8	0.2	0.2	0.5	-0.0	-0.2	-0.9	-3.6
width	1.87	2.41	2.91	3.06	3.00	3.00	3.12	3.28	3.50	4.25
$(\alpha, \beta) = (0.95, 0.95)$										
Group Estimator										
bias	1.6	2.3	2.5	2.0	1.2	0.3	-0.3	-1.0	-2.4	-7.5
width	1.73	2.16	2.68	2.99	3.11	3.12	3.21	3.37	3.84	5.33
Individual Estimator										
bias	0.2	0.2	0.5	0.1	0.1	-0.0	-0.0	-0.3	-0.8	-2.8
width	1.56	2.11	2.28	2.22	2.44	2.34	2.45	2.56	2.70	3.30
$(\alpha, \beta) = (0.99, 0.99)$										
Group Estimator										
bias	1.7	1.9	2.0	1.2	1.0	0.6	-0.1	-0.6	-1.6	-6.2
width	1.66	2.20	2.56	2.72	2.80	3.04	3.01	3.22	3.51	5.23
Individual Estimator										
bias	0.2	0.3	0.4	0.2	0.3	0.3	0.3	0.2	-0.1	-1.2
width	1.16	1.43	1.53	1.74	1.73	1.84	1.84	2.04	2.14	2.59

Table 2.3: Results from the expectation-maximization-pool-adjacent-violators algorithm in groups of size ten and the pool-adjacent-violators algorithm for individuals in a setting with n =10,000 individuals on 10 fixed censoring times for several fixed, equal and constant specificity and sensitivity. Grouping was done across the Cs, although there were 1,000 individuals at each C. Here we present the bias of each estimator (compared to $F(T)$) multiplied by 1,000 (for example, a reported bias of 1.4 corresponds to a true bias of 0.0014) and the width of the 95% Monte Carlo quantile intervals, multiplied by 100 (e.g. a reported width of 1.53 corresponds to an actual width of 0.0153).

Results from Simulations with Fixed Censoring Times and Grouping Within Grid Points

To calculate the prevalence of a disease from grouped test results with sensitivity α and specificity β, Tu et al. suggest calculating $p = 1 - \{(\alpha - P)(\alpha + \beta - 1)^{-1}\}^{1/k}$, where k is the group size and P is the maximum likelihood estimate of the prevalence from the grouped test results. They also suggest getting inference for this parameter using

$$Var(p) = \frac{(\alpha - P)^{2/k-2} P(1-P)}{nk^2(\alpha + \beta - 1)^{2/k}}$$

In the case of grided current status data, we can adapt this formula to estimate the prevalence at each point on the grid. First, we use the pool-adjacent-violators algorithm on the grouped test results to get an estimate of $P(t)$ for each t in the grid. We transform back to the individual test results using $F(t) = 1 - [\{\alpha - P(t)\}(\alpha + \beta - 1)^{-1}]^{1/k}$. We can calculate confidence intervals using

$$Var\{f(t)\} = \frac{\{\alpha - P(t)\}^{2/k-2} P(t)\{1 - P(t)\}}{nk^2(\alpha + \beta - 1)^{2/k}}$$

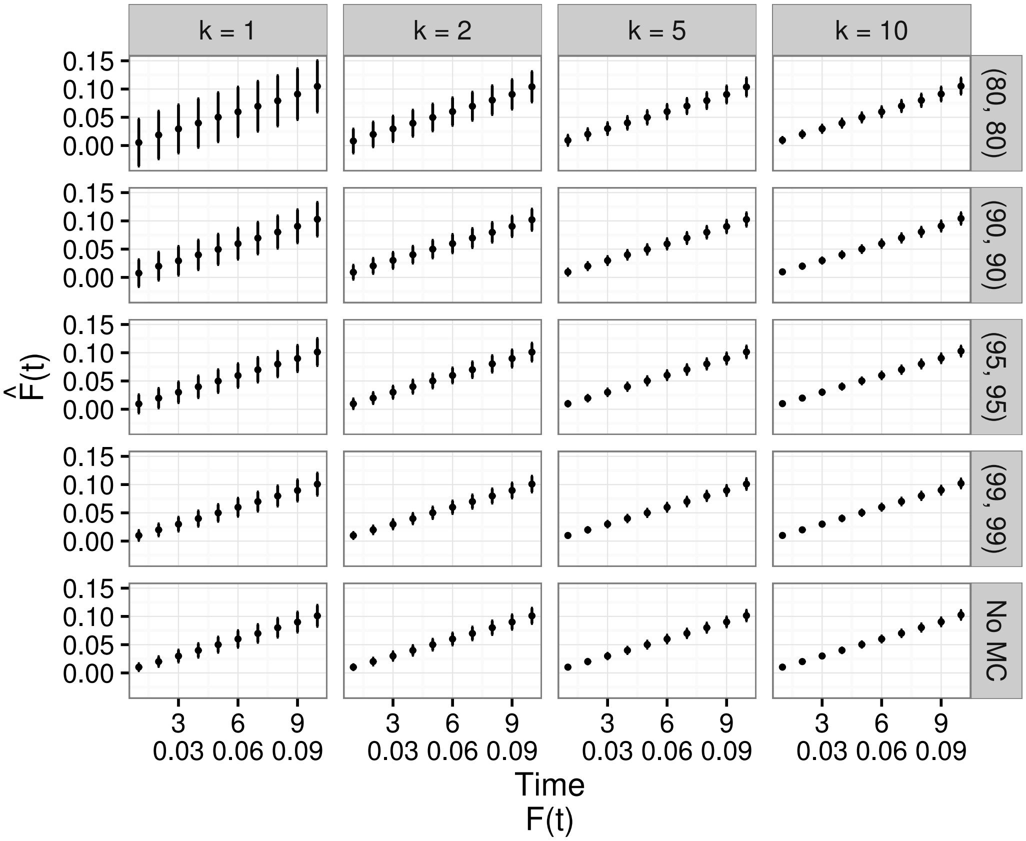


Figure 2.7: Results from simulations done on a grid of ten possible censoring times with $n = 1,000$ individuals observed at each censoring time. Grouping was done within the grid points. Groups of sizes 1, 2, 5, and 10, and misclassification rates of 0, 1%, 5%, 10%, and 20% were examined. Estimates and variances were calculated as described in Section 4 of the supplemental information. Points represent the estimates and confidence bars represent asymptotically normal 95% confidence intervals.

Code to Replicate Simulations

Code to replicate the simulations presented in this chapter is available at

www.github.com/lpetito/groupedcs.

Chapter 3

An Application of Marginal Structural Modeling

3.1 Gentle Introduction to Marginal Structural Models

Introduction

Often in medical studies, researchers aim to isolate the effect of an exposure or treatment A on an outcome Y. They may ask questions like “Does drinking coffee cause pancreatic cancer?”, “Does prepregnancy hypertension impact risk of preeclampsia?”, or “Does a particular treatment decrease AIDS-related morbidity/mortality?” These questions can be difficult to answer due to the presence of confounding factors, W, variables that are related both to the exposure A and the outcome Y. Failing to account for such factors can produce biased estimates of the desired effect. For instance, maternal age (say, being over 40) is a confounding factor when studying the relationship between prepregnancy hypertension and preeclampsia (see figure 3.1). This is a classic example of a confounder because older women are at greater risk of both preeclampsia and prepregnancy hypertension. We are more likely to see hypertension in older women, which in turn implies that we are more likely to see more cases of preeclampsia since it is more common in older women. This conclusion can be reached independently of prepregnancy hypertension - confounding the relationship between prepregnancy hypertension and preeclampsia.

Randomized controlled trials (RCTs) are experimental studies that aim to estimate the effect of an intervention A on an outcome Y by randomly allocating individuals to the intervention $A = 1$ or to the control group (generally a placebo accompanied by standard of care) $A = 0$. By randomly assigning individuals to their treatment group, RCTs address confounding from both measured and unmeasured factors by essentially breaking any association that may exist between those factors and the intervention. However, it is worth noting

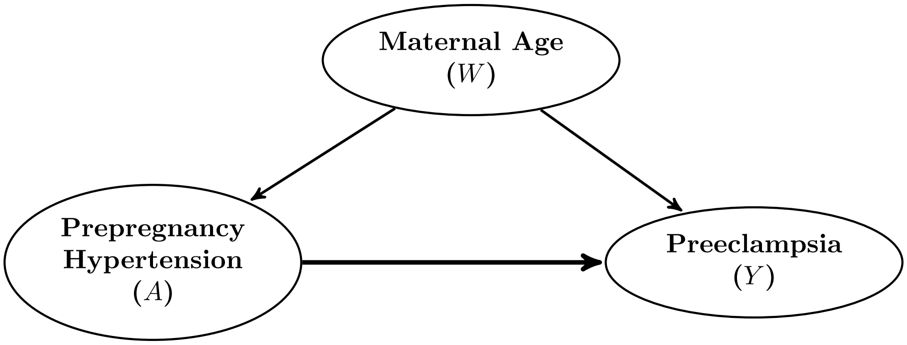


Figure 3.1: Directed acyclic graph representing the confounding process assumed in the prepregnancy hypertension - preeclampsia example.

that due to random chance, it is possible that some measured or unmeasured confounders may not be balanced between the intervention and control groups. Despite the advantages of conducting RCTs, in certain situations they can be prohibitively costly, not practical, or unethical. For example, RCTs cannot be used to study of the effects of childhood abuse on later-in-life morbidities for ethical reasons, as well as the very large cost of running a trial for decades. RCTs also cannot address questions such as "Does high cholesterol increase the risk of heart attack?", as an individual's cholesterol level cannot be randomly assigned.

Instead, researchers have turned to observational data sources to answer these kinds of questions. As individuals in observational data sources have chosen their own treatment, as opposed to randomly being assigned to an intervention, these data must be analyzed more carefully, as there are almost certainly factors present that have affected an individual's treatment choice as well as their resulting outcome. In the case of an RCT, the effect of A on Y is usually estimated simply, as $E[Y \mid A = 1] - E[Y \mid A = 0]$. Confounding factors are not considered as they are considered to be balanced between the treatment groups, though this practice of presenting unadjusted statistics is still debated in the statistics community (cite some people like Mark, Michael Rosenbloom). However, most statisticians agree that in observational studies, adjusting for confounding factors is needed. There are two classes of estimator that can adjust for these confounders: conditional estimators and marginal estimators.

The conditional approach is the more common approach taken in statistical practice. It works by first estimating $E(Y \mid A, W)$ for every possible combination of A and W, and then for a fixed level $A = a$, averaging those predicted values over the observed distribution of W. (Note that this is just amounts to the tower property of expectations: $E(Y \mid A) = E[E_W(Y \mid A, W)])$.) The estimation of $E(Y \mid A, W)$ can be done either nonparametrically or parametrically. The most common nonparametric technique is called stratification (cite Mark I think?), where the data are divided into subsets based on the observed confounders W, and then the relationship $E(Y|A)$ is assessed within each strata.

Going back to our preeclampsia example, this amounts to estimating the relationship between prepregnancy hypertension and preeclampsia separately in women who are over or under 40 years old, and then creating a weighted average of our two estimates based on the proportion of women in the study who are over or under 40 years old. The largest barrier to using this technique is as the number of confounding factors W increases, the number of strata can become prohibitively large. Other nonparametric techniques, such as random forest and SuperLearner (add citations), are also available to use, and can be more adept at handling larger sets of covariates.

Parametric modeling is more commonly used to estimate the conditional quantity $E(Y \mid A, W)$, as the set of confounding factors W is often too large for nonparametric techniques, or it contains too many continuous variables. Parametric modeling estimates the joint distribution directly, most simply as a main-terms linear regression model $E[Y|A,W] = \alpha + \beta A + \gamma W$, though the model can be specified in many more ways (adding quadratic terms, cubic splines, interactions, etc.). There are two major limitations to this technique: model specification and model overfitting. The most difficult aspect to directly modeling this joint distribution is that the model must be correctly specified in order for the effect of interest β to be interpreted as the true effect of A on Y. This property is almost certainly untestable in practice, though there are theoretical properties of some model-fitting techniques that allow statisticians to be more sure of their final model (cite oracle inequality for SuperLearner). Additionally, as the number of covariates W approaches the sample size in the data, the final model can become increasingly overfit to the particular dataset used in the modeling process. Overfitting models is not desirable because it reduces the external generalizability of the estimates. k-fold cross-validation is a tool that has been developed to prevent the overfitting of a model, though it is not yet widely used in epidemiologic practice.

When estimating the effect of a treatment A on the outcome Y, the marginal approach is less commonly used than the conditional approach. The guiding philosophical principle for randomized controlled trials is the same principle that grounds the marginal approach to calculate effect estimates from observational data: if the treated and control groups are the same in every way except their treatment regimen, any differences researchers see in the outcome between the two groups can be attributed to the treatment. Matched selection is a type of marginal approach, which can be found in perfectly balanced RCTs as well as matched case-control studies (cite Jack and Paul's Impact Eval book), but here we focus on data reweighting, a technique first studied in the context of surveys done with non-random sampling (cite something about surveys - Thomas Lumley). This data reweighting approach works by creating a set of weights such that when a dataset is reweighted, the observed covariate distribution in each treatment group is the same, and then estimating the effect of the treatment A on the outcome Y in the reweighted sample.

A major advantage of the marginal approach is the ability of the model to incorporate information from many confounding factors without worrying about model overfitting. Incorporating all covariates will overfit the model in when conditionally estimating $E[Y \mid A, W]$, whereas the reweighting of the sample allows covariates to be included in the model for the weights but not the final weighted average, $E[Y \mid A]$. This advantage is especially helpful in situations where the outcome of interest is rare. Nonetheless, there are several limitations to the marginal approach. First, we can only balance the data on measured confounding factors; unmeasured confounding cannot be addressed. Second, the model for the creation of the weights must be properly specified. The same concerns stated about model fitting in conditional estimation also apply here. Last, there can be large variability in the weights, which may lead to certain individuals having undue influence on the effect estimate.

Inverse Probability Weighting

Inverse probability weighting (IPW) is a marginal approach to estimating the effect of an exposure A on an outcome Y while still accounting for potential confounding factors W by weighting the data to balance the representation of each strata of W within each exposure group. It is most commonly used in observational studies, as an attempt to approximate the conditions of an RCT by breaking any association between observed confounding factors and the exposure. Practically, each observation is weighted by the reciprocal of the predicted probability of being in the covariate subgroup that was observed for each patient. The process is briefly outlined below using our maternal age/prepregnancy hypertension/preeclampsia example.

1. Create the weight model. Use a logistic regression model to estimate the probability of prepregnancy hypertension A in each maternal age strata W (< 40 vs. ≥ 40):

   $$logit[Pr(A = 1)|W] = \beta_0 + \beta_1 W.$$

   [Note in this case, a logistic regression model will provide the same answer as stratified estimation.]

2. Create individual weights. From the model in step 1, predict the probability of prepregnancy hypertension for each individual, as

   $$p(w) = Pr(A = 1|W = w) = \frac{1}{1 + exp[-(\beta_0 + \beta_1 w)]}.$$

   Then the weight for each individual who experienced prepregnancy hypertension becomes $\omega_i = 1/p(w_i)$, and the weight for each individual who did not experience prepregnancy hypertension becomes $\omega_i = 1/(1 - p(w_i))$, where w_i is that individual's age.

3. Calculate the effect estimate by simply taking the weighted average

   $$E[Y \mid A = 1] - E[Y \mid A = 0] = \sum_{i=1}^{n} \omega_i Y_i [I(A_i = 1) - I(A_i = 0)].$$

There are three primary concerns throughout the weight-creation process: is "balance" achieved, what is the population of interest, and how big are the weights. First, achieving balance is important because it limits bias in the effect estimate. Balance in this scenario refers to the similarity in the distribution of confounding factors needed between treatment and control groups. For example, once the exposed and unexposed groups have been weighted, if we have achieved balance we would expect to see the same (or similar) proportion of females in each group. It is worth noting that when comparing the distributions of confounding factors between strata of A that the magnitude of the difference seen is more important than the statistical significance. Second, defining the population of interest, and specifically knowledge about the distribution of confounding factors in that population, is essential to the weight-creation process because it speaks to the generalizability of the effect estimates. Last, the size of the weights directly impacts how much influence an individual has on the final effect estimate. For example, an individual with a weight of 5 is 4 times less important than a person with a weight of 20.

The presence of multiple confounding factors makes the development of weights more difficult. Instead of being able to estimate the probability of exposure in each strata of a covariate, we have to develop models (generally linear or logistic regression) to smooth over the parameter space so we can extrapolate weights from each individual's unique covariate pattern. To ensure that the weights are really taking care of potential confounding only, statisticians decided to stabilize the weights by dividing the weight defined in step 2 above by the probability of exposure. This process reduces the number of individuals who have extreme weights, but these weights are not directly interpretable anymore. They no longer have to be ≥ 1, which implies that they are no longer simply the inverse of the probability of exposure. When using IPW, users should perform checks to ensure that the mean of the weights is close to 1; if not, the model specification for the weight model (step 2) is probably incorrect, or another essential assumption has been violated.

Another method to deal with extreme weight values is truncation. Commonly, the extreme weights (<1st percentile or >99th percentile) are reset them to the 1^{st} percentile or 99^{th} percentile, although some people advocate for using the 5^{th} and 95^{th} percentiles as the cutoffs. This practice is a classic example of bias-variance trade off, as resetting the weights could induce bias into the point estimate, but will reduce the variability in the final results.

Marginal Structural Modeling

Now that we understand IPW in the context of a cross-sectional study, we can begin to conceptualize how we might approach questions of a longitudinal nature. These are inherently more difficult to answer, as defining the question can be half the battle. Are the researchers interested in the effect of a time-varying treatment on an outcome at a specific point in time, e.g. 5 year survival? Are they interested in the average effect of a time-varying treatment on a time-varying outcome?

Marginal structural models (MSM) were developed to address the presence of time-dependent confounding when studying the effects of a time-varying exposure on an outcome. A time-dependent confounder behaves slightly differently than a traditional confounder in a cross-sectional analysis. Because the both the exposure and the confounder vary with time, future levels of the confounder are affected by prior levels of the exposure, meaning that the time-dependent confounder is a part of the causal pathway between the exposure and the outcome. Simply adjusting for this time-varying confounder would not allow us to detangle the direct effect of the exposure on the outcome. However, not adjusting for this time-varying confounder would also yield invalid results, as the confounder could be indicative of a preference in future levels of the exposure.

Using IPW to estimate the parameters in MSMs allows us to re-weight the data at each point in time during the study follow-up to balance the covariate distributions in the treatment groups. This approach breaks the effect that the time-varying covariate has on the exposure and the outcome by essentially randomizing the covariate at each point in time. We can consider this a "marginal" approach because the balancing of the sample at each time point happens before the estimation of the parameter of interest. IPW can also be used in MSMs to address issues in follow-up through the study duration. These "censoring" weights can be used to re-weight the data so that it approximates what the data would have been had each individual been followed through the whole study period.

In order to interpret the results from an MSM as causal, several assumptions must be either empirically verified or taken on faith. The first and most important assumption is called exchangeability, which is also referred to as the assumption of no unmeasured confounding. This assumption cannot be verified from available data, but rather must be well thought-out before the study begins, so researchers can collect the relevant data.

Another required assumption is called consistency. This assumption comes from the counterfactual framework. Each individual has a set of counterfactual outcome that corresponds to a possible treatment level. A particular counterfactual outcome, Y_a for an individual is their outcome in a world where, possibly contrary to fact, the exposure has been set to a particular level, $A = a$. Consistency mandates that we must assume that the world of observed data is "consistent" with the counterfactual world. This means that for a given individual who experiences exposure at level a, the outcome we see for them in the real world is the same outcome we would have seen for them in their counterfactual outcome set.

The experimental treatment assumption, or positivity, is an assumption that requires a non-zero probability of being in a particular treatment group for every possible covariate pattern observed. An example of a theoretical violation of positivity is including women in a prostate cancer trial. Because women do not have a prostate, their probability of being treated for prostate cancer is 0. Although this violation is contrived for purposes of demonstration, often theoretical positivity violations can be addressed through more careful study design.

The positivity assumption can also be violated practically through the existence of "random zeros." A random zero occurs when there are many available covariates. In this scenario, the number of strata within which to estimate weights approaches the number of people in the study, so there exist strata where all individuals have been assigned to one particular treatment group. This assumption can be checked empirically, although it may be time-consuming. If positivity is found to be practically violated, there are several ways to correct it. First, the number of strata can be reduced by collapsing covariate levels. Additionally, covariates that have a "weak" association with the exposure and the outcome can be excluded, as the likelihood of a random zero decreases with the number of covariates. Finally, parametric models can be imposed to smooth over the choppy covariate space.

Although MSMs provide a suitable way to address time-dependent confounding in longitudinal studies, they have a few drawbacks. MSMs can be very time consuming. Fitting each of the models for the exposure, censoring, and outcome can be prohibitive, especially if the user chooses machine-learning methods to create each model. Additionally, when using IPW to estimate the parameters in a MSM, the variances that are estimated are typically 20% higher than the unadjusted estimated variances. This implies that MSMs require a larger sample size to have the same power as unadjusted analyses.

In section 3.2, we apply marginal structural modeling to a dataset to assess the joint effect of three time-varying exposures on an outcome in the context of time-dependent confounders. We then further discuss the marginal structural modeling framework in the context of counterfactuals in section 3.3. Material presented in section 3.1 summarizes and reframes work presented in [36], [37], [38], and [39].

3.2 Using marginal structural modeling to evaluate the roles of lipid levels, inflammatory markers, and rheumatoid arthritis disease activity in assessing risk of cardiovascular events

Introduction

Rheumatoid arthritis (RA) is an autoimmune disorder affecting approximately 1.5 million people in the United States that primarily causes swelling, warmth, and pain in the membranes surrounding joints [40–42]. Moreover, in moderate to severe cases, it can cause inflammation around the heart and lungs [42]. RA patients are at increased risk of myocardial infarction (MI), cardiovascular disease, stroke, and cardiovascular disease-related mortality compared with the general population; however, this increased risk may not be due solely to established risk factors (family history of CVD, diabetes, hypertension, dyslipidemia), but rather to other comorbidities associated with RA (higher erythrocyte sedimentation rates (ESRs), vasculitis, pulmonary disease) [43, 44]. Contrary to expectation, researchers previously found that low cholesterol patients were at increased risk of cardiovascular disease in a cohort of RA patients with high ESRs [45].

Characteristics of rheumatoid arthritis disease activity have been found to affect cardiovascular event occurrence. Studies have reported that patient response to treatment may be indicative: cardiovascular events occurred less frequently in 1) patients who received nonbiologic disease-modifying antirheumatic drugs (DMARDs) and 2) those who responded to a biologic DMARD compared with those who did not [46, 47]. Moreover, interleukin-6 (IL-6), one of the primary drivers of RA-related inflammation, has been linked with the development of coronary heart disease in both patients with RA and the general population [48–53].

Tocilizumab (Actemra) is a humanized monoclonal antibody against the IL-6 receptor that was developed by Genentech/Roche/Chugai and first approved by the Food and Drug Administration (FDA) to treat moderate to severe rheumatoid arthritis (RA) in January 2010 [54]. Although tocilizumab has an approved regulatory label for inadequate responders to one or more DMARDs (DMARD-IR), it came at a time where tumor necrosis factor (TNF)-α inhibitors had already been on the market for a decade, thus relegating it to second or third line use. Tocilizumab aims to reduce inflammation and decelerate the rate of articular damage in patients with RA.

When patients with chronic inflammation are placed on anti-inflammatory drug regimens, an elevation in their lipid levels is typically seen that persists through their treatment. This elevation could be due to the regulatory integration the drugs into metabolic and inflammatory molecular networks [55–57]. Patients taking tocilizumab specifically see a greater magnitude of elevation in their lipid levels, particularly low-density lipids, than patients on other anti-inflammatory agents, e.g. TNF-α blockers, although the implications of these elevations are still under study [51, 52, 56, 57].

In patients with moderate to severe RA who are taking tocilizumab, the independent contributions of pre-treatment versus post-treatment levels of RA disease activity, HDL cholesterol, and LDL cholesterol to the risk of major adverse cardiac events (MACE) are unknown. Detangling the effects of these measures before treatment and post-treatment will allow us to attribute the risk of MACE to measures that are representative of chronic levels versus acute changes in these levels. Furthermore, it is important to evaluate whether inflammatory markers simply confound the relation between RA disease activity, HDL cholesterol, LDL cholesterol, and MACE or if they are an integral part of the causal pathway, partially mediating these relationships as a time-dependent confounder. Time-dependent confounders (TDC) behave as traditional confounders - the TDC affects both the outcome and the exposure of interest, but additionally, previous levels of the exposure influence future levels of the TDC.

In this paper, we address the problems of time-dependent confounding implicit in this analysis by applying marginal structural modeling to 1) estimate the effect of pre-treatment levels of RA disease activity and HDL and LDL cholesterol on the risk of MACE and 2) estimate the effect of post-treatment levels of RA disease activity and HDL and LDL cholesterol independent of pre-treatment levels on risk of MACE.

Methods

Study population

This study population has been described in detail previously (see [58] for more details). Between 2007 and 2011, Roche conducted five clinical trials (OPTION, AMBITION, RADIATE, TOWARD, and LITHE) and enrolled participants into corresponding long-term extension (LTE) studies (GROWTH95, GROWTH96, and LITHE) to study tocilizumab. For this post-hoc analysis, we used the available data on 3,986 individuals who participated in both a clinical trial and LTE study. These individuals received at least 1 dose of tocilizumab intravenously every 4 weeks, either as their only RA treatment or in combination with other drugs, through April 1, 2011. Each individual was followed for up to 60 months after beginning their tocilizumab treatment (for clinical trial participants who were controls, this period began during the long-term extension study). MACE and censoring times were recorded as they occurred through the 60-month follow-up period. Demographic and general health status data were collected at the beginning of the study, as was information on concomitant medication usage. Laboratory values were collected three times: pre-tocilizumab treatment initiation, six months after treatment initiation, and twelve months after treatment initiation.

Study variables

The occurrence of a major adverse cardiac event (MACE), defined as at least one of definite nonfatal MI, nonfatal stroke, or death caused by CVD, was the primary outcome in this study. All MACE events were confirmed by an independent cardiologist with experience serving on MACE adjudication panels. More details on this process can be found in [58].

The exposures of interest were RA disease activity and HDL and LDL cholesterol levels, all patient characteristics that are used to determine treatment courses for RA. The RA Disease Activity Score in 28 joints (RA-DAS28) was used to assess the severity of rheumatoid arthritis in each patient. Although subscales of the RA-DAS28, swollen joint count and tender joint count, were also available, they were not used in analyses due to their high correlation with RA-DAS28 (0.69 and 0.82 respectively).

Both time-independent and time-dependent covariates were chosen a priori from theory-based causal diagrams (figures 3.2 and 3.3) [59]. They included demographic and health status characteristics (age, sex, body mass index, family history of cardiac disorders, personal history of cardiac disorders, blood pressure, smoking status, and RA duration), laboratory values that included inflammatory markers (interleukin-6 receptor, erythrocyte sedimentation rate, C3, C4, C-reactive protein, albumin, haptoglobin, absolute neutrophil count, platelet count) and other lipid levels (total cholesterol, apolipoprotein A-I, apolipoprotein B, Lp(a), serum amyloid A), and use of concomitant medications (oral steroids, statins, NSAIDs, methotrexate). Baseline demographic and health status characteristics and concomitant medication usage were considered time-independent confounders (figure 3.3). All inflammatory markers and lipid levels were considered potential time-dependent confounders. Triglycerides were excluded as a covariate as it is a linear combination of total, HDL, and LDL cholesterol.

Statistical Analysis

Exploratory analysis

We first described the characteristics of the sample at each time point they were measured, as well as the distribution of the exposures and prevalence of MACE. We then assessed the censoring pattern observed in the data.

Cross-sectional analysis

We used a weighted logistic regression model to estimate the (not causal) association between pre-treatment levels of RA disease activity, HDL cholesterol, and LDL cholesterol on the 5-year odds of MACE. Because our outcome was so rare and the number of potential confounders was so large, traditional methods to adjust for all measured confounders would have substantially over-fit the model. The inverse probability weighting (IPW) [36, 37, 60] estimator allows us to re-weight our sample to balance covariate distributions across our exposures, thereby approximating conditions in a randomized controlled trial as closely as possible.

IPW estimators require the specification of two models: a summary model and a model to create weights. As our exposures of interest are continuous, we cannot enumerate all possible exposures, and are forced to pre-specify an un-saturated summary model. In this analysis, the summary model was chosen to be a logistic regression model that included as main terms each exposure, as well as age and personal history of cardiac disorders, which were the covariates that were most strongly associated with MACE [58].

Another consequence of having continuous exposures of interest was modeling conditional densities instead of modeling the probability of observing a particular exposure. We opted to estimate each of the three conditional densities of interest via linear regression, assuming a normal kernel. All covariates measured at or before treatment initiation (demographic and health status, concomitant medications, and laboratory values), as well as pre-treatment initiation levels of the exposures of interest, were incorporated as covariates in the three linear regression models used to create the weights. To address loss to follow-up, we used a logistic regression model to estimate the probability of completing all follow-up time (five years), adjusting for all covariates measured at or before treatment initiation, as well as the three exposures of interest. Each of the four weights was stabilized [61], and their product was used to fit the summary weighted logistic regression model used to estimate the odds of MACE. The mean weight was 1.13, and the range was truncated to 0.01-30. As confidence intervals reported from standard software are known to be too conservative [36], we used a nonparametric bootstrap with 1,000 resamples to estimate robust standard errors.

We also ran an unweighted main-terms logistic regression model, to provide a comparison to our weighted model. Robust standard errors are reported to address model misspecification.

Longitudinal analysis

We implemented a logistic marginal structural model (MSM) to estimate the *causal* effects of post-treatment changes in RA disease activity, HDL cholesterol, and LDL cholesterol on the MACE outcome described above. Since many of the inflammatory markers are part of the postulated causal pathway between RA disease activity and HDL/LDL cholesterol and MACE (see figure 3.3), adjusting for these factors by simply adding them as time-dependent variables in a classic logistic regression model would produce non-interpretable, non-causal effect estimates [36, 37]. Instead, we opted to use a logistic MSM to model the exposure-specific effects on MACE conditional on the history of exposures and covariates outlined in the **Study variables** section. We used the IPW estimator to estimate the parameters in our MSM, so time-dependent confounders and other factors on the causal pathway were only incorporated in the process of creating the exposure model weights.

There were two parts to the modeling process: choosing a summary model, and weight creation. We chose to use the following summary model:

$$logit(Pr[Y(t) = 1|A(t), V]) = \alpha + \beta A(0) + \gamma[A(0) - A(t)] + \delta V,$$

which incorporates age and personal history of CHD as covariates (V). Here, $\beta = (\beta_1, \beta_2, \beta_3)$ represent the effects of pre-treatment levels of RA disease activity, HDL cholesterol, and LDL cholesterol on the log odds of occurrence of a MACE at time t, $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ represent the effects of post-treatment changes in RA disease activity, HDL cholesterol, and LDL cholesterol on the log odds of occurrence of a MACE at time t, and $\delta = (\delta_1, \delta_2)$ represent the effects of age and personal history of CHD, respectively, on the log odds of occurrence of a MACE at time t.

To create an appropriate set of weights, we first needed to determine which time-varying variables acted as time-dependent confounders. Thus, all laboratory values (inflammatory markers and lipid levels) were examined to determine if they satisfied the criteria for time-dependent confounding by separately examining the unadjusted association between each exposure at time $(t - 1)$ and each the confounder at time (t). As all potential confounders were continuous, linear regression models were used, with robust (sandwich) standard errors to account for repeated measurements on each individual. All variables for which the association had P < 0.2 were retained for use as time-dependent confounders in the creation of weights to use in the marginal structural model, as shown in Table 3.1.

To create the exposure weights, we estimated each of the three conditional densities using linear regression assuming a normal kernel. These models controlled for all pre-treatment variables as well as the time-varying confounders listed in Table 3.1. We also created a set of inverse probability of censoring weights using main-terms logistic regression models, controlling for all time-independent and time-dependent confounders as listed in **Study Variables**, as individuals who were censored might not have been representative of the whole cohort. Each of these weights was stabilized, and their product was used to fit the summary weighted regression model to estimate the risk of MACE. Again, the mean weight was 1.13, and the range was truncated to $0.01 - 30.0$. We used a nonparametric bootstrap with 1,000 resamples to estimate robust standard errors for the MSM.

After completing our MSM, we also performed an associational unweighted logistic regression model in the same form as the summary model detailed above for practical comparison. Robust standard errors were computed using a sandwich estimator in this unweighted model.

For more details about the assumptions underlying and implementation of the MSM, please see section 3.3.

Further details

Missing data at baseline was imputed 50 times using multiple imputation by chained equations [62], with a burn-in period of 20 iterations for each imputed dataset [63]. Please see table 3.2 for details about data missingness and truncation values used. Any missingness seen in time-varying covariates after the pre-treatment measurement was imputed using carry-forward imputation. All estimates were adjusted for the imputation process according to rules presented in [64]. All statistical analyses were conducted in R version 3.3.2.

Results

Sample description

Our study included 3,986 patients with moderate to severe RA (average duration 9.3 years upon study entry) taking a course of tocilizumab (Table 3.3). A majority of patients were female, and were taking a course of oral steroids, NSAIDs, and/or methotrexate concomitantly with their tocilizumab treatment. About half of patients had abnormal blood pressure at treatment initiation. Fewer than twenty percent of patients were current smokers, had a personal or family history of cardiac disorders, or were currently using statins.

Once patients began using tocilizumab, they saw a sharp decline in their DAS28 in the first 6 months (mean decrease of 2.8 points), and then a tapering in the following 6 months (mean decrease of 0.5 points) (Table 3.4). The LDL and total cholesterol levels rose in the first six months of treatment (mean increases of 0.4 and 0.7 respectively), and then stabilized. In contrast, there was little variation in the mean HDL cholesterol level over the first year of tocilizumab treatment. Most other laboratory values saw a more substantial change in the first six months of treatment, and then stabilized.

Fifty individuals (1.25%) had major adverse cardiac events during the 5-year follow-up (Table 3.5). Of those events, 5 occurred between 0 and 6 months post-treatment initiation, 9 occurred between 6 and 12 months post-treatment initiation, and 36 occurred between 12 months and 5 years post-treatment initiation. Although there was a fair amount of censoring during follow-up (10% during the first year, and an additional 60% in the next four years before the study end), most of those remained in follow-up at least through 48 months.

Cross-sectional results

The estimates of the association between pre-treatment levels of RA disease activity, HDL cholesterol, LDL cholesterol, and five-year odds of MACE, adjusted only for other pre-treatment measurements, are presented in table 3.6. The unweighted regression analysis suggests a 36% increase (95% CI: 1.07, 1.74) in five-year odds of MACE for each 1-point increase in RA disease activity, adjusting for HDL cholesterol, LDL cholesterol, age, and personal history of cardiac disorders. However, in the results from the weighted regression analysis, this association is attenuated to a non-statistically significant increase of 11% (95% CI: 0.58, 2.11). The association between LDL cholesterol and five-year odds of MACE is also attenuated from the unweighted model to the weighted model, though in both models the association is not statistically significant. The association between HDL cholesterol and five-year odds of MACE changes direction when the weights are added - the association for a one-unit increase in HDL cholesterol decreases the odds of MACE in the unweighted model (OR: 0.61), but not in the weighted model where a one-unit increase in HDL increases the odds of MACE (OR: 1.22). All reported confidence intervals are wider in the weighted analysis than in the unweighted analysis.

Longitudinal results

Table 3.7 presents odds ratios to summarize the effects of post-treatment initiation changes in levels of RA disease activity score, HDL cholesterol, and LDL cholesterol on MACE. The unweighted model suggests a 14% increase (95% CI: 1.00, 1.31) in odds of MACE for each 1-point increase in change in RA disease activity scores, controlling for pre-treatment RA disease activity score, pre- and post-treatment initiation levels of HDL and LDL cholesterol, age, and personal history of cardiac disorders. (Note: this implies that patients who see greater decreases in their RA disease activity score are at lower risks of MACE.) The pre-treatment RA disease activity score and LDL cholesterol are also associated with 29% and 32% increases in odds of MACE.

However, once we applied the marginal structural model to estimate causal odds ratios, reweighting our sample so both pre-treatment and post-treatment covariates were randomized with respect to all three exposures, we did not identify any statistically significant effects. In fact, the effects of each of the post-treatment initiation changes in each exposure flipped direction. Most notably, the effect of each 1-point increase in change in RA disease activity score on odds of MACE changed from a 14% increase to a 23% decrease (95% CI: 0.50, 1.21). Additionally, the point estimates for the pre-treatment levels of each exposure were substantially farther from the null in the MSM than in the unweighted model; however the confidence intervals in the MSM were much wider rendering all of the estimates not statistically significant, despite the smaller effects identified by the unweighted model being significant.

Overall, the models that did not adjust for time-dependent confounding showed a statistically significant increase in risk of MACE for every unit-change post-treatment initiation in RA-DAS28. However, the MSM that accounted for time-dependent confounding showed no significant effects, but the point estimates for post-treatment initiation levels of each exposure trended towards a decrease in risk of MACE.

Discussion

Implications and Comparison to Previous Work

In this methodological study, we did not find evidence that pre- and post-treatment levels of RA disease activity, HDL cholesterol, and LDL cholesterol were associated with increased risk of major adverse cardiac events in patients with moderate to severe rheumatoid arthritis taking tocilizumab, once time-dependent confounding from inflammatory markers and other lipid levels was taken into account.

Previously, researchers explored the relationship between post-treatment levels of RA disease activity, lipid levels, and inflammatory markers in this dataset; however, time-dependent confounding was not able to be captured by their models [58]. After applying a MSM to take into account potential time dependent confounding, we found that the results presented in [58] were similar to our estimate of the association of post-treatment changes in HDL and LDL cholesterol with odds of MACE (HR of 0.56 versus OR of 0.63, and HR of 1.07 versus OR of 1.10, respectively), though all estimates were not statistically significant. The variability surrounding these estimates was also similar, though the variability was higher in the MSM, which is to be expected due to the variability added by the weighting process. However, the point estimate from our MSM for RA disease activity is substantially different from that in [58] (HR of 1.29 versus OR of 0.77). Although the confidence intervals overlap a bit, the point estimates from each model are not contained within the confidence interval from the other model. This change in direction can almost certainly be attributed to the presence of at least one time-dependent confounder. Time-dependent confounders can bias the effects of time-varying exposures on risk of MACE, as they provide alternate pathways between the exposure and the outcome that are not accounted for in traditional analyses. We therefore conclude that this result in [58] is artificially inflated, and that there is not evidence to conclude that post-treatment changes in RA disease activity score are causally related to MACE.

It is worth noting that the effects of pre-treatment levels of RA disease activity, HDL cholesterol, and LDL cholesterol estimated from the cross-sectional model as in Table 3.6 are not directly comparable to those seemingly same effects in the marginal structural model, as in Table 3.7. The marginal structural model separates the pre-treatment effects from the post-treatment changes, thereby approximating true cross-sectional and longitudinal effects. As the effects of post-treatment initiation changes are different from the pre-treatment effects seen in Table 3.7, we can conclude that the associations seen in the cross-sectional analysis

3.6 may be a muddled combination of both the cross-sectional and longitudinal effects. Of course, as all weighted results are not statistically significant, directionality of all effect estimates should not be interpreted with too much certainty.

The observed lipid profiles of the patients in this study did change over the study duration, but the changes were not extreme, as has been found in previous studies [57]. Our results suggest that these acute changes were not large enough in this study to be influential on the risk of MACE. However, here we found that the pre-treatment LDL levels were more strongly associated with risk of MACE, agreeing with past literature that suggests that chronic levels of LDL cholesterol are associated with risk of MACE.

Strengths

A major strength of this study is the data quality. Clinical trial data, though not often available for secondary data analyses, are a rich resource for tracked clinical changes over time, especially as collection is exhaustive. A wide variety of information is collected repeatedly over a substantial period of time, which could be invaluable in post hoc longitudinal studies of other patient characteristics in populations that are on a particular drug regimen.

Other major strengths of this analysis were the ability of marginal structural models to account for confounding by indication and additional time-dependent confounding [36, 37]. Because safety outcomes are so rare, statistically addressing confounding by indication becomes more difficult as we are concerned about over-fitting - that is, if we include as many covariates in the model as there are individuals in our dataset, our regression model will be able to perfectly predict all of the outcomes for the individuals in the dataset, but will not be generalizable to external datasets. As MSMs rebalance the covariate distribution across exposure groups, we can incorporate all of the available covariate data into the exposure models, while limiting the covariate inclusion in the summary model.

Limitations

This study had several limitations. First, as there were only 50 individuals who experienced a MACE in this dataset, we worry about the generalizability of our findings to the patient population taking tocilizumab. Should another study on tocilizumab be undertaken to evaluate the different patient characteristics between those who experience a MACE versus those who do not, we recommend a retrospective case-control study, to ensure sufficient power to detect differences.

Second, all three of our exposures of interest were continuous, which poses several methodological concerns. We are forced to use a non-saturated summary model in our marginal structural modeling, which extrapolates from the observed data when there is an exposure history in the model space that has not been observed. Additionally, although the summary model does not need to be correctly specified, to interpret the coefficients as causal effects, the model used to generate the weights must be correctly specified [39]. As we used a main-terms regression model, it is highly possible that our weight model was incorrectly specified. Future work could involve more non-parametric, data-adaptive methods to estimate these conditional densities; however, they involve prohibitive amounts of intensive computation [65]. Finally, despite the wealth of information these data offer, we almost surely are in violation of one of the assumptions needed to interpret these results as causal: the exchangeable treatment assumption (ETA). The ETA requires that each exposure history was observed for each type of individual in the dataset. Because our exposures are continuous, the ETA is never possible in practice. Nonetheless, because we have satisfied this assumption theoretically, we can be assured that we have approximated the causal effects as closely as possible.

Conclusions

In the presence of time-varying confounders, marginal structural models should be considered as an analytic option, even if the exposures of interest are not binary. Though the assumptions required to make them causally interpretable seem arduous, the same assumptions (no unmeasured confounding, correctly specified model, and noninformative censoring) are needed in standard statistical models to causally interpret the effects of time-independent exposures on outcomes.

Table 3.1: Variables included in exposure models used to create weights for longitudinal marginal structural model

Exposure (RA Disease Activity)	Exposure (HDL Cholesterol)	Exposure (LDL Cholesterol)
Exposures of Interest		
RA disease activity $(t-1, t_0)$	RA disease activity (t, t_0)	RA disease activity (t, t_0)
HDL cholesterol (t, t_0)	HDL cholesterol $(t-1, t_0)$	HDL cholesterol (t, t_0)
LDL cholesterol (t, t_0)	LDL cholesterol (t, t_0)	LDL cholesterol $(t-1, t_0)$
Laboratory Values		
Apolipoprotein A–I (t, t_0)	Apolipoprotein A–I (t, t_0)	Apolipoprotein A–I (t, t_0)
Apolipoprotein B (t_0)	Apolipoprotein B (t, t_0)	Apolipoprotein B (t, t_0)
C3 (t, t_0)	C3 (t, t_0)	C3 (t_0)
C4 (t, t_0)	C4 (t, t_0)	C4 (t, t_0)
Total Cholesterol (t, t_0)	Total cholesterol (t, t_0)	Total cholesterol (t, t_0)
C-reactive protein (t_0)	C-reactive protein (t_0)	C-reactive protein (t_0)
ESR (t_0)	ESR (t_0)	ESR (t_0)
Interleukin-6 receptor (t, t_0)	Interleukin-6 receptor (t, t_0)	Interleukin-6 receptor (t, t_0)
Lp(a) (t, t_0)	Lp(a) (t_0)	Lp(a) (t, t_0)
Serum amyloid A (t_0)	Serum amyloid A (t_0)	Serum amyloid A (t_0)
Albumin (t_0)	Albumin (t_0)	Albumin (t_0)
Haptoglobin (t, t_0)	Haptoglobin (t, t_0)	Haptoglobin (t, t_0)
Absolute neutrophil count (t, t_0)	Absolute neutrophil count (t, t_0)	Absolute neutrophil count (t, t_0)
Platelet count (t, t_0)	Platelet count (t, t_0)	Platelet count (t, t_0)

Other variables included in all three models were pre-treatment (t_0) measurements of age, sex, body mass index, personal history of cardiac disorders, family history of cardiovascular disease, smoking status, abnormal blood pressure, RA duration, statin use, steroid use, NSAID use, and methotrexate use.

RA: Rheumatoid Arthritis; HDL: high-density lipoprotein; LDL: low-density lipoprotein

Table 3.2: Description of imputed data at baseline

Variable	N (%)	Truncation [min, max]
Demographics and health status		
Family history of CHD	4 (0.1)	− − −
Body mass index	21 (0.5)	− − −
Disease status		
Disease activity score (DAS28)	19 (0.5)	[0, 10]
Laboratory values		
HDL cholesterol	262 (6.6)	[0, 4]
LDL cholesterol	285 (7.2)	[0, 11]
Total cholesterol	203 (5.1)	[2, 14]
Apolipoprotein A−I	27 (0.7)	[0, 4]
Apolipoprotein B	27 (0.7)	[0, 4]
C3	125 (3.1)	[0, 4]
C4	213 (5.3)	[0, 2]
Haptoglobin	45 (1.1)	[0, 7.5]
Interleukin-6 receptor	295 (7.4)	[10, 1000]
Lp(a)	26 (0.7)	[5, 300]
Absolute neutrophil count	5 (0.1)	[0.25, 20]
Platelet	14 (0.4)	[3.5, 91]

CHD: cardiovascular disease; HDL: high-density lipoprotein; LDL: low-density lipoprotein
All imputations were truncated to observed ranges of covariate values.

Table 3.3: Description of pre-treatment patient characteristics

	N (%)
Demographics and Health Status	
Duration of RA, mean (SD)	9.3 (8.5)
Body mass index, mean (SD)	27.6 (6.3)
Age, mean (SD)	52 (12.5)
Female	3,286 (82.4)
Family history of CHD	528 (13.2)
History of cardiac disorders	349 (8.8)
Abnormal blood pressure	1,901 (47.7)
Current smoker	715 (17.9)
Concomitant medication use	
Oral steroids	2,307 (57.9)
Statins	356 (8.9)
NSAIDs	2,864 (71.9)
Methotrexate	3,165 (79.4)

RA: rheumatoid arthritis; SD: standard deviation; CHD: coronary heart disease; NSAIDs: non-steroidal anti-inflammatory drugs

Table 3.4: Description of time-varying patient characteristics

	t_0 (n = 3,986) $\bar{X}$ (SD)	t_1 (n = 3,899) $\bar{X}$ (SD)	t_2 (n = 3,558) $\bar{X}$ (SD)
Disease Characteristics			
RA DAS28	6.4 (1.3)	3.6 (1.6)	3.1 (1.5)
Laboratory values			
HDL cholesterol	1.5 (0.4)	1.6 (0.5)	1.6 (0.5)
LDL cholesterol	3.0 (0.9)	3.4 (1.1)	3.5 (1.1)
Total cholesterol	5.1 (1.1)	5.8 (1.3)	5.8 (1.3)
Apolipoprotein A−I	1.5 (0.3)	1.7 (0.3)	1.6 (0.3)
Apolipoprotein B	1.1 (0.3)	1.2 (0.4)	1.2 (0.4)
Lp(a)	33.8 (33.5)	22.5 (24.3)	21.4 (24.3)
C3	1.4 (0.3)	1.1 (0.3)	1.1 (0.3)
C4	0.3 (0.1)	0.2 (0.1)	0.2 (0.1)
C-reactive protein	2.4 (2.9)	0.5 (1.3)	0.4 (1.3)
Erythrocyte sedimentation rate	46.1 (26.8)	13.3 (16.0)	11.0 (16.0)
Interleukin-6 receptor	45.7 (57.2)	429.6 (238.4)	472.4 (238.4)
Albumin	38.1 (3.8)	41.8 (3.5)	41.8 (3.5)
Haptoglobin	2.1 (0.9)	0.7 (0.6)	0.6 (0.6)
Absolute neutrophil count	5.8 (2.4)	4.2 (2.2)	4.0 (2.2)
Platelet count	33.3 (10.0)	24.2 (6.8)	23.4 (6.8)

t_0 corresponds to pre-treatment measurements, t_1 corresponds to measurements taken 6 months post-treatment initiation, and t_2 corresponds to measurements taken 12 months post-treatment initiation.

SD: Standard deviation; RA DAS28: Rheumatoid Arthritis Disease Activity Score out of 28; HDL: high-density lipoprotein; LDL: low-density lipoprotein

Table 3.5: Description of censoring observed in dataset

	$[t_0, t_1)$	$[t_1, t_2)$	$[t_2, t_3)$
N at risk at beginning of interval	3,986 (100%)	3,899 (97.8%)	3,558 (89.3%)
Number of MACE in interval	5 (0.1%)	9 (0.2%)	36 (0.9%)
N censored during interval	82 (2.1%)	332 (8.3%)	2,364 (59.3%)

t_0 is treatment initiation, t_1 is 6 months post-treatment initiation, t_2 is 12 months post-treatment initiation, and t_3 is the end of the study period, 5 years post-treatment initiation.

Table 3.6: Odds Ratios (95% Confidence Intervals) from weighted and unweighted logistic regression models for the pre-treatment effects of rheumatoid arthritis disease activity score, HDL-cholesterol, and LDL-cholesterol on major adverse cardiac events in patients with moderate to severe RA taking tocilizumab.

	Weighted Model OR (95% CI)	Unweighted Model OR (95% CI)
RA DAS28 (Pre-treatment)	1.11 (0.58, 2.11)	**1.36 (1.07, 1.74)**
HDL Cholesterol (Pre-treatment)	1.22 (0.24, 6.19)	0.61 (0.30, 1.24)
LDL Cholesterol (Pre-treatment)	1.08 (0.68, 1.71)	1.29 (0.97, 1.70)
Age (Pre-treatment)	1.02 (0.98, 1.06)	**1.06 (1.04, 1.09)**
Personal History of CD (Pre-treatment)	3.33 (0.56, 19.79)	**2.03 (1.06, 3.90)**

RA DAS28: rheumatoid arthritis disease activity score out of 28; HDL: high density lipoprotein; LDL: low density lipoprotein; CD: cardiac disorder

Statistically significant ($p<0.05$) results are denoted in bold.

Weights incorporate all demographic and health status, concomitant medication, and laboratory values.

Table 3.7: Odds Ratios (95% Confidence Intervals) from marginal structural model and unweighted regression model for the post-treatment effects of RA Disease Activity Score, HDL Cholesterol, and LDL Cholesterol on Major Adverse Cardiac Events in patients with moderate to severe RA taking tocilizumab.

	Marginal Structural Model OR (95% CI)	Unweighted Model OR (95% CI)
RA DAS28 (Post-treatment change)	0.77 (0.50, 1.21)	**1.14 (1.00, 1.31)**
HDL Cholesterol (Post-treatment change)	0.63 (0.21, 1.86)	1.13 (0.86, 1.49)
LDL Cholesterol (Post-treatment change)	1.10 (0.65, 1.84)	0.95 (0.87, 1.04)
DAS28 (Pre-treatment)	1.60 (0.62, 4.15)	**1.29 (1.02, 1.65)**
HDL Cholesterol (Pre-treatment)	0.51 (0.09, 2.91)	0.56 (0.28, 1.13)
LDL Cholesterol (Pre-treatment)	1.77 (0.98, 3.21)	**1.32 (1.00, 1.74)**
Age (Pre-treatment)	1.05 (0.99, 1.11)	**1.07 (1.04, 1.09)**
Personal History of CD (Pre-treatment)	2.57 (0.32, 20.37)	**2.14 (1.14, 4.03)**

RA DAS28: rheumatoid arthritis disease activity score out of 28; HDL: high density lipoprotein; LDL: low density lipoprotein; CD: cardiac disorder
Statistically significant (p<0.05) results are denoted in bold.

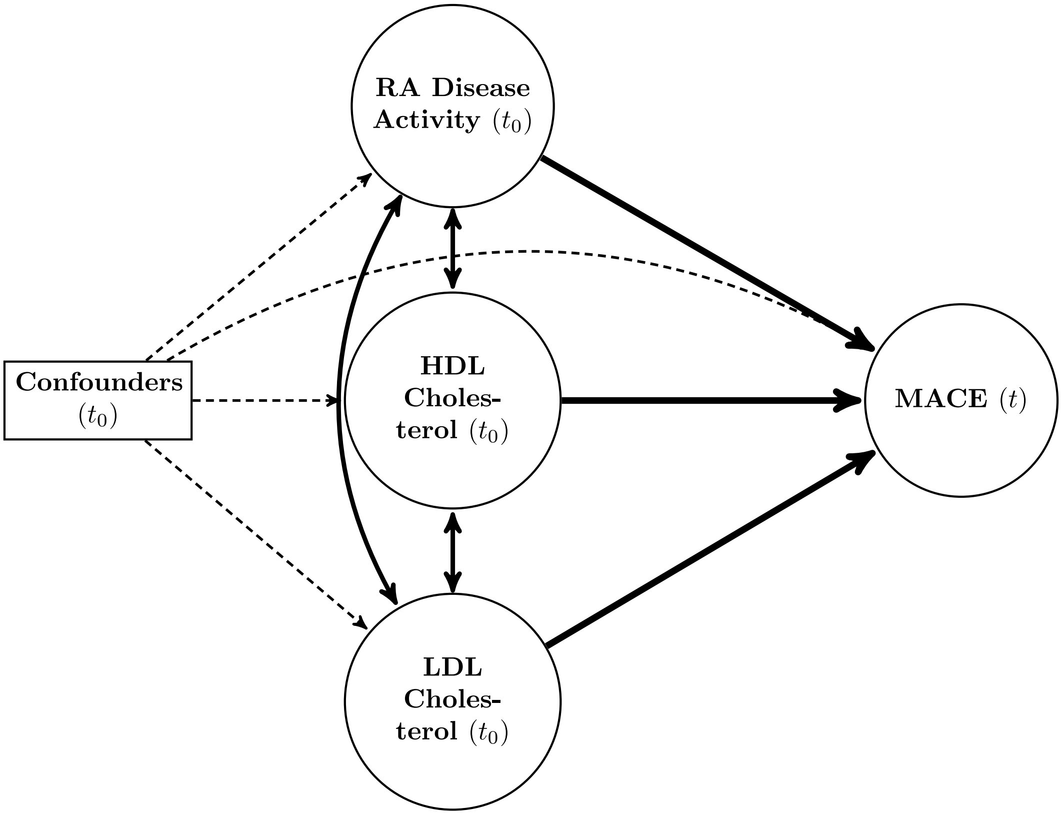
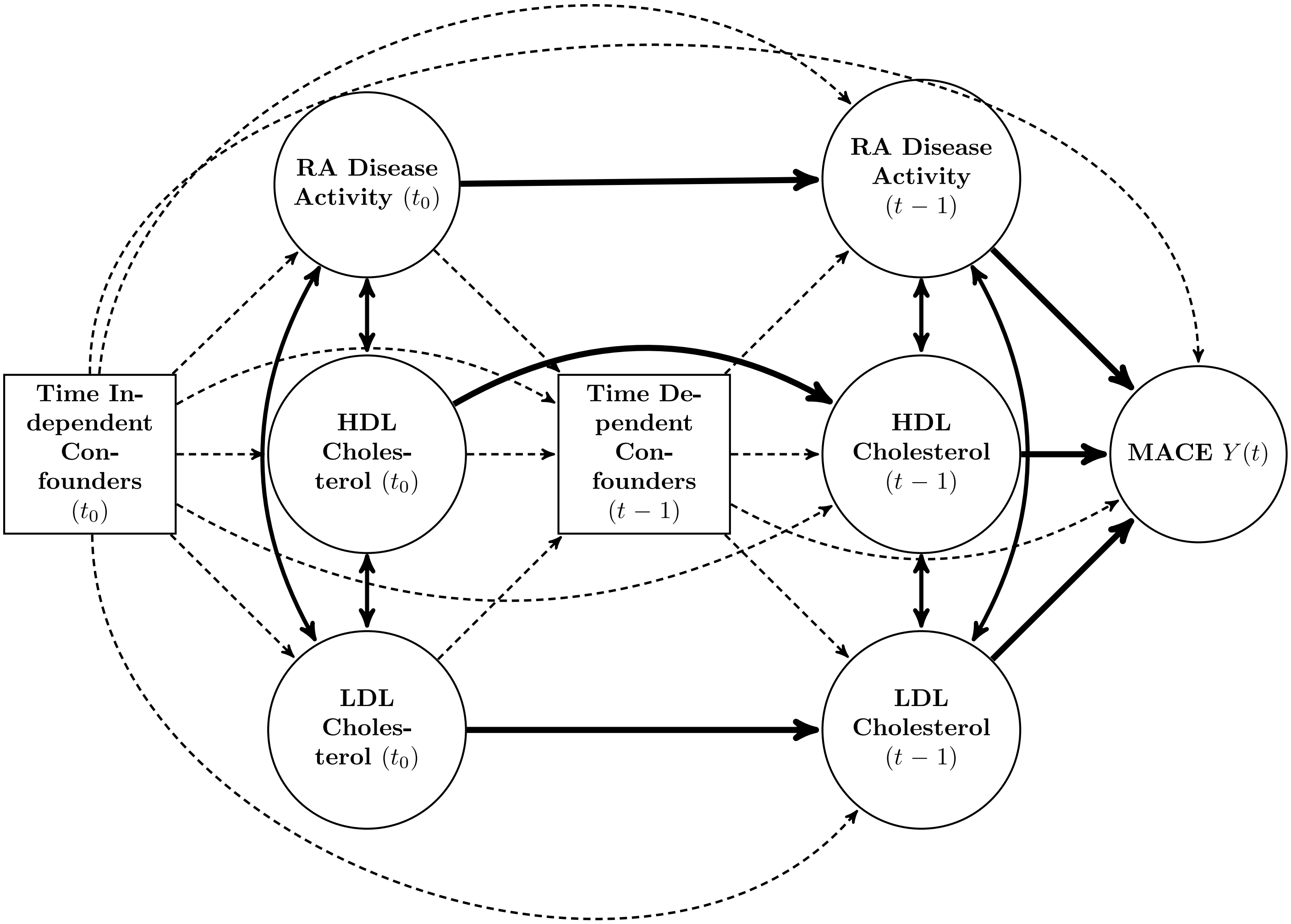

RA Disease Activity (t_0)
Confounders (t_0)
HDL Choles-terol (t_0)
MACE (t)
LDL Choles-terol (t_0)

Figure 3.2: Directed acyclic graph representing the process assumed (when ignoring repeated measures) for studying the risk of having a major adverse cardiovascular event before the end of the study. The principal exposures of interest are shown in circles. Confounders (demographics, health status, concomitant medication use, and pre-treatment initiation levels of all laboratory values) are indicated by a square. Solid thick arrows denote the causal paths of interest. Dashed thin arrows indicate potential pathways for confounding.



RA Disease Activity (t_0)
HDL Cholesterol (t_0)
LDL Cholesterol (t_0)
Time Independent Confounders (t_0)
Time Dependent Confounders ($t-1$)
RA Disease Activity ($t-1$)
HDL Cholesterol ($t-1$)
LDL Cholesterol ($t-1$)
MACE $Y(t)$

Figure 3.3: Directed acyclic graph representing the hypothesized longitudinal process for major adverse cardiovascular events. The principal exposures of interest are shown in circles. Time independent confounders (e.g. demographics, health status, and concomitant medication use) and time dependent confounders (e.g. laboratory values) are shown in squares. Solid thick arrows denote the causal paths of interest. Dashed thin arrows indicate potential pathways for confounding. All confounders used in analyses are listed in table 3.1.

3.3 Implementation of the Marginal Structural Model from Longitudinal Data

Marginal structural models (MSMs) are based in a counterfactual framework, wherein each individual has a set of potential outcomes – one for each possible exposure history. Given a specific exposure history, a potential outcome is defined as a subject's hypothetical outcome if, possibly contrary to fact, they experienced that particular exposure history. Here, we describe the implementation of a MSM to assess the effect of a particular exposure on an outcome in non-randomized data structures.

Notation

Our observed data is $O = \{(V, L(t), A(t), Y(t) : t \in \{0, 1, 2\}\}$. $A(t)$ is a vector of observed rheumatoid arthritis disease activity score (DAS28), HDL cholesterol (HDL-c), and LDL cholesterol (LDL-c) at time t. $L(t)$ is a vector of time-dependent covariates including inflammatory markers, other cholesterol measures, and other disease activity measures. $L(0)$ is the vector of covariates measured at baseline, that includes both time-independent and time-dependent covariates. V is a subset of the covariates measured at baseline, $V \subset L(0)$. $Y(t)$ is an indicator of whether MACE occurred in the interval $[t, t+1)$. Time 0 corresponds with baseline measurements, time 1 corresponds with measurements 6 months from baseline, and time 2 corresponds with measurements 12 months from baseline. The outcome $Y(2)$ is an indicator of MACE in the interval 12 months to 60 months from baseline.

The time ordering of the data are: $\{L(0), A(0), Y(0), L(1), A(1), Y(1), L(2), A(2), Y(2)\}$. Specifically, we are assuming that the covariates L and the exposures A occur in time before Y. We are also assuming that the covariates L are predictors of the exposure A.

Marginal Structural Model

We aim to evaluate a time-dependent process, (L(t), A(t), Y(t), observation times t = 0, 1, 2). We assume that the following theoretical model summarizes the relationship between A(t) and Y(t):

$$logit(Pr[Y(t) = 1|A(t), L(t)]) = \alpha_0 + \beta A(0) + \gamma[A(0) - A(t)] + \delta L(t),$$

where $\beta = (\beta_1, \beta_2, \beta_3)$ are the parameter estimates of the adjusted effects of baseline rheumatoid arthritis DAS28, HDL cholesterol, and LDL cholesterol on the log odds of occurrence of a MACE at time t, $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ are the parameter estimates of the adjusted effects of change from baseline rheumatoid arthritis DAS28, HDL cholesterol, and LDL cholesterol on the log odds of occurrence of a MACE at time t, and δ is a vector of nuisance parameter estimates for time-dependent covariates.

To obtain causal estimates of the effects of the exposure variables on the log-odds of a MACE in a time-dependent setting, we use the following counterfactual marginal structural model:

$$logit(Pr[Y_{\bar{a}}(t) = 1|V]) = \beta_0 + \sum_{j=1}^{3} [\beta_{1j} a_j(0) + \beta_{2j} cum(\bar{a}_j)] + \beta_3 V.$$

Here, $\bar{a} = \{\bar{a}_j\} : j \in \{1, 2, 3\}$ is a particular exposure history $(a_j(0), a_j(1), a_j(2))$ associated with DAS28 ($j = 1$), HDL-c ($j = 2$), and LDL-c ($j = 3$). Each exposure history is associated with a counterfactual outcome, $Y_{\bar{a}}$. $cum(\bar{a}_j)$ is a summary measure of the counterfactual cumulative exposure history, defined here to be the cumulative change from baseline for each exposure. β_{21}, β_{22}, and β_{23} are our causal parameter estimates of the effects of the cumulative change from baseline of DAS28, HDL-c, and LDL-c respectively on the log-odds of having a major adverse cardiac event within strata of V, controlling for baseline levels of DAS28, HDL-c, and LDL-c. We are interested in the marginal effect of the change in DAS28, HDL-c, and LDL-c after beginning a course of tocilizumab, so therefore the βs represent the average effect over time since enrolling in the study, controlling for pre-treatment levels.

Assumptions needed for Identifiability of MSM

- **Positivity.** Each treatment assignment at each time t, conditional on current covariate, past exposure, and past outcome status, must exist with non-zero probability for all individuals in the study. Formally,

  $$Pr[A(t) = a \mid \bar{A}(t-1), \bar{L}(t), \bar{Y}(t-1)] > 0 \;\; \forall a \in \{a(t) : a \in A, \bar{a}(t-1) = \bar{A}(t-1)\},$$

  where A is the set of all possible exposures.

- **Temporality.** The time ordering of our data are:

  $$\{L(0), A(0), Y(0), L(1), A(1), Y(1), L(2), A(2), Y(2)\}.$$

  Most importantly, we assume that our covariates L and exposures A occur earlier in time than the outcome Y, and that covariates L influence the exposure A.

- **Link observed data to counterfactuals.** We must assume that our observed data can be linked to counterfactual data, that is data observed in individuals corresponds to the counterfactual data we would have seen given the exposure observed. More precisely,

  $$(A(t), L(t), Y(t)) \equiv (A(t), L_A(t), Y_A(t))$$

  We note that needing covariates L to be predictors of A implies that $L_{\bar{a}}(t) = L_{\bar{a}(t-1)}(t)$, whereas A predicts Y so $Y_{\bar{a}}(t) = Y_{\bar{a}(t)}(t)$.

- **Sequential randomization.** We assume no unmeasured confounding of $A(t)$, which corresponds to
$$A(t) \perp\!\!\!\perp (Y_{\bar{a}}, L_{\bar{a}} : \bar{a} \in A) \mid \bar{L}(t-1), \bar{A}(t-1)$$

Weight Creation

Probability of Exposure Weights

The stabilized subject- and time-specific probability of exposure weights are calculated as the joint probability of exposure, where each exposure is indexed by $\{1, 2, 3\}$. For our analyses, $A_1(\cdot)$ refers to RA disease activity score, $A_2(\cdot)$ refers to HDL cholesterol, and $A_3(\cdot)$ refers to LDL cholesterol.

$$\begin{aligned} Sw_T(t) = &\prod_{k=0}^{t} \frac{Pr(\mathbf{A}(k) \mid \bar{\mathbf{A}}(k-1), V)}{Pr(\mathbf{A}(k) \mid \bar{\mathbf{A}}(k-1), \bar{L}(k-1))} = \prod_{k=0}^{t} \frac{Pr(A_1(k) \cap A_2(k) \cap A_3(k) \mid \bar{\mathbf{A}}(k-1), V)}{Pr(A_1(k) \cap A_2(k) \cap A_3(k) \mid \bar{\mathbf{A}}(k-1), \bar{L}(k-1))} \\ &= \prod_{k=0}^{t} \frac{Pr(A_1(k) \mid \bar{\mathbf{A}}(k-1), V)}{Pr(A_1(k) \mid \bar{\mathbf{A}}(k-1), \bar{L}(k-1))} \times \frac{Pr(A_2(k) \mid A_1(k), \bar{\mathbf{A}}(k-1), V)}{Pr(A_2(k) \mid A_1(k), \bar{\mathbf{A}}(k-1), \bar{L}(k-1))} \\ &\times \frac{Pr(A_3(k) \mid A_1(k), A_2(k), \bar{\mathbf{A}}(k-1), V)}{Pr(A_3(k) \mid A_1(k), A_2(k), \bar{\mathbf{A}}(k-1), \bar{L}(k-1))} \end{aligned}$$

Because each of these exposures is on a continuous scale, we estimate the conditional density of each exposure separately assuming a normal kernel, the mean of which is estimated using main terms linear regression with fixed effects for time and study. The standard deviation is calculated using the residuals from the fitted model.

Probability of Censoring Weights

The stabilized subject- and time-specific probability of censoring weights are calculated as:

$$Sw_C(t) = \prod_{k=0}^{t} \frac{Pr(C(k) = 0 \mid C(k-1) = 0, \bar{\mathbf{A}}(k-1), V)}{Pr(C(k) = 0 \mid C(k-1) = 0, \bar{\mathbf{A}}(k-1), \bar{Y}(k-1), \bar{L}(k-1))}$$

where we note that each of the numerator and the denominator is the complement of the probability of being lost to follow-up given a specific set of treatment and covariate history. The numerator and denominator were modeled separately using pooled logistic regression, again including fixed effects for time and study identifiers.

Final weights used in analysis

We first calculated for each individual the product of the stabilized exposure and censoring weights, or
$$Sw_i(t) = Sw_{iT}(t) \times Sw_{iC}(t).$$

We truncated the weights in the range $[0.1, 30]$. Results appeared to be fairly stable to other weight ranges.

Chapter 4

Weight gain in pregnancy and preterm birth

4.1 Introduction

Preterm birth is commonly defined as a birth occurring between viability (usually 22 weeks gestation) and 37 weeks gestation [66], and affects more than one in ten infants born worldwide [67]. It is indicative of the health of the infant at birth - infants born after 37 weeks' gestation are more likely to have fully matured (especially their heart and lungs) than those born before 37 weeks. As such, it remains the leading cause of perinatal and infant morbidity and mortality in developed countries [67]. Prematurity is also associated with lifelong health problems, including but not limited to neurological impairment (from mild executive function disorders to moderate-severe developmental delay), visual and hearing impairment, chronic lung and cardiovascular issues, and childhood obesity [68].

Biologically, the mechanisms that cause preterm birth are fairly well established [66, 69]. The best-understood risk factor for preterm birth is multiple births (twins, triplets, etc.), which can be attributed to uterine overdistension and/or higher rates of pre-eclampsia or other maternal or fetal distorders [69, 70]. Intrauterine infections may be responsible for 25-40% of preterm births, though this may be a conservative estimate because these infections are difficult to diagnose [69, 71, 72].

However, in most cases a precise mechanism cannot be identified, which has led researchers to identify other risk factors with the goal of targeting possible interventions. Women who have a history of prior preterm birth and/or miscarriages, or an interpregnancy interval of <6 months are at greater risk of preterm birth. A current mechanistic hypothesis is that the uterus requires sufficient time to return to its normal state before becoming occupied again [69, 73, 74]. Risk of preterm birth also varies by race/ethnicity. In the U.S., black women experience preterm birth at much higher rates than white women [67, 69, 75]. Other

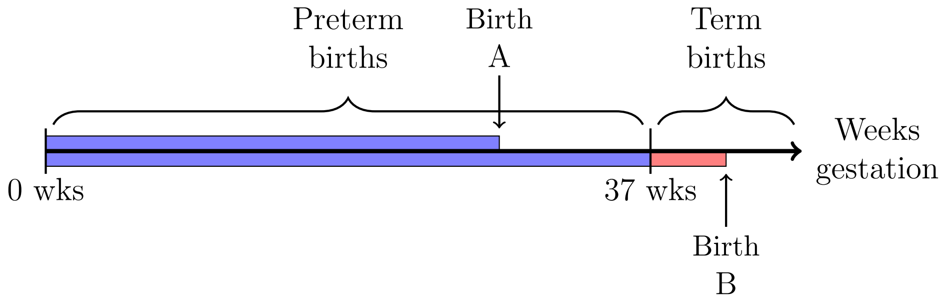


Figure 4.1: An illustration of when term versus preterm births occur. The blue rectangles indicate weight gain that happens while women are still at risk for preterm birth, while the red rectangle indicates weight gain that happens after a woman is not at risk of preterm birth.

sociodemographic characteristics, such as high and low maternal age, low education, and low socioeconomic status, are also associated with preterm birth, though the pathways through which they operate are not well understood [67, 69]. Smoking tobacco during pregnancy is also a risk factor for preterm birth. This mechanism is harder to detangle as there are many chemicals present in smoke, but researchers hypothesize that it could operate through an inflammatory pathway and/or through restricted uteroplacental blood flow, as both nicotine and carbon dioxide are vasoconstrictors [67, 69, 76, 77]. Both extremes of prepregnancy weight are also associated with preterm birth. In particular, one study found that underweight women are nearly four times more likely to have a preterm birth as compared to heavier women [78]. In a Swedish cohort, extreme obesity (BMI $\geq$ 35) more than doubled the odds of extremely early preterm birth compared to normal weight women [79].

Although prepregnancy underweight and obesity are known risk factors for preterm birth, the relation between weight gain during pregnancy and preterm birth is still contested in the literature. This confusion is partly due to difficulties measuring weight gain. Prepregnancy weight, the baseline weight measure that is used as the referent weight for weight change in pregnancy, is often measured by recall, as pregnancies can be unplanned. Additionally, studies of gestational weight gain often do not have the resources to measure weight gain at multiple points through pregnancy. Even if the study is sufficiently funded, the first prenatal visit often occurs between 8 and 14 weeks gestation, so early pregnancy weight gain is often unknown. Delivery weight is also poses difficulties, as weighing a woman who is in labor is generally unadvisable. Instead, delivery weight is often collected by recall a day or two after delivery, or a recorded weight from a prenatal visit within two weeks of delivery is used.

Aside from any unreliability in the measurement of gestational weight gain, there are statistical concerns associated with assessing the relation between weight gain in pregnancy and preterm birth. Weight gain in pregnancy and gestational duration form a joint process, so the question "does weight gain in pregnancy cause preterm birth?" must be carefully specified. For example, studies have used total weight gain during pregnancy as their gestational weight gain measure. These studies are predisposed to finding that low weight gain is associated with preterm birth, as their weight gain measure for term births incorporates weight that was gained after a women stopped being at risk for preterm birth (see Figure 4.1 for an illustration). Other studies have attempted to address how women gain weight in pregnancy by using rate of weight gain, defined as total weight gain in pregnancy divided by gestational duration. However, this measure assumes the rate of weight gain is the same in all three trimesters, which has been established to be false - women gain weight differently in the first trimester than in the later trimesters.

The best measure of total gestational weight gain proposed to date is weight gain-for-gestational age z-scores [80, 81]. These z-scores represent the relation between a woman's total gestational weight gain at a particular week of gestation and a population-average gestational weight gain trajectory. Although these z-scores attempt to remove the bias in the relationship between GWG and gestational duration, this dependency is only removed when the weight gain trajectory model is correctly specified.

Although it is expensive to measure weight at each week of pregnancy while conducting a study, incorporating data on the whole pattern of weight gain is arguably the best way to assess the relationship between gestational weight gain and gestational duration. A recent paper proposes incorporating the repeated measurements of weight gain through pregnancy as a time-varying covariate in a Cox model, to assess the overall hazard of birth for each additional pound of weight [82]. Other people have proposed using latent class analysis to group the women into classes that represent similar weight gain trajectories, and using those classes to predict preterm birth (need reference).

This chapter includes two studies of weight gain in pregnancy and preterm birth. Section 4.2 addresses the black-white disparity in rates of preterm birth in the United States. Although preterm birth rates in the U.S. have been decreasing for the past decade, there are still approximately 5 more preterm births per 100 live births born to black mothers compared to any other race [83]. In our study, we use data from the 2011 to 2015 U.S. Birth Data Files to determine if the association between gestational weight gain and preterm birth differs by race using a novel measure of weight gain, weight gain-for-gestational age z-scores [80, 81]. Section 4.3 concerns the different ways gestational weight gain is measured and used in epidemiological studies. Specifically, we incorporate measures of GWG that attempt to address how and when women gain weight, in addition to the measures that only consider total GWG. In a cohort of Swedish women, we create models to predict the occurrence of preterm birth using seven measures of gestational weight gain separately, assess the predictive ability of each model, and compare the agreement between models.

4.2 Weight gain during pregnancy and the black-white disparity in preterm birth

Introduction

In 2015, one in every eight non-Hispanic (NH) black infants in the U.S. was born preterm (< 37 weeks' gestation) compared with one in every 11 NH white infants [83]. This difference was even greater for early preterm birth; one in every 32 NH black infants was born < 32 weeks' gestation compared with one in every 79 NH white infants [83]. The magnitude of this black-white disparity in preterm birth has persisted throughout the past decade and is not explained by socioeconomic factors [84–86].

Prepregnancy body mass index (BMI) and weight gain during pregnancy may be modifiable contributors to preterm birth and racial disparities in preterm birth rates. [87–89]. Higher rates of preterm birth have been consistently observed in women who are underweight or obese before pregnancy as well as in women who gain little or no weight during pregnancy [88–92]. Additionally, a higher proportion of women who identify as NH black or African American begin pregnancy obese and, if underweight or normal weight, gain below the Institute of Medicine (IOM) gestational weight gain recommendations, compared with NH white women [93, 94]. Previous studies of weight gain and preterm birth have been limited by the methods available to account for the fact that women who deliver preterm have less time to gain weight during pregnancy, and these studies have not focused on racial/ethnic differences except in very limited study samples [95–97].

In this national study, we used gestational weight gain-for-gestational age z-scores to assess, by prepregnancy BMI category: (1) if gestational weight gain is associated with early and late preterm birth, and (2) if such associations differ between NH black and NH white women.

Materials and Methods

Data and Study Design

We used data from the 2011 to 2015 U.S. Birth Data Files created by the National Center for Health Statistics (NCHS). Of the 19,849,688 available birth records, we included those that used the 2003 Revision of the U.S. Standard Certificate of Live Birth (n = 18,445,524) and were live singletons born to NH black or NH white women (n = 11,770,568). We set extreme values of gestational age (< 22 or > 42 week, n = 168,102), birthweight-for-gestational age (following Alexander et al. [98], n = 38,999), and gestational weight gain z-scores (< -4 or > 4 standard deviations (SD), n = 63,342) to missing. [80, 81].

NCHS previously edited values outside the ranges of 100-400 lb. for maternal weight at delivery, 75-375 lb. for prepregnancy weight, and 30-78 in. for height as missing. Additional exclusions include: those missing information on prepregnancy weight or height (n = 425,474), delivery weight (n = 213,790), and infant birthweight (n = 141,660). The final study sample included 10,718,449 births. Because the data are completely de-identified and publicly available, the study did not require ethics approval from an Institutional Review Board.

Measures

We used obstetric estimates to categorize gestational age at delivery as early preterm (< 32 weeks), late preterm (≥ 32 weeks to < 37 weeks), and term (≥ 37 weeks to < 42 weeks), following the World Health Organization guidelines.

Maternal race was self-reported as black (black or African American) or white and ethnicity was self-reported as non-Hispanic (not Spanish/Hispanic/Latino) or other. Prepregnancy BMI was calculated from self-reported prepregnancy weight and height (kg/m^2) and categorized as underweight (< 18.5), normal weight (≥ 18.5 to < 25), overweight (≥ 25 to < 30), obese class I (≥ 30 to < 35), obese class II (≥ 35 to < 40), and obese class III (≥ 40). Prepregnancy BMI and race (NH black and NH white) were effect modifiers of interest [88].

Gestational weight gain was the difference between maternal delivery weight and prepregnancy weight. We converted maternal weight gain into weight gain-for-gestational age z-scores using charts previously created by our team using serial prenatal weight measurements from women in Pennsylvania with healthy, term, singleton births [80, 81]. The charts are prepregnancy BMI-specific (separate charts for underweight through obesity class III categories) because optimal weight gain varies by prepregnancy BMI [88]. We categorized the continuous z-scores as $<$ -1 standard deviation (SD) (low), -1 SD to +1 SD (moderate), and $>$ +1 SD (high) weight gain. For a normal-weight woman at 40 weeks' gestation, these categories are equivalent to $<$ 11.2 kg, 11.2-23 kg, and $>$ 23 kg, respectively [81]. The weight gain equivalents to the categorized z-scores by week of pregnancy are provided in Figure 4.2.

Confounders were selected a priori using theory-based causal graphs [99] and included sociodemographic characteristics (education level, age, marital status, acknowledgment of paternity, method of payment, and calendar year) and health characteristics (pregestational hypertension, pregestational diabetes, maternal height, trimester of entry to prenatal care, interpregnancy interval $<$ 6 months, average number of cigarettes smoked per day in pregnancy, and parity).

Data Analysis

First, we descriptively analyzed prenatal characteristics and calculated crude rates of early and late preterm birth by maternal race. We also calculated crude rates of early and late preterm birth within each subgroup of maternal race, prepregnancy BMI, and pregnancy weight gain.

Next, we estimated associations of categorized weight gain z-scores with early and late preterm birth using multiple logistic regression models. Risk of early preterm birth was calculated over all live births and risk of late preterm birth was calculated over live births $\geq$ 32 weeks' gestation. We used inverse probability weighting with stabilized weights to correct for missing covariate data because 11.2% of the sample was missing information on at least one covariate [100]. Covariate missingness ranged from 0-4% (median 0.7%). We carried out analyses in Stata version MP 14.1. Population attributable risks (PAR) were calculated with the final multiple logistic regression models by the module "regpar" and PAR percentages were calculated with the same models by the module "punaf" [101, 102]. Because of the large sample size, clinical significance was determined by the magnitude of effect estimates and width of confidence intervals in addition to p-values [103].

Results

The crude early preterm birth rate was 2.5 per 100 live births among NH black women, and 0.9 among NH white women. The crude late preterm birth rate was 11.6 per 100 live births in NH black women and 7.2 in NH white women. Two-thirds of NH black women and nearly three-quarters of NH white women gained weight within the moderate range (Table 4.1). A higher proportion of NH black women were overweight or obese prepregnancy, delivered as adolescents, did not have private health insurance, and had a high school education or less, compared to NH white women. A higher proportion of NH white women smoked during pregnancy.

Tables 4.2 and 4.3 display the crude rates of early and late preterm birth, respectively, stratified by race, prepregnancy BMI, and weight gain. The early preterm birth rate was lowest (0.6%) in NH white women who were normal weight prepregnancy and gained weight within the moderate range (equivalent to 7-15 kg at 28 weeks and 11-23 kg at 40 weeks) and highest (5.8%) in NH black women with class III obesity and high weight gain (equivalent to $>$ 13 kg at 28 weeks and $>$ 21 kg at 40 weeks) (Table 4.2). The late preterm birth rate was also lowest (6.2%) in normal-weight, NH white women with moderate weight gain and highest (17.7% and 18.5%, respectively) in obese class III, NH black women with high weight gain and underweight, NH black women with low weight gain (equivalent to $<$ 7 kg at 28 weeks and at 40 weeks) (Table 4.3). For every combination of prepregnancy BMI and gestational weight gain, early and late preterm birth rates were higher in NH black women than in NH white women.

In adjusted analyses, the PAR and PAR percentages for early preterm birth were significant for high - but not low - weight gain during pregnancy among both NH black and NH white women who were obese prepregnancy (Table 4.4). High and low weight gain conferred increased population risks of early preterm birth in non-obese women, but low weight gain was most important among underweight women and high weight gain was most important among overweight women. For high weight gain in all BMI groups, the PAR and PAR percentages were higher in NH black women than in NH white women. Among NH black women, the PAR for high gain ranged from 1.9 to 3.0 and the PAR percentage ranged from 7.8% to 10.2%. That is, if 100 NH black women gained moderate pregnancy weight instead of gaining high, a reduction of 2 to 3 - or 8% to 10% - of early preterm births could be expected, depending on prepregnancy BMI. Among NH white women, the PAR for high weight gain and early preterm birth ranged from 0.5 to 0.9 and the PAR percentage ranged from 5.9% to 7.8%. For low weight gain and early preterm birth, the PAR percentage in underweight women was higher in NH black than NH white women, but the PAR percentages in normal-weight and overweight women were higher in NH white than NH black women.

Table 4.5 shows that low pregnancy weight gain in non-obese women and high weight gain in all women were associated with late preterm birth. For all of these associations, the PAR was higher in NH black women than in NH white women, but the black-white differences in PAR percentages were small or nonexistent. The PAR for high weight gain and late preterm birth ranged from 3.2 to 7.9 in NH black women and from 2.5 to 5.5 in NH white women. The PAR percentage ranged from 2.4% to 6.5% in NH black women and from 2.8% to 6.8% in NH white women.

Discussion

In this national study of more than 10 million recent live births to NH black and NH white women in the U.S., moderate weight gain during pregnancy was associated with a significantly lower risk of both early and late preterm birth in all BMI groups. Preventing high pregnancy weight gain was estimated to have a greater impact on the population risk percentage of early, but not late, preterm birth in NH black women than in NH white women.

Our findings on the associations between gestational weight gain and preterm birth are difficult to compare with prior evidence because almost every study has defined weight gain differently [88, 89]. Most studies measure total weight gain, but using total gain in analysis induces a biased association between low gain and preterm birth because women who deliver preterm have less time to gain weight [97]. Some studies have used various approaches to address this bias, the most common being the division of total gain by gestational age to approximate weekly rate of gain [88–91]. However, this approach is still biased because it over-attributes gain to the first trimester, when the rate of gain is lower [88, 97]. The weight gain-for-gestational age z-scores used in this study improve upon these past approaches by standardizing weight gain for gestational age, similar to fetal weight z-scores [80, 81, 104]. Overall, our results using the z-scores in the U.S. population of NH black and NH white women corroborate substantial evidence of an association between low gain and preterm birth in underweight and normal-weight women [88, 89]. In addition, our study clarifies prior mixed evidence that high weight gain is a risk factor for preterm birth in women of all BMI groups and that low weight gain is not a risk factor in obese women [88–92, 105].

There is very limited previous knowledge on how weight gain during pregnancy may be related to differences in preterm birth between NH black and NH white women [95, 96]. Previously, one study found low weight gain was more strongly associated with preterm birth in NH white women than in NH black women [96], while another found the association to be stronger in NH black women and that high weight gain was only associated with preterm birth in NH black women [95]. However, both of these studies were conducted among relatively small samples of non-obese women who delivered more than 15 years ago and did not separately evaluate early and late preterm birth. Using population attributable risk percentages, our study did not find evidence of the impact of weight gain on late preterm birth differing between NH black and NH white women. We did, however, find that the estimated impact of high weight gain on early preterm birth was highest in NH black women. Early, compared to late, preterm birth may be etiologically distinct, represents a much larger black-white disparity, and accounts for a substantial proportion of infant deaths and acquired developmental disabilities [83, 84]. Multiple hypotheses have been proposed, but current evidence suggests that multiple proximal and distal factors - driven by interpersonal and institutionalized racial discrimination -may interact to cause the black-white disparity in early preterm birth [84–86]. In particular, chronic stress induced by social and economic adversity throughout the life-course may cause dysfunction of the hypothalamic-pituitary-adrenal axis, alter immune functions, and increase oxidative stress, leading to higher rates of preterm birth in NH black women than in NH white women [84–86]. These factors could compound effects of high weight gain during pregnancy or independently affect both weight gain and preterm birth. In our study, we also found PAR percentages to be higher in NH white women than in NH black women for early preterm birth in normal-weight and overweight women, which is unexplained but supports a similar finding by Hickey et al. [96] in non-obese black and white women.

There are several limitations of this observational study. Because we only included NH black and NH white women to focus on the black-white disparity in preterm birth, our results are not generalizable to other races and ethnicities. Maternal BMI and weight gain may have different effects on spontaneous and medically indicated preterm births [92, 95], but it was not possible to validly distinguish these births using vital records [106]. We were also unable to assess potential mediators of the association between weight gain and preterm birth, such as pre-eclampsia, or exclude infants with congenital anomalies because of inconsistent reporting on the birth certificates [106]. Missing data and reporting errors are also weaknesses of the birth certificate data and self-reported prepregnancy weight and height may have resulted in some misclassification of prepregnancy BMI and gestational weight gain. Results from a recent systematic review suggest that misclassification due to self-report may have minimal impact on study findings, but we did not test this assumption in the current study [107].

Birth certificate data allow the calculation of total gestational weight gain, but precluded us from assessing pattern of weight gain or using time-to-event analysis [82, 96, 108]. Previous studies have reported mixed findings on the pattern of weight gain and preterm birth, which will require further study in other datasets [82, 96, 108]. Recently, a multi-country, prospective study found that weight gain z-scores in healthy pregnancies were similar across populations [109], which suggests that the z-scores applied in our study - although developed in a U.S. subpopulation - may be appropriate to use in other populations. However, we could not test this assumption [110].

The limitations of this study are counterbalanced by a number of strengths. The study sample was drawn from all births in the U.S. from 2011 to 2015, making the results highly generalizable to the contemporary U.S. population of NH black and NH white women. The size of the study sample also provided sufficient statistical power to study the associations between a wide range of weight gain amounts and early and late preterm birth across six categories of prepregnancy BMI in NH black and NH white women. Our use of weight gain-for-gestational age z-scores accounted for the dependence of gestational weight gain on gestational duration, which has previously prevented valid assessment of weight gain and preterm birth [80].

In 2009, the IOM revised guidelines for weight gain during pregnancy to no longer recommend higher weight gain in black women, and to include a specific, relatively narrow range of recommended gain in obese women [88]. There were limited data to guide these recommendations and the IOM report called for additional research. Our study responds to this need and may be useful for future weight gain guidelines by investigating the relationships between weight gain, maternal race, and prepregnancy BMI. Additionally, our results add to evidence that improving maternal weight gain could reduce the risk of preterm birth [88, 89]. Although behavior change is challenging, there is growing evidence that moderate weight gain can be promoted effectively [87, 111]. As a recent example, Herring et al. [112] found in a pilot randomized clinical trial that socioeconomically disadvantaged, African-American women who received a behavioral intervention were half as likely to gain weight above the IOM guidelines as women who received usual prenatal care. In a similar study population, Schulman and Kottke [113] reported that a pregnant woman's accurate knowledge of her recommended range of weight gain was associated with appropriate weight gain, but only 27% of women were aware of their recommended gain. Provider advice about gestational weight gain has been strongly associated with actual weight gain, but the majority of pregnant women - particularly those with a low BMI - do not receive accurate advice on weight gain [114, 115]. Moreover, the built and social environments in which women live and work are likely contributors to their ability to gain a healthy amount of weight in pregnancy [88, 115, 116]. Our study findings highlight the need for continued research to determine how moderate weight gain during pregnancy can best be promoted at individual, community, and policy levels. Although our results are not from an intervention, they suggest that women achieving moderate weight gain could help meet the Healthy People 2020 goal of reducing preterm birth in the U.S. by 10%, as well as reduce the gap in early preterm birth rates between NH black and NH white women.

Tables and Figures

Table 4.1: Characteristics of the study sample and rates of preterm birth by maternal race, U.S. singleton pregnancies in non-Hispanic black and non-Hispanic white women (2011-2015)

Characteristic	Non-Hispanic Black (n = 2,179,626)			Non-Hispanic White (n = 8,538,823)		
	Column %	Preterm births per 100 live births		Column %	Preterm births per 100 live births	
		Early	Late		Early	Late
Preterm birth		2.5	11.6		0.9	7.2
Pregnancy weight gain (z-score)						
Low (<-1 SD)	22.5	2.5	12.4	17.1	1.1	8.0
Moderate (-1 to 1 SD)	66.4	2.3	10.9	72.6	0.8	6.6
High (>1 SD)	11.2	4.0	15.0	10.4	1.4	10.0
Prepregnancy BMI (kg/m^2)						
Underweight (<18.5)	3.6	2.9	14.5	4.0	1.3	9.6
Normal weight (18.5-24.9)	36.2	2.4	11.9	50.3	0.8	6.8
Overweight (25-29.9)	26.9	2.4	11.1	23.8	0.8	7.0
Obese class I (30-34.9)	16.8	2.7	11.3	12.0	1.0	7.5
Obese class II (35.0-39.9)	8.9	2.8	11.5	5.9	1.1	7.9
Obese class III ($\geq$ 40)	7.6	3.0	11.6	4.1	1.3	8.5
Maternal age (y)						
10-19	10.8	2.6	12.6	5.1	1.4	9.2
20-24	31.3	2.3	11.2	20.2	1.0	7.7
25.29	26.7	2.4	10.9	30.5	0.8	6.7
30-34	19.3	2.6	11.5	29.1	0.7	6.5
$\geq$ 35	11.9	3.3	13.5	15.2	1.0	8.0
Maternal education						
Less than high school	17.5	2.8	13.9	8.3	1.4	9.9
High school or equivalent	33.0	2.6	12.2	21.4	1.1	8.4
Some college	34.4	2.5	10.9	31.4	0.9	7.4
College graduate	15.2	2.1	9.2	38.9	0.6	5.8
Average cigarettes smoked smoked per day in pregnancy						
None	93.1	2.5	11.4	87.8	0.8	6.7

1-9	5.2	3.4	14.3	6.6	1.5	9.4
10-19	1.2	3.5	15.7	4.2	1.5	11.0
$\geq$ 20	0.5	3.8	17.2	1.5	1.8	12.2
Marital status						
Married	28.5	2.1	10.2	71.1	0.7	6.5
Not married	71.5	2.7	12.2	28.9	1.3	8.9
Source of payment						
Private insurance	26.0	2.4	10.2	61.6	0.7	6.4
Other	74.0	2.6	12.1	38.5	1.1	8.5
Trimester of entry to prenatal care						
First	65.3	2.5	11.2	80.7	0.8	6.9
Second	25.4	2.3	11.9	15.3	1.0	7.7
Third or never	9.3	2.7	12.7	4.0	1.3	9.3

Table 4.2: Crude rates of early preterm birth by maternal race, prepregnancy BMI, and pregnancy weight gain z-score categories, U.S. singleton pregnancies in non-Hispanic black and non-Hispanic white women (2011-2015)

Prepregnancy BMI	Weight gain category	Weight gain (kg) equivalent at 28 wk (kg)	Weight gain (kg) equivalent at 28 wk (kg)	Non-Hispanic Black			Non-Hispanic White		
				Total	Cases	Per 100 live births	Total	Cases	Per 100 live births
Underweight (<18.5 kg/m^2)	Low	< 6.6	< 6.8	16,151	691	4.28	50,422	1,028	2.04
	Moderate	6.6-14.0	6.8-21.4	50,662	1,207	2.38	240,240	2,654	1.10
	High	> 14.0	> 21.4	11,909	383	3.22	51,724	829	1.60
Normal weight (18.5-24.9 kg/m^2)	Low	<6.7	<11.2	215,441	5,800	2.69	797,776	9,116	1.14
	Moderate	6.7-14.9	11.2-23.0	475,353	9,696	2.04	3,028,308	19,500	0.64
	High	>14.9	> 23.0	98,684	3,194	3.24	467,598	5,136	1.10
Overweight (25.0-29.9 kg/m^2)	Low	< 4.8	< 9.0	145,517	3,302	2.27	352,545	3,415	0.97
	Moderate	4.8-16.2	9.0-24.6	387,488	8,571	2.21	1,499,404	10,793	0.72
	High	> 16.2	> 24.6	53,648	2,127	3.96	181,207	2,390	1.32
Obese class I (30.0-34.9 kg/m^2)	Low	< 2.3	< 5.6	74,020	1,708	2.31	167,912	1,593	0.95
	Moderate	2.3-14.8	5.6-22.3	250,451	6,315	2.52	743,843	6,973	0.94
	High	> 14.8	> 22.3	40,744	1,768	4.34	108,450	1,696	1.56
Obese class II (30.0-34.9 kg/m^2)	Low	< -0.8	< 1.9	26,557	426	1.60	61,309	495	0.81
	Moderate	-0.8-12.9	1.9-21.1	146,045	3,897	2.67	393,549	4,237	1.08
	High	> 12.9	> 21.9	22,276	1,137	5.10	48,450	1,002	2.07
Obese class III (30.0-34.9 kg/m^2)	Low	< -3.9	< -2.3	12,848	341	2.65	26,752	360	1.35
	Moderate	-3.9-12.8	-2.3-20.8	136,208	3,712	2.73	293,887	3,557	1.21
	High	> 12.8	> 20.8	15,624	913	5.84	25,447	648	2.55

Weight gain categories are defined using z-score values as low (<-1 standard deviation), moderate (-1 to 1 SD), and high (>+1 SD).

Table 4.3: Crude rates of late preterm birth by maternal race, prepregnancy BMI, and pregnancy weight gain z-score categories, U.S. singleton pregnancies in non-Hispanic black and non-Hispanic white women (2011-2015)

Prepregnancy BMI	Weight gain category	Weight gain (kg) equivalent at 28 wk	Weight gain (kg) equivalent at 28 wk	Non-Hispanic Black			Non-Hispanic White		
				Total	Cases	Per 100 live births	Total	Cases	Per 100 live births
Underweight	Low	< 6.6	< 6.8	16,151	2,980	18.45	50,422	6,282	12.46
(<18.5 kg/m^2)	Moderate	6.6-14.0	6.8-21.4	50,662	6,685	13.20	240,240	21,015	8.75
	High	> 14.0	> 21.4	11,909	1,722	14.46	51,724	5,464	10.56
Normal	Low	<6.7	<11.2	215,441	29,231	13.57	797,776	65,708	8.24
weight	Moderate	6.7-14.9	11.2-23.0	475,353	51,540	10.84	3,028,308	187,133	6.18
(18.5-24.9 kg/m^2)	High	>14.9	> 23.0	98,684	13,053	13.23	467,598	40,111	8.58
Overweight	Low	< 4.8	< 9.0	145,517	16,479	11.32	352,545	25,830	7.33
(25.0-29.9 kg/m^2)	Moderate	4.8-16.2	9.0-24.6	387,488	40,640	10.49	1,499,404	97,143	6.48
	High	> 16.2	> 24.6	53,648	8.014	14.94	181,207	18,926	10.44
Obese	Low	< 2.3	< 5.6	74,020	8,087	10.93	167,912	12,257	7.30
class I	Moderate	2.3-14.8	5.6-22.3	250,451	27,063	10.81	743,843	53,245	7.16
(30.0-34.9 kg/m^2)	High	> 14.8	> 22.3	40,744	6,046	14.84	108,450	11,206	10.33
Obese	Low	< -0.8	< 1.9	26,557	2,650	9.98	61,309	4,108	6.70
class II	Moderate	-0.8-12.9	1.9-21.1	146,045	16,068	11.00	393,549	29,438	7.48
(30.0-34.9 kg/m^2)	High	> 12.9	> 21.9	22,276	3,733	16.76	48,450	6,140	12.67
Obese	Low	< -3.9	< -2.3	12,848	1,396	10.87	26,752	2,122	7.93
class III	Moderate	-3.9-12.8	-2.3-20.8	136,208	14,965	10.99	293,887	23,591	8.03
(30.0-34.9 kg/m^2)	High	> 12.8	> 20.8	15,624	2,769	17.72	25,447	3,676	14.45

Weight gain categories are defined using z-score values as low (<-1 standard deviation), moderate (-1 to 1 SD), and high (>+1 SD).

Table 4.4: Population attributable risks and population attributable risk percentages for early preterm birth (< 32 wk) due to low and high pregnancy weight gain, stratified by race and prepregnancy BMI and adjusted for confounders, in U.S. singleton pregnancies in non-Hispanic black and non-Hispanic white women (2011-2015)

	Population Attributable Risks per 100 Live Births		Population Attributable Risk Percentages (95% CI)	
Pregnancy Weight Gain Category	Non-Hispanic Black	Non-Hispanic White	Non-Hispanic Black	Non-Hispanic White
Underweight (BMI < 18.5 kg/m^2)				
Low	3.7	1.3	13.7	11.1
	(2.8, 4.6)	(1.1, 1.6)	(10.4, 16.8)	(9.1, 13.0)
Moderate	Reference	Reference	Reference	Reference
High	1.9	0.7	7.8	5.9
	(1.2, 2.6)	(0.5, 0.9)	(4.9, 10.6)	(3.9, 7.8)
Normal weight (BMI 18.5-24.9 kg/m^2)				
Low	1.2	0.8	5.7	11.6
	(1.0, 1.5)	(0.8, 0.9)	(4.5, 7.0)	(10.8, 12.3)
Moderate	Reference	Reference	Reference	Reference
High	2.0	0.5	9.3	7.3
	(1.8, 2.2)	(0.5, 0.5)	(8.3, 10.3)	(6.7, 8.0)
Overweight (BMI 25.0-29.9 kg/m^2)				
Low	0.1	0.4	0.5	5.0
	(-0.1, 0.4)	(0.3, 0.4)	(-0.7, 1.8)	(4.0, 5.9)
Moderate	Reference	Reference	Reference	Reference
High	2.1	0.5	8.8	7.2
	(1.8, 2.3)	(0.4, 0.6)	(7.9, 9.7)	(6.4, 8.0)
Obese Class I (BMI 30.0-34.9 kg/m^2)				
Low	-0.2	0.0	-0.9	0.4
	(-0.5, 0.1)	(-0.1, 0.1)	(-2.2, 0.3)	(-0.7, 1.5)
Moderate	Reference	Reference	Reference	Reference
High	2.3	0.6	8.5	6.5
	(2.0, 2.6)	(0.5, 0.7)	(7.4, 9.7)	(5.4, 7.5)
Obese Class II (BMI 35.0-39.9 kg/m^2)				
Low	-1.2	-0.2	-5.1	-2.2
	(-1.5, -0.9)	(-0.3, -0.1)	(-6.2, -3.9)	(-3.3, -1.1)
Moderate	Reference	Reference	Reference	Reference
High	2.9	0.9	10.2	7.8
	(2.5, 3.4)	(0.7, 1.0)	(8.7, 11.7)	(6.5, 9.1)

	Obese Class III (BMI $\geq$ 40.0 kg/m^2)			
Low	0.1 (-0.2, 0.4)	0.1 (0.0, 0.3)	0.3 (-0.7, 1.3)	1.1 (6.5, 9.1)
Moderate	Reference	Reference	Reference	Reference
High	3.0 (2.5, 3.4)	0.9 (0.7, 1.1)	10.0 (8.6, 11.4)	7.1 (5.8, 8.4)

BMI, body mass index; CI, confidence interval
All results are adjusted for maternal age, education level, marital status, acknowledgement of paternity, method of payment for delivery, calendar year, pregestational hypertension, pregestational diabetes, height, trimester of entry to prenatal care, interpregnancy interval $<$ 6 months, cigarettes smoked per day in pregnancy, and parity.

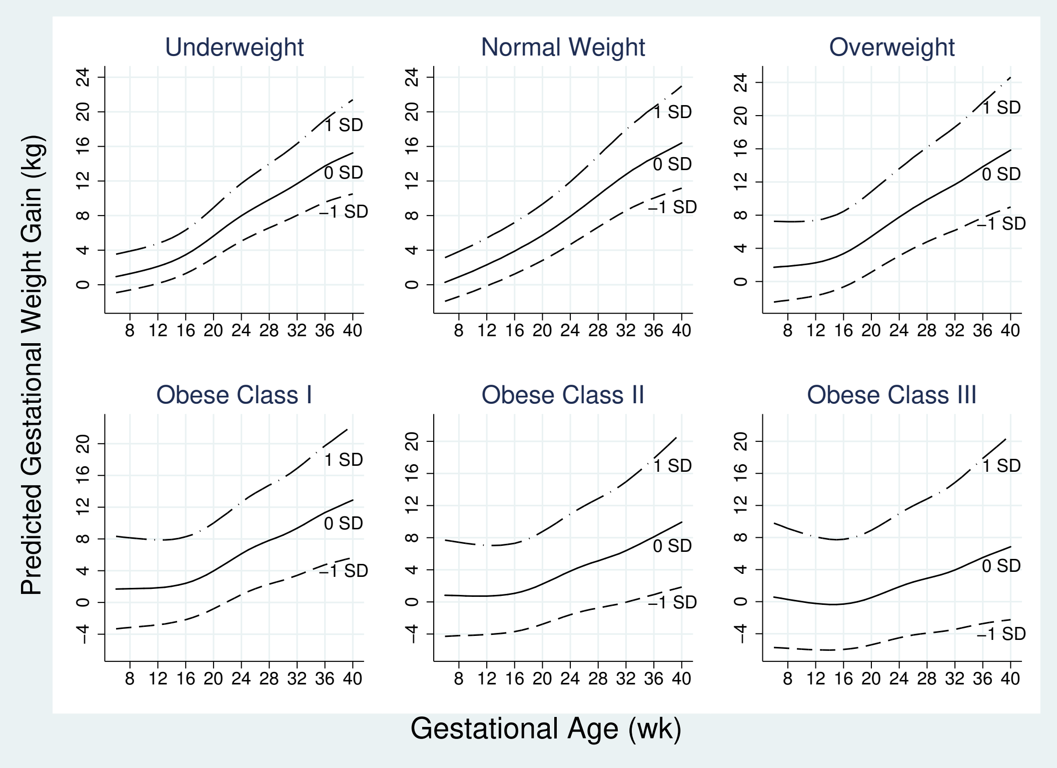
Table 4.5: Population attributable risks and population attributable risk percentages for late preterm birth ($\geq$ 32 wk and $<$ 37 wk) due to low and high pregnancy weight gain, stratified by race and prepregnancy BMI and adjusted for confounders, in U.S. singleton pregnancies in non-Hispanic black and non-Hispanic white women (2011-2015)

	Population Attributable Risks per 100 Live Births		Population Attributable Risk Percentages (95% CI)	
Pregnancy Weight Gain Category	Non-Hispanic Black	Non-Hispanic White	Non-Hispanic Black	Non-Hispanic White
	Underweight (BMI < 18.5 kg/m^2)			
Low	11.2	5.7	7.6	6.1
	(9.4, 13.0)	(5.1, 6.3)	(6.4, 8.8)	(5.5, 6.7)
Moderate	Reference	Reference	Reference	Reference
High	3.2	2.5	2.4	2.8
	(1.8, 4.7)	(2.0, 3.1)	(1.3, 3.4)	(2.2, 3.4)
	Normal weight (BMI 18.5-24.9 kg/m^2)			
Low	6.5	3.2	5.5	4.9
	(5.9, 7.1)	(3.0, 3.3)	(5.0, 6.0)	(4.6, 5.1)
Moderate	Reference	Reference	Reference	Reference
High	4.5	2.7	4.0	4.2
	(4.1, 5.0)	(2.6, 2.8)	(3.6, 4.4)	(4.0, 4.4)
	Overweight (BMI 25.0-29.9 kg/m^2)			
Low	1.2	0.7	1.1	1.1
	(0.6, 1.7)	(0.6, 0.9)	(0.5, 1.6)	(0.8, 1.4)
Moderate	Reference	Reference	Reference	Reference
High	5.9	4.1	5.3	6.0
	(5.5, 6.3)	(4.0, 4.3)	(4.9, 5.6)	(5.8, 6.3)
	Obese Class I (BMI 30.0-34.9 kg/m^2)			
Low	-0.1	-0.1	-0.1	-0.1
	(-0.7, 0.6)	(-0.3, 0.2)	(-0.6, 0.5)	(-0.5, 0.3)
Moderate	Reference	Reference	Reference	Reference
High	6.1	4.0	5.3	5.3
	(5.5, 6.7)	(3.7, 4.3)	(4.8, 5.8)	(4.9, 5.6)
	Obese Class II (BMI 35.0-39.9 kg/m^2)			
Low	-1.5	-1.0	-1.4	-1.4
	(-2.2, -0.9)	(-1.3, -0.7)	(-2.0, -0.8)	(-1.8, -1.0)
Moderate	Reference	Reference	Reference	Reference
High	7.9	5.5	6.5	6.8
	(7.1, 8.6)	(5.1, 5.9)	(5.9, 7.1)	(6.4, 7.3)

	Obese Class III (BMI $\geq$ 40.0 kg/m^2)			
Low	-0.1	-0.1	-0.1	-0.1
	(-0.7, 0.4)	(-0.4, 0.2)	(-0.6, 0.4)	(-0.5, 0.3)
Moderate	Reference	Reference	Reference	Reference
High 7.1	5.1	5.9	5.9	
	(6.3, 7.8)	(4.7, 5.4)	(5.3, 6.5)	(5.5, 6.3)

BMI, body mass index; CI, confidence interval
All results are adjusted for maternal age, education level, marital status, acknowledgement of paternity, method of payment for delivery, calendar year, pregestational hypertension, pregestational diabetes, height, trimester of entry to prenatal care, interpregnancy interval $<$ 6 months, cigarettes smoked per day in pregnancy, and parity.

Figure 4.2: Predicted gestational weight gain at each week of gestation at z-scores of -1 standard deviation (SD), 0 SD, and +1 SD



4.3 Weight gain during pregnancy and preterm birth in a Swedish Cohort

Biologically, gestational weight gain (GWG) and gestational duration (GD) are inextricably linked. The longer a woman is pregnant, the greater opportunity she has to gain weight. This dependence has provided many challenges, both practical and statistical, in quantifying the effect GWG has on GD, resulting in an inconclusive body of literature [104, 108].

Statistically, studying gestational duration poses analytic challenges. As previously discussed in section 4.2, preterm birth is associated with infant mortality and infant and childhood morbidities. The interest of the medical community in this condition has turned a continuous process, time until delivery, into a binary process, wherein instead of modeling the whole "time-to-event" process, we model the chance a women carries to a certain week. This binary approach has a major advantage - results from the models are more interpretable - but at the same time, this approach does not incorporate all of the information available. Other researchers advocate for using a survival model [82], where the hazard of giving birth at each week of gestation is modeled, as opposed to the risk at one week of gestation.

Studying gestational weight gain in relation to gestational-time dependent outcomes requires some thought. In an ideal scenario, data would be collected that contained serial weight measurements on each woman throughout her pregnancy, thereby collecting the whole trajectory. However, many studies do not have the resources to collect serial weight measurements throughout pregnancy, and only contain the cumulative weight change. Historically, cumulative weight gain in pregnancy has been used to study the relation between GWG and preterm birth in two ways: 1) raw total weight gain, and 2) rate of weight gain = total weight gain / total gestational duration. However, there are flaws with each of these measures. Raw total weight gain predisposes studies to finding that low gestational weight gains are associated with preterm birth. Total rate of weight gain does not accurately capture the pattern of GWG in most women - it assumes a linear rate of weight gain in the first trimester, which we know to be not true.

In 2013, Hutcheon et al. proposed another approach to studying GWG in relation to preterm birth: prepregnancy BMI-specific weight gain-for-gestational age z-scores, developed in a reference population [81]. A z-score represents how many standard deviations away from the population average weight gain trajectory a woman is for a given gestational age. However, these z-scores are correct only if the population average weight gain trajectory has been correctly modeled; otherwise, the relationship between GWG and gestational duration will still be biased. To date, these z-scores have been developed in the Swedish population [117], a sample in Pennsylvania [81], and in an international cohort [109].

Mitchell et al. object to all of these approaches in their 2015 paper "It's About Time"

[82]. Their main objection is with the timing of the measurement of weight during pregnancy. They state that women are only at risk for preterm birth before 37 weeks' gestation, so in order to establish an association between weight gain in pregnancy and preterm birth, the weight measurement used should also be from when the women are at risk for preterm birth. Although they allege that repeated measurements of weight gain over pregnancy should be the "gold standard," they also propose a method for using cumulative weight gain before 37 weeks to predict preterm birth.

Another part of their objection is statistical. If we view gestational duration as a random variable, T, preterm birth can be viewed as a summary statistic:

$$PTB = 1(T < 37).$$

Mitchell et al. claim that we should be using a Cox model to estimate the instantaneous hazard of delivering at any particular week of gestation due to a one-unit increase in gestational weight gain. To approximate this quantity, the following model can be fit:

$$log - log[Pr(PTB|GWG, GA)] = \alpha + \beta GWG + \gamma \sum_{i=28}^{36} 1(i = GA),$$

where the GWG measurement used is from the third trimester, taken before a woman leaves the risk set (before 37 weeks' gestation). If the relative risk or odds of preterm birth are desired, then a log or logit model can be used instead of the complementary log-log model.

Is there a most valid measure of weight gain in pregnancy to use when studying the relationship between gestational weight gain and risk of preterm birth? In this study, we 1) describe the pattern of weight gain observed in a subset of the Swedish National Birth Registry Data; 2) create a model of preterm birth using several previously proposed gestational weight gain measures; and 3) compare the predictive ability of each these models. As the relationship between gestational weight gain and preterm birth differs by prepregnancy BMI, and there are other covariate data we wish to include in our model, we use discrete SuperLearner, a data-adaptive machine learning technique, to develop each model by minimizing the cross-validated mean squared error while avoiding overfitting.

Methods

Sample description

Our study sample is a subset of all women who had singleton pregnancies born between January 2008 and October 2014 whose prenatal clinic visits, delivery admissions, and postpartum admissions data from electronic medical records were collected in the Stockholm-Gotland Obstetric Database (n = 175,522) [117]. The cohort was then restricted to live births with available early pregnancy (<14 wks) BMI with plausible gestational duration

(22-42wk) and birthweight for gestational age [118] (n=148,667). We required women to have at least 1 weight measurement in both the second and third trimesters for term births, and at least 1 weight measurement in the second trimester for preterm births to minimize interpolation for analyses that use pattern of weight gain. We further excluded women without available covariate information, producing a final sample of 103,661 pregnancies. Figure 4.3 flow chart further detailing sample selection. This study did not require institutional review board consent as the data are completely de-identified.

Study variables

Preterm birth, defined as any delivery occurring after 22 wks and before 37 wks gestation, was the primary outcome in this study. Gestational age at birth was assessed using obstetric estimates.

Prepregnancy weight measurements unfortunately were not available in this data. To circumvent this problem, we instead used the earliest available first trimester weight measurement (before 14 wks gestation), thereby assuming that there was negligible weight gain the first trimester. This first weight measurement was used in conjunction with measured height at the first visit to calculate early pregnancy BMI (kg/m^2). Early pregnancy BMI was then categorized as underweight (< 18.5), normal weight(18.5 - 24.9), overweight (25-29.9), and obese (≥ 30).

Measurements of gestational weight gain were the primary exposure in this study. In Sweden, women are weighed as part of routine prenatal care, so the electronic medical records contain serial weight gain measurements throughout pregnancy. The first prenatal visit typically occurs between 8 and 12 weeks, with follow-up appointments occurring around weeks 24, 28, 31-32, and then every other week until delivery [117]. Pregnancy weight gain at each visit was calculated as the difference between a woman's current weight and the measured weight from her first prenatal visit (kg).

The following measures of weight gain were of interest in this study:

1. **Overall Rate.** The rate of gestational weight gain after the first trimester, defined as kilograms gained after week 12 divided by the weeks of gestation after week 12 [90]

2. **Trimester-specific rates.** The rate of gestational weight gain in the second and third trimesters separately. Second trimester rate was calculated as kilograms gained after week 12 and before week 28 divided by 16 (total number of weeks gestation in second trimester). Third trimester rate was defined as kilograms gained after week 27 divided by the weeks of gestation after week 27.

3. **Z-scores.** Total GWG, measured as weight gain-for-gestational age z-scores. Total GWG used the weight measurement from the visit closest to delivery (either at delivery or up to two weeks before) [81, 117]

4. **Range of z-scores.** Range of z-score was calculated by taking the difference between the smallest measured z-score and the largest measured z-score (this is my idea: could use this to assess the stability of the weight gain trajectory)

5. **"Comparable" weight.** In term pregnancies, use the most recent cumulative weight gain measurement (kg) recorded before 37 weeks gestation that occurred in the third trimester (after 27 weeks). In preterm pregnancies, use the total GWG measurement (kg). [82]

6. **Pattern.** Used latent class analysis to group overall weight gain trajectories using all weight gain measurements

7. **Pattern28.** Used latent class analysis to group weight gain trajectories using weight gain measurements only up to 28 weeks gestation.

The rate measures of weight gains required interpolation of the weights between visits to get measures of weight gain at exactly 28 weeks. These were assumed to be unchanging in the first trimester, and then to change linearly between visits in the second and third trimesters. The study sample was further restricted to births occurring between 28 and 41 weeks for the **Trimester-specific rates** GWG measure analyses.

Covariate data were prospectively collected in the electronic medical records, and included maternal age at delivery, maternal height (cm), parity, cohabitation, prepregnancy diabetes, prepregnancy hypertension, prepregnancy snus use, prepregnancy smoking status, and child sex. As weight gain z-scores were developed separately in each prepregnancy BMI class, early pregnancy BMI was treated as a necessary effect modifier for those analyses [117]. For all other analysis, early pregnancy BMI was treated as a potential effect modifier.

Statistical Analysis

Exploratory analysis

To ensure we had not systematically excluded pregnancies, we compared the demographic characteristics of our eligible sample to our final sample. We then began by describing the demographic characteristics of the sample stratified by preterm birth. We then described the probability of delivery at each week of gestation. (survival curve) We also graphed the average weight gain at each week of gestation stratified by early pregnancy BMI and preterm birth. To better understand the differences between these curves, we plotted their difference.

Model Building

We built seven models to predict the probability that a woman gave birth preterm (gestational age < 37 weeks), one for each measure of gestational weight gain described above, using the following process.

1. First, we determined the best fit for the relationship between the GWG measure and preterm birth. We used the Latent Class Analysis Stata plugin (cite LCA) to identify the weight gain trajectory groups for measures **Pattern** and **Pattern28**. These trajectories are visualized in figures x and y. For the continuous measures (**Overall rate**, **Trimester-specific rates**, **Z-scores**, **Range of z-scores**, and **"Comparable" weight**), we fit logistic regression models incorporating the GWG measure as linear, quadratic, piecewise linear splines with knots placed at 14w and 28w gestation, and restricted cubic splines with 3, 4, and 5 knots terms, and then chose the model that minimized the 10-fold cross-validated mean squared error (MSE).

2. Once the shape of the relation between the GWG measure and preterm birth had been established, we then determined which covariates were confounders by adding each covariate separately to the GWG-preterm model. To account for multiple testing, covariates were considered significant at $p < 0.006$ (cite Bonferroni).

3. We then allowed for interaction between each confounder and the GWG measure, and searched the covariate space for the combination of variables that minimizes the 10-fold cross-validated MSE. This process was achieved by:

   a) Splitting the data into 10 subsets.

   b) Fitting a best subset logistic regression model in $9/10^{th}$s of the data, excluding the data in the j^{th} fold. This algorithm selects the logistic regression model that maximizes the model R^2. The covariate space this model searched contained the GWG measure, each confounder determined in step 2, and the interaction between the GWG measure and each confounder. The algorithm was further required to keep the GWG measure

   c) Predicting the probability of preterm birth $\hat{y}$ in the remaining $1/10^{th}$ of the data (from fold j) using the model selected in the previous step. We then calculated the squared error in fold j as

   $$SSE(j) = \sum_{i \in \text{Fold}(j)} (y_i - \hat{y}_i)^2$$

   d) Repeating steps b) and c) for each fold of the data.

   e) The final 10-fold cross-validated MSE was calculated as

   $$MSE = \frac{1}{10} \sum_{j=1}^{10} SSE(j)$$

Model Comparison

Once we had built all seven models, we assessed the comparative validity of the models with measures of calibration, discrimination, and classification agreement.

Calibration quantifies the accuracy of a prediction model by comparing the number of observed cases to the number of expected cases predicted by the model. We assessed each model's calibration overall and within risk quintiles - five equally-sized groups of the validation cohort ordered by risk. A score of 1 for the expected-to-observed ratio implies a perfectly calibrated model, less than 1 is an underestimating model, and greater than 1 is an overestimating model. To quantify the calibration in terms of number of over- or under-estimated preterm births, we also reported the expected-to-observed difference (0 is a perfectly calibrated model, greater than 0 is an over-estimating model, less than 0 is an under-estimating model). Confidence intervals for the expected-to-observed ratio and expected-to-observed difference were computed based on Normal approximations to Poisson distributions, as demonstrated below.

$$\begin{aligned}\text{95\% CI for } \frac{E}{O}: &\quad \frac{E}{O} \times \exp\left[\pm 1.96 * \sqrt{\frac{1}{O}}\right] \\ \text{95\% CI for } E - O: &\quad (E - O) \pm 1.96 \times \sqrt{O}\end{aligned}$$

For discrimination, we calculated the area under the receiver operating characteristic curve (AUC, also referred to as the c-statistic), which represents the probability that a model will assign a higher risk to a randomly selected preterm birth than a randomly selected term birth. Nonparametric binomial confidence intervals were calculated using the Stata command "roctab".

Classification agreement between models was assessed with a weighted Cohen's κ statistic, which takes a value between -1 and 1 (<0=no agreement, 1=perfect agreement) and indicates moderate agreement above 0.4, substantial agreement above 0.6, and almost perfect agreement above 0.8. The κ statistic was applied to every pairwise model comparison using the groups formed by each model's risk quintiles, assigning an integer value to each quintile group (1=lowest risk, 5=highest risk). We first assessed overall agreement between the risk quintiles, and then agreement in only in the highest risk quintile.

Results

Descriptive Results

The rate of preterm birth was substantially higher in pregnancies included (5.2%) in our study sample than in pregnancies we excluded (2.4%) (Table 4.6). The most striking difference was that excluded women had fewer weight measurements through pregnancy (76% of excluded women only had 3 weight measurements versus 13% of included women), although they appeared to both have their first prenatal visit during their 9^{th} week of pregnancy on average. Women excluded from our sample had higher rates of snus use prepregnancy and were more likely to be normal weight. A higher proportion of women included in our sample were primiparous, had prepregnancy diabetes and/or hypertension, and smoked prepregnancy. However, on average, early pregnancy weight, gestational age at the first visit, GWG z-scores at delivery, maternal age, maternal height, and child birthweight were not substantially different between included and excluded pregnancies.

In our final sample of 103,661 pregnancies, the differences between term and preterm births were not extreme (table 4.7). A slightly higher proportion of women who had preterm births were underweight, overweight, or obese, primiparous, had prepregnancy diabetes and/or hypertension, and smoked prepregnancy. On average, the age of the mothers was the same in the two groups, though mothers of preterm babies tended to be slightly shorter than mothers of term babies. Although on average the GWG z-score at delivery was the same between the two groups, the GWG (kg) at the last visit before 37 weeks was higher on average in term births (11.8kg) than in preterm births (9.4kg). A greater proportion of preterm births were male (54.2% versus 50.7%), and on average infants carried to term weighed 1kg more than infants born prematurely.

Figure 4.4 illustrates the probability of not delivering at each week of gestation in our sample. Overall, 5.2% of our sample gave birth before 37 weeks. 0.2% were extremely preterm (< 28w), 0.4% were very preterm (28 to <32w), and 4.6% were moderate to late preterm (32 to < 37w).

Figures 4.5 and 4.6 illustrate the average weight gain at each week of gestation stratified by early pregnancy BMI and preterm birth. Figure 4.5 illustrates that the average weight gain trajectories in underweight, normal weight, and overweight women are approximately the same, while obese women gain less weight overall at a slower rate. This pattern is apparent in both term and preterm births, although the trajectories in preterm births were generally less stable than in term births. This instability is probably due to smaller sample sizes. Figure 4.6 demonstrates that the average weight gain trajectories in term and preterm births are approximately the same through 36 weeks' gestation. However, figure 4.7 illustrates that there may be a trend in these small differences (the absolute range of difference is <4kg). In overweight and obese women, as gestational age increases, preterm women tend to gain more weight on average than term women. In normal weight women, there is no substantial difference. In underweight women, preterm women tended to gain less weight than term women in the second and third trimesters.

The results from the model building and model agreement analyses will be reported and discussed at a later date.

Conclusions

In this preliminary study of electronic medical record data from 103,661 Swedish women, we found that there is not a strong relationship between gestational weight gain and preterm birth. Specifically, the pattern of weight gain between term and preterm births appears to be very similar. This preliminary finding confirms the results presented in Sharma et al. (2015) [119]; however, we will still finish the project as proposed here.

Future work in this dataset will include the following:

- We will directly compare using z-scores as specified in Johansson et al. (2016) [117] to the survival method proposed in Mitchell et al. (2015) [82]. We will do this separately in spontaneous and indicated preterm births, and account for the measured covariates available in this dataset.

- We will expand the model building portion of this project to include additional commonly used measures of gestational weight gain, such as overall rate of gestational weight gain, trimester-specific rates, area-under-the-gestational-weight-gain-curve, and latent class analysis.

- Gestational duration (GD) can be thought of as a random variable. We can then formulate preterm birth as a current status random variable, where the only information we have about a birth is whether GD happened before the censoring time C (GD $\leq$ C) or that it could happen after the censoring time (GD $>$ C). This formulation opens us up to the body of current status data literature, where both parametric and nonparametric methods have been developed to assess the relation of a time-varying covariate with a censoring variable. Specifically, we will investigate both proportional and additive hazard models in this context.

Tables and Figures

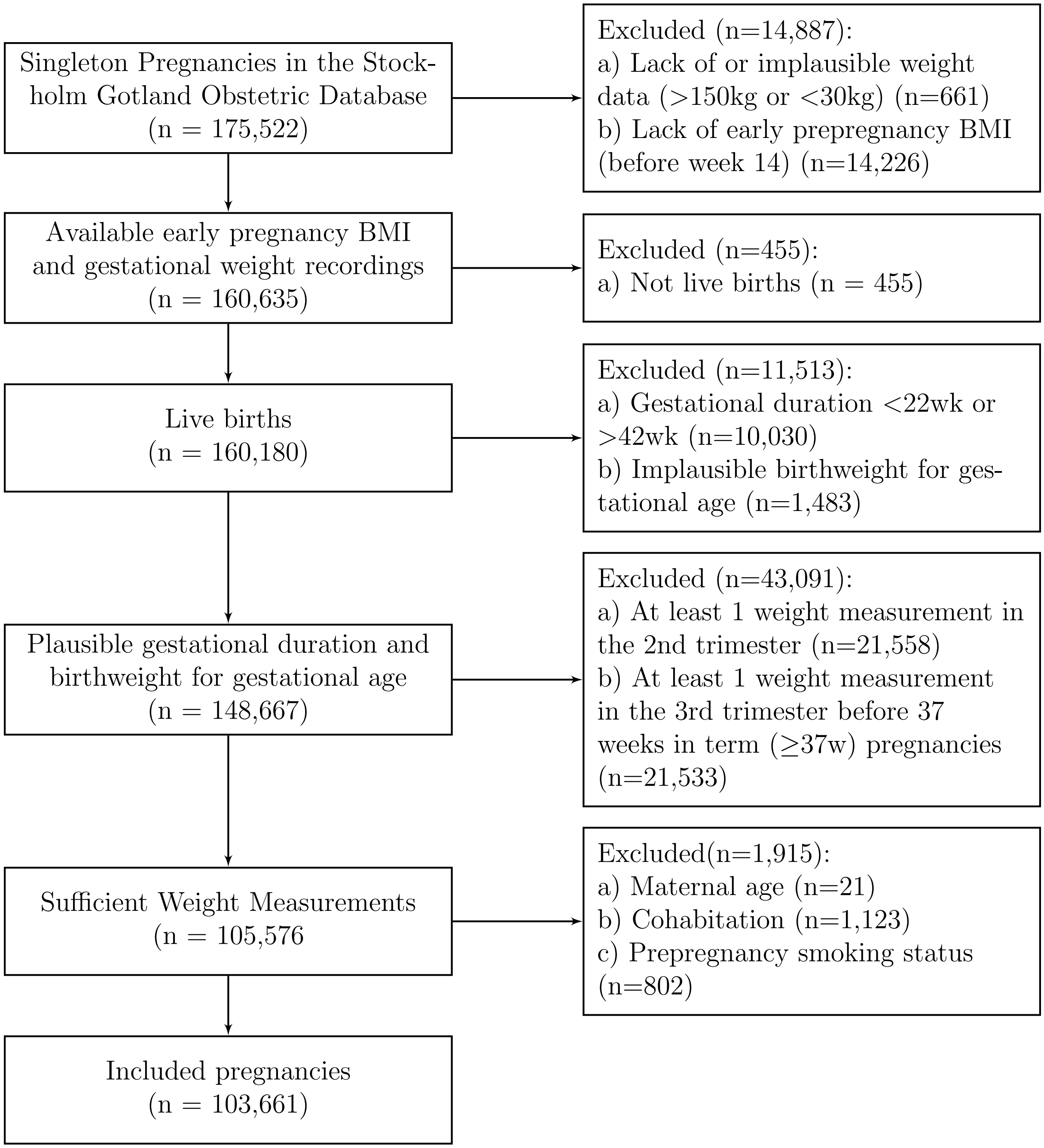

Singleton Pregnancies in the Stockholm Gotland Obstetric Database (n = 175,522)
Excluded (n=14,887):
a) Lack of or implausible weight data (>150kg or <30kg) (n=661)
b) Lack of early prepregnancy BMI (before week 14) (n=14,226)
Available early pregnancy BMI and gestational weight recordings (n = 160,635)
Excluded (n=455):
a) Not live births (n = 455)
Live births (n = 160,180)
Excluded (n=11,513):
a) Gestational duration <22wk or >42wk (n=10,030)
b) Implausible birthweight for gestational age (n=1,483)
Plausible gestational duration and birthweight for gestational age (n = 148,667)
Excluded (n=43,091):
a) At least 1 weight measurement in the 2nd trimester (n=21,558)
b) At least 1 weight measurement in the 3rd trimester before 37 weeks in term (≥37w) pregnancies (n=21,533)
Sufficient Weight Measurements (n = 105,576)
Excluded(n=1,915):
a) Maternal age (n=21)
b) Cohabitation (n=1,123)
c) Prepregnancy smoking status (n=802)
Included pregnancies (n = 103,661)

Figure 4.3: Sample selection from all women who delivered singleton infants in the Stockholm and Gotland regions of Sweden, 2008-2014

Table 4.6: Comparison of characteristics of included and excluded women who delivered live, singleton infants in the Stockholm and Gotland regions of Sweden, 2008-2014

	Excluded births $n = 45,006$	Included births $n = 103,661$	p
Preterm	1,073 (2.4)	5,419 (5.2)	< 0.001
Weight characteristics			
Early pregnancy weight, kg	65.7 ± 11.4	66.0 ± 12.3	< 0.001
Early pregnancy BMI, kg/m^2			< 0.001
Underweight	1,272 (2.8)	3,330 (3.2)	
Normal weight	31,298 (69.5)	68,745 (66.3)	
Overweight	9,337 (20.8)	22,173 (21.4)	
Obese	3,099 (6.9)	9,413 (9.1)	
GWG Z-score at delivery	0.0 ± 0.9	0.0 ± 1.0	0.006
GWG at last visit before 37w	10.9 ± 4.8	11.7 ± 4.4	< 0.001
Maternal characteristics			
Age, y	32.0 ± 4.9	31.4 ± 5.1	< 0.001
Height, cm	166.8 ± 6.5	166.1 ± 6.5	< 0.001
Primiparous	19,436 (43.2)	47,633 (46.0)	< 0.001
Living with baby's father	40,945 (94.4)	98,197 (94.7)	0.025
Prepregnancy diabetes	128 (0.3)	604 (0.6)	< 0.001
Prepregnancy hypertension	294 (0.7)	1,040 (1.0)	< 0.001
Prepregnancy snus use	1,397 (3.1)	2,777 (2.7)	< 0.001
Prepregnancy smoker	4,558 (10.4)	13,844 (13.4)	< 0.001
Gestational age at first visit, wk	9.7 ± 1.9	9.4 ± 2.0	< 0.001
Number of weight measurements through pregnancy			< 0.001
3	34,204 (76.0)	13,501 (13.0)	
4	6,767 (15.0)	19,225 (18.6)	
5	2,146 (4.8)	18,516 (17.9)	
6	918 (2.0)	15,670 (15.1)	
7	437 (1.0)	13,136 (12.7)	
8	264 (0.6)	10,387 (10.0)	
≥ 9	270 (0.6)	13,226 (12.8)	
Child characteristics			
Child sex			0.869
Female	22,934 (51.0)	52,775 (50.9)	
Male	22,072 (49.0)	50,886 (49.1)	
Child birthweight, kg	3.5 ± 0.5	3.5 ± 0.5	< 0.001

Table 4.7: Characteristics of 103,661 women who delivered live, singleton infants in the Stockholm and Gotland regions of Sweden, 2008-2014

	Term births (GA 37-42w) $n = 98,242$	Preterm births (GA $<$ 37w) $n = 5,419$
Weight characteristics		
Early pregnancy weight, kg	66.0 ± 12.2	66.1 ± 13.0
Early pregnancy BMI, kg/m^2		
Underweight (BMI <18)	3,126 (3.2)	204 (3.8)
Normal weight (BMI 18-24.9)	65,349 (66.5)	3,396 (62.7)
Overweight (BMI 25.0-29.9)	20,973 (21.35)	1,200 (22.1)
Obese (BMI $\geq$ 30)	8,794 (9.0)	619 (11.4)
GWG Z-score at delivery	0.0 ± 1.0	0.0 ± 1.1
GWG at last visit before 37w	11.8 ± 4.4	9.9 ± 4.9
Gestational age at first visit, wk	9.4 ± 2.0	9.4 ± 2.0
Number of weight measurements through pregnancy		
3	11,269 (11.5)	2,232 (41.2)
4	18,246 (18.6)	979 (18.1)
5	17,674 (18.0)	842 (15.5)
6	15,016 (15.3)	654 (12.1)
7	12,728 (13.0)	408 (7.5)
8	10,234 (10.4)	153 (2.8)
$\geq$9	13,075 (13.3)	151 (2.8)
Maternal characteristics		
Age, y	31.4 ± 5.1	31.4 ± 5.3
Height, cm	166.1 ± 6.5	165.1 ± 6.6
Primiparous	44,649 (45.5)	2,984 (55.1)
Living with baby's father	93,118 (94.8)	5,079 (93.7)
Prepregnancy diabetes	512 (0.5)	92 (1.7)
Prepregnancy hypertension	892 (0.9)	148 (2.7)
Prepregnancy snus use	2,635 (2.7)	142 (2.6)
Prepregnancy smoker	13,038 (13.3)	806 (14.9)
Child characteristics		
Child sex		
Female	49,838 (50.7)	2,937 (54.2)
Male	48,404 (49.3)	2,482 (45.8)
Child birthweight, g	3.5 ± 0.5	2.5 ± 0.6

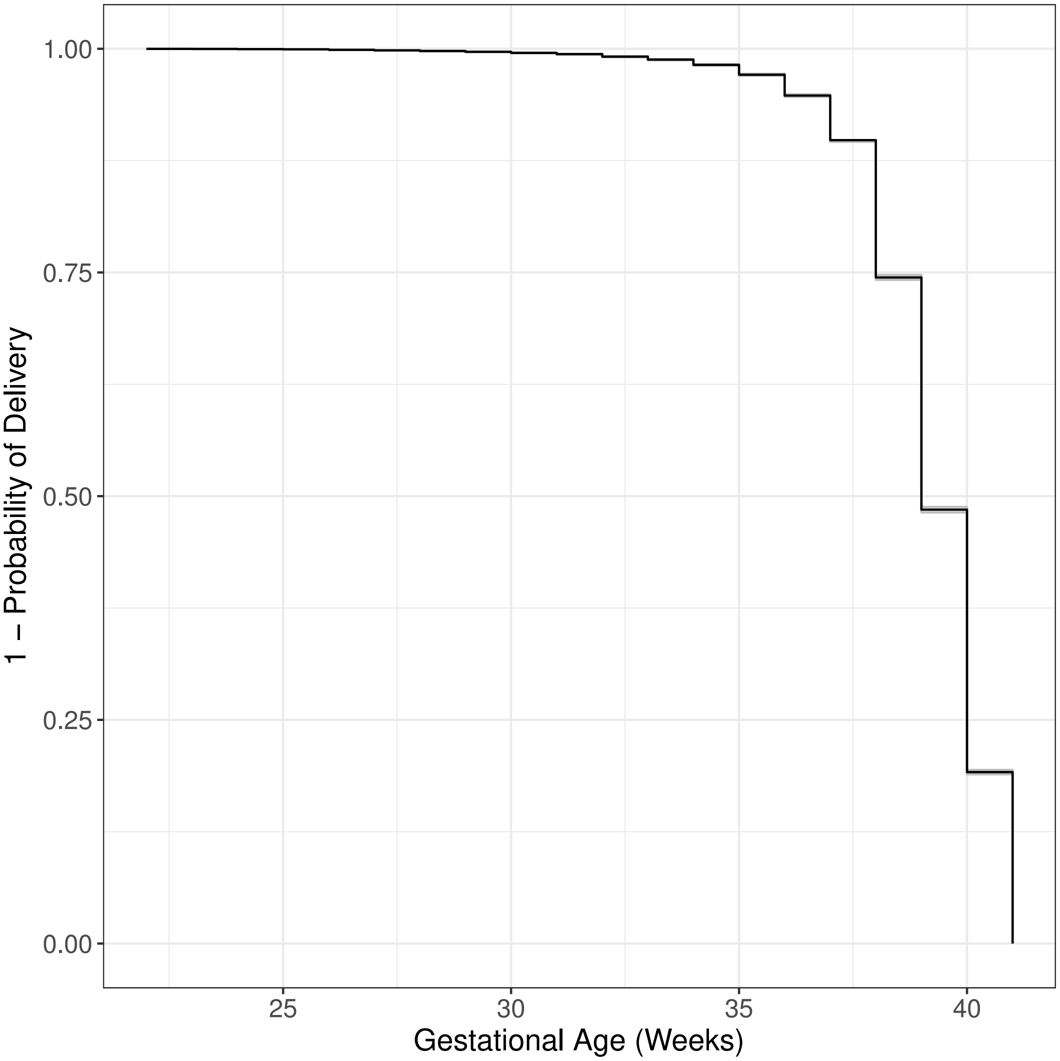


Figure 4.4: Kaplan-Meier curve describing probability of women remaining pregnant at each week of gestation in 103,661 women who delivered live, singleton infants in the Stockholm and Gotland regions of Sweden, 2008-2014. Note confidence intervals are present, but so small they become negligible.

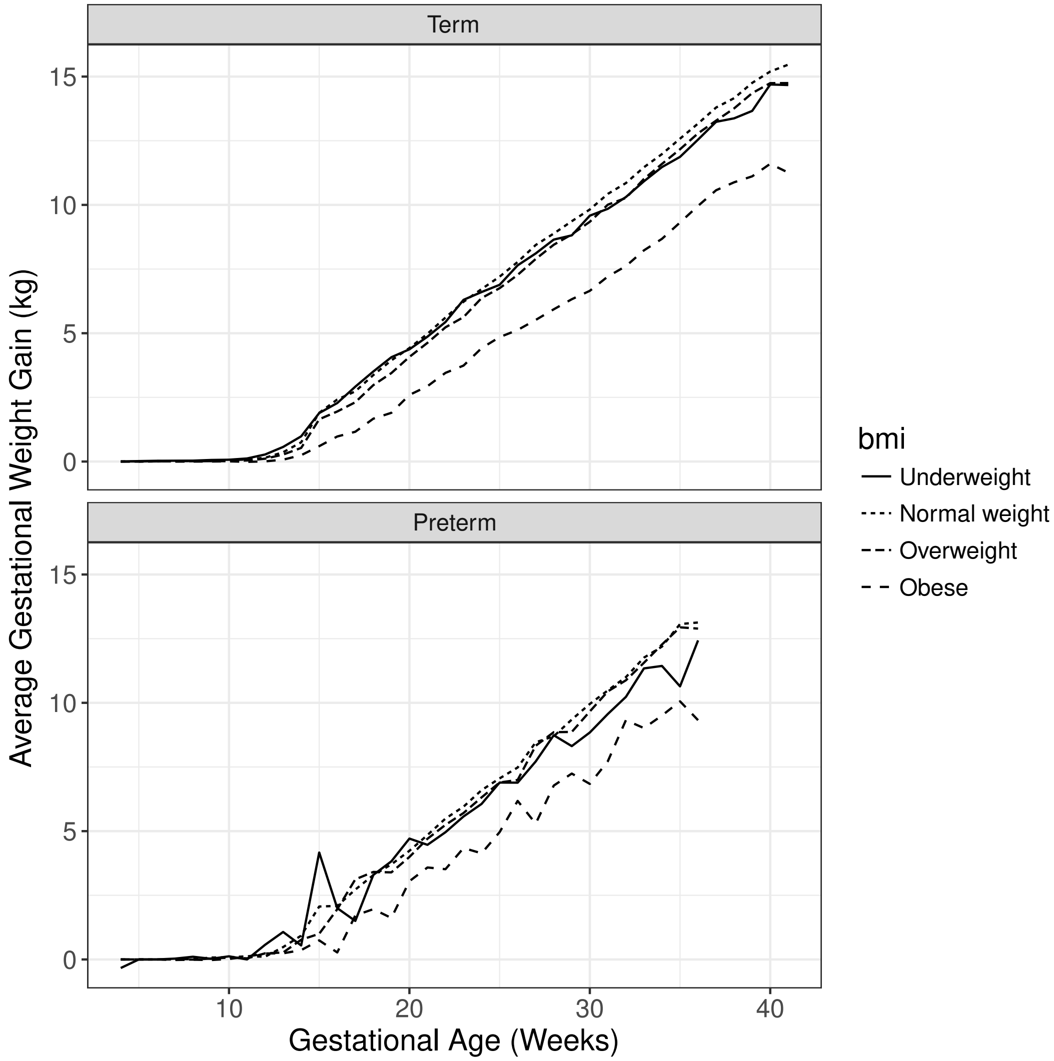


Figure 4.5: Average weight gain at each week of pregnancy in term and preterm pregnancies, stratified by early pregnancy BMI .

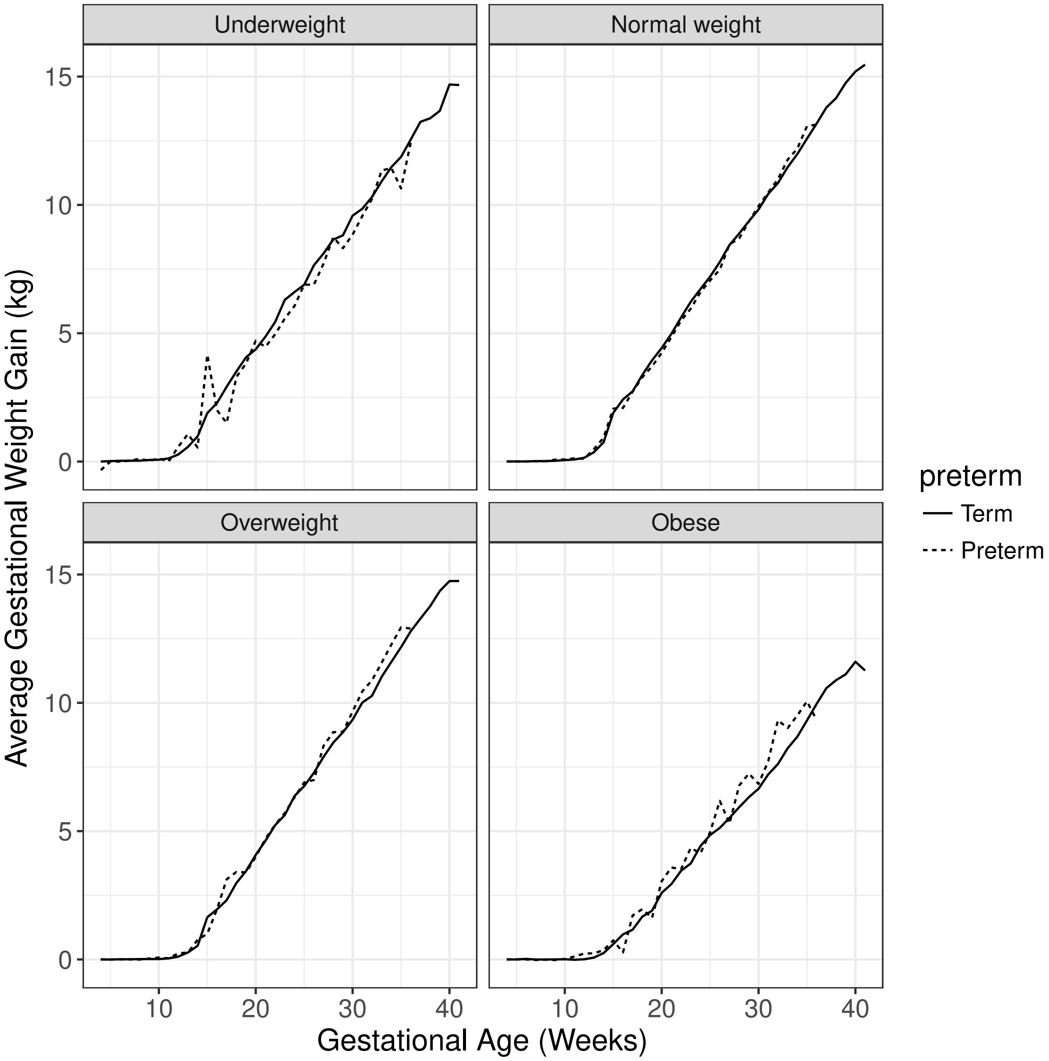


Figure 4.6: Average weight gain at each week of pregnancy in each early pregnancy BMI group, stratified by gestational duration

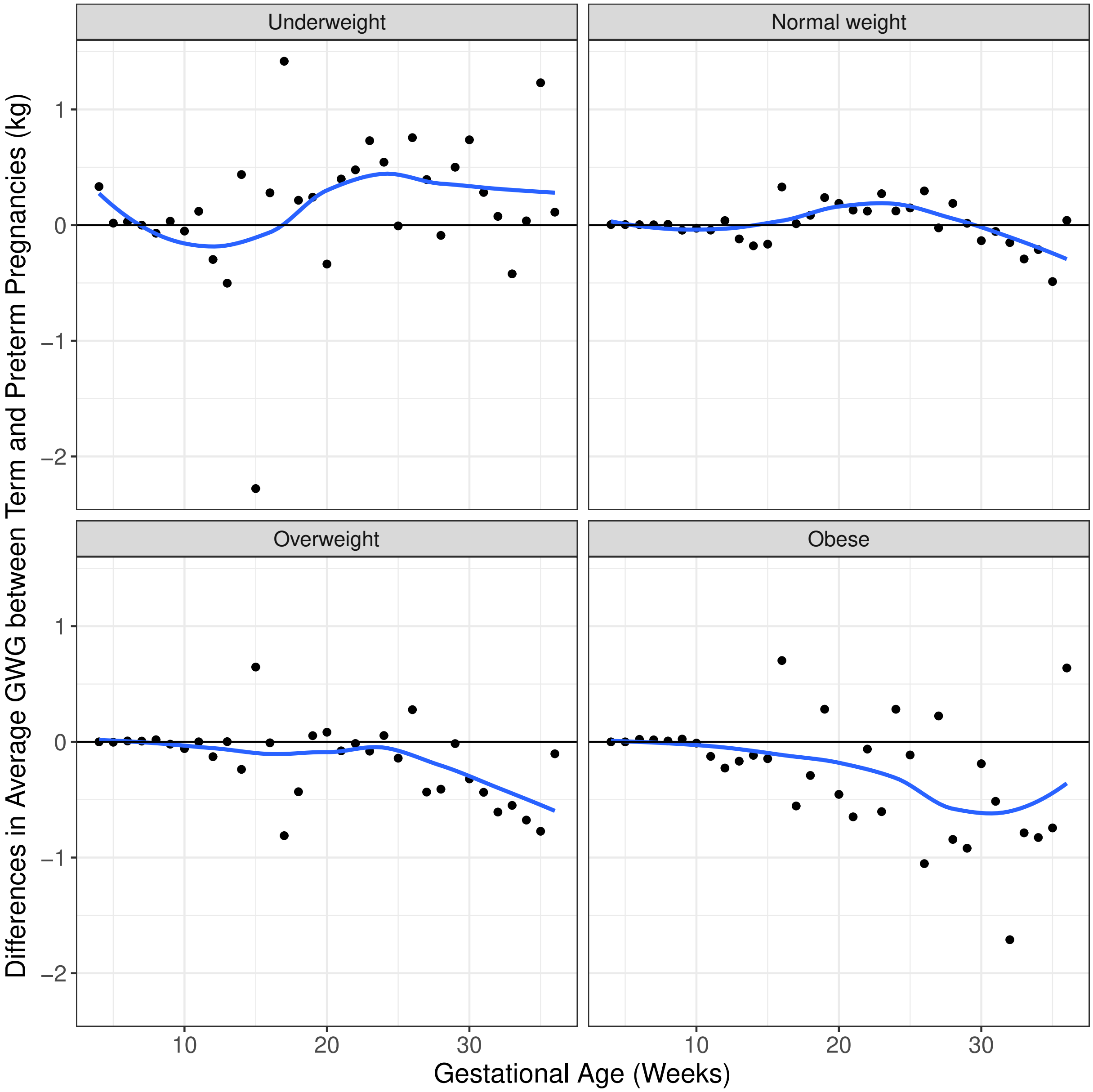


Figure 4.7: The points represent differences in average weight gain between term and preterm pregnancies at each weeks of gestation in each early pregnancy BMI group. The black line at 0kg represents no difference in average weight gain between term and preterm pregnancies. The blue line is a lowess curve showing the smoothed trend in these residuals. Values below 0kg indicate higher average weight gain in preterm pregnancies.

Bibliography

Chapter 1

[1] Miller Jr., R. G. *Survival Analysis*. Vol. 66. New York: John Wiley & Sons, 2011.

[2] Kleinbaum, D. G. and Klein, M. *Survival Analysis: A Self-Learning text*. Statistics for Biology and Health. DOI: 10.1007/978-1-4419-6646-9. New York, NY: Springer New York, 2012.

[3] Kaplan, E. L. and Meier, P. "Nonparametric Estimation from Incomplete Observations". *Journal of the American Statistical Association* 53.282 (June 1958), pp. 457–481. DOI: 10.1080/01621459.1958.10501452.

[4] Rich, J. T. et al. "A practical guide to understanding Kaplan-Meier curves". *Otolaryngology - Head and Neck Surgery* 143.3 (Sept. 2010), pp. 331–336. DOI: 10.1016/j.otohns.2010.05.007.

[5] Cox, D. "Regression Models and Life-Tables". *Journal of the royal statistical society. Series B (methodological)* 34.2 (1972), pp. 187–220.

[6] Andersen, P. and Gill, R. "Cox's regression model for counting processes: a large sample study". *The Annals of Statistics* 10.4 (1982), pp. 1100–1120.

[7] Wei, L.-J. "The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis". *Statistics in medicine* 11.14-15 (1992), pp. 1871–1879.

[8] Aalen, O. O. "A linear regression model for the analysis of life times". *Statistics in medicine* 8.8 (1989), pp. 907–925.

Chapter 2

[9] Wahed, M. A. et al. "A modified routine analysis of arsenic content in drinking-water in Bangladesh by hydride generation-atomic absorption spectrophotometry". *Journal of Health, Population and Nutrition* (2006), pp. 36–41.

[10] Dhand, N. K., Johnson, W. O., and Toribio, J.-A.L.M. L. “A Bayesian Approach to Estimate OJD Prevalence From Pooled Fecal Samples of Variable Pool Size”. *Journal of Agricultural, Biological and Environmental Statistics* 15.4 (Dec. 2010), pp. 452–473. DOI: `10.1007/s13253-010-0032-8`.

[11] Remlinger, K. S., Hughes-Oliver, J. M., Young, S. S., and Lam, R. L. “Statistical Design of Pools Using Optimal Coverage and Minimal Collision”. *Technometrics* 48.1 (Feb. 2006), pp. 133–143. DOI: `10.1198/004017005000000481`.

[12] Dorfman, R. “The detection of defective members of large populations”. *The Annals of Mathematical Statistics* 14.4 (1943), pp. 436–440.

[13] Hwang, F. K. “Group testing with a dilution effect”. *Biometrika* 63.3 (1976), pp. 671–680.

[14] Wein, L. M. and Zenios, S. A. “Pooled testing for HIV screening: capturing the dilution effect”. *Operations Research* 44.4 (1996), pp. 543–569.

[15] Liu, A., Liu, C., Zhang, Z., and Albert, P. S. “Optimality of group testing in the presence of misclassification”. *Biometrika* 99.1 (Mar. 2012), pp. 245–251. DOI: `10.1093/biomet/asr064`.

[16] Delaigle, A. and Hall, P. “Nonparametric regression with homogeneous group testing data”. *The Annals of Statistics* 40.1 (Feb. 2012), pp. 131–158. DOI: `10.1214/11-AOS952`.

[17] Tu, X. M., Litvak, E., and Pagano, M. “Screening tests: Can we get more by doing less?” *Statistics in Medicine* 13.19-20 (Oct. 1994), pp. 1905–1919. DOI: `10.1002/sim.4780131904`.

[18] Jewell, N. P. and Laan, M. J. van der. “Current status data: Review, recent developments and open problems”. *Handbook in statistics: advances in survival analysis.* Ed. by Balakrishnan, N. and Rao, C. R. Vol. 23. Elsevier, 2004, pp. 625–42.

[19] Jewell, N. P. and Emerson, R. “Current status data: an illustration with data on avalanche victims”. *Handbok of Survival Analysis.* Ed. by Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. Boca Raton: Chapman and Hall/CRC, 2013, pp. 391–412.

[20] Ayer, M. H., Brunk, D., Ewing, G. M., Reid, W. T., and Silverman, E. “An empirical distribution function for sampling with incomplete information”. *The Annals of Mathematical Statistics* 26.4 (1955), pp. 641–647.

[21] Groeneboom, P. and Wellner, J. A. *Nonparametric maximum likelihood estimators for interval censoring and deconvolution.* Boston: Birkhauser-Boston, 1992.

[22] Banerjee, M. “Current status data in the 21st century: Some interesting developments.” *Interval-censored time-to-event data: methods and applications.* Ed. by Chen, D.-G., Sun, J., and Peace, K. E. Boca Raton: Chapman and Hall/CRC, 2012, pp. 45–90.

[23] Banerjee, M. and Wellner, J. A. "Likelihood ratio tests for monotone functions". *Annals of Statistics* (2001), pp. 1699–1731.

[24] Banerjee, M. and Wellner, J. A. "Confidence intervals for current status data". *Scandinavian Journal of Statistics* (2005), pp. 405–424.

[25] Yu, Q., Schick, A., Li, L., and Wong, G. Y. C. "Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times". *Canadian Journal of Statistics* 26.4 (Dec. 1998), pp. 619–627. DOI: `10.2307/3315721`.

[26] Maathuis, M. H. and Hudgens, M. G. "Nonparametric inference for competing risks current status data with continuous, discrete or grouped observation times". *Biometrika* 98.2 (June 2011), pp. 325–340. DOI: `10.1093/biomet/asq083`.

[27] Tang, R., Banerjee, M., and Kosorok, M. R. "Likelihood based inference for current status data on a grid: A boundary phenomenon and an adaptive inference procedure". *The Annals of Statistics* 40.1 (Feb. 2012), pp. 45–72. DOI: `10.1214/11-AOS942`.

[28] Rosas, V. G. Sal y and Hughes, J. P. "Nonparametric and semiparametric analysis of current status data subject to outcome misclassification". *Statistical Communications in Infectious Diseases* (2010).

[29] Barlow, R. E., Bartholomew, D. J., Bremner, J., and Brunk, H. D. *Statistical inference under order restrictions: The theory and application of isotonic regression.* New York: Wiley, 1972.

[30] McKeown, K. and Jewell, N. P. "Misclassification of current status data". *Lifetime Data Analysis* 16.2 (Apr. 2010), pp. 215–230. DOI: `10.1007/s10985-010-9154-0`.

[31] McMahan, C. S., Tebbs, J. M., and Bilder, C. R. "Regression models for group testing data with pool dilution effects". *Biostatistics* 14.2 (Apr. 2013), pp. 284–298. DOI: `10.1093/biostatistics/kxs045`.

[32] Vansteelandt, S., Goetghebeur, E., and Verstraeten, T. "Regression models for disease prevalence with diagnostic tests on pools of serum samples". *Biometrics* 56.4 (2000), pp. 1126–1133.

[33] Xie, M. "Regression analysis of group testing samples". *Statistics in medicine* 20.13 (2001), pp. 1957–1969.

[34] Chen, P., Tebbs, J. M., and Bilder, C. R. "Group Testing Regression Models with Fixed and Random Effects". *Biometrics* 65.4 (Dec. 2009), pp. 1270–1278. DOI: `10.1111/j.1541-0420.2008.01183.x`.

[35] Delaigle, A. and Meister, A. "Nonparametric Regression Analysis for Group Testing Data". *Journal of the American Statistical Association* 106.494 (June 2011), pp. 640–650. DOI: `10.1198/jasa.2011.tm10520`.

Chapter 3

[36] Robins, J. M. “Association, causation, and marginal structural models”. *Synthese* 121.1 (1999), pp. 151–179.

[37] Hernán, M. A., Brumback, B., and Robins, J. M. “Marginal structural models and causal inference in epidemiology”. *Epidemiology* 11.5 (2000), pp. 550–560.

[38] Crowson, C. S., Schenck, L. A., Green, A. B., Atkinson, E. J., and Therneau, T. M. *The basics of propensity scoring and marginal structural models.* Technical Report #84. Rochester, MN: Mayo Clinic, Aug. 2013.

[39] Hernán, M. A. and Robins, J. M. *Causal Inference.* Boca Raton: Chapman and Hall/CRC, 2017.

[40] Helmick, C. G. et al. “Estimates of the prevalence of arthritis and other rheumatic conditions in the United States: Part I”. *Arthritis & Rheumatism* 58.1 (Jan. 2008), pp. 15–25. DOI: 10.1002/art.23177.

[41] Lawrence, R. C. et al. “Estimates of the prevalence of arthritis and other rheumatic conditions in the United States: Part II”. *Arthritis & Rheumatism* 58.1 (Jan. 2008), pp. 26–35. DOI: 10.1002/art.23176.

[42] Marcia Vital, O. o.C.a.P. L. *Handout on Health: Rheumatoid Arthritis.* 2016. URL: https://www.niams.nih.gov/health_info/rheumatic_disease/ (visited on 05/02/2017).

[43] Solomon, D. H. et al. “Patterns of cardiovascular risk in rheumatoid arthritis”. *Annals of the Rheumatic Diseases* 65.12 (May 2006), pp. 1608–1612. DOI: 10.1136/ard.2005.050377.

[44] Gonzalez, A. et al. “The widening mortality gap between rheumatoid arthritis patients and the general population”. *Arthritis & Rheumatism* 56.11 (Nov. 2007), pp. 3583–3587. DOI: 10.1002/art.22979.

[45] Myasoedova, E. et al. “Lipid paradox in rheumatoid arthritis: the impact of serum lipid measures and systemic inflammation on the risk of cardiovascular disease”. *Annals of the Rheumatic Diseases* 70.3 (Mar. 2011), pp. 482–487. DOI: 10.1136/ard.2010.135871.

[46] Suissa, S., Bernatsky, S., and Hudson, M. “Antirheumatic drug use and the risk of acute myocardial infarction”. *Arthritis & Rheumatism* 55.4 (Aug. 2006), pp. 531–536. DOI: 10.1002/art.22094.

[47] Dixon, W. G. et al. “Reduction in the incidence of myocardial infarction in patients with rheumatoid arthritis who respond to anti?tumor necrosis factor ? therapy: Results from the British Society for Rheumatology Biologics Register”. *Arthritis & Rheumatism* 56.9 (Sept. 2007), pp. 2905–2912. DOI: 10.1002/art.22809.

[48] Danesh, J. et al. "Long-term interleukin-6 levels and subsequent risk of coronary heart disease: two new prospective studies and a systematic review". *PLoS Med* 5.4 (2008), e78.

[49] Kishimoto, T. "IL-6: from its discovery to clinical applications". *International Immunology* 22.5 (May 2010), pp. 347–352. DOI: `10.1093/intimm/dxq030`.

[50] Boekholdt, S. M. and Stroes, E. S. "The interleukin-6 pathway and atherosclerosis". *The lancet* 379 (Mar. 2012), pp. 1176–1178.

[51] Collaboration, I. G.C.E.R. F. and others. "Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies". *The Lancet* 379.9822 (2012), pp. 1205–1213.

[52] Consortium, I.-.R.M.R.A.I. M. and others. "The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis". *The Lancet* 379.9822 (2012), pp. 1214–1224.

[53] Lee, J. K. et al. "Association between Serum Interleukin-6 Concentrations and Mortality in Older Adults: The Rancho Bernardo Study". *PLoS ONE* 7.4 (Apr. 2012). Ed. by Herder, C., e34218. DOI: `10.1371/journal.pone.0034218`.

[54] *Actemra (tocilizumab) injection for intravenous infusion [prescribing information]*. South San Francisco: Genentech, Inc., 2012.

[55] Hurt-Camejo, E. et al. "Elevated Levels of small, low-density lipoprotein with high affinity for arterial matrix components in patients with rheumatoid arthritis". *Arthritis & Rheumatism* 44.12 (2001), pp. 2761–2767.

[56] Choy, E and Sattar, N. "Interpreting lipid levels in the context of high-grade inflammatory states with a focus on rheumatoid arthritis: a challenge to conventional cardiovascular risk actions". *Annals of the Rheumatic Diseases* 68.4 (Apr. 2009), pp. 460–469. DOI: `10.1136/ard.2008.101964`.

[57] McInnes, I. B. et al. "Effect of interleukin-6 receptor blockade on surrogates of vascular risk in rheumatoid arthritis: MEASURE, a randomised, placebo-controlled study". *Annals of the Rheumatic Diseases* 74.4 (Apr. 2015), pp. 694–702. DOI: `10.1136/annrheumdis-2013-204345`.

[58] Rao, V. U. et al. "An Evaluation of Risk Factors for Major Adverse Cardiovascular Events During Tocilizumab Therapy: Risk Factors for Cardiovascular Events With Tocilizumab". *Arthritis & Rheumatology* 67.2 (Feb. 2015), pp. 372–380. DOI: `10.1002/art.38920`.

[59] Pearl, J. *Causality*. Cambridge university press, 1995.

[60] Horvitz, D. G. and Thompson, D. J. "A Generalization of Sampling Without Replacement from a Finite Universe". *Journal of the American Statistical Association* 47.260 (1952), pp. 663–685. DOI: `10.1080/01621459.1952.10483446`.

[61] Cole, S. R. and Hernán, M. A. “Constructing Inverse Probability Weights for Marginal Structural Models”. *American Journal of Epidemiology* 168.6 (July 2008), pp. 656–664. DOI: 10.1093/aje/kwn164.

[63] Buuren, S. van. “Multiple imputation of discrete and continuous data by fully conditional specification”. *Statistical Methods in Medical Research* 16.3 (June 2007), pp. 219–242. DOI: 10.1177/0962280206074463.

[64] Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons, 1987.

[65] Díaz Muñoz, I. and Laan, M. J. van der. “Super Learner Based Conditional Density Estimation with Application to Marginal Structural Models”. *The International Journal of Biostatistics* 7.1 (Jan. 2011), pp. 1–20. DOI: 10.2202/1557-4679.1356.

Chapter 4

[66] Buhimschi, C. S. and Norman, J. E. “Pathogenesis of Spontaneous Preterm Birth”. *Maternal-Fetal Medicine: Principles and Practice*. Ed. by Creasy, R. K. et al. 7th ed. Philadelphia: Elsevier, 2014, pp. 599–623.

[67] Simhan, H. N., Berghella, V., and Iams, J. D. “Preterm Labor and Birth”. *Maternal-Fetal Medicine: Principles and Practice*. Ed. by Creasy, R. K. et al. 7th ed. Philadelphia: Elsevier, 2014, pp. 624–653.

[68] Saigal, S. and Doyle, L. W. “An overview of mortality and sequelae of preterm birth from infancy to adulthood”. *The lancet*. Preterm Birth 371 (2008), pp. 261–69.

[69] Goldenberg, R. L., Culhane, J. F., Iams, J. D., and Romero, R. “Epidemiology and causes of preterm birth”. *The lancet*. Preterm Birth 371.9606 (2008), pp. 75–84.

[70] Goldenberg, R. L. and Tamura, T. “Prepregnancy weight and pregnancy outcome”. *Journal of the American Medical Association* 275.14 (1996), pp. 1127–1128.

[71] Goldenberg, R. L., Hauth, J. C., and Andrews, W. W. “Intrauterine infection and preterm delivery”. *New England journal of medicine* 342.20 (2000), pp. 1500–1507.

[72] Relman, D. A., Loutit, J. S., Schmidt, T. M., Falkow, S., and Tompkins, L. S. “The agent of bacillary angiomatosis”. *New England Journal of Medicine* 323.23 (1990), pp. 1573–1580.

[73] Smith, G. C., Pell, J. P., and Dobbie, R. “Interpregnancy interval and risk of preterm birth and neonatal death: retrospective cohort study”. *Bmj* 327.7410 (2003), p. 313.

[74] Conde-Agudelo, A., Rosas-Bermudez, A., and Kafury-Goeta, A. C. “Birth spacing and risk of adverse perinatal outcomes: a meta-analysis”. *Journal of the American Medical Association* 295.15 (2006), pp. 1809–1823.

[75] Brett, K. M., Strogatz, D. S., and Savitz, D. A. "Employment, job strain, and preterm delivery among women in North Carolina." *American Journal of Public Health* 87.2 (1997), pp. 199–204.

[76] Bermudez, E. A., Rifai, N., Buring, J. E., Manson, J. E., and Ridker, P. M. "Relation between markers of systemic vascular inflammation and smoking in women". *American Journal of Cardiology* 89.9 (2002), pp. 1117–1119.

[77] Tracy, R. P. et al. "Lifetime smoking exposure affets the association of C-reactive protein with cardiovascular disease risk factors and subclinical disease in healthy elderly subjects". *Arteriosclerosis* 17.10 (1997), pp. 2167–2176.

[78] Moutquin, J. "Classification and heterogeneity of preterm birth". *BJOG: An International Journal of Obstetrics and Gynaecology* 110 (Apr. 2003), pp. 30–33. DOI: 10.1016/S1470-0328(03)00021-1.

[79] Cnattingius, S. et al. "Maternal obesity and risk of preterm delivery". *Journal of the American Medical Association* 309.22 (2013), pp. 2362–2370.

[80] Hutcheon, J. A. et al. "Pregnancy weight gain charts for obese and overweight women: Pregnancy Weight Gain Charts". *Obesity* 23.3 (Mar. 2015), pp. 532–535. DOI: 10.1002/oby.21011.

[81] Hutcheon, J. A. et al. "A weight-gain-for-gestational-age z score chart for the assessment of maternal weight gain in pregnancy". *The American Journal of Clinical Nutrition* 97.5 (May 2013), pp. 1062–1067. DOI: 10.3945/ajcn.112.051706.

[83] Martin, J. A., Hamilton, B. E., Osterman, M. J., Driscoll, A. K., and Matthews, T. *Births: Final Data for 2015*. Tech. rep. Volume 66, Number 1. U.S. Department of Health, Human Services, Centers for Disease Control, and Prevention: National Center for Health Statistics, National Vital Statistics System, Division of Vital Statistics, Jan. 2017.

[84] Kramer, M. R. and Hogue, C. R. "What Causes Racial Disparities in Very Preterm Birth? A Biosocial Perspective". *Epidemiologic Reviews* 31.1 (Nov. 2009), pp. 84–98. DOI: 10.1093/ajerev/mxp003.

[85] Lu, M. C. and Halfon, N. "Racial and ethnic disparities in birth outcomes: a life-course perspective". *Maternal and Child Health Journal* 7.1 (Mar. 2003), pp. 13–30.

[86] Braveman, P. A. et al. "The role of socioeconomic factors in black-white disparities in preterm birth". *American journal of public health* 105.4 (2015), pp. 694–702.

[87] Flynn, A. C. et al. "Dietary interventions in overweight and obese pregnant women: a systematic review of the content, delivery, and outcomes of randomized controlled trials". *Nutrition Reviews* 74.5 (May 2016), pp. 312–328. DOI: 10.1093/nutrit/nuw005.

[88] IOM (Institue of Medicine) and NRC (National Research Council). *Weight Gain during Pregnancy: Reexamining the Guidelines*. Washington, D.C.: National Academies Press, 2009.

[89] Viswanathan, M. et al. *Outcomes of Maternal Weight Gain*. Evidence Report/ Technology Assessment 168. Research Triangle Park, North Carolina: Agency for Healthcare Research, Quality, U.S. Department of Health, and Human Services; prepared by RTI International, May 2008.

[90] Oken, E., Kleinman, K. P., Belfort, M. B., Hammitt, J. K., and Gillman, M. W. "Associations of Gestational Weight Gain With Short- and Longer-term Maternal and Child Health Outcomes". *American Journal of Epidemiology* 170.2 (July 2009), pp. 173–180. DOI: `10.1093/aje/kwp101`.

[91] Savitz, D. A., Stein, C. R., Siega-Riz, A. M., and Herring, A. H. "Gestational Weight Gain and Birth Outcome in Relation to Prepregnancy Body Mass Index and Ethnicity". *Annals of Epidemiology* 21.2 (Feb. 2011), pp. 78–85. DOI: `10.1016/j.annepidem.2010.06.009`.

[92] Bodnar, L. M., Siega-Riz, A. M., Simhan, H. N., Himes, K. P., and Abrams, B. "Severe obesity, gestational weight gain, and adverse birth outcomes". *American Journal of Clinical Nutrition* 91.6 (June 2010), pp. 1642–1648. DOI: `10.3945/ajcn.2009.29008`.

[93] Headen, I. E., Davis, E. M., Mujahid, M. S., and Abrams, B. "Racial-Ethnic Differences in Pregnancy-Related Weight". *Advances in Nutrition: An International Review Journal* 3.1 (Jan. 2012), pp. 83–94. DOI: `10.3945/an.111.000984`.

[94] Headen, I., Mujahid, M. S., Cohen, A. K., Rehkopf, D. H., and Abrams, B. "Racial/ Ethnic Disparities in Inadequate Gestational Weight Gain Differ by Pre-pregnancy Weight". *Maternal and Child Health Journal* 19.8 (Aug. 2015), pp. 1672–1686. DOI: `10.1007/s10995-015-1682-5`.

[95] Stotland, N. E., Caughey, A. B., Lahiff, M., and Abrams, B. "Weight gain and spontaneous preterm birth: the role of race or ethnicity and previous preterm birth". *Obstetrics & Gynecology* 108.6 (2006), pp. 1448–1455.

[96] Hickey, C. A., Cliver, S., McNeal, S. F., Hoffman, H. J., and Goldenberg, R. L. "Prenatal weight gain patterns and spontaneous rpeterm birth among nonobese black and white women". *Obstetrics & Gynecology* 85.6 (June 1995), pp. 909–914.

[97] Hutcheon, J. A. et al. "The bias in current measures of gestational weight gain: Bias from gestational weight gain measures". *Paediatric and Perinatal Epidemiology* 26.2 (Mar. 2012), pp. 109–116. DOI: `10.1111/j.1365-3016.2011.01254.x`.

[98] Alexander, G. R., Himes, J. H., Kaufman, R. B., Mor, J., and Kogan, M. "A United States national reference for fetal growth". *Obstetrics & Gynecology* 87.2 (1996), pp. 163–168.

[99] Hernán, M. A., Hernández-Díaz, S., Werler, M. M., and Mitchell, A. A. "Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology". *American journal of epidemiology* 155.2 (2002), pp. 176–184.

[100] Seaman, S. R. and White, I. R. "Review of inverse probability weighting for dealing with missing data". *Statistical Methods in Medical Research* 22.3 (June 2013), pp. 278–295. DOI: `10.1177/0962280210395740`.

[101] Newson, R. *REGPAR: Stata module to compute population attributable risks from binary regression models.* Nov. 2011. URL: `https://ideas.repec.org/c/boc/bocode/s457361.html`.

[102] Newson, R. *PUNAF: Stata module to compute population attributable fractions for cohort studies.* Oct. 2010. URL: `https://ideas.repec.org/c/boc/bocode/s457193.html`.

[103] Wasserstein, R. L. and Lazar, N. A. "The ASA's Statement on p -Values: Context, Process, and Purpose". *The American Statistician* 70.2 (Apr. 2016), pp. 129–133. DOI: `10.1080/00031305.2016.1154108`.

[104] Bodnar, L. M., Hutcheon, J. A., Parisi, S. M., Pugh, S. J., and Abrams, B. "Comparison of Gestational Weight Gain z-Scores and Traditional Weight Gain Measures in Relation to Perinatal Outcomes". *Paediatric and Perinatal Epidemiology* 29.1 (Jan. 2015), pp. 11–21. DOI: `10.1111/ppe.12168`.

[105] Bodnar, L. M. et al. "Low Gestational Weight Gain and Risk of Adverse Perinatal Outcomes in Obese and Severely Obese Women:" *Epidemiology* 27.6 (Nov. 2016), pp. 894–902. DOI: `10.1097/EDE.0000000000000535`.

[106] Dietz, P. et al. "Validation of selected items on the 2003 US standard certificate of live birth: New York City and Vermont". *Public Health Reports* 130.1 (2015), pp. 60–70.

[107] Headen, I., Cohen, A., Mujahid, M., and Abrams, B. "The accuracy of self-reported pregnancy-related weight: a systematic review: Accuracy of self-reported pregnancy-related weight". *Obesity Reviews* 18.3 (Mar. 2017), pp. 350–369. DOI: `10.1111/obr.12486`.

[108] Sharma, A. J. et al. "Rate of Second and Third Trimester Weight Gain and Preterm Delivery Among Underweight and Normal Weight Women". *Maternal and Child Health Journal* 20.10 (Oct. 2016), pp. 2030–2036. DOI: `10.1007/s10995-016-2032-y`.

[109] Ismail, L. C. et al. "Gestational weight gain standards based on women enrolled in the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project: a prospective longitudinal cohort study". *BMJ* 352 (Feb. 2016), p. i555. DOI: `10.1136/bmj.i555`.

[110] Hinkle, S. N., Mitchell, E. M., Grantz, K. L., Ye, A., and Schisterman, E. F. "Maternal Weight Gain During Pregnancy: Comparing Methods to Address Bias Due to Length of Gestation in Epidemiological Studies". *Paediatric and Perinatal Epidemiology* 30.3 (May 2016), pp. 294–304. DOI: `10.1111/ppe.12284`.

[111] Muktabhant, B., Lawrie, T. A., Lumbiganon, P., and Laopaiboon, M. “Diet or exercise, or both, for preventing excessive weight gain in pregnancy”. *Cochrane Database of Systematic Reviews*. Ed. by The Cochrane Collaboration. Chichester, UK: John Wiley & Sons, Ltd, June 2015.

[112] Herring, S. J. et al. “Preventing excessive gestational weight gain among African American women: A randomized clinical trial: Weight Control in Pregnancy”. *Obesity* 24.1 (Jan. 2016), pp. 30–36. DOI: `10.1002/oby.21240`.

[113] Shulman, R. and Kottke, M. “Impact of maternal knowledge of recommended weight gain in pregnancy on gestational weight gain”. *American Journal of Obstetrics and Gynecology* 214.6 (June 2016), 754.e1–754.e7. DOI: `10.1016/j.ajog.2016.03.021`.

[114] Phelan, S. “Pregnancy: a teachable moment for weight control and obesity prevention”. *American Journal of Obstetrics and Gynecology* 202.2 (Feb. 2010), 135.e1–135.e8. DOI: `10.1016/j.ajog.2009.06.008`.

[115] Reid, A. E. et al. “Discrimination and excessive weight gain during pregnancy among Black and Latina young women”. *Social Science & Medicine* 156 (May 2016), pp. 134–141. DOI: `10.1016/j.socscimed.2016.03.012`.

[116] Papas, M. A. et al. “The Built Environment and Obesity”. *Epidemiologic Reviews* 29.1 (May 2007), pp. 129–143. DOI: `10.1093/epirev/mxm009`.

[117] Johansson, K., Hutcheon, J. A., Stephansson, O., and Cnattingius, S. “Pregnancy weight gain by gestational age and BMI in Sweden: a population-based cohort study”. *The American Journal of Clinical Nutrition* (Mar. 2016), ajcn110197. DOI: `10.3945/ajcn.115.110197`.

[118] Marˇ sál, K. et al. “Intrauterine growth curves based on ultrasonically estimated foetal weights”. *Acta paediatrica* 85.7 (1996), pp. 843–848.

[119] Sharma, A. J. et al. “Associations of Gestational Weight Gain with Preterm Birth among Underweight and Normal Weight Women”. *Maternal and Child Health Journal* 19.9 (Sept. 2015), pp. 2066–2073. DOI: `10.1007/s10995-015-1719-9`.