ANALYTICS AND MACHINE LEARNING FOR HEALTHCARE DATA

A Dissertation
Presented to
The Academic Faculty

by

Toyya A. Pujol

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology
December 2020

ANALYTICS AND MACHINE LEARNING FOR HEALTHCARE DATA

Approved by:

Dr. Nicoleta Serban, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Sherri Rose
School of Medicine
Stanford University

Dr. Julie Swann
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Gregory Gibson
School of Biology
Georgia Institute of Technology

Dr. Branislav Vidakovic
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: [August 06, 2020]

To do well enough to be able to do the next thing…

ACKNOWLEDGEMENTS

Six years is a long to commit to anything in one's life. When I look back at my doctoral journey, I want to acknowledge the many people who have helped me along the way. I begin with my family. My mother has been on this journey with me from day one, listening to my daily triumphs and failures, sharing in my anxieties and joys. Her support was invaluable. My godmother, Barbara, I thank her for the endless encouragement and unwavering faith in my success. Bud, my fiancé, though only in the last 2 years of my degree, you've seen some of the hardest. Thank you for your patience, flexibility, and true investment in my happiness and satisfaction in life. Diane for always being a 2nd mother and knowing that I could always count on you. Alex, I thank you for all the pep talks and calming me on all my bouts of anxiety

Next, I want to thank my friends. Swetha, my best friend, I thank you for helping me manifest the life I want every day. Jerry, I want for being my longest friend and always knowing when I need a good laugh. Sabrina, thank you for being my number one cheerleader and always making me feel I had someone proud in my corner. I thank all my friends in the program: Jovan, Ethan, Ben, Erin, Daniel, Josh, and Simon. I couldn't have made it through without your advice, guidance and support. For my friends outside the program including Evan, Will, Tammy, and Nikki, I appreciate you all for helping make GT feel like home to me.

Finally, I want to acknowledge each person in my committee. Nicoleta, the endless hours you have invested in helping me to become a better researcher and getting me to this finish line. Sherri, thank you for taking a chance on me and investing in my like one of

your own students. You have helped me grow my confidence and discern the type of researcher I want to be. Julie, thank you for always being available for advice. Brani, I appreciate your support and the chats that lifted my spirits. Greg, thank you for supporting my training and helping integrate into the biology department. In addition, I want to thank Alan for when given the choice, always choosing the option that was best for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ATT	Average Treatment Effect Among the Treated
CDC	Centers for Disease Control and Prevention
CMS	Centers for Medicare & Medicaid Services
CPT	Current Procedural Terminology
CRG	Critical Risk Grouping
DOB	Date of Birth
DID	Difference-in-Differences
EBP	Episode-Based Payment
ED	Emergency Department
FPM	Family Planning Management
GLM	Generalized Linear Model
HEM	Highest Efficacy Method
ICD	International Classification of Disease
IP	Inpatient Table
IPW	Inverse Propensity Score Weighting
IUD	Intrauterine Device(s)
LBW	Low Birth Weight
MAX	Medicaid Analytic Extract
MEC	US Medical Eligibility Criteria for Contraceptive Use
MLE	Maximum Likelihood Estimate
NPI	National Provider Index
OT	Other Therapy Table

PS	Personal Summary Table
RUCC	Rural-Urban Continuum Codes
SE	Standard Error
SD	Standard Deviation
SUTVA	Stable Unit Treatment Value Assumption
TMLE	Targeted Maximum Likelihood Estimation

SUMMARY

The volume of health care data is expected to grow faster than any other industry. This creates a demand for the development of rigorous analytics and machine learning methods for applications to large health data sets. These data sets, which contain personally identifiable information, come with privacy protections that place limitations on data visibility and its release. In addition, patient data often contains complex relationships such as non-linear relationships and heterogeneity. These characteristics can cause unique complications for analysis of health care data and restrict the use of out-of-the-box solutions. Notably, healthcare research has incredibly high stakes, it can be the difference between life and death and have a major impact on an individual's quality of life and medical treatment. For these reasons, the development of rigorous solutions are that much more critical. This thesis focuses on the application of analytics and machine learning to solve applied research problems based on healthcare data.

Chapter 1 is an introduction to the thesis. It presents the research objectives and contributions for each research study. This chapter also discusses the value of the methods used in each study and the benefits of using administrative claims data.

In chapter 2, we discuss the study that determined the impact of a new CDC recommendation on contraception medical claims for women with high-risk chronic health conditions. The study included Medicaid-enrolled women two years prior to the release of the recommendations and two years following the release for 14 states. We focused on two outcome measures: (1) overall contraception use and (2) the use of CDC recommended contraception (i.e. those of the highest efficacy). We evaluated each outcome for the entire

study population and by health condition. The ratio of the after-recommendation rate over the before-recommendation rate was used to determine statistical significance in the uptake of the new recommendations. The results found that there had been an increase in the overall use of contraception methods among women with these health conditions and for each condition individually. However, the results also showed that the use of the highest-efficacy methods increased overall but not for every condition. The chapter concludes with suggestions for further increasing the use of the highest-efficacy methods within this population.

In chapter 3, we assess the health and wellness outcomes of infants born to adolescent mothers. Our nationwide study assesses the impact of adolescent pregnancy on the health and wellness of infants within their first year of life. Each infant in the study group (infants born to adolescent mothers) is matched with the control group (infants born to adult mothers) based on their mother's demographics. The outcomes assessed are: low birth weight, substance exposure, foster care, health status, mortality, emergency department visits, and wellness visits. The results suggested differences between the two groups, especially for low birth weight and emergency departments visits. However, the differences are not as drastic as previous research has found -- suggesting a promising result that the gap between these two groups may be closing. The chapter also includes recommendations on how to support adolescent mothers.

In chapter 4, we apply a statistical learning method to a difference-in-differences (DID) study setting. Commonly used DID methods rely on parametric statistical models that make strong assumptions about the unknown underlying functional form of the data. In this study, we extend existing statistical machine learning methods to target a DID

parameter, defined nonparametrically, while considering a larger nonparametric model space that makes fewer assumptions. We develop a general framework for DID designs that allow researchers to estimate causal or statistical effect quantities using machine learning while providing statistical inference. We demonstrate its performance through a simulation in which we compare it to more traditional methods. The study's motivating example estimates the effects of episode-based bundle payment on perinatal spending. The chapter concludes with limitations of the proposed estimator and suggestions for future work.

Chapter 5 applies machine learning to the problem of edge weight estimation for social networks. Social network analysis is used to visualize, quantify, and assess relationships between two entities. Within healthcare, social networks can be helpful in a variety of settings to quantify the impact of a relationship on healthcare outcomes, interventions, or physician treatment decisions. Algorithms have been used to predict information on social networks, such as edge existence, or similarity measures, such as common neighbors. However, little research focuses on weighted graphs and even less work on the estimation of their edge weights. Accurate weight estimation can serve as a data quality tool to check if the weights in the data are correct and where we would expect new stronger (or weaker) relationships to occur next. This study evaluates the performance of three estimators, including an ensemble machine learning approach, to predict the edge weights of a weighted social network. We use a faculty hiring example to compare the three methods' accuracy and finish with suggestions for future work.

Chapter 6 is the conclusion of the thesis. It includes a discussion of the overall impact of the research with respect to health care policy and developed techniques for

administrative claims data. Future work is proposed along with additional health care applications.

CHAPTER 1. INTRODUCTION

The volume of healthcare data is expected to continue growing faster than any other industry [1]. This creates a demand for the development of rigorous analytics and machine learning methods for applications to large health data sets These data sets, which contain personally identifiable information, come with privacy protections that place limitations on data visibility and its release. In addition, patient data often contains complex relationships such as non-linear relationships and heterogeneity. These characteristics can cause unique complications for analysis of health care data and restrict the use of out-of-the-box solutions. Notably, healthcare research has incredibly high stakes, it can be the difference between life and death and have a major impact on an individual's quality of life and medical treatment. For these reasons, the development of rigorous solutions is that much more critical. This thesis focuses on the application of analytics and machine learning to solve applied research problems based on healthcare data.

Section I, which includes Chapters 2 and 3 of the thesis, concentrates on health analytics with applications to child and maternal health. The data source for this section is Medicaid claims data from the Centers for Medicare and Medicaid Services (CMS). For our analysis we utilized three tables: Personal Summary (PS), Inpatient (IP), and Other Therapy (OT). The PS table contains the demographic and administrative information about the Medicaid enrollee. It has 105 data elements including items such as date of birth (DOB), gender, ethnicity/race, number of months the enrollee was insured in the current year, residential zip code, state of residency, and Medicaid eligibility. The Medicaid eligibility element identifies the reason the enrollee qualifies for public insurance. The

personal table is unique in that it is the only table that does not consist of administrative claims. The IP table includes all inpatient claims. An inpatient claim is one in which the Medicaid enrollee was admitted into the hospital or care facility. It contains 55 data elements and information about the stay within the facility. Examples of IP data elements we used in CHAPTER 3 are admission date, discharge date, diagnosis codes and procedural codes. Diagnosis codes specify the condition that is being treated in the claim and the procedural codes identify the physical action that is performed by the clinician. For example, the diagnosis code would state "influenza vaccine" and the procedural code would be for administration of the shot. The OT table includes all outpatient claims, such as doctor visits, physical therapy services etc. It contains 50 data elements and information very similar to that of the IP table. The main difference between IP and OT is that OT contains less elements for diagnosis and procedure codes. The CMS data is a rich data source for a variety of reasons. First, the data covers a non-trivial percentage of Americans on one insurance program; enabling us to get a reasonable snapshot of the American population. For example, recent enrollments numbers from CMS show as many as one in five American are enrolled in Medicaid [2]. Second, the data allows for longitudinal studies, which can be valuable when evaluating the impact of a policy on physician behavior or health care outcomes. Third, the data is provided in a centralized database and contains demographic information of the enrollees, such as age, gender, race, and state. The demographic information permits for analysis of health disparities across various subpopulations.

Chapter 2 evaluates contraception claims for women on Medicaid with chronic health conditions before and after the release of new contraceptive guidelines by the

Centers for Disease Control and Prevention (CDC). The purpose of the guidelines, titled the US Medical Eligibility Criteria for Contraceptive Use (MEC), is to assist health care providers in making evidence-based decisions on contraception. In the event of an unplanned pregnancy, the MEC identified 20 health conditions that posed a level of unacceptable risk to the health of the mother and/or fetus. Therefore, it recommended that these women use contraceptive methods with the highest level of efficacy. Over 13 million reproductive-aged women are included in the analysis with about 4% of the population having one of the health conditions. Women were identified as having one of the 22 conditions if there was a minimum of three Medicaid claims for that condition within a two-year period. A one-sided Poisson test on the contraception rates was used to determine if there was an increase in (1) overall contraception use and (2) the use of CDC recommended contraception (i.e. those of the highest efficacy). The results provide information on the impact of aggressive publication of a new federal health care policy on physician behavior, by evaluating if the dissemination methods used for the CDC were effective.

Chapter 3 assesses the health and wellness outcomes of infants born to adolescent mothers. Seven outcomes are evaluated: foster care, health risk level, infant mortality, low birth weight, substance dependency, number of emergency department (ED) visits and number of wellness visits. Over 65,000 infants, born in the year 2011, are included in the study. The chapter also includes the development of a sequential process for implementing the casual inference method of matching on medical claims data. The process involves a data pipeline that iteratively pairs infants to their mothers and then matches each adolescent mother to an adult mother based on demographic information. Outcomes of the infants born

to adolescent mothers are compared to those of adult mothers and evaluated using a proportion test for the rate outcomes (foster care, health risk level, infant mortality, low birth weight) and a two-sided Poisson test for the count outcomes (number of ED visits and number of wellness visits). In addition to finding the impact of adolescent pregnancy on infants, the study also evaluates differences by ethnicity/race and urbanicity – identifying additional health disparities. The chapter contains recommendations to address these disparities and ensure mothers are better informed on the healthcare needs of their infants.

Section II, which includes Chapters 4 and 5 of the thesis, focuses on machine learning applications to research areas that originated in disciplines outside of industrial engineering and statistics. These include casual inference, which began in economics, and social networks, which commonly found in computer science. This section of the thesis demonstrates how ensemble machine learning and statistical methods can be used in these research areas with applications to important health care problems.

Chapter 4's motivating example is the impact of episode-based payments (EBP) on perinatal costs in Arkansas [3]. Under this payment policy, physicians are rewarded for remaining below a pre-determined spending cap and penalized for spending above it. To establish causality, the costs in Arkansas are compared to those of nearby southern states using a difference-in-differences (DID) study design. DID takes the difference of the difference between the intervention and comparison group pre-intervention and the difference between of the two groups post-intervention. A popular parameter, the average treatment effect among the treated (ATT), is used to determine the average causal impact on a randomly selected unit within the intervention group. Parametric algorithms, such as regression, are traditionally used to predict the outcome for the ATT. However, these

algorithms are limited by the outcome being correctly specified, which is rarely the case in practice. Our proposed super learner algorithm does not assume a functional form of the outcome. The super algorithm is therefore more flexible and less sensible to bias due to misspecification. In this study we determine if by using machine learning to relax these strict parametric assumptions to improve estimation of the ATT. We develop a simulation study to evaluate each algorithm's performance. The simulation is based on commercial medical claims data from IBM Truven Marketscan. The variables used in the simulation study are generated based on the parameters derived from perinatal claims for women in Arkansas and comparison states. The performance of super learner is evaluated at three different effect sizes (-$250, -$400, and -$750) and two different sample sizes (2,000 and 5,000). The chapter discusses the implications of the results and proposes future work.

Chapter 6 focuses on the application of machine learning for the estimation of the edge weight for social networks. Social network analysis can be used to visualize, quantify, and assess relationships between two entities. The edge weights of social networks represent the strength of the social interaction. Social networks can be a valuable tool in health care by assessing how social interactions can impact health outcomes through social influence, physician behavior, etc. The estimator's performance for the edge weights is evaluated for three algorithms: generalized linear model (glm), Poisson mixture model and super learner. The predictors for estimation are metrics of the nodes, such as common neighbors or community membership, as well as a similarity measure for the metadata (covariates) of the node. We test each algorithm using a faculty hiring network. In the network, each node is an American institution. An edge exists between the two nodes *i* and *j* if an individual who received their PhD from node *i* is

hired as a professor by node *j*. The edge weight is how many individuals have been hired from node *i* to node *j*. Health care applications as well as future work is discussed in the conclusion of this chapter.

CHAPTER 2. EVALUATING CONTRACEPTION CLAIMS FOR COMPLIANCE TO CDC GUIDELINES

2.1 Introduction

In 2010, the Centers for Disease Control and Prevention (CDC) released the Medical Eligibility Criteria for Contraceptive Use (MEC) to guide providers to evidence-based medical decision making regarding contraceptive provision. The MEC highlighted 20 medical conditions that present an increased risk for adverse outcomes during pregnancy, stating that long-acting, highly-effective contraceptive methods may be the best choice for women with these medical conditions [4]. These contraceptive methods include reversible options, like intrauterine devices (IUDs) and implants, and permanent options, like sterilization. Sole use of behavioral-based methods, like condoms, were not recommended due to their higher typical-use failure rates.

The CDC disseminated MEC guidelines through mobile applications, publications and presentations [5]. Nevertheless, a recent survey identified that providers' knowledge of the MEC is low [6]. Some studies exploring individual medical conditions have identified low levels of highly-effective contraceptive use, high levels of unintended pregnancy, and provider-imposed limitations to effective contraceptive options for women with these conditions [7-10].

The MEC guidelines may be particularly relevant for providers who serve women of low income and those enrolled in Medicaid. These women are more likely to experience unintended pregnancies [11] and medical comorbidities [12]. Furthermore, in 2016 over

20% of reproductive- aged women in the US were insured by Medicaid [13] and Medicaid covered just under half of all births in 2010 [14]. However, patterns regarding contraceptive provision for Medicaid-insured women with high-risk medical conditions before and after the release of the US MEC are unknown.

The objective of this paper is to investigate contraceptive provision for women enrolled in Medicaid with one or more of the 20 highlighted medical conditions. We quantified the provision of any family planning management for these women and determined if the use of the highly effective methods had increased before and after the release of the MEC.

2.2 Methods

2.2.1 Data Description

We used Medicaid Analytical Extract (MAX) medical claims acquired from the Centers of Medicare and Medicaid Services (CMS) for the years 2009-2012. The MAX dataset consists of individual-level claims data for all Medicaid-enrolled beneficiaries. We examined enrollees from 14 states, accounting for over 50% of all Medicaid enrollees in the United States. The 14 states included 10 southeast states (Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, and Texas) as well as states from different regions of the country (California, Minnesota, New York, and Pennsylvania). All data was derived from the MAX files and meet a minimum cell size of 11 patients, according to the Data Use Agreement with CMS.

2.2.2 Identifying the Study Population

The *overall population* consists of all reproductive-aged women enrolled in Medicaid in one of the 14 states in the years 2008, 2009, 2011, and 2012. We investigated two time periods: the first time-period spans two years prior to the MEC release (2008 and 2009) and the second time period spans two years after the MEC release (2011 and 2012).

The *study population* is a subset of the overall population. It includes reproductive-aged women with one or more of the 20 conditions listed in the MEC (Table 1). The study population was stratified by:

- Age group of the women (15-24; 25-34; 35-44) [15]
- Medical condition
- State

We obtained the age of each woman using the date of birth in the Personal Summary (PS) table of the MAX data. A woman was assigned to an age group based on her age at the beginning of each time period (years 2008 and 2011).

Medical condition is defined as one of the 20 conditions identified by the MEC. A woman was identified as having a specific non-surgical condition if she had at least three Medicaid claims for that specific condition recorded on three different days within a time period (2008-2009 or 2011-2012) [16]. The Medicaid claim could be a claim from the Other Therapy (OT) table or the Inpatient (IP) table. International Classification of Disease-Ninth Revision (ICD-9) diagnosis codes were used to identify non-surgical conditions. Different approaches were needed to identify women with surgical conditions. The two surgical conditions identified by the MEC were bariatric surgery and solid organ transplant. These women were identified using the corresponding surgery condition

procedure codes in the IP table of the MAX data. We screened for the procedure codes in the IP claims that occurred within the two time periods, and a woman was assigned to the time period in which the surgery occurred. When identifying patients, we considered each condition separately to account for comorbidities. See Table 1 for the ICD-9 and CPT codes used to identify each of the medical conditions.

We identified a woman's state by the state listed on her claim. This ensured that a woman was counted in each state in which she received service.

2.2.3 Outcome Analysis

We considered two outcome measures as described below. We documented the number of women for both outcome measures for each time period and medical condition.

2.2.3.1 Outcome 1: Family Planning Management (FPM)

We defined a family planning management (FPM) claim as one containing a diagnosis code that began with "V25", the overarching code for "Encounter for contraceptive management" [17]. The FPM measure includes many forms of contraceptive claims ranging from discussion of contraceptive options with the clinician to extensive procedures such as IUDs and sterilization. We aggregated the number of women with V25 claims for each time period and each condition, with comparisons between the study population and the overall population.

2.2.3.2 Outcome 2: Highest-efficacy Methods (HEM)

The highest-efficacy methods (HEM) outcome included contraceptive claims for IUDs, contraceptive implants, and sterilizations. The MEC recommends HEM methods for women with high-risk conditions. We used the diagnosis codes for IUD insertion (V25.1), IUD surveillance (V25.42), and implant surveillance (V25.43) and searched in both IP and OT claims. Due to the nature of the procedure, we searched for sterilizations (V25.2) in IP claims only.

2.2.3.3 Rate Analysis

FPM and HEM utilization rates had increased nationally in the years of the study [18]. Therefore, we used utilization rates of the overall population as a scaling factor for the study population. The scaling factor was applied to the study population utilization rate to accurately determine the change in rates before and after the MEC.

A one-sided exact Poisson test was used to determine if there was a statistically significant increase in contraceptive provision for the study population. The alternative hypothesis was defined as the before-MEC rate was smaller than the after-MEC rate. A ratio greater than 1 indicates an increase in provision; a ratio of 1.1 would indicate a 10% increase in the rate.

The test statistic comparing before-MEC and after-MEC outcome measures is scaled by the rates of each time period (2008-2009 and 2011-2012) by the corresponding outcome measure of the overall population.

Setting up the problem, we define:

C_i^k: Total number of women in overall population in phase i for outcome k

$C_i = \sum_{\forall k} C_i^k =$ Total overall population in phase i

$\pi_i^k =$ rate of outcome k in phase i for overall population $= \frac{C_i^k}{C_i}$

M_i^k: Total number of women in study population in phase i for outcome k

$M_i = \sum_{\forall k} M_i^k =$ Total study population in phase i

Given that M_i is a subset of the overall population, the expected number of women in the study population with outcome k is $E[M_i^k] = \pi_i^k * M_i^k$. Furthermore, let

μ_{ij}^k : scaled proportion ofthe study population for time period i and medical condition j where: time period $i \in (0,1)$, medical condition $j \in (1,2, \dots 20)$ and outcome $k \in (0,1)$.

We determine the rate for each time-period:

Outcome 1: FPM ratio which includes counseling, insertions, and surveillance for contraceptive methods.

$$\mu_{ij}^1 = \frac{scaled\ number\ of\ women\ from\ study\ population\ with\ FPM\ claim}{total\ study\ population}$$

$$= \frac{M_i^1 - E[M_i^1]}{M_i}$$

Outcome 2: HEM ratio which includes (1) insertion and surveillance of intrauterine devices (IUD) and (2) implants.

$$\mu_{ij}^2 = \frac{scaled\ number\ of\ women\ from\ study\ population\ with\ HEM\ claim}{total\ study\ population}$$

$$= \frac{M_i^2 - E[M_i^2]}{M_i}$$

The ratio of rates in time period 0 and time period 1 were assessed using a one-sided exact Poisson test.

$$H_0: \frac{(\mu_{1j}^k)}{(\mu_{0j}^k)} = 1$$

$$H_1: \frac{(\mu_{1j}^k)}{(\mu_{0j}^k)} > 1$$

The test procedure was applied to all conditions together and each individual condition separately. For the condition-level analysis, we corrected for the testing of multiple outcomes simultaneously using Bonferroni's correction.

2.3 Results

2.3.1 Study Population

Table 2 shows the number of women covered by Medicaid in the selected 14 states. The study included more than 12 million women in both time periods. Most reproductive-aged women enrolled in Medicaid did not have claims for these conditions; less than five percent were identified as having one of the 20 high-risk conditions. Though low, there was an increase from 3.5% in the before-MEC time period to 3.9% in the after-MEC time period. More than half of the women with high-risk conditions were older, with 53.5% and 66.9% (before and after MEC time period, respectively) in the 35-44 age group (Figure 1). Four conditions made up 83% of the study population (Figure 2). The most common conditions were hypertension and diabetes, followed by epilepsy and HIV. Most of the study population came from the states in the southeast -- 60% of the before-MEC study population and 57% of the after-MEC study population. For summary for study population by count and percentage see Table 2.

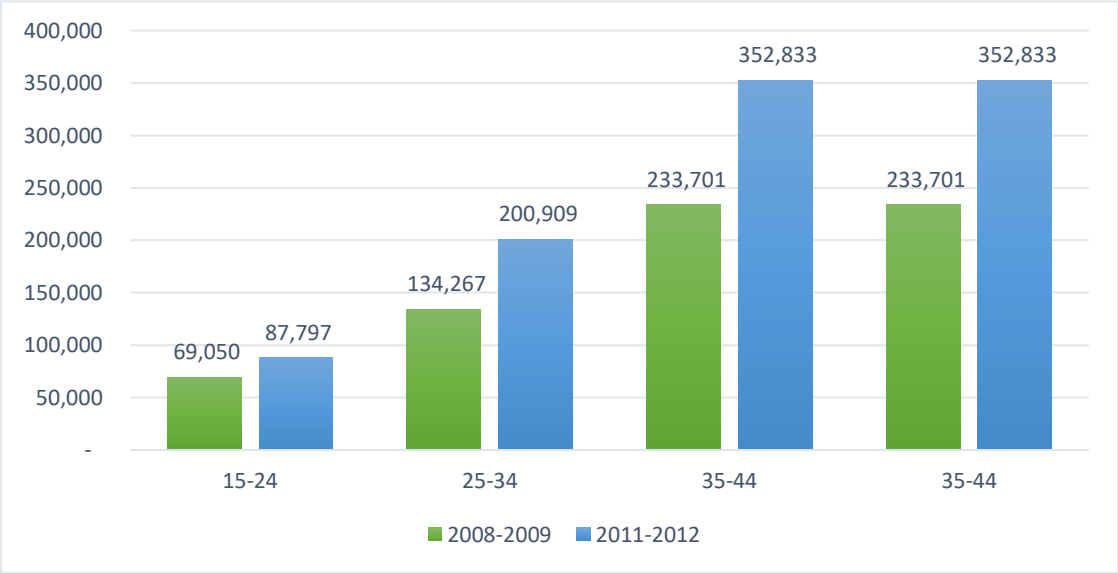
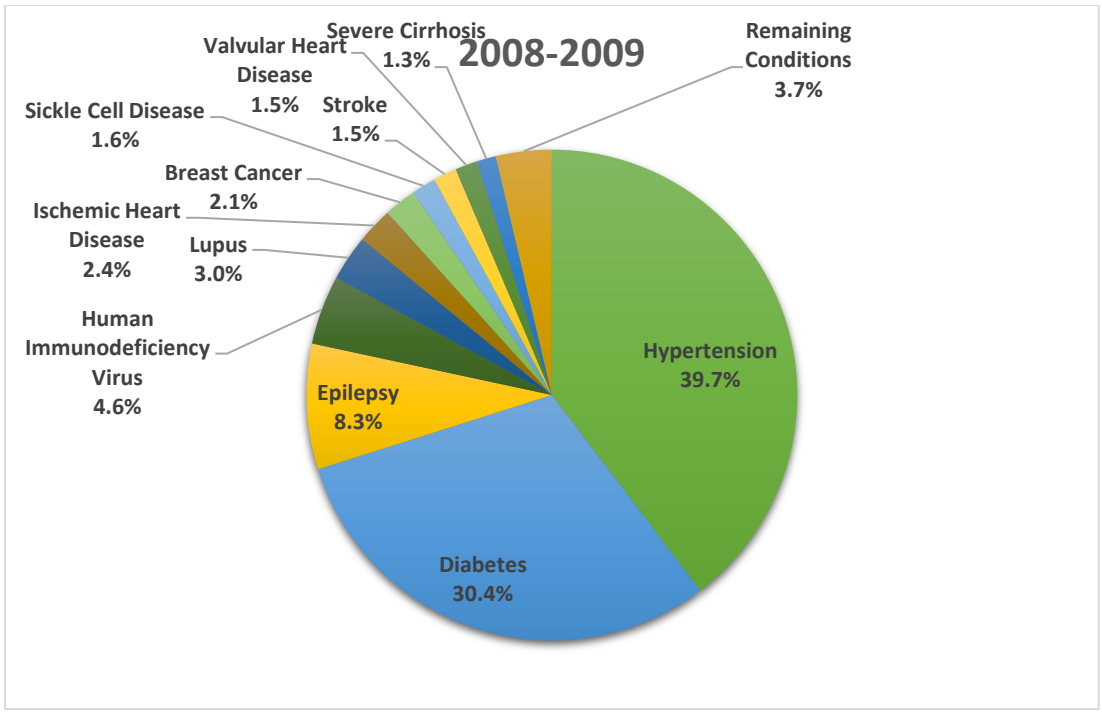


Figure 1: Age Count of Women with Chronic Medical Condition by Time Period



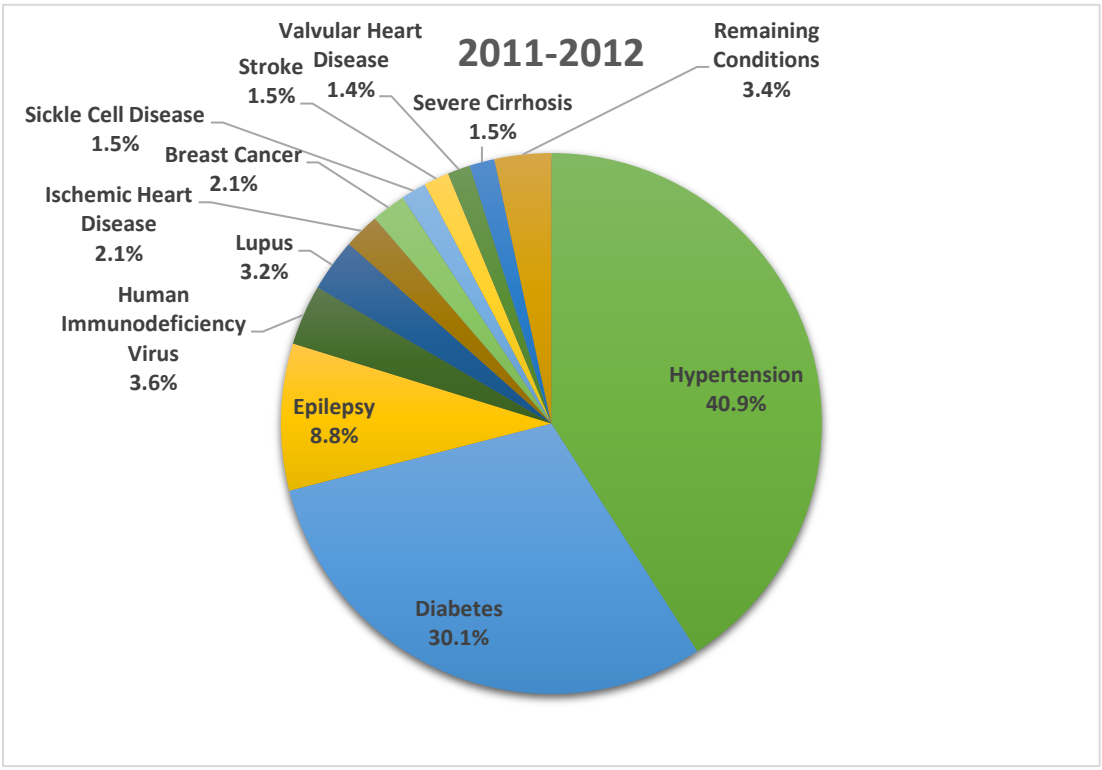


Figure 2: Medical Conditions by Time Period

2.3.2 Outcome 1: FPM - Aggregate and Condition Level Analysis

Family planning management (FPM) provision is detailed in Table 3 for the overall population, the study population, and individual medical conditions. It provides the number of women with a FPM claim for each condition, the *percentage rates* of FPM [defined as the ratios of the number of women (values from Table 3) over the total number of women for the corresponding condition and time period (values from Table 2)], and the results of a one-sided test for significance. We considered 19 conditions for the FPM outcome; schistosomiasis condition was excluded because of low counts (cell size smaller than 11 patients cannot be reported).

There was an increase in FPM provision from 17.9% to 18.2% for all reproductive-aged women in Medicaid. We saw a comparable increase for women in the study population from 16.7% to 17.8%. There was variability of FPM by medical condition in both time periods, ranging from 4.4% for those with liver cancer to 46.6% for those with peripartum cardiomyopathy. The conditions with the highest rates of FPM provision in both time periods were peripartum cardiomyopathy, sickle cell disease (SCD), and thrombogenic heart disease (Figure 3).

Before and after the MEC release, 12 of the 19 conditions showed a statistically significant increase at the 1% significance level; the Bonferroni adjusted p-value threshold was 0.0005. After accounting for the increase at the overall population level, five conditions showed a greater than 30% increase in FPM including solid organ transplant, endometrial and ovarian cancer, liver cancer, bariatric surgery and HIV (Table 3). The 7 conditions that did not show a significant increase in FPM were epilepsy, GTD, peripartum

cardiomyopathy, sickle cell disease, thrombogenic heart disease, tuberculosis, and valvular heart disease (Table 3).

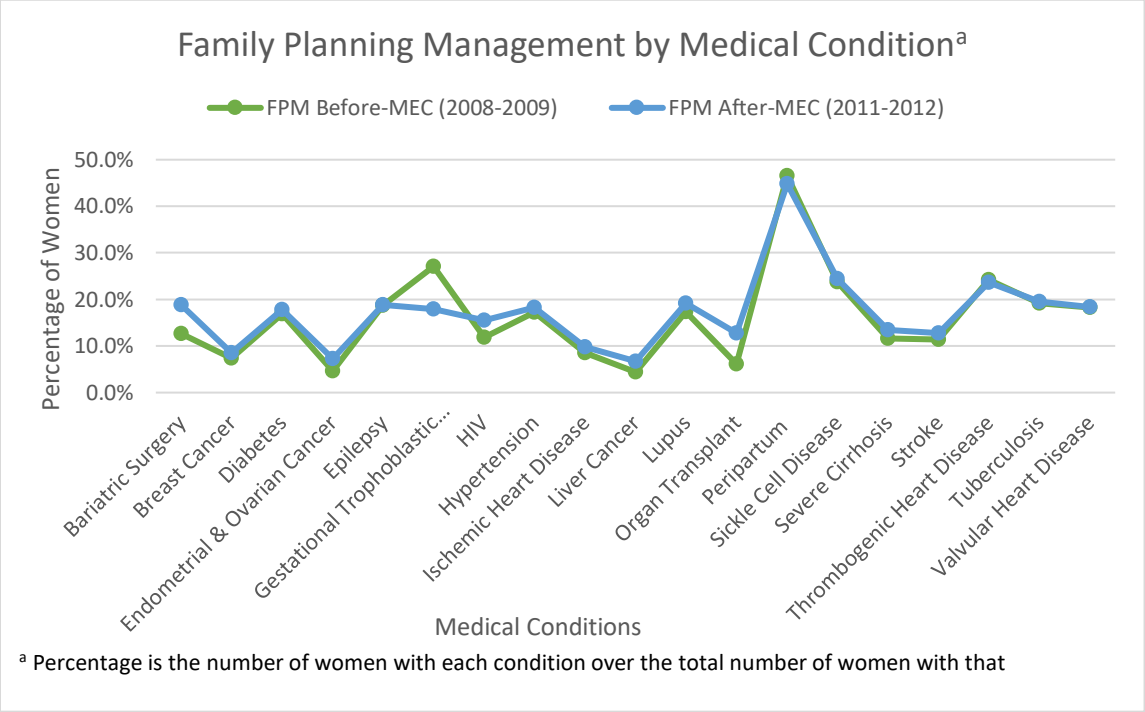


Figure 3: Percentage of Women with a Family Planning Management (FPM) Claim by Medical Condition

2.3.3 Outcome 2: HEM - Aggregate and Condition Level Analysis

Highest-efficacy method (HEM) provision is detailed in Table 4 for the overall population, the study population, and individual medical conditions. It provides the number of women in the HEM outcome for each condition, the *percentage rates* of HEM (as defined above) and the results of a one-sided test for significance. We considered 16 of

the 20 medical conditions. GTD, liver cancer, schistosomiasis, and solid organ transplant were excluded due to counts below 11.

There was an increase in HEM provision from 3.5% to 5.0% for all reproductive-aged women in Medicaid. We saw a comparable increase for women in the study population from 4.1% to 5.7%. There was variability of HEM by medical condition in both time periods, ranging from <1% for those with endometrial or ovarian cancer to greater than 25% for those with peripartum cardiomyopathy (Figure 4). After accounting for the increase at the overall population level, all of the 16 conditions showed a statistically significant increase at the 1% significance level. The Bonferroni adjusted p-value threshold was 0.0006. Two conditions saw a doubling of HEM provision between the time points: bariatric surgery and endometrial and ovarian cancer.

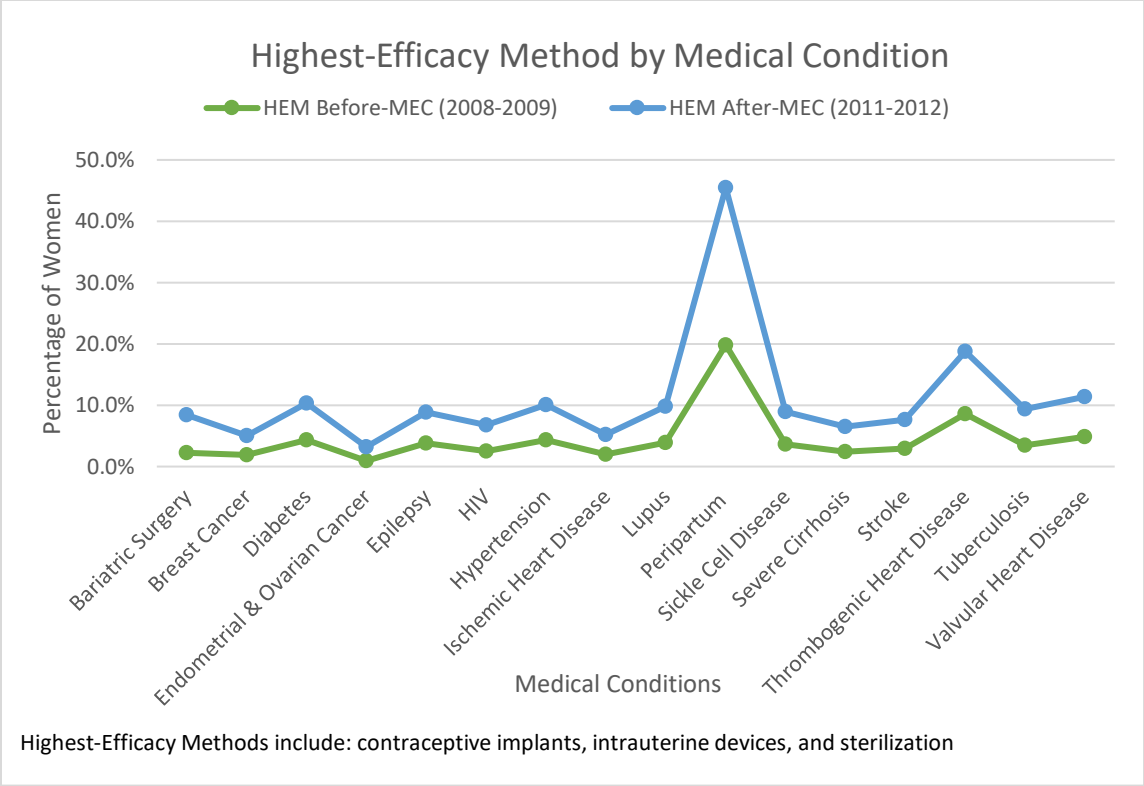


Figure 4: Percentage of Women with a Highest-Efficacy Method (HEM) Claim by Medical Condition

2.4 Discussion

This study showed an increase in the provision of family planning, generally, and in the provision of contraceptive methods with the highest efficacy (IUDs, implants and sterilization). The study specifically showed this increase for women with one or more of the 20 medical conditions identified as high risk by the MEC in the two years after its release. When all conditions were considered together, the difference was found to be statistically significant for both outcomes. For individual conditions, statistical significance was found for the majority of medical conditions for FPM and for all medical conditions

for HEM. The increase in provision of highly effective contraceptive methods mirrors the trends that were happening nationally. According to an analysis from the National Survey of Family Growth, the use of IUDs and contraceptive implants among reproductive-aged American women increased from 6% in 2008 to 12% in 2012 [18] . While our study accounted for the increase seen in the overall population and documented an increase across medical conditions, the HEM rates for women with one or more of the 20 conditions were below the national average. Champaloux and colleagues had a similar finding in their review of claims of women with medical conditions from a privately insured population [19].

The uptake of the highest-efficacy methods is particularly important for women with the identified medical conditions because pregnancy can lead to severe health outcomes for this population. The physiologic changes of pregnancy affect nearly every organ system in the body. Normal pregnancy is a state of anemia, increased oxygen demand and cardiac output, hypercoaguability, immune compromise and insulin resistance, to name a few. These necessary changes support gestation and are generally well tolerated by healthy women. However, women with underlying medical conditions may experience amplification of their condition or predisposition to complications and morbidities, including maternal death [20]. The maternal mortality rate in the US is the worst of the developed world and climbing [21]. A recent review of maternal mortalities from nine states identifies hemorrhage, cardiovascular and coronary conditions, infection and cardiomyopathy as the most common causes of maternal mortality. The review identified both racial and age-related differences underlying the cause of mortality. It also estimates that 63.2% of these deaths are preventable [22]. One step proximal to preventing maternal

mortality is to prevent maternal morbidity. As a prevention strategy for serious morbidity, a woman with a high-risk medical condition should have ready access to the most effective methods of contraception until she desires pregnancy. Then when wanting to conceive, a woman should have access to pre-conceptual care to optimize her health, medication management and transition her to and through pregnancy. This strategy will help women with high-risk conditions to attain their reproductive goals while decreasing their health risk [23].

This study found that FPM and HEM provisions varied by medical condition. For example, we found that cancer conditions (breast, endometrial and ovarian) showed lower rates for FPM and HEM, relative to other medical conditions. One explanation is nature of these conditions and the methods used to treat them. For example, hysterectomies or bilateral oophorectomies are common forms of treatment for endometrial and ovarian cancer. After such procedures, contraception is no longer needed. In addition, the nature of these cancers can limit women's contraceptive options. IUDs are contraindicated in women with endometrial cancer and hormonal IUDs and implants for women with breast cancer [4]. On the other hand, peripartum cardiomyopathy had the highest rates for FPM and HEM. One possible reason for this is condition is associated with very high morbidity and mortality as high as 19% with a subsequent pregnancy [24]. Second, by definition, peripartum cardiomyopathy is diagnosed the last month of pregnancy or first few months after delivery. The peripartum timing of the diagnosis may create the opportunity for an OB/GYN provider to facilitate a conversation on the importance of contraception due to the high-risk of a subsequent pregnancy. However, the medical conditions impacting the most women fall in between these extremes. Hypertension, diabetes, epilepsy and HIV

affected more than 430,000 women in our study. These conditions put them at high-risk for adverse health outcomes with pregnancy. Therefore, additional focus should be placed on these conditions for highest impact.

There are several limitations to consider in this study. The CMS data restricts us to diagnosis and procedure coding by the clinician during the visit. Therefore, we may not have captured women using methods that did not require a clinician or for un-coded services. Similarly, we were only able to reliably capture sterilization procedures that occurred within the years of the study. Hence, there are women who use tubal sterilization and partner vasectomy as a form of birth control that were not captured. For these reasons, we believe our findings to be underestimates. Because claims data do not include sexual or relationship history, we were unable to ascertain whether a woman was at risk for pregnancy by being sexually active with a male partner, nor were we able to assess if her medical condition precluded sexual activity or fertility. Medicaid eligibility criteria for women varies by state, and women who become pregnant may only be eligible for a limited time. For the two very common conditions, hypertension and diabetes, the MEC guidelines apply to more severe cases; our analysis was more inclusive. Finally, statistical analysis shows an association but cannot directly address causality or reasons for the change.

Overall, this study found a limited, but encouraging, change in clinical practice shortly after release of the MEC guidelines. The relatively low rate of family planning and highly effective method provision we found suggests that access to highly effective contraceptives was a barrier. Access issues for contraception can arise from financial and systems issues as well as from provider bias [25]; these may also present opportunities for ongoing and next steps for full implementation of the MEC guidelines.

Historically, access to contraception methods have been limited, especially for women with low income [26]. Fortunately, there have been several efforts to lessen financial and system barriers to accessing contraception since release of the MEC. After the mandate for contraceptive coverage from the Affordable Care Act went into effect, the percentage of women using IUDs and implants increased, while the usage of oral contraception remained flat among sexually active women [27, 28] . The 6|18 Initiative by the CDC and its partners outline four interventions for reducing financial and logistic barriers for public and private payers and providers. For women with no insurance coverage, family planning services can be obtained from the Federal Title X grant [29]. These multi-level and collaborative approaches to reducing barriers may serve to increase the uptake of the MEC guidelines [30].

In addition to these interventions, parallel programs have been working to ensure provider knowledge and application of the MEC in practice. This has included endorsement and implementation support of the MEC by several medical associations including the American College of Obstetricians and Gynecologists, the American Academy of Family Physicians and the American Academy of Pediatricians [31-33]. Focusing future efforts on subspecialty providers may help ensure that women with high-risk medical conditions receive evidence-based care and referrals to provision, as needed. Additionally, to increase demand, grassroots and community-based interventions could be explored to educate women with these conditions about the risks of unintended pregnancy and the contraceptive options uniquely suited for them.

Table 1: ICD-9 Codes for 20 High Risk Medical Conditions Identified by the MEC

#	Condition	ICD-9 Code
1	breast cancer	174
2	diabetes	250
3	endometrial & ovarian cancer	179, 182, 183
4	epilepsy	345
5	History of bariatric surgery (last 2 years)[a]	V45.86
6	HIV	042
7	hypertension	401-405
8	ischemic heart	410, 412-414
9	malignant gestational trophoblastic disease	181
10	malignant liver tumors and hepatocellular carcinoma of the liver	155
11	peripartum cardiomyopathy	674.5
12	schistosomiasis with fibrosis of the liver	120.9
13	severe cirrhosis	571
14	sickle cell	282.6
15	solid organ transplant in the last 2 years[a]	V42.0, V42.1, V42.6, V42.7, V42.83, V42.9
16	stroke	430-434, 436-438
17	systemic lupus erythematosus	710.0
18	thrombogenic mutations	286.
19	tuberculosis	010-018
20	valvular heart disease	424

[a] Current Procedural Terminology (CPT) Code; used to identify surgical medical conditions (Source: www.icd9data.com)

Table 2: Reproductive-aged Women enrolled in Medicaid by Age, State, and Condition

Time-Period		2008-2009		2011-2012	
Overall Population		12,422,899		13,597,612	
Study Population[a]		437,018	3.5%	527,660	3.9%
Age[b]	15-24	69,050	15.8%	87,797	16.6%
	25-34	134,267	30.7%	200,909	38.1%
	35-44	233,701	53.5%	352,833	66.9%
State[c]	Alabama	16,312	3.7%	18,721	3.5%
	Arkansas	10,310	2.3%	11,750	2.2%
	California	84,653	19.1%	96,830	18.1%
	Florida	41,298	9.3%	54,755	10.3%
	Georgia	31,543	7.1%	32,481	6.1%
	Louisiana	23,031	5.2%	25,568	4.8%
	Minnesota	11,389	2.6%	16,844	3.2%
	Mississippi	17,284	3.9%	18,888	3.5%
	New York	70,602	16.0%	97,243	18.2%
	North Carolina	40,180	9.1%	41,878	7.8%
	Pennsylvania	10,374	2.3%	18,615	3.5%
	South Carolina	15,134	3.4%	20,439	3.8%
	Tennessee	32,329	7.3%	34,784	6.5%
	Texas	37,985	8.6%	44,823	8.4%
Medical Conditions[c]	Bariatric Surgery	5,158	1.0%	6,726	1.1%
	Breast Cancer	11,072	2.1%	13,016	2.1%
	Diabetes	159,042	30.4%	190,648	30.1%
	Endometrial & Ovarian Cancer	2,259	0.4%	2,557	0.4%
	Epilepsy	43,213	8.3%	55,666	8.8%
	Malignant Gestational Trophoblastic Disease	118	0.0%	123	0.0%
	Human Immunodeficiency Virus	23,865	4.6%	22,894	3.6%
	Hypertension	207,286	39.7%	259,571	40.9%
	Ischemic Heart Disease	12,357	2.4%	13,577	2.1%
	Liver Cancer	273	0.1%	342	0.1%
	Lupus	15,750	3.0%	20,014	3.2%
	Schistosomiasis	120	0.0%	NA	
	Solid Organ Transplant	588	0.1%	578	0.1%
	Peripartum Cardiomyopathy	2,817	0.5%	3,024	0.5%
	Sickle Cell Disease	8,395	1.6%	9,564	1.5%
	Severe Cirrhosis	6,626	1.3%	9,451	1.5%
	Stroke	8,090	1.5%	9,612	1.5%
	Thrombogenic Heart Disease	4,944	0.9%	5,645	0.9%
	Tuberculosis	2,938	0.6%	2,469	0.4%
	Valvular Heart Disease	7,645	1.5%	8,630	1.4%

[a] Percentage is study population (women with high-risk conditions) relative to the overall population
See Appendix A for list of conditions highlighted by the MEC

[b] Percentage is stratification group relative to study population

[c] Percentage is stratification group relative to sum of women in that strata
Sum of all categories in stratification group may be greater than total study population; women can belong to more than one category in same stratification

Table 3: Family Planning Management: Number of Reproductive-aged Women with Medical Conditions Enrolled in Medicaid and Statistics for Poisson Test Ratio

Family Planning Management (FPM)[a]	2008-2009	2011-2012	Estimate[b]	p-value[c]	Lower Bound CI[d]
Overall Population	2,221,325 (17.9%)	2,477,023 (18.2%)			
Study Population	87,115 (16.7%)	112,851 (17.8%)	1.06	<0.001	1.05
Medical Conditions					
Bariatric Surgery	650 (12.6%)	1265 (18.8%)	1.49	<0.001	1.42
Breast Cancer	822 (7.4%)	1117 (8.6%)	1.16	<0.001	1.13
Diabetes	26915 (16.9%)	33928 (17.8%)	1.05	<0.001	1.04
Endometrial & Ovarian Cancer	105 (4.6%)	187 (7.3%)	1.58	<0.001	1.48
Epilepsy	8104 (18.8%)	10469 (18.8%)	1.00	0.55	0.98
Malignant Gestational Trophoblastic Disease	32 (27.1%)	22 (17.9%)	0.66	1	0.49
HIV	2816 (11.8%)	3549 (15.5%)	1.31	<0.001	1.28
Hypertension	35681 (17.2%)	47465 (18.3%)	1.07	<0.001	1.06
Ischemic Heart Disease	1049 (8.5%)	1331 (9.8%)	1.15	<0.001	1.12
Liver Cancer	12 (4.4%)	23 (6.7%)	1.52	<0.001	1.25
Lupus	2731 (17.3%)	3830 (19.1%)	1.10	<0.001	1.07
Peripartum Cardiomyopathy	1312 (46.6%)	1355 (44.8%)	0.96	0.95	0.9
Severe Cirrhosis	769 (11.6%)	1275 (13.5%)	1.16	<0.001	1.12
Sickle Cell Disease	1996 (23.8%)	2337 (24.4%)	1.03	0.02	1.00
Solid Organ Transplant	36 (6.1%)	74 (12.8%)	2.14	<0.001	1.86
Stroke	922 (11.4%)	1224 (12.7%)	1.12	<0.001	1.08
Thrombogenic Heart Disease	1200 (24.3%)	1335 (23.6%)	0.97	0.94	0.93
Tuberculosis	564 (19.2%)	481 (19.5%)	1.02	0.24	0.96
Valvular Heart Disease	1399 (18.3%)	1584 (18.4%)	1.00	0.52	0.96

Bolded conditions were statistically significant

[a] Percentage is number of women with a FPM claim relative to women in that category (See Table 1 for values)
FPM claim includes all claims with an ICD-9 code that begins with "V25"

[b] The estimate is the ratio of the after-MEC scaled rate over the before-MEC scaled rate

[c] p-values based on one-sided Poisson test at 99% confidence level
Bonferroni adjustment for p-value threshold: 0.005

[d] One-sided 99% Confidence Interval

Table 4: Highest Efficacy Methods: Number of Reproductive-aged Women with Medical Conditions Enrolled in Medicaid and Statistics for Poisson Test Ratio

Highest Efficacy Methods (HEM)[a]	2008-2009	2011-2012	Estimate[b]	p-value[c]	Lower Bound CI[d]
Overall Population	437,036 (3.5%)	679,230 (5.0%)			
Study Population	21,413 (4.1%)	36,176 (5.7%)	1.37	<0.001	1.36
Medical Conditions					
Bariatric Surgery	114 (2.2%)	416 (6.2%)	2.8	<0.001	2.68
Breast Cancer	214 (1.9%)	400 (3.1%)	1.59	<0.001	1.55
Diabetes	6892 (4.3%)	11377 (6.0%)	1.38	<0.001	1.37
Endometrial & Ovarian Cancer	21 (0.9%)	58 (2.3%)	2.43	<0.001	2.27
Epilepsy	1658 (3.8%)	2813 (5.1%)	1.32	<0.001	1.3
HIV	602 (2.5%)	976 (4.3%)	1.69	<0.001	1.65
Hypertension	8902 (4.3%)	15072 (5.8%)	1.35	<0.001	1.34
Ischemic Heart Disease	242 (2.0%)	439 (3.2%)	1.65	<0.001	1.6
Lupus	615 (3.9%)	1187 (5.9%)	1.52	<0.001	1.48
Solid Organ Transplant	NA	24 (4.2%)		NA	
Peripartum Cardiomyopathy	559 (19.8%)	775 (25.6%)	1.29	<0.001	1.21
Sickle Cell Disease	302 (3.6%)	511 (5.3%)	1.49	<0.001	1.44
Severe Cirrhosis	159 (2.4%)	391 (4.1%)	1.72	<0.001	1.66
Stroke	237 (2.9%)	452 (4.7%)	1.6	<0.001	1.55
Thrombogenic Heart Disease	423 (8.6%)	576 (10.2%)	1.19	<0.001	1.14
Tuberculosis	102 (3.5%)	146 (5.9%)	1.71	<0.001	1.61
Valvular Heart Disease	371 (4.9%)	563 (6.5%)	1.35	<0.001	1.3

[a] Percentage is number of women with a HEM claim relative to population for that category (See Table 1 for values)

HEM claims includes IUDs, contraceptive implants and sterilization

[b] The estimate is the ratio of the after-MEC scaled rate over the before-MEC scaled rate

[c] p-values based on one-sided Poisson test at 99% confidence level

Bonferroni adjustment for p-value threshold: 0.006

[d] One-sided 99% lower bound of confidence interval

CHAPTER 3. HEALTH AND WELLNESS OUTCOMES OF INFANTS BORN TO ADOLESCENT MOTHERS

3.1 Introduction

While teenage pregnancies in the US have dropped nationwide, the US rate still remains high compared to other developed countries [34], particularly in lower income populations [35]. Previous studies have found that infants born to teenage mothers are at risk for lower educational attainment [36], lower income, and higher rates of unemployment, increased risk for incarceration, increased health care costs, increased rate of entering into foster care, and increased risk of teenage pregnancy for the child [37]. Infants born to adolescent mothers also have worse health outcomes. For example, infants born to adolescent mothers were found to be associated with higher rates of low birth weight, preterm delivery, low APGAR score, postpartum hemorrhage, and neonatal mortality [38-40]. This is particularly relevant because the health status of an infant is linked with health outcomes later in life. For instance, previous studies have shown that birth conditions, such as low birth weight, are linked to developing obesity [41] , diabetes [42], and cardiovascular disease [43] in adults. In addition, emotional and/or physical stressors in the infant's first year of life, such as foster care or abuse, can be detrimental to cognitive, language, and behavioral function due to impact on brain development [44, 45].

Our study focuses on the health and wellness outcomes for infants born to Medicaid-enrolled mothers from birth through the first year of life in order to capture long-term effects on the infant's health. We selected the Medicaid population for our study because

it covers 48% of all births in the US [46]. Additionally, the nature of Medicaid eligibility suggests infants born to mothers in this population are at a higher risk of low income and/or low health status (i.e. medically needy) [47]. We choose to investigate outcomes that have been linked to long-term health and wellness of the infant and are trackable within Medicaid claims data. We included the following outcomes in the study: substance exposure (such as neonatal drug or alcohol exposure), health risk status, entry into foster care, infant mortality, preterm birth/low birth weight (LBW), emergency department (ED) visits, and wellness visits. This study adds value in several ways. To our knowledge, there is no study that focuses on the Medicaid population and multiple outcomes within the first year of the infant's life. We also found no study that tracks infants throughout their first year of life across several outcomes. For example, we assess infant mortality beyond birth and included deaths that occur within those first 12 months as well. In addition, the data we use provides the ability to stratify the results and discover new differences across race and urbanicity.

The objective of this study is to investigate the impact of adolescent pregnancy on the health and wellness of infants enrolled in Medicaid within their first year of life. The questions addressed in this study are:

- What is the frequency of occurrence of the outcome measures among Medicaid births?
- Are the health and wellness outcomes of these infants born to adolescent mothers on Medicaid significantly different than those of infants born to adult mothers on Medicaid?

- How do these differences change when evaluated by urbanicity level of the infant's residence and his or her race/ethnicity?

3.2 Methods

To isolate the impact of an adolescent mother from other factors, we used the causal inference method of exact matching. We matched adolescent mothers to adult mothers and compared the outcomes of the two infants' subpopulations. We developed algorithms based on geographic and demographic variables to find the mother-infant pairs as well as the adolescent-adult mother matching. This section will provide additional details on this process.

3.2.1 Data Description

The data source for this study consists of 2011-2012 Medicaid Analytic Extract (MAX) claims data acquired from the Centers of Medicare and Medicaid Services (CMS) and several publicly-available datasets: International Classification of Diseases, Ninth Revision (ICD-9); National Plan and Provider Enumeration System's National Provider Index (NPI); and the US Department of Agriculture's Economic Research Service rural-urban continuum codes (RUCC). Our study focuses on 42 states (Table 6) covering over 70% of all Medicaid enrollment nationwide [48]. The remaining eight states were not included for the following reasons: (1) data was not available (Colorado and Idaho), (2) data had quality issues (Alabama, California, and Mississippi), and (3) data was not large enough (Arizona, Hawaii, and New Jersey). The research was approved by the authors'

Institute's Institutional Review Board. In accordance with the Data Use Agreement, all data derived from the MAX files meet a minimum count of 11 patients.

3.2.2 Identifying the Study Population (Pairing of Adolescent Mothers and Infants)

The study population consists of Medicaid-enrolled infants born in 2011 to Medicaid-enrolled adolescent mothers. The age of the adolescent mothers was obtained from the Personal Summary (PS) table of the MAX data; ages 10 to 19 (per the World Health Organization definition of adolescence [49]), was based on the mother's final age in 2011.

3.2.2.1 Identification of mother-infant pairs from MAX data

The MAX data does not directly link mothers to their infants (denoted herein as *pairing*). Therefore, we used an algorithm similar to that in Palmsten et al [50]. First, we paired mothers based on two criteria: (1) the mother and infant state case ID number matched and (2) the infant's date of birth was between the hospital admission and discharge dates for the mother's delivery. For the adolescent mothers that were not paired by this approach, we used demographics of the mother and infant to find additional pairs. The process of pairing the mothers to their infants is discussed in this section.

Mothers and infants were identified from the MAX PS table using the recipient delivery code. The recipient delivery code equals 1 to indicate the enrollee gave birth that year, i.e. mothers, or 2 to indicate the enrollee was born that year, i.e. infants [51]. A maternal delivery claim is an IP claim with a recipient delivery code equal to 1, which specifies the claim of the mother's delivery. We removed cases of multiple maternal

delivery claims with the same state case ID number if they were partial duplicates (i.e. all the data elements in the duplicate claim did not match the original claim). These claims were removed because it is not clear which claim contained the correct information.

A newborn delivery claim is an IP claim with a recipient delivery code equal to 2, which specifies the claim of the infant's birth. We removed all cases of multiple newborn delivery claims with the same state case ID number. If multiple newborn delivery claims had the same state case ID number and demographic information (i.e. only hospital admission dates and NPI were different), we kept the claim with the earliest hospital admission date. Otherwise, all the claims with that same state case ID were removed from the study. We also excluded multiple gestation newborn delivery claims since infants of multiple births are more prone to adverse birth outcomes [52]. We identified the multiple gestation claims to remove using the ICD-9 diagnosis codes beginning with "651".

We utilized maternal and newborn delivery claims for the pairing algorithm and completed three iterations to achieve the highest pairing rate. In iteration 1, we paired adolescent mothers and infants based on three criteria.

1. The adolescent mother and infant have the same state case ID number
2. The adolescent mother and infant have the same zip code
3. The date of birth (DOB) of the infant is between the mother's hospital admission and discharge dates

In iteration 2, the remaining unpaired mothers were paired with the remaining unpaired infants via demographics. We assigned mother-infant pairs if the following criteria were met.

1. Mother and infant have the same ethnicity
2. Mother and infant have the same zip code
3. Mother and infant have the NPI on their delivery claim
4. The DOB of the infant is between the mother's hospital admission and discharge dates

We removed all duplicate pairs (mothers paired to multiple infants). In iteration 3, the remaining unpaired infants and mothers were paired using the demographics and DOB criteria of the second iteration but the NPI constraint was relaxed. Instead of the NPI directly, the mother and infant were paired when they have the same health provider address given by the NPPES database. The provider addresses were clustered using the Google's Open Refine application. We again removed all duplicate pairs found in this iteration. We validated iterations 2 and 3 by evaluating the number of pairs in iteration 1 that the process of iteration 2 and 3 was able to capture (88%).

3.2.3 Identifying the Control Population (Matching of Adolescent Mother-Infant Pairs with Adult Mother-Infant Pairs)

The control population consists of infants born to adult mothers that share similar demographics with the paired adolescent mothers (i.e. mothers of the study population). We used exact matching versus traditional propensity scoring because the size of the data allowed for it. We selected mothers aged 20 to 44 who gave birth in 2011 and were actively enrolled in Medicaid that year. The age of the adult mother was obtained from the waPS table of the MAX data and based on their final age in 2011.

To obtain the control population, first we identified adult mother-infant pairs by using the same pairing algorithm that was used for the study population. Second, we matched each adolescent mother-infant pair with an adult mother-infant pair based on factors included in the MAX data. These include: (1) mother's race/ethnicity; (2) mother's state of residency; (3) mother's basis for qualifying for Medicaid, i.e. Medicaid eligibility; (4) mother's health status; and (5) urbanicity of the mother's reported home residence.

3.2.3.1 Matching of adolescent mother-infant pairs with adult mother-infant pairs from MAX data

The five factors used to match the adolescent mother pair to the adult mother pair are discussed in this section. Mother's ethnicity was included because it is a determinant for pregnancy outcome disparities [53]. Home state is a factor for matching because of variations in Medicaid policies and reimbursement rates by state. The Medicaid eligibility provides information on if the mother was on Medicaid due to low income, medical disability, or foster care. Medicaid eligibility can serve as a proxy for environment and social health determinants.

As a proxy for the mother's health risk status, we considered the mother's Critical Risk Grouping (CRG) [54], which was determined using the 3M™ Core Grouping Software with the CRG version 1.12. CRG describes the burden of illness of the mother based on twelve months of administrative claims. All diagnosis codes, procedures and prescription drug codes found in a patient's medical claims are used to classify the patient's CRG into 10 categories. The classification scale ranges from 1 (healthy) to 9 (catastrophic condition). We further categorize health risk status as low risk (CRG 1 to CRG 3), medium

risk (CRG 4 to CRG 5a), and high risk (CRG 5b to CRG 9). See Table 7 for more details on the classification levels.

The urbanicity of the mother was determined by her county residence. Each county is categorized based on the corresponding RUCC [55]: urban (1-3), suburban (4-6), or rural (7- 9). The RUCC level is based on the population of the area. See Table 8 for more details on the RUCC definitions.

In the first iteration, we matched adolescent mother pairs to adult mother pairs using all five demographic criteria. The second iteration, we removed the Medicaid eligibility because foster care was one of the eligibility criteria and no adult mothers would satisfy that eligibility. For the final iteration, we removed eligibility and relaxed ethnicity. An match could be made between a mixed ethnicity and mother who shared one of the mixed ethnicities. All remaining unmatched adolescent mother-infant pairs were removed from the study. Figure 5 shows the percentage adolescent mother-infant pairs matched to adult mother-infant pairs at each iteration.

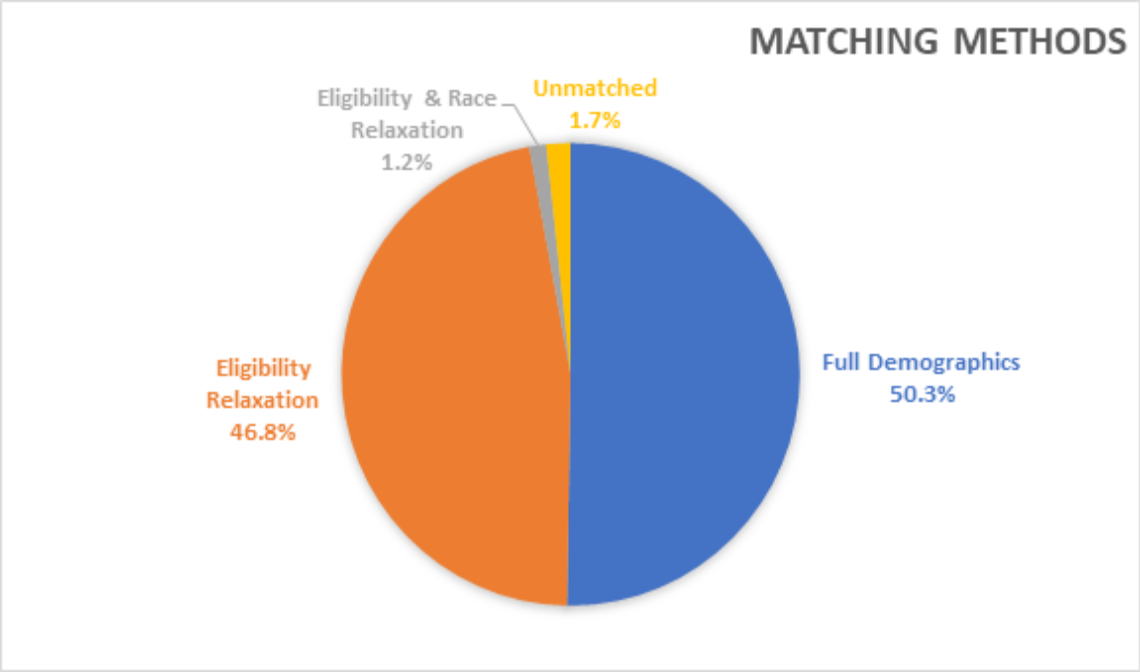


Figure 5: Percentage of Matched Adolescent Mother-Infant Pairs by Matching Criteria

3.2.4 Outcome Measures

We analyze seven outcomes for infants in the study and control population cohorts. The outcomes are derived from the IP and Other Therapy (OT), i.e. outpatient, claims of the infants, and are limited to the first year after birth. The categorical outcome measures include: substance exposure, health risk status (i.e. CRG), foster care entrance, infant mortality, and preterm birth/low birth weight (LBW). The count outcome measures include: emergency department (ED) visits, and wellness visits. A description of all the codes used for identifying the outcomes can be found in Table 5 .

3.2.4.1 Categorical Outcome Measures

The categorical outcome measures and the corresponding ICD-9 codes are discussed in this section.

Substance Exposure Outcome: Neonatal substance exposure is associated with a higher risk for negative health problems such as congenital anomalies, behavioral problems, and neurological development issues [56]. We determined an infant had in utero exposure to an addictive substance if one or more of the ICD-9 codes for this outcome was listed in the newborn delivery claim. The ICD-9 codes that are used for substance exposure are: '291', '292.2', '303.9', '648.4', '655.5', and '760.7' .

Health Risk Level Outcome: Health risk is based on the CRG metric, a proxy for the health of the infant within the first year of life. The health risk status of an infant is derived using the same methods as used for the mothers and described in 3.2.3.1.

***Foster Care Outcome*:** An infant being placed in foster care can be an indicator of the mother's health and/or parental resources to care for the infant. We identified an infant as being in foster care if the ICD-9 code 'V6081' is listed in any claims in the IP or OT tables for the first year or if the Medicaid eligibility code '48' was in the PS file.

***Infant Mortality Outcome*:** Infant mortality is an indicator of infant health. Causes for infant mortality include birth defects, low birth weight, sudden infant death syndrome, maternal pregnancy complications, and injuries such as suffocation [57]. Infants who died after delivery within the first year of life were identified using the PS date of death field. Still births were excluded from this study.

***Low Birthweight Outcome (LBW)*:** Low birthweight and preterm birth serve as indicators of adverse birth outcomes and can directly relate to an infant's health. LBW is often a result of preterm birth or restricted fetal growth. LBW greatly increases the rate of infant mortality and associated development problems. It also may factor into chronic diseases later in the infant's life [42, 43] LBW in an infant was identified if the birth claim had an ICD-9 code that began with '765'.

3.2.4.2 Count Outcome Measures

The count outcomes measures and the corresponding ICD-9 codes are discussed in this section.

***Number of Emergency Department Visits*:** While some may be unavoidable, the total number of emergency department (ED) visits can serve as an indicator of poor health, accidents, neglect, and/or abuse [58]. We exclude hospitalizations from this outcome since those are captured indirectly in the CRG metric described above. We identified ED visits using the type of service and place of service reported in the claims. We considered a claim to be an ED visit only if the ED claim was not associated with an inpatient stay (e.g. claim on the same day in the IP file) or hospitalization (determined by place of service).

***Number of Wellness Visits*:** Regular wellness visits determine if the infant is receiving adequate health monitoring and preventive care. This is because wellness visits include essential monitoring done by the pediatrician such as checking the growth, development, and vision of the infant. The number of wellness visits also serves as a proxy for the infants receiving vaccinations according to the recommended CDC schedule [59]. We identified wellness visits using the ICD-9 code 'V202' in any IP or OT claim.

Table 5: Description of Diagnosis and Procedure Codes for Identification of Outcomes in MAX Data

	Code*	Description
Substance Exposure	291	Alcohol induced disorders
	292.2	Pathological drug intoxication
	303.9	Other and unspecified alcohol dependence
	648.4	Mother mental disorders
	655.5	Suspected damage to fetus from drugs
	760.7	Drug affecting fetus or newborn via placenta or breast milk
Foster Care	V608.1	Foster care (status)
	48	Medicaid eligibility code for foster care child in any MAX file
LBW	765	Disorders relating to extreme immaturity of infant
Wellness Visits	V202	Routine infant or child health check
ED Visits	11	Code for outpatient hospital visit for type of service
	23	Code for emergency room for place of service
	21	Code for hospitalization claim for place of service
*ICD-9 code unless otherwise specified. Code represents the beginning of all the codes that were queried. Source: www.icd9data.com & MAX 2011 OT Dictionary		

3.2.5 Statistical Analysis

For our study, we were interested in the differences between the adolescent-mother infants (study population) and the adult-mother infants (control population). To assess if the differences between outcomes were statistically significant, we used a two-sided proportion test for categorical outcome measures and a two-sided Poisson test for count outcome measures. We performed both tests at the 95% confidence level. Since the counts of ED visits and wellness visits are dependent on the length of the infant's enrollment in Medicaid, we normalized the count data by enrollment months. Enrollment months were pulled from the PS file for each infant. The following normalization was done for each infant's visit count.

We define:

$T_{i,j}$: $number\ of\ visits\ of\ adolescent - mother\ infant\ i\ for\ count\ outcome\ j$

E_i: $number\ of\ enrollment\ months\ for\ adolescent - mother\ infant\ \ i$

$A_{i,j}$: $number\ of\ visits\ of\ adult - mother\ infant\ i\ for\ count\ outcome\ j$

F_i: $number\ of\ enrollment\ months\ for\ adult - mother\ infant\ i$

$Scale_i$: $average\ number\ of\ enrollment\ months\ for\ the\ home\ state\ of\ infant\ i$

Therefore, the normalized count for the adolescent and adult infants respectively are:

$$\tilde{T}_{i,j} = \left(\frac{T_{i,j}}{E_i}\right) * Scale_i$$

$$\tilde{A}_{i,j} = \left(\frac{A_{i,j}}{F_i}\right) * Scale_i$$

The ratio of visits for adolescent mother infants and adult mother infants were assessed using a two-sided exact Poisson test.

$$H_0: \frac{\tilde{T}_{i,j}}{\tilde{A}_{i,j}} = 1$$

$$H_1: \frac{\tilde{T}_{i,j}}{\tilde{A}_{i,j}} \neq 1$$

Statistical analysis is also performed to compare the outcomes by urbanicity and race/ethnicity. For the stratification results, the Bonferroni correction was used to adjust the p-values based on the number of categories.

To determine potential clinical relevance, we assessed the odds ratio for the categorical outcome measures and the rate ratio for the count outcome measures. We used a threshold of 1.2 to determine the significance of the effect.

3.3 Results

3.3.1 Study and Control Populations

For the study population, we identified 134,784 adolescent mothers of which 70,942 (52.6%) were paired with infants. Of the adolescent mother-infant pairs, we matched 68,562 with adult mother-infant pairs. Table 6 shows the counts of paired infants included in the study, the percentage of the paired mothers in each ethnicity/race group and each urbanicity group by state.

To identify potential pairing bias, we also observed the frequency chart distribution of paired and unpaired adolescent mothers by urbanicity, ethnicity and Medicaid eligibility. The distribution charts for the paired groups are in

Figure **6** and Figure 7.

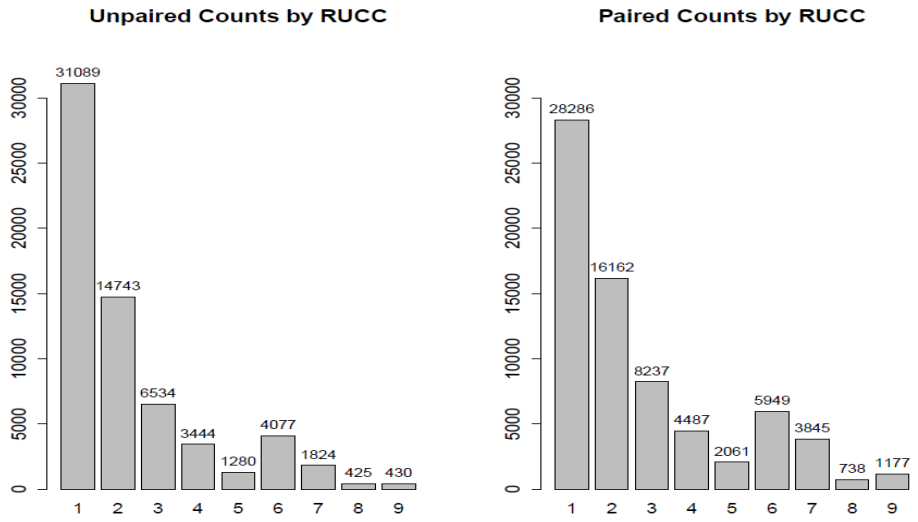


Figure 6: Paired Adolescent Mothers by Urbanicity

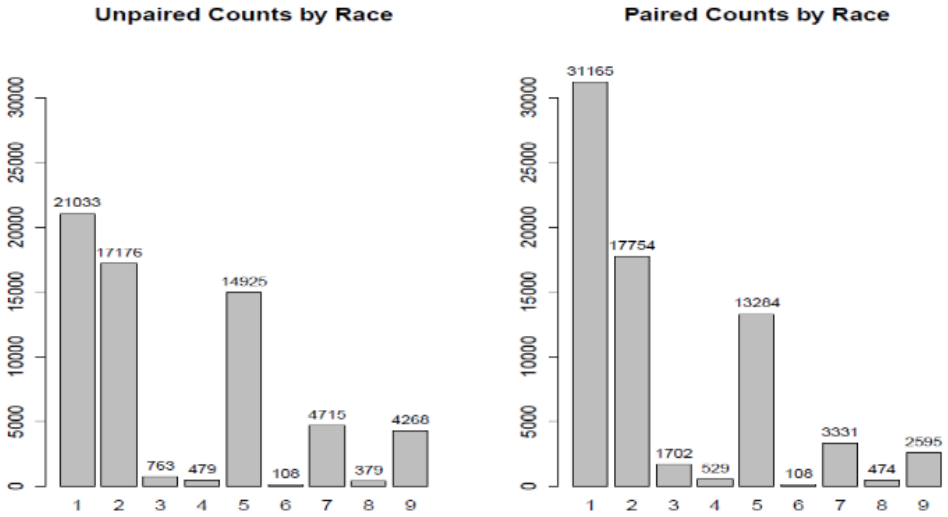


Figure 7: Paired Adolescent Mothers by Race

To identify potential matching bias, we observed the frequency chart distribution of (1) adolescent-adult mother matches based on all five factors (i.e. full match) versus (2) adolescent-adult mother matches with less restrictive criteria by urbanicity, ethnicity, and

health status (partial match). The distribution charts for the matching are in Figure 8 Figure 9.

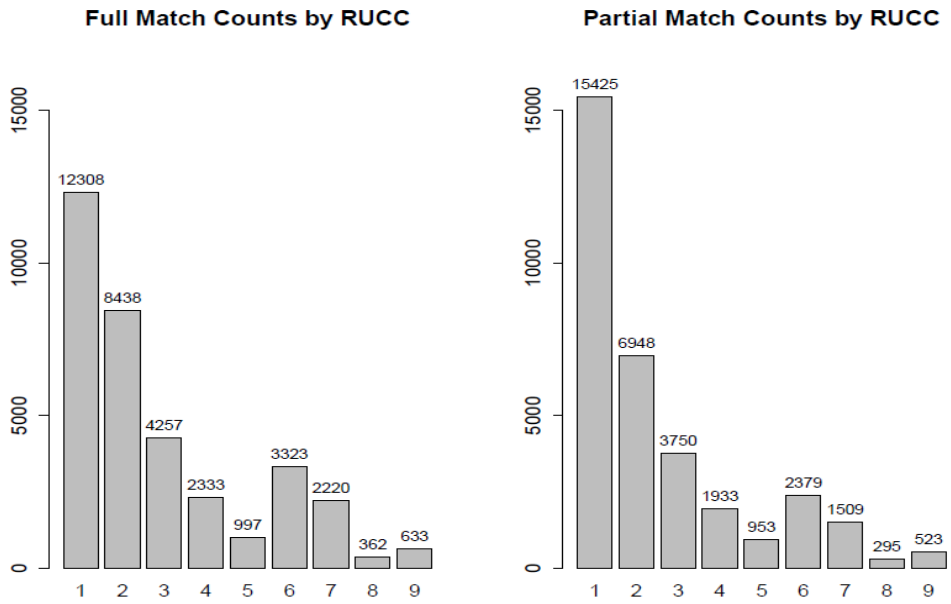


Figure 8: Matched Adolescent Mothers by Urbanicity

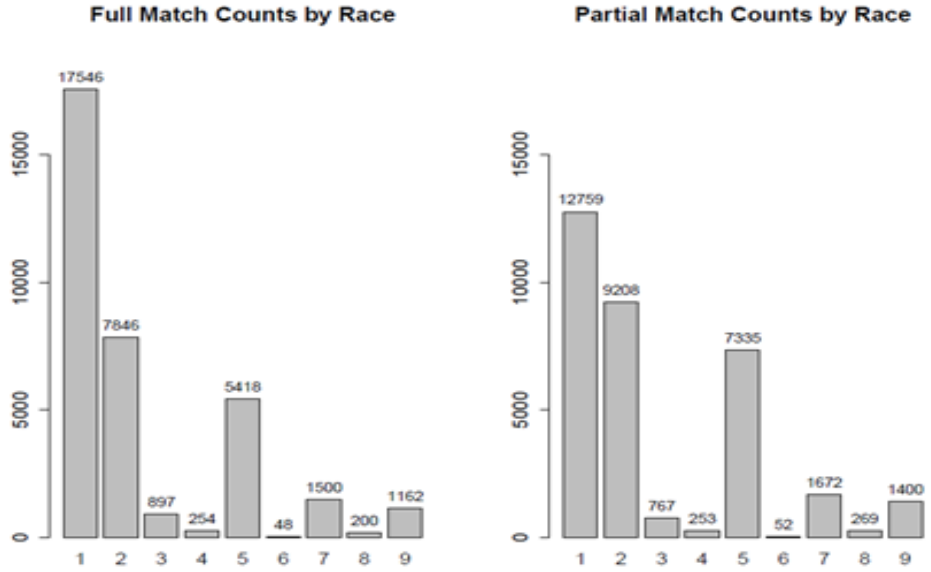


Figure 9: Matched Adolescent Mothers by Race

3.3.2 Overall Outcomes

Figure 10 shows the counts for each of the outcomes. Table 9 provides counts, proportions, and p-values for the seven outcome measures. Overall, the percentage of infants born with substance exposure was 0.7% and 0.8% for adolescent mother infants and adult mother infants, respectively. Adolescent mother infants had statistically significant higher rates of low risk health status (90.3% vs 90% with p-value: <0.04), LBW (8.4% vs 6.0% with a p-value: <0.005), and ED visits (ratio of 1.311 and p-value: <0.005). The average number of wellness visits per infant was 4.4, and the average number of ED visits per an infant was 1.3. There was no statistically significant difference for substance exposure, entry into foster care, high risk health status and wellness visits. Table 6 presents the outcomes by state.

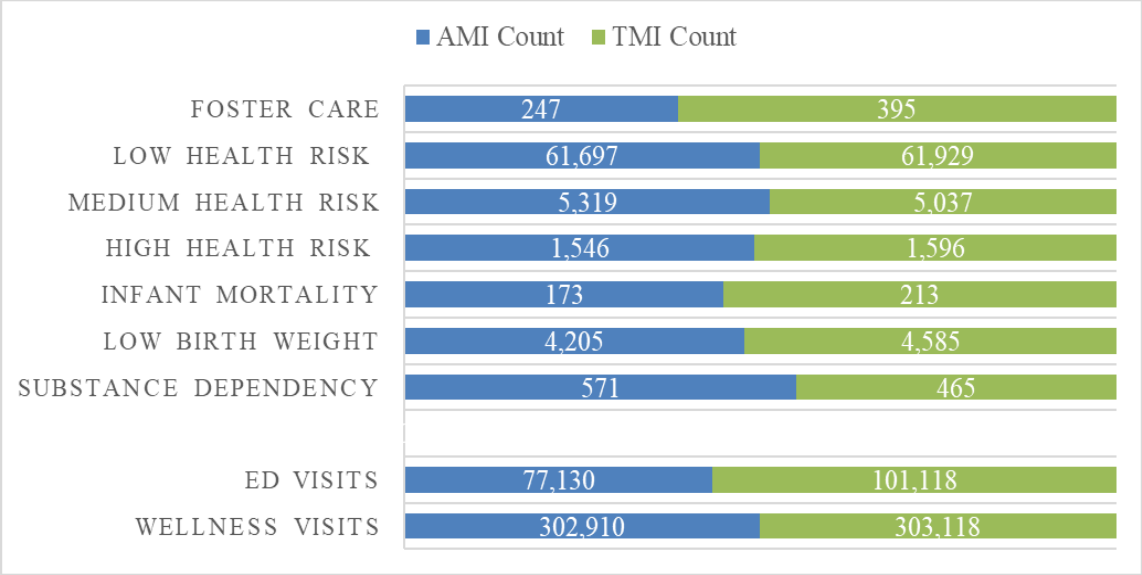


Figure 10: Total Results for Outcome Measures for Infants Born in 2011

3.3.3 Outcome Measures by Urbanicity

Figure 11 and Figure 12 show highlights of the outcomes by urbanicity. Table 10 presents the results of the seven outcome measures by urbanicity and includes proportions and p-values. The statistical results by urbanicity showed that substance exposure and foster care are significantly different in urban areas. Urban adolescent mother infants had a lower rate of substance exposure (0.7% vs 0.9%) and a higher rate of foster care (0.6% and 0.3%) compared to urban adult mothers' infants. The odds ratio suggested potential clinical relevance with values of 1.22 and 1.82 for substance exposure and foster care, respectively. In addition, urban adolescent mother infants had a statistically higher rate than urban adult mother infants for low birth weight (6.9% vs 6.2% with p-value= <0.005). As before, adolescent mother infants had statistically higher level of ED visits. ED visits stayed significant at each urbanicity level (p-value: <0.005) with suburban mothers having the highest discrepancy with adolescent mothers having 38% more ED visits than their adult counterparts. There was no statistically significant difference at any urbanicity level for all three health status levels or for infant mortality. Similar to the aggregate level, wellness visits showed no statistically significant differences.

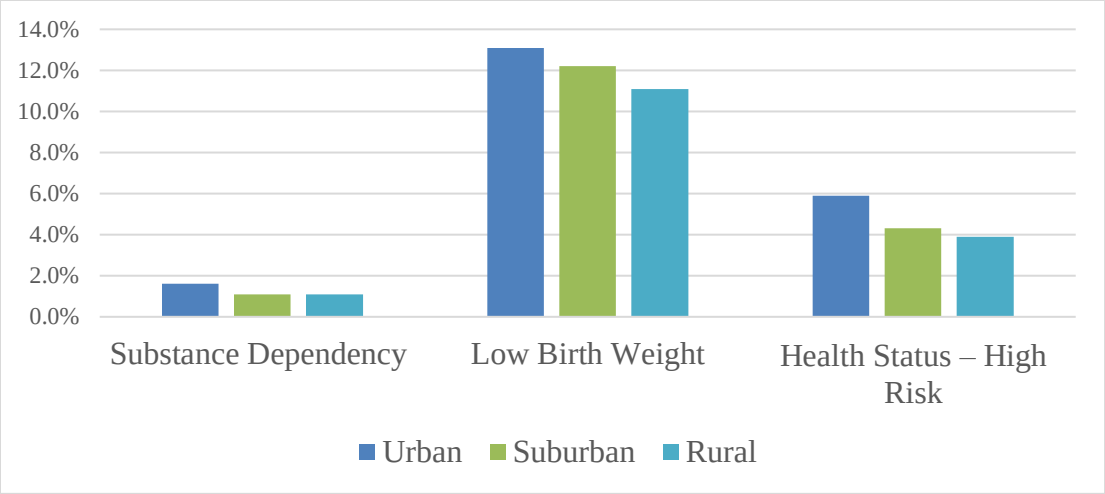


Figure 11: Selected Percentages of Total of the Health and Wellness Outcomes via Urbanicity

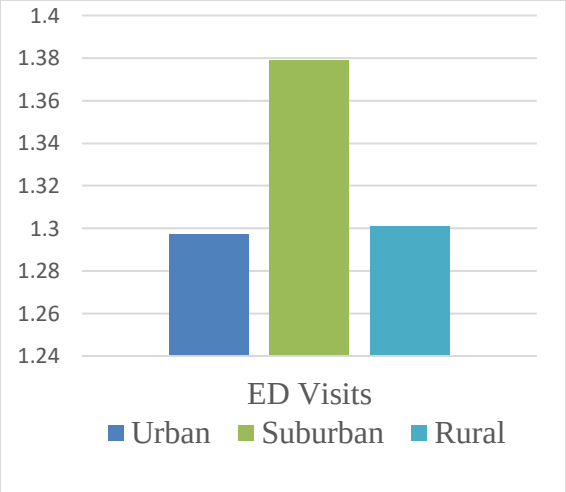


Figure 12: Ratio of Emergency Department Visits via Urbanicity

Outcome Measures by Race/Ethnicity

Table 11 presents the results of the seven outcome measures by race/ethnicity and includes proportions and p-values. Unlike at the aggregated level, substance exposure and foster care were statistically significant for white infants. Infants of white adolescent mothers had a lower rate of substance exposure (0.6% and 0.9%) and a higher rate for foster care (0.7% and 0.4%) than white adult mother infants. The odds ratio of 1·37 and 1·54 for substance exposure and foster care, respectively, demonstrated potential clinical relevance. In addition, black adolescent mother infants had a statistically significant higher rate for foster care (0·8% vs 0·4% with p-value: <0·005) with an odds ratio of 2·06. Similarly to the aggregate level, adolescent mother infants had a statistically higher level of ED visits at each race/ethnicity (p-value: <0·005). Infants of white mothers had the highest ratio for ED visits (1·37). Infants of white adolescent mother had more wellness visits (ratio 1·02 with p-value: <0·005) than infants of white adult mothers while Hispanic (ratio is 0·977 with p-value: <0·005) and other race adolescent mother infants (ratio is 0·985 with p-value: 0·006) had fewer wellness visits than infants of mothers with their respective race. The total percentage of infants (study and control population combined) with substance exposure with mothers who were white, black, or Hispanic, was 0·77%, 1·1%, and 0·45%, respectively. The rate of entry into foster care was 0·42%, 0·26% , and 0·12% for white, black, Hispanic infants respectively.

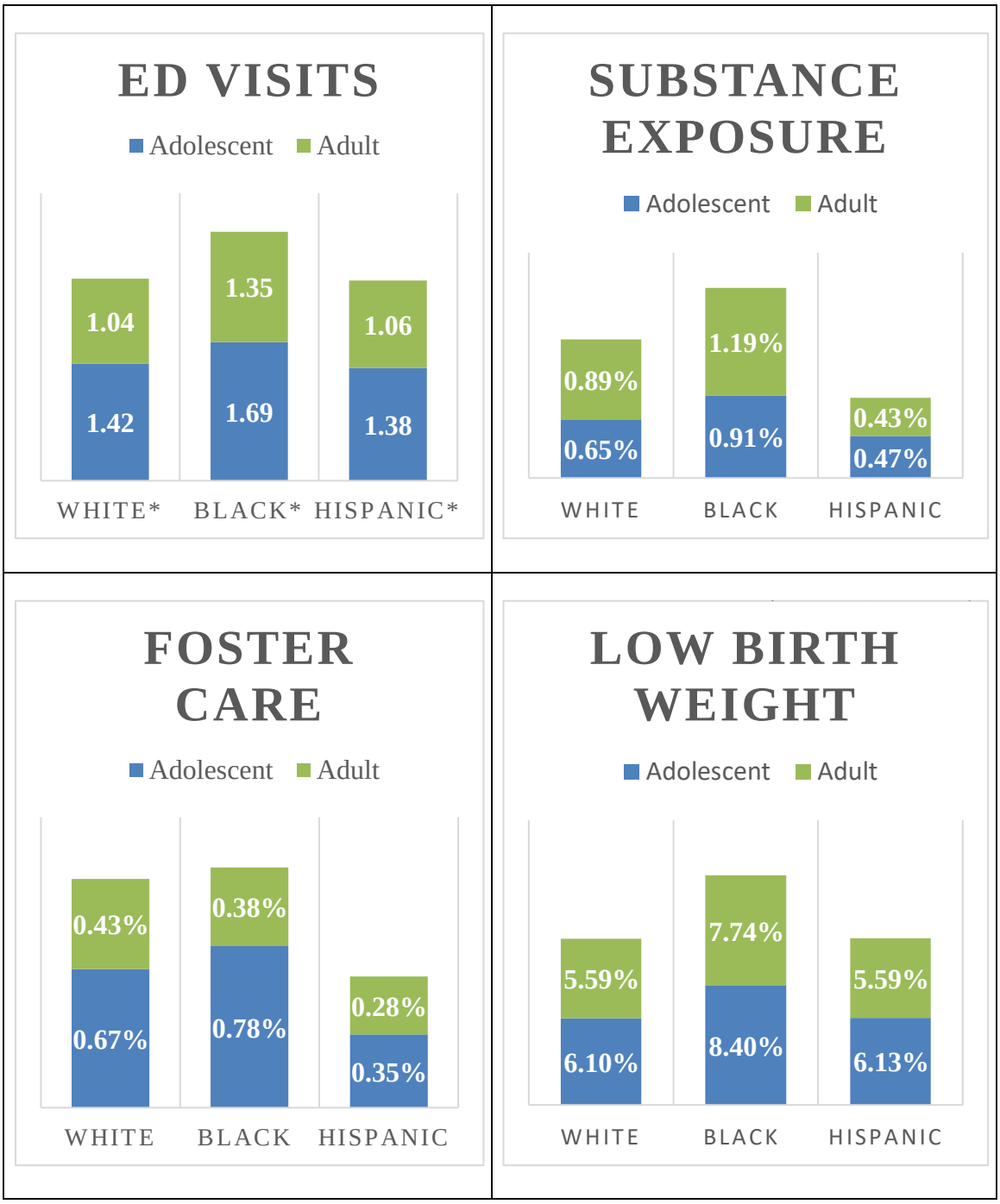


Figure 13: Highlights of Health and Wellness Outcomes via Race/Ethnicity

3.4 Discussion

Our study provides new insights into healthcare utilization in several ways. To our knowledge, we found no study that has addressed all seven outcomes in one study (substance exposure, foster care, low birth weight, infant mortality, emergency department visits, wellness visits, and health risk status) nor one that tracks infants throughout their first year of life. In addition, there is little research on the number of wellness visits infants from low-income families are receiving. The Medicaid population is of particular interest because it is a population vulnerable to health disparities [60], given that the beneficiaries are often low-income, and may be less likely to receive prenatal care than others [61].

While the rates of LBW and mortality in this study are compatible with existing research [57, 62], the difference between Medicaid-enrolled adolescent and adult mothers is not as extreme as previously assumed. In a national study Roth and colleagues conclude LBW rates of 9.4% for infants born to adolescent mothers compared to 7.0% for infants born to adult mothers [29]. While the difference is statistically significant in our study as well, rates of 6.7% and 6.1% respectively paint a different picture. Similarly, the gap between the difference in mortality rate in the study is small compared to previous literature. [63]. One potential explanation to this difference is the nature of Medicaid enrollment itself; young adults have limited access to employer-based insurance and are limited financially. According to the Kaiser Family Foundation, approximately 48% of individual aged 0-18 are enrolled in Medicaid compared to 18% of adults aged 18-44 [64].

Our study found new insights into the wellness of these infants; most notable, the number of wellness visits. This outcome was chosen to assess if the infant was receiving

the recommended preventive care (i.e., health and developmental check-ups). Importantly, the average wellness visits per infant (both those born to adolescent and adult mothers) was 4·4 visits, well below the seven recommended number of visits by the American Academy of Pediatrics [65]. Wellness visits are important for vaccinations as well as ensuring the infant's physical and cerebral development are normal. This may present an opportunity for intervention: Medicaid programs could create policies and practices to support wellness visits for all infants of Medicaid-enrolled mothers. For example, in-home visits or education initiatives to the clinicians can assist in improving infant wellness care. Assigning case workers, providing performance feedback, phone call reminders and providing clinicians information on best practices have been proven to improve the rate of wellness visits [66].

There were the greatest differences between infants of adolescent mothers and those of adult mothers in the number of emergency department (ED) visits, in aggregate, across all levels of urbanicity, and across all race/ethnicity groups. While this could reflect a greater frequency of illness, accidents, neglect, or abuse, it may also point to insufficient access to primary care or lack of experience that adolescent mothers may have to recognize a true emergency [67, 68]. A higher rate of ED visits for infants of adolescent mothers is consistent with some results in the existing literature; it has been shown that low maternal age, non-white race, enrollment in Medicaid [69] and low health literacy [70] are associated with an increase in the ED visit frequency. Thus, this is an important area of intervention in which adolescent mothers can learn about signs of fatal or emergency health conditions of their infants [70]. Further supporting optimal wellness visit frequency, as above, may help to reduce unnecessary ED visits.

The study found that urban adolescents were driving the increase in LBW. There are many factors that have been previously associated with LBW [71], including age less than 18, stress and environmental exposures. One proposed solution is group-based prenatal care, which has been associated with decreased rates of LBW for adolescents in New York City [72]. It has also shown promise of improving other health behaviors during and after delivery that may have positive outcomes for infants and their adolescent mothers [73].

There are several limitations to our study. There are many aspects of adolescent motherhood that our data cannot address, such as breastfeeding. We also acknowledge that using Medicaid claims data has limitations, since there can be missing or incomplete data. For example, if a person has a claim but no eligibility information, mortality cannot be measured. Also, previous research has found that MAX data has only been able to capture 73% of foster care cases [74]. We paired 55% of the Medicaid-enrolled adolescent mothers with infants; leaving 45% not included in the study. The bias analysis showed some minor bias within the study population from both the pairing and the matching steps. We found the white, suburban and rural populations (study and control) were more likely to be paired and matched based on the full criteria. However, we found that adolescent mothers whose Medicaid eligibility is low income are more likely to pair (i.e. find the mother's corresponding infant). Urban infant pairs were more likely to have a partial match. Hence, the populations included in the study may have higher representation than that of the full Medicaid population.

Our study highlights the importance of analyzing outcome measures across different stratifications of the study population. This analysis may help shed light on

differences by race and urbanicity, as well as, inform possible policies and programs that could impact them. The findings of this study point to the implementation of targeted interventions to support Medicaid-insured infants of adolescent mothers in achieving optimal and equitable health. For example, group-based prenatal care has been shown to provide a lower-cost option for prenatal risk screenings and patient education with high levels of patient satisfaction. In addition, community outreach or programming, particularly in urban areas with for adolescent mothers may also help to reduce stress and provide support for young mothers. This model could be expanded to postpartum care and education with joint mother-infant visits to help reduce unnecessary ED visits, increase utilization of wellness visits and reduce cost.

Further, since multiple outcomes are interrelated such as low birth weight and mortality, decreasing one negative outcome can have a compounding effect. Finally, additional research efforts to uncover the factors or determinants that underlie these differences is needed. To have the greatest potential for impact, this should be done in partnership with people from the affected populations.

Table 6: Total Counts, Ethnicity, Urbanicity and Average Enrollment Months of Study and Control Populations by State

State	Total_Paired Infants	White	Black	Hispanic	Other	Urban	Suburban	Rural	Avg Enrollment Months Teen Baby	Avg Enrollment Months Adult Baby
Alaska	370	20.27%	3.78%	2.97%	72.97%	47.80%	2.97%	49.19%	11.31	11.64
Arkansas	1375	60.44%	24.80%	7.05%	7.71%	49.31%	28.65%	22.04%	10.95	11.02
Connecticut	478	31.27%	21.86%	43.53%	2.30%	94.18%	3.48%	2.30%	11.37	11.46
Delaware	218	21.78%	31.80%	40.94%	5.05%	89.91%	5.05%	5.05%	11.41	11.72
Florida	565	26.02%	40.35%	17.88%	15.75%	93.19%	4.86%	1.95%	11.56	11.6
Georgia	4120	40.63%	45.61%	1.12%	12.65%	75.95%	17.40%	6.65%	10.22	9.63
Illinois	2406	26.48%	45.05%	23.61%	4.86%	56.10%	24.68%	19.23%	11.51	11.58
Indiana	3217	63.69%	20.70%	13.89%	1.71%	90.23%	7.65%	2.12%	11.47	11.67
Iowa	697	16.07%	3.44%	4.59%	75.90%	75.88%	21.57%	2.55%	11.67	11.71
Kansas	1060	47.36%	14.53%	31.98%	6.13%	61.79%	26.98%	11.23%	5.91	6.36
Kentucky	2129	86.52%	8.41%	3.48%	1.60%	35.56%	23.06%	41.38%	11.59	11.51
Louisiana	1402	35.02%	60.41%	1.57%	3.00%	86.59%	9.63%	3.78%	9.99	10.02
Maine	174	57.56%	6.32%	6.32%	29.56%	87.36%	6.32%	6.32%	10.52	11.08
Maryland	2477	23.29%	49.13%	15.54%	12.03%	96.38%	3.18%	0.44%	11.75	11.76
Massachusetts	739	25.03%	8.80%	11.37%	54.80%	65.12%	23.84%	11.05%	11.64	11.7
Michigan	4430	32.21%	34.81%	5.85%	27.13%	82.44%	11.20%	6.37%	11.39	11.5
Minnesota	915	42.62%	17.70%	13.33%	26.34%	64.26%	17.81%	17.98%	11.45	11.37
Missouri	2353	62.77%	25.88%	4.16%	7.18%	63.20%	20.82%	15.98%	11.62	11.57
Montana	227	58.43%	4.85%	4.85%	31.68%	28.63%	33.04%	38.33%	11.05	11.38
Nebraska	684	48.25%	9.06%	25.29%	17.40%	71.18%	25.95%	2.87%	11.77	11.63
Nevada	126	56.75%	8.73%	25.80%	8.73%	44.15%	11.17%	44.68%	11.23	11.33
New Hampshire	222	79.91%	4.95%	9.88%	4.95%	57.42%	21.00%	21.59%	11.39	11.71
New Mexico	1538	13.85%	1.30%	68.40%	16.45%	57.47%	33.03%	9.50%	11.36	11.64
New York	1208	45.61%	15.81%	28.73%	9.85%	59.56%	33.62%	6.83%	10.71	11.07
North Carolina	3102	35.72%	45.55%	10.83%	7.90%	18.40%	67.20%	14.40%	10.24	10.53
North Dakota	188	46.08%	5.85%	5.85%	41.94%	78.31%	20.45%	1.24%	11.77	11.7
Ohio	3435	61.31%	33.89%	4.37%	0.44%	76.01%	21.89%	2.10%	11.57	11.64
Oklahoma	2983	46.50%	11.00%	24.44%	18.07%	58.08%	28.28%	13.64%	11.67	11.74
Oregon	1535	45.21%	2.61%	35.70%	16.48%	74.92%	27.35%	2.74%	11.36	11.49
Pennsylvania	1459	67.10%	10.83%	17.55%	4.52%	66.28%	27.48%	6.24%	11.55	11.61
Rhode Island	88	12.5%	12.5%	12.5%	62.50%	75.00%	12.50%	12.50%	9.38	9.27
South Carolina	1367	41.92%	40.60%	12.14%	30.82%	77.54%	21.95%	0.51%	11.67	11.68
South Dakota	439	41.46%	2.51%	4.33%	52.16%	36.07%	19.18%	44.75%	11.41	11.49
Tennessee	1623	62.66%	32.10%	4.25%	0.99%	70.19%	23.46%	6.36%	11.75	11.77
Texas	12806	10.88%	8.68%	42.79%	37.65%	82.96%	12.57%	4.47%	11.56	11.65
Utah	1185	83.02%	0.93%	13.54%	2.34%	86.15%	4.48%	9.38%	11.18	11.36
Vermont	53	53.69%	NA	NA	44.43%	77.54%	9.83%	12.63%	11.59	11.55
Virginia	1496	27.74%	20.52%	19.92%	31.82%	15.04%	41.51%	43.40%	11.53	11.00
Washington	1188	26.18%	2.78%	14.56%	56.48%	87.79%	11.70%	51.00%	11.68	11.62
West Virginia	188	82.45%	5.85%	5.85%	5.85%	77.86%	16.57%	5.57%	11.56	11.55
Wisconsin	1992	29.42%	20.23%	11.70%	38.65%	66.85%	24.46%	8.70%	11.37	11.08
Wyoming	305	71.35%	3.61%	14.27%	10.47%	30.82%	22.62%	46.56%	11.25	11.27

*Percentages are adjusted to account for counts less than 11

Table 7: Critical Risk Grouping Classifications

Critical Risk Grouping (CRG) Classifications*
CRG1: Healthy
CRG2: Recent History of Significant Acute Disease
CRG3: Single Minor Chronic Disease
CRG4: Minor Chronic Disease in Multiple Organ Systems
CRG5a: Single Moderate Chronic Disease
CRG5b: Single Dominant Chronic Disease
CRG6: Significant Chronic Disease in Multiple Organ Systems
CRG7: Dominant Chronic Disease in Three or More Organ Systems
CRG8: Dominant, Metastatic and Complicated Malignancies
CRG9: Catastrophic Conditions
*Patients were aggregated into three groups: non-chronic (CRG1 and CRG2), minor (CRG3-CRG5a), and severe (CRG5b-CRG9)

Table 8: RUCC Classifications

RUCC Code	Description
Metropolitan Counties*	
1	Counties in metro areas of 1 million population or more
2	Counties in metro areas of 250,000 to 1 million population
3	Counties in metro areas of fewer than 250,000 population
Nonmetropolitan Counties	
4	Urban population of 20,000 or more, adjacent to a metro area
5	Urban population of 20,000 or more, not adjacent to a metro area
6	Urban population of 2,500 to 19,999, adjacent to a metro area
7	Urban population of 2,500 to 19,999, not adjacent to a metro area
8	Completely rural or less than 2,500 urban population, adjacent to a metro area
9	Completely rural or less than 2,500 urban population, not adjacent to a metro area

Source: https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx

Table 9: Total Results for Outcome Measures for Infants Born in 2011 for the Study Population (Adolescent-Mother Infants) and the Control Population (Adult-Mother Infants). Includes Total Counts, Point Estimates, and the P-value for the Statistical Test

Outcomes for All Adolescent Mother Infants (TMI) and Adult Mother Infants (AMI)					
			Proportion Test		
Categorical Outcome	**TMI Count**	**AMI Count**	**TMI Estimate**	**AMI Estimate**	**p-value**
Substance Exposure	465	571	0.007	0.008	0.57
Health Status – Low Risk*	61,929	61,697	0.903	0.900	0.04
Health Status – Medium Risk*	5,037	5,319	0.073	0.078	<0.005
Health Status – High Risk	1,596	1,546	0.023	0.023	0.37
Foster care	395	247	0.006	0.004	0.95
Infant Mortality	213	173	0.003	0.003	0.05
Low Birth Weight*	4,585	4,205	0.067	0.061	<0.005
			Poisson Test		
Count Outcome**	**TMI Count**	**AMI Count**	**Rate Ratio**		**p-value**
ED Visits	101,118	77,130	1.311		<0.005
Wellness Visits	303,118	302,910	1.001		0.79

*Outcome significant at the 95% confidence interval based on the 68,562 count of mother-infant pairs

**Count outcomes are normalized by enrollment months

Health status is based on one-year healthcare utilization (i.e. Critical Risk Grouping)

Table 10: Total Results for Outcome Measures for Infants Born for 2011 for the Study Population(Adolescent-Mother Infants) and the Control Population (Adult-Mother Infants). Point Estimates and the P-Value for the Statistical Test by Urbanicity

Outcomes by Urbanicity for Adolescent Mother Infants (TMI) and Adult Mother Infants (AMI)			
Categorical Outcome	**TMI Estimate**	**AMI Estimate**	**p-value**
Substance Exposure			
Urban*	0.007	0.009	0.005
Suburban	0.005	0.006	0.722
Rural	0.004	0.007	0.023
Health Status Low Risk			
Urban	0.719	0.718	0.889
Suburban	0.729	0.729	0.942
Rural	0.735	0.747	0.180
Health Status Medium Risk			
Urban	0.252	0.252	0.960
Suburban	0.249	0.249	1.000
Rural	0.243	0.234	0.317
Health Status High Risk			
Urban	0.029	0.030	0.631
Suburban	0.022	0.021	0.718
Rural	0.021	0.018	0.376
Foster care			
Urban*	0.006	0.003	<0.005
Suburban	0.005	0.005	0.525
Rural	0.004	0.004	0.885
Infant Mortality			
Urban	0.003	0.002	0.128
Suburban	0.003	0.003	1.000
Rural	0.004	0.002	0.042
Low Birth Weight			
Urban*	0.069	0.062	<0.005
Suburban	0.061	0.061	0.914
Rural	0.058	0.053	0.263
	Poisson Test		
Count Outcome**	**Rate Ratio**		**p-value**
ED Visits			
Urban*	1.297		<0.005
Suburban*	1.379		<0.005
Rural*	1.301		<0.005
Wellness Visits			
Urban	1.000		0.915
Suburban	1.003		0.642
Rural	0.999		0.948
*Outcome significant at the 95% confidence interval or Bonferroni adjusted p-value less than 0.017 **Counts for count outcomes are normalized by enrollment months Health status is based on one year of healthcare utilization (i.e. Critical Risk Grouping)			

Table 11: Total Results for Outcome Measures for Infants Born in 2011 for the Study Population (Adolescent-Mother Infants) and the Control Population (Adult-Mothers Infants) Point Estimates and the P-Value for the Statistical Test by Race/Ethnicity

Race/Ethnicity for Adolescent Mother Infants (TMI) and Adult Mother I			
Categorical Outcome	**TMI Estimate**	**AMI Estimate**	**p-value**
Substance Exposure			
White*	0.006	0.009	<0.005
Black*	0.009	0.012	0.018
Hispanic	0.005	0.004	0.689
Other	0.007	0.007	0.916
Health Status Low Risk			
White	0.723	0.727	0.304
Black	0.715	0.709	0.286
Hispanic	0.718	0.716	0.651
Other	0.731	0.736	0.388
Health Status Medium Risk			
White	0.255	0.251	0.296
Black	0.259	0.265	0.212
Hispanic	0.245	0.245	0.936
Other	0.238	0.236	0.604
Health Status High Risk			
White	0.022	0.022	0.775
Black	0.026	0.025	0.762
Hispanic	0.037	0.039	0.394
Other	0.030	0.028	0.379
Foster care			
White*	0.007	0.004	<0.005
Black*	0.008	0.004	<0.005
Hispanic	0.004	0.003	0.333
Other	0.004	0.003	0.213
Infant Mortality			
White	0.003	0.003	0.130
Black	0.003	0.002	0.148
Hispanic	0.002	0.002	0.783
Other	0.003	0.003	0.861
Low Birth Weight			
White*	0.061	0.056	0.014
Black*	0.084	0.077	0.033
Hispanic	0.061	0.056	0.066
Other*	0.064	0.058	0.044
	Poisson Test		
Count Outcome**	**Rate Ratio**		**p-value**
ED Visits			
White*	1.366		<0.005
Black*	1.255		<0.005
Hispanic*	1.295		<0.005
Other*	1.313		<0.005
Wellness Visits			
White*	1.024		<0.005
Black	0.998		0.747
Hispanic*	0.977		<0.005
Other*	0.985		0.006
Counts	**Ratio**		
White	1.017		
Black	0.981		
Hispanic	1.029		
Other	0.962		

*Outcome significant at the 95% confidence interval or Bonferroni adjusted p-value less than 0.012

**Counts for count outcomes are normalized by enrollment months Health status is based on one year healthcare utilization (i.e. Critical Risk Grouping)

CHAPTER 4. STATISTICAL LEARNING METHODS FOR DIFFERENCE-IN-DIFFERENCES DESIGN

4.1 Introduction

When trying to understand the impact of policies in health care, rarely are the policy implementations conducted in randomized settings. Difference-in-differences (DID) study designs are one option to leverage observational data to estimate causal effects in the absence of experimentation. The DID framework is typically structured as two time periods relative to the time of intervention (pre and post) and the intervention status of the subject (intervention or comparison). Therefore, DID considers longitudinal data, often with multiple time points before and after a policy intervention. The comparison group remains unexposed to the policy intervention in both pre- and post- time periods. In contrast, the intervention group is unexposed in the pre-time periods and exposed in the post-time periods. The goal is then to estimate the average change in outcome trends between the two groups. We use the average treatment effect among the treated (ATT) as the parameter to measure this change. The ATT measures the treatment effect for a randomly selected individual within the intervention group. It quantifies the difference in the outcome between an individual within the intervention group who was exposed to the policy and an individual within the intervention group who remained unexposed. Additionally, a set of causal assumptions specific to difference-in-differences study designs are required for an enriched causal interpretation of the target parameter.

Health care policy data has specific traits that are distinctive from the attributes of much of the data traditionally used in previous DID research. This is seen in the parametric approaches that are largely developed for and applied to scenarios with decades of longitudinal data. However, this would be an atypical setting in health care policy. The volume of observational health care data has drastically increased due to the expansion of electronic health records, clinical registries, and other sources. However, despite these data sources, most health care policy studies are limited to only a few years of data. Thus, many existing approaches may have differential performance in these settings. These databases are typically fraught with multiple complexities, including informative missingness, patients clustered within hospitals, and high-dimensional confounder sets. Consequently, there is a need for flexible and robust methodology to analyze various forms of observational health care data.

Parametric models are the most commonly used methods to estimate this average change in outcome trends between the intervention and comparison groups. One common method is a fixed-effects outcome regression. In this approach, the outcome is specified as a linear function based on the observations' baseline covariates, treatment assignment, and pre/post time period [75, 76]. The disadvantage of this technique is it assumes additive effects of the covariates, which may not be true. Parametric techniques also restrict the model space of the estimator, requiring that the functional form is known up to a finite number of parameters, which is unlikely in many practical applications within health care, as well as others.

Another popular approach is inverse propensity score weighting (IPW). These techniques rely on consistent estimation of the propensity score function; the conditional

probability of exposure to the intervention given baseline covariates. The propensity score is a balancing score, such that subjects with the same propensity score value have the same distribution of the covariates between both intervention groups. For example, Abrevaya used IPW as part of a conditional average treatment effect estimator [77]. Callaway and Sant'Anna's paper [78] showed that normalizing the propensity score weights when using IPW can lead to finite sample improvements. However, inverse propensity score methods are generally known to suffer from substantial bias due to misspecification of the propensity score function and can have large variance in empirical samples.

A few newer studies have developed double robust methods for difference-in-differences to reduce bias due to model misspecification. Zimmert introduced nonparametric technique for group average treatment effects. The nonparametric approach provides protection against misspecification and lifts the restriction of parametric models that the number of predictors need to be less than the number of observations. Also, by using group average treatment effects, the estimates are less sensitive to heterogeneity [79]. Han, Yu, & Friedberg used a double robust variant of the IPW approach by Lunceford and Davidian [80]. The technique incorporates the expected value of the change in the outcome, which is estimated using linear regression, into the estimator [81]. This double robust estimator shows improved performance when compared to the traditional IPW approach, original least squares, and the generalized linear model. Sant'Anna created a double-robust estimator that weights the regression errors instead of the observed outcomes; as is done in IPW [82]. The approach has the additional property that it accounts for heterogeneity, one potential problem that arises when using the fixed-effects regression for policy evaluation. This is an important feature since the fixed effect regression assumes

a constant treatment effect across all covariates which may not be true. The estimator's results had smaller standard errors and bias compared to a two-way fixed effects regression and IPW. Li and Li developed a double robust estimator that augments an IPW estimator with regression [83]. This double robust estimator has the advantage that it only depends on estimating the outcome of the comparison group instead of both intervention groups. However, it is also more sensitive to the misspecification of the outcome regression than the propensity score model in the case of estimating additive treatment effects, such as the average difference between the two intervention groups. Lu proposed two double robust nonparametric methods for DID estimation [84]. For homogenous treatment effects, a transformed regression, based on an orthogonal decomposition of the outcome, is used. For heterogenous treatment effects, a balancing estimation method using minimax to estimate linear weights is presented. The methods emphasize the importance of nonparametric approaches in order to have lower mean squared errors as shown in their performance when compared to IPW, original least squares, and a double robust augmented IPW [85].

Even with this previous work, there remains a need for flexible methodology to analyze various forms of observational health care data. For difference-in-differences studies in health care, bringing machine learning to handle the common design where we have few pre-intervention and post-intervention time periods in billing claims data is a gap in the literature we aim to address here. Thus, in this chapter, we extend the ensemble machine-learning method, super learner, to the DID study setting, developing an estimator and realistic simulation study.

The motivating study for our research assessed the impact of episode-based payments (EBP) on perinatal costs in Arkansas [3]. Under this payment policy, physicians are rewarded for keeping the aggregated costs of a designated health care event (i.e. an episode) below a pre-determined spending cap and penalized for spending above it. Arkansas costs were compared to that of nearby states that had not implemented the policy. These comparison states included: Alabama, Kentucky, Louisiana, and Oklahoma. The study found a 3.8% decrease in perinatal spending in Arkansas under the policy. It used a parametric regression analysis approach to estimate the ATT and incorporated maternal demographics and clinical characteristics as covariates. The study contributed to the understanding of financial incentives of physician behavior. Fee-for-service reimbursement has been linked to higher levels of healthcare utilization as compared to EBP [86] and conflicting incentives for the physician and the patient. Additional research has found that episode-based payments had reduced cost for outpatient hospital care within the Medicaid population [87, 88]. However, the Arkansas study found that it also reduced cost for a longer-term episode (such as prenatal through post-partum care) and under private insurance.

This chapter is organized as the following. In section 2, we introduce the model, target parameter of interest, and required assumptions. Next, we present the simulation study and the application of the super learner algorithm for the difference-in-differences setting. Lastly, we present concluding remarks and discuss future work including double robust approaches.

4.2 Study Framework

We design our estimator and simulation study based on the prior 2018 analysis on Arkansas perinatal costs during the years of 2009 to 2014 to access the impact of the episode-based payment policy. The Episode-based Payment (EBP) implementation was fully implemented in 2014 with 80% of births in the large market being covered under EBP. Blue Cross Blue Shield, the largest commercial insurer in Arkansas , QualChoice Arkansas, the third largest commercial insurer, and Baptist Health, the third largest employer in Arkansas , had all implemented EBP by this time. Perinatal care claims from Arkansas were considered exposed to the intervention, and claims from the four comparison states: Alabama, Kentucky, Louisiana, and Oklahoma, were considered unexposed. The total perinatal costs included the associated costs of insurance claims for prenatal care in the 40 weeks prior to the live birth, delivery, and postnatal care up to 60 days after the birth. Each episode's aggregated costs are assigned to the quarter of the episode's inception (i.e. live birth). In addition to costs, several relevant characteristics of the mother are available from the claims data and considered in the study. These included: maternal age (under 25, 25-29, 30-34, 35+), policy holder status (policy holder, spouse, dependent), clinical characteristics of the delivery (fetal malpresentation, fetal distress, multiple gestation, preterm birth, and previous cesarean) and insurance plan type (HMO, PPO, POS, HDHP).

4.2.1 Data Description

The data, O, consist of the outcome, Y; intervention status indicator, A; and eight baseline covariates $\mathbf{X}$ = (maternal age, maternal policy holder, maternal plan type, fetal malpresentation, multiple gestation, previous cesarean, preterm birth, and fetal distress) observed at intervention time-period indicator T. The value of $A = 1$ indicates the

observation is in the intervention group and the value $T = 1$ indicates the observation is in the post-intervention time period. We will use lowercase letters to signify specific values for random variables. Let *P(O)* be the data generating distribution of *O* and can be factorized based on the time-ordering of the data *P(O)=P(Y(t)|A,X(t))*P(A|X(t)) *P(X(t)).*

The outcome, *Y*, is the spending for an episode and a continuous variable that depends on the baseline covariates of the subjects, the intervention status, and the intervention time-period. The underlying data-generating mechanism for intervention status, *A*, is a function of the baseline covariates.

4.2.2 Parameter of Interest

We are interested in assessing the effect of a policy, and in our motivating study this is the impact of EBP policy on perinatal costs in Arkansas. That is, we are interested in the ATT or the perinatal costs in Arkansas under EBP compared to the perinatal costs in Arkansas without EBP.

We now define Y_t^a as the counterfactual, i.e. potential outcomes, in which $A=a$ and $T=t$. The four potential outcomes are: comparison group pre-intervention (Y_0^0), comparison group post-intervention (Y_1^0), intervention group pre-intervention (Y_0^1) and then intervention group post-intervention (Y_1^1). We write our target parameter, Ψ, as: $\Psi = E[Y_1^1 - Y_1^0|\text{A=1}]$. However, though $(Y_1^1|A = 1)$ is observed, the term $(Y_1^0|A = 1)$ is not. To estimate $(Y_1^0|A = 1)$, we need make some assumptions. The first assumption, parallel trends, states the difference in the outcomes of the comparison group pre- and post-intervention is a good proxy for the difference in the outcomes for the intervention group, if the group were not treated. In our example, it would be said that we expect the Arkansas

perinatal costs to increase at the same rate as the comparison states given that EBP was not implemented. We demonstrate this assumption mathematically as:

$$\mathrm{E}[Y_1^0 - Y_0^0 | A = 1] = \mathrm{E}[Y_1^0 - Y_0^0 | A = 0]$$

$$=> \quad E[Y_1^0 | A = 1] = E[Y_0^0 | A = 1] + \mathrm{E}[Y_1^0 - Y_0^0 | A = 0] \tag{1}$$

This assumption is not testable because the left-hand side of the equation includes unobserved counterfactuals. However, we use the right-hand side of the equation as the first step to identify the target parameter. The second assumption, stable unit treatment value (SUTVA), assumes no interference (i.e. no spillover effects) and consistency. Under the no inference assumption, the treatment of one unit does not "spillover" and impact the outcome of another. For example, we assume that Arkansas' use of EBP does not impact how clinicians in the control states deliver care. The consistency assumption states we can only observe one of the two potential post-intervention outcome (i.e. the outcome of the actual treatment). More simply, a live birth could have only occurred in Arkansas or in one of the comparison states. Therefore, we are always missing data for one of the two potential post-intervention outcomes. However, since the intervention does not impact outcomes retroactively, the two potential pre-intervention outcome are equal to the observed value for that group. Applying this to our example, the actual perinatal cost for an episode in Arkansas represents both the actual outcome for the episode that occurred in Arkansas as well as the counterfactual outcome for an episode that occurred in a comparison state given it was in Arkansas. Precisely,

$$(Y_0^0|A=0) = (Y_0^0|A=1) = Y(0,1) \quad (2)$$

With these assumptions in place, we can now apply them to the target parameter. We begin with:

$\Psi = E[Y_1^1 - Y_1^0|A=1]$ (target parameter: ATT)

$= E[Y_1^1] - E[Y_1^0|A=1]$

$= E[Y_1^1] - E[Y_0^0|A=1] - \mathrm{E}[Y_1^0 - Y_0^0|A=0]$ (using equation (1) from the parallel trends assumption)

$= E[Y(1,1) - Y(1,0)] - \mathrm{E}[Y(0,1) - Y(0,0)]$ (using equation (2) from the SUTVA assumptions)

By making this set of key causal assumptions, we were able to identify the parameter (i.e. all terms are observed values). Even though the assumptions cannot be tested, when applied, they can also provide an enriched causal interpretation of the parameter. Hence, the difference-in-differences (DID) study setting, where the target parameter is ATT can be estimated using $\Psi = E[(Y_1^1 - Y_0^1) - (Y_1^0 - Y_0^0)]$.

4.3 Super Learner Estimator for DID

We follow the steps of G-computation to estimate the ATT, the sample data is presented in

Table **12**.

First, the outcome algorithm that will used to predict the outcome is selected. The selected algorithm is trained on the sample and then the trained algorithm is used to predict *Y* (column 4) for all observations in the sample. Next, the trained algorithm of interest is used to predict the four potential outcomes (columns 5-8): comparison group pre-intervention (Y_0^0), the comparison group post-intervention (Y_1^0), the intervention group pre-intervention (Y_0^1) and then intervention group post-intervention (Y_1^1). The potential outcomes are predicted for each observation by using the covariates for each observation and then choosing the corresponding values for *A* and *T*. Explicitly, for the prediction of Y_0^0 for observation *i*, *A = 0* and *T = 0* is used in conjunction with the observed values for the other covariates for observation *i*. The four values can then be substituted into: $(Y_1^1 - Y_0^1)$ - $(Y_1^0 - Y_0^0)$ for each observation in the sample (column 9). The value of the column labeled ATT is the estimated ATT for that sample. We repeat this process 500 times and take the mean of the ATT across the 500 samples.

Super learner is an ensemble machine learning algorithm used for prediction. Ensemble techniques incorporate several machine learning algorithms into the prediction. The advantage of these techniques is they out perform better using any single algorithm. Super learner implements a library of prediction algorithms and then creates a prediction based on the weighted average of the algorithms in the library. We incorporate a library of three algorithms: generalized linear model (glm), penalized regression with a lasso (glmnet), and random forests (randomForest). Glmnet fits a generalized linear model via penalized maximum likelihood. The penalty can be a weighted sum of L1 (lasso) and L2

(ridge) regularization which help with variable selection and shrinkage. Random Forests create an ensemble of decision trees, trained on different randomly selected samples. Each observation's final predicted value is the mode value (i.e. majority vote) across all the decision trees. The same steps for G-computation are used but super learner is being used as the prediction algorithm instead of regression. In this case, we input all six variables into the algorithm: *age, fetal.distress, fetal.malpresentation, multiple.gestation, preterm.birth,* and *previous.cesarean.*

Table 12: Sample of Simulated Data with Predicted Outcomes Based on Correctly Specified Regression Model. Data has True Effect Size = -250

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ID	A	T	Y	Y_0^1	Y_0^0	Y_1^1	Y_1^0	ATT
36044	0	0	5,719.69	5,013.34	6,498.98	6,757.91	8,485.09	(241.54)
802089	1	1	10,575.68	10,414.67	11,900.32	12,159.24	13,886.42	(241.54)
77387	0	1	21,554.49	17,542.05	19,027.69	19,286.61	21,013.79	(241.54)
55939	0	1	9,377.47	4,907.42	6,393.06	6,651.99	8,379.17	(241.54)
342480	1	0	12,079.57	12,537.81	14,023.45	14,282.38	16,009.56	(241.54)
987954	1	0	5,361.09	6,832.07	8,317.71	8,576.64	10,303.82	(241.54)

4.4 Simulation Study Design

To determine the performance of our proposed super learner estimator, we tested the method under a set of simulated data with different characteristics. The simulated data was based on the data from the Arkansas study; we incorporated the distributions and

relationships of the variables. For the simulation, we generated a population of 1 million observations.

4.4.1 Baseline Covariates

Maternal age (age) was drawn from a truncated normal distribution with the following parameters: $min = 18, max = 45, mean = 28.4,$ and $sd = 8$. Maternal plan type (*plan.type*) considers the type of health insurance plan and serve as a proxy for unobserved characteristics of the mother, such as household income. Maternal plan type was also generated from a multinomial distribution with four categories: Health Maintenance Organization (HMO), Preferred Provider Organization (PPO), Point of Service (POS), and High Deductible Health Plan (HDHP) and respective probability vector $p_{plan.type} = (0.02, 0.71, 0.14, 0.13)$. The indicator variables for each of these categories are referred to as (*plan.type_HMO, plan.type_PPO,* $plan.type\_POS$, $plan.type\_HDHP$). Maternal policy holder ($policy$) identifies the main person who is covered by the insurance and responsible for paying the premiums, also referred to as the primary policy holder. The variable, *policy*, was generated from a multinomial distribution with three categories (policy holder, spouse, and dependent) and probability vector, $p_{policy} = (0.37, 0.4, 0.23)$, respectively. The indicator variables for each of these categories are referred to as (*policy_PH, policy_SP,* and $policy\_DEP$), respectively. Maternal policy holder serves as a proxy to incorporate unobserved characteristics of the mother. The remaining covariates, fetal distress (*fetal.distress*), fetal malpresentation (*fetal.malp*), multiple gestation (*multi.gestation*), preterm birth (*preterm.birth*), previous cesarean (*prev.cesarean*), and metropolitan statistical area (*msa*) are modeled as truncated normal distribution between 0

and 1. The higher the value for each is interpreted as a higher probability of the individual having the characteristic. For example, the higher the value of *msa* the higher probability the observation lives in a rural area. Each was generated with a mean and a standard deviation. The mean vector $p_b = (0.07,\ 0.02,\ 0.02,\ 0.07,\ 0.18,\ 0.3)$ for each covariate respectively. The standard deviation vector $s_b = (0.04,\ 0.01,\ 0.01,\ 0.04,\ 0.1,\ 0.15)$. Figure 14 shows the simulated data of the baseline covariates.

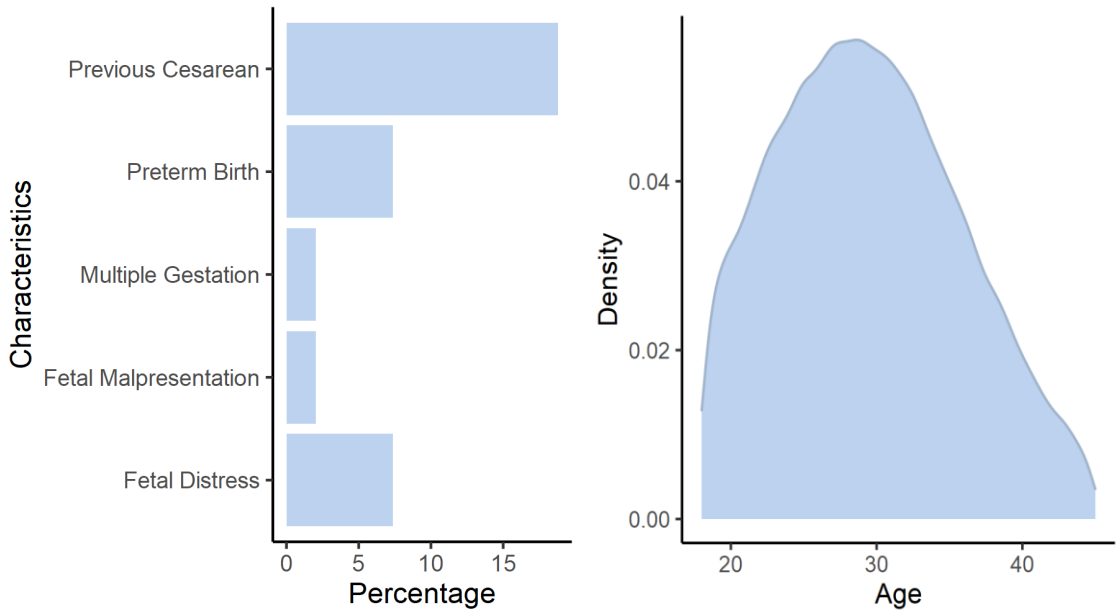


Figure 14: Simulated Data of Maternal Characteristics

4.4.2 Policy Intervention Status

The time-period variable, *T*, is an indicator for the post-intervention time period and generated by a series of binomial distributions and subsequent assignments to avoid

time conflicts. The intervention assignment variable, A, was generated from a binomial distribution with a probability, $p_A = \exp(0.65 + 0.1 * msa - 0.05 * policy_{PH} - 0.08 * plan.type_{HMO})^{-1}$. The intervention assignment variable is dependent on *msa*, *policy*, and *plan.type* to reflect the relationships observed in the study data. Figure 15 shows the simulated data of the baseline covariates that influence the intervention assignment.

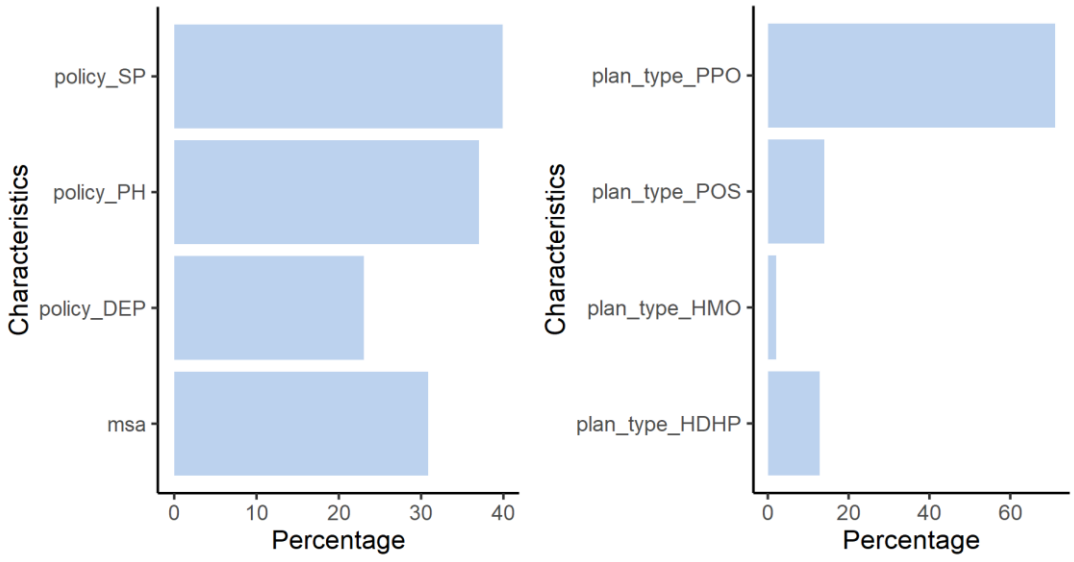


Figure 15: Simulated Data of Maternal Policy, MSA, and Insurance Plan Type

4.4.3 Perinatal Episode Spending

We introduce an unmeasured confounder, U_Y , for such considerations as physician price sensitivity. U_Y is generated from a normal distribution with a mean = 0 and sd = 250. The outcome variable, Y, is the perinatal spending for an episode. For the simulation, Y was generated to produce three different effect sizes (-250, -400 and -750) referred to as Y_I

, Y_{II} , and Y_{III} respectively. The effect size is the cost difference, in dollars, for a randomly selected individual within the treatment group. Therefore, a negative sign would indicate cost savings. Y is generated from a normal distribution.

Y_I has a mean of : $6200 - 1500 * \mathrm{A} - 250 * (T * \mathrm{A}) + \exp\left(\frac{age}{4}\right) + 800 * fetal.malp + 1200 * multi.gestation + 500 * prev.cesarean + (age^2) * fetal.distress + 1800 * preterm.birth + (10^5) * multi.gestation * preterm.birth + U_Y$ and SD = 750.

Y_{II} has a mean of : $6200 - 1500 * \mathrm{A} - 400 * (T * \mathrm{A}) + \exp\left(\frac{age}{4}\right) + 800 * fetal.malp + 1200 * multi.gestation + 500 * prev.cesarean + (age^2) * fetal.distress + 1800 * preterm.birth + (10^5) * multi.gestation * preterm.birth + U_Y$ and SD = 1200.

Y_{III} has a mean of: $6200 - 1500 * \mathrm{A} - 750 * (T * \mathrm{A}) + \exp\left(\frac{age}{4}\right) + 800 * fetal.malp + 1200 * multi.gestation + 500 * prev.cesarean + (age^2) * fetal.distress + 1800 * preterm.birth + (10^5) * multi.gestation * preterm.birth + U_Y$ and SD = 2250.

The distribution for the Y_I , Y_{II} , and Y_{III} can be see in Figure 15.

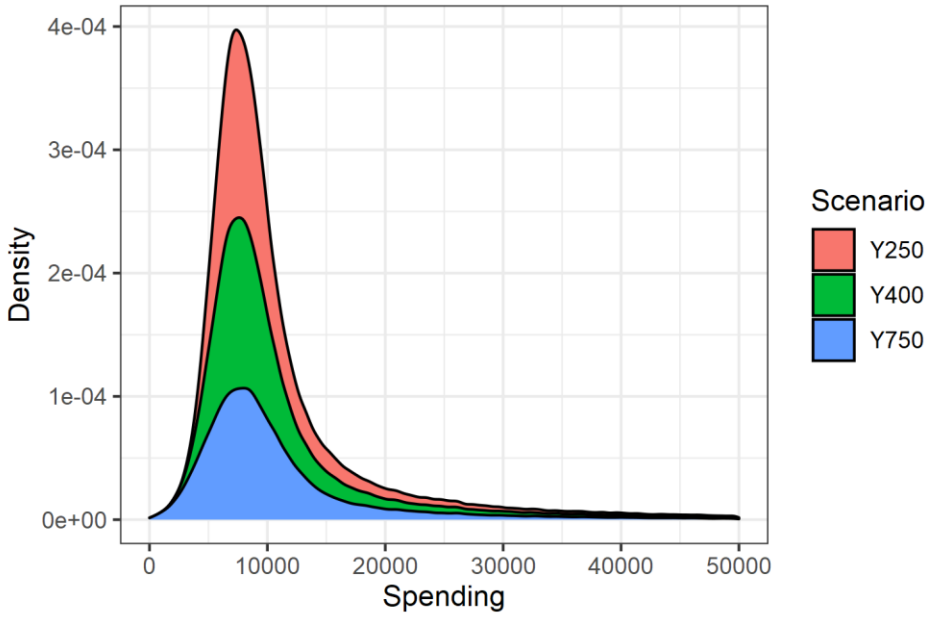


Figure 16: Density of Simulated Perinatal Spending

4.4.4 Simulation Scenarios and Results

We use the simulation to demonstrate the following. First, we know if we knew the true data-generating distribution of the outcome, we could correctly predict the outcome values. However, this rarely happens in practice. Hence, the goal is to get as close as possible to the true value. To show this, we estimate the ATT using a correct specification for the outcome regression. This represents the ideal scenario. Next, we use a regression that is misspecified, the common scenario in real-world practice. Third, we use super learner which predicts the outcome without needing to make strict assumptions about its functional form. The bias for two sample sizes (2,000 and 5,000) is calculated for all three effect sizes (Y_I , Y_{II} , and Y_{III}) and for each of the three estimators.

4.4.4.1 Regression Outcome

In our first example, the fit will be based on the correctly specified regression model where $\mathrm{E}(\mathrm{Y}|\mathrm{A},\mathbf{X}) = \mathrm{A} + T + (T * \mathrm{A}) + \exp\left(\frac{age}{4}\right) + fetal.malp + multi.gestation + prev.cesarean + (age^2) * fetal.distress + preterm.birth + multi.gestation * preterm.birth$. The results of the completed procedures for all six scenarios are listed in Table 13.

Table 13: Average Treatment Effect Among the Treated (ATT) Estimates and the Standard Error by Effect Size and Sample Size for Correctly Specified Regression Model

True Effect Size	Sample = 2,000	Sample = 5,000
-252.39	-252.6 (SE =78.92)	-251.6 (SE =46.06)
-403.82	-398.4 (SE =112.19)	-400.2 (SE =72.36)
-757.16	-749.2 (SE =201.6)	-750.7 (SE =127.37)

In the second scenario, we use a regression algorithm that is misspecified by leaving out the variable *age* and all the interaction terms. Hence, the regression model is: $\mathrm{E}(\mathrm{Y}|\mathrm{A},\mathrm{T},\mathbf{X}) = A + T + \mathrm{A} * T + fetal.malp + multi.gestation + prev.cesarean + fetal.distress + preterm.birth$. The results are listed in Table 14.

Table 14: Average Treatment Effect Among the Treated (ATT) Estimates and the Standard Error by Effect Size and Sample Size for Misspecified Regression Model

True Effect Size	Sample = 2,000	Sample = 5,000
-252.39	-152.9 (SE =922.69)	-197.5 (SE =616.15)
-403.82	-268.0 (SE =915.87)	-334.8 (SE =595.29)
-757.16	-617.7 (SE =932.61)	-689.4 (SE =604.24)

4.4.4.2 Super Learner

In the third scenario, we use super learner. For super learner we include all the covariates used in the generation of *Y*: *T, A, age, fetal.malp, multi.gestation, prev.cesarean, fetal.distress,* and *preterm.birth.* However, we do not assume a functional form for *Y*. The results for super learner are listed in Table 15.

Table 15: Average Treatment Effect Among the Treated (ATT) Estimates and the Standard Error by Effect Size and Sample Size for Super Learner

True Effect Size	Sample = 2,000	Sample = 5,000
-252.39	-190.2 (SE =150.96)	-203.3 (SE =92.36)
-403.82	-342.1 (SE =120.1)	-356.5 (SE =76.28)
-757.16	-593.6 (SE =211.2)	-651.5 (SE =146.85)

In this simulation study, we used G-computation with correctly specified and misspecified parametric regressions and super learner to estimate the ATT. The results in Table 13 show, as expected, that the correctly specified regression has the lowest bias across all effect sizes and sample sizes. We also see from all three tables (Table 13, Table

14 and Table 15) that larger sample sizes assist in decreasing bias. The misspecified regression produced (Figure 17 and Figure 18) the highest percentage bias for the lower effect sizes (-250 and -400). Super learner shows significant bias reduction over the misspecified regression for the lower effect sizes. Super learner's performance gains are even greater at the smaller size. For example, we see for the effect size of 400 at the sample size = 2,000, there is over a 50% reduction in bias. This is an important finding, that super learner had comparable bias at a sample size of 2,000 as the misspecified regression at a sample size of 5,000. In practice, where we often have small sample sizes relative to the full population, super learner allows us to obtain estimates with the bias percentage of a far larger sample size. However, super learner has the highest percentage bias for the higher effect size (-750). G-computation showed minimal bias when the outcome was correctly specified, demonstrating its value as an estimator. Though it is important to note, for both the misspecified outcome regression and the super learner with G-computation there is significant bias, with most of the simulation scenarios yielding over 10% bias. This highlights the limitations of G-computation, that when the prediction algorithm for *Y* is misspecified, we can expect bias.

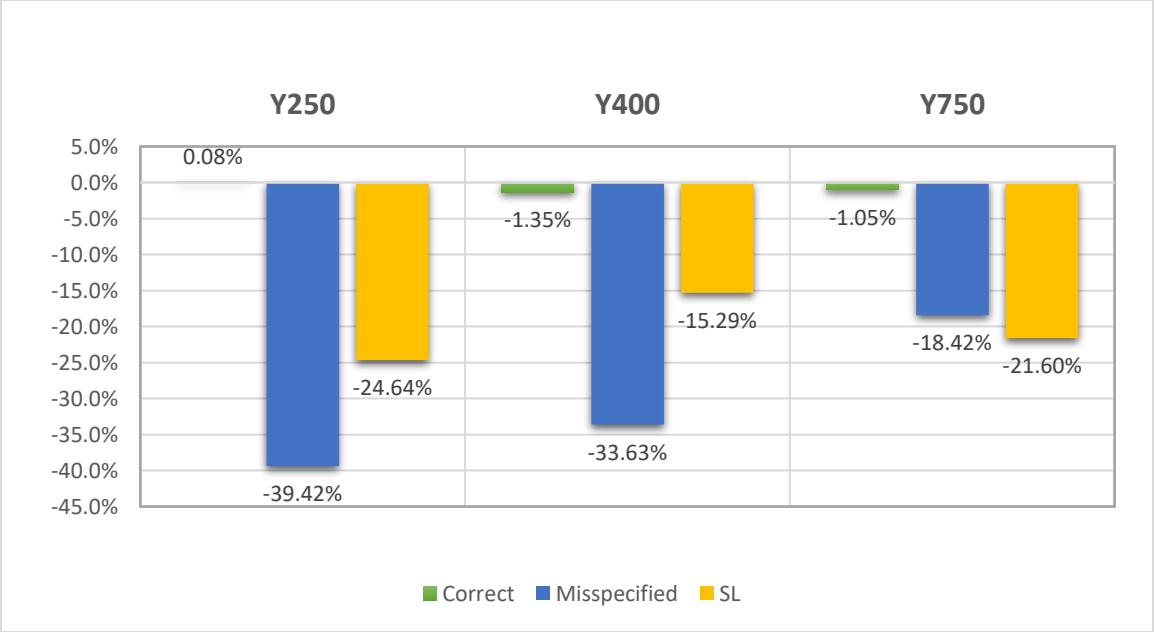


Figure 17: Percentage Bias by Estimator - Sample Size = 2,000

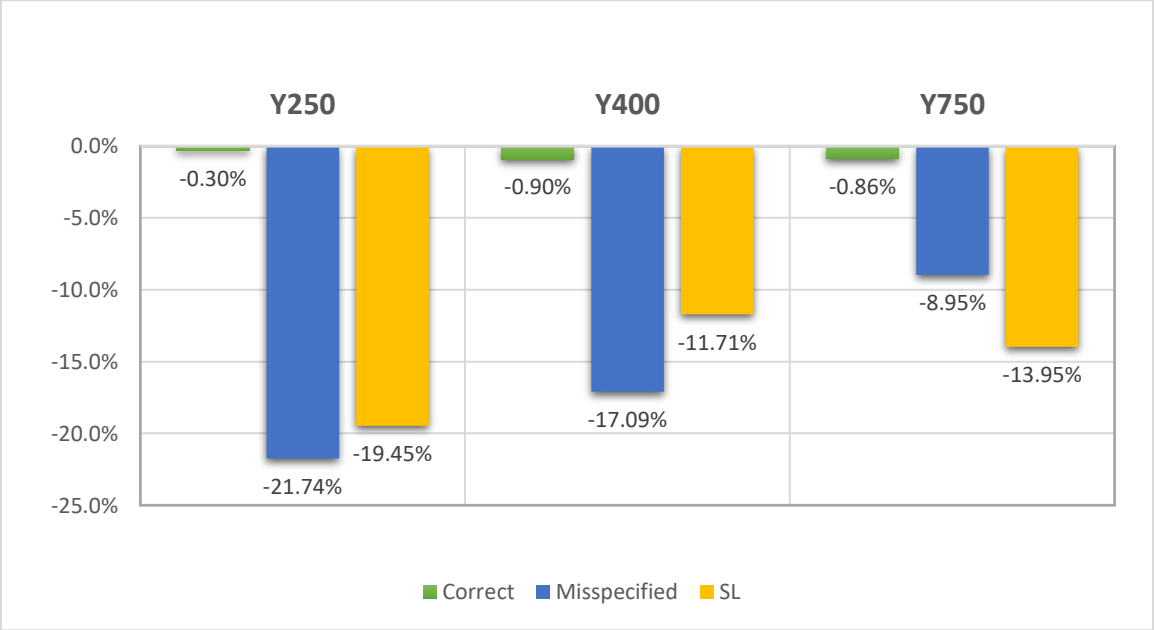


Figure 18: Percentage Bias by Estimator - Sample Size = 5,000

4.5 Discussion

G-computation is a popular method to estimate causal effects in observational studies, although it has been infrequently used in DID studies. Some previous research has shown G-computation's effectiveness for estimating ATT with parametric regression [89]. However, other research has shown that G-computation is sensitive to model misspecification in DID [83, 90]. In practice, researchers rarely know the correct specification of the data which can lead to estimates with high bias. In this chapter, we attempt to address this limitation by implementing a machine learning algorithm for the relationship between the outcome and the covariates. We created a simulation study to evaluate the G-computation estimator with super learner to estimate the ATT, which is a novel contribution to the literature. We base the simulation on a previous study that assessed the impact of episode-based payment policy on perinatal costs. Machine learning methods, such as random forest, can account for complex data structures and do not require the researcher to choose the functional form of the data. Furthermore, by using an ensemble machine learning method, we can implement multiple algorithms providing additional flexibility to best model the data. Given that observational data can often be complex, these methods can greatly reduce bias. In the simulation, we saw bias reductions of over 50%. An additional benefit we found in our work, was there was substantially lower standard deviation in the estimates with super learner. This can be an important feature if sample sizes are smaller and could explain the larger performance gains we see in the smaller sample sizes. Nevertheless, our proposed method did not reduce bias for the largest effect size. Therefore, we propose future work to implement double robust targeted learning with super learner to the simulation data. The advantages of the double-

robust estimator may provide further bias reduction than that of the G-computation with super learner. If such is the result, we also suggest a replication study in which we apply TMLE to the real data of the original study.

4.6 Targeted Learning

Targeted maximum likelihood estimation (TMLE) is a nonparametric estimator that has been shown to reduce bias and improve the accuracy over commonly used methods. The double robust feature protects the estimator from misspecification of either the outcome mechanism or the exposure mechanism. This is due to TMLE's approach of incorporating both the outcome and the exposure mechanism into its estimation sequence. These features of TMLE allow for a more realistic and flexible estimator within a health care setting.

Previous studies have used TMLE to estimate a variety of parameters and in a wide range of applications. Weber used TMLE for three estimands (outcome post-intervention, difference of community level means and DID) and demonstrated the bias that occurs when a target parameter's identification assumptions aren't held [91]. Schuler and Rose presented an overview of TMLE for estimating the average treatment effect and discussed its benefits over IPW and G-computation [90]. Van der Laan et al. showed a one-step TMLE for the ATT [92]. Rose and Normand used TMLE to estimate the patient outcome for coronary artery stents; further developing the approach to accommodate multiple unordered treatment options and cluster observations [93]. TMLE has also been used for longitudinal data to determine the impact of an exposure that occurs over time [94]. This study implemented longitudinal TMLE to estimate the risk of ischemic heart disease.

Nevertheless, TMLE has not been implemented for the DID study design like the one described here, in which there is a pre-intervention and a post-intervention time period. The methodological contribution of this proposed future work is to extend TMLE to this setting and assess its performance compared to G-computation.

CHAPTER 5. EDGE WEIGHT ESTIMATION FOR SOCIAL NETWORKS

5.1 Introduction

Social networks have been demonstrated to be valuable tools in visualizing relationships and assessing interdependencies. Hence, they can provide powerful insight for heath care problems. Social networks have been used to assess the impact of social influence on health outcomes such as smoking [95-97], obesity [98, 99], and diabetes [100, 101]. They have be used to track disease spread such as HIV transmissions [102] and are increasingly used in a range of epidemiological applications [103]. Health insurance claims have also served as a data source to create social networks as well. These networks offer unique connections that can be used to evaluate physician behavior. Appel, et al used health insurance claims to track physician activities across 18 months [104]. The paper demonstrated the value of claims data for social network analysis by creating networks to evaluate three types of physician relationships: physician-patient, physician-physician and physician-health care provider. Herrin et al showed how physician peer groups can be identified with high accuracy using a patient-sharing network developed from claims data [105]. Landon et al showed how social networks derived from claims data can identify naturally occurring networks of physicians. The results can be used to inform the structure of Accountable Care Organizations (ACOs), a group of health care providers that offer coordinated care for the chronically ill [106]. In addition patient-sharing social networks from claims data were able to identify professional networks of physicians and evaluate the varying network characteristics of the professional network [107]. Medical claims data

has proven to be a rich source for social network analysis research and to be of good quality [108]. However, there can be missing elements in the data [109]. When choosing to use claims data to create social networks, the weight of the edges contain pertinent information regarding the strength of the relationship between the two entities. Hence, it is important that the information regarding the weight of the edges is as complete as possible. Accurate weight estimation can be useful for social networks as they can serve as a data quality tool to check if the weights provided in the social network are correct. Edge weight prediction can also help determine which relationships will strengthen (i.e. increase in weight) or weaken (decrease in weight) in the next few time periods. In this chapter, we address this issue by presenting three methods to estimate weights of existing edges.

Network Science research related to estimation of edge characteristics has primarily focused on edge existence. For example, Wasserman used a range of logit and logistic regression models to estimate edge existence on a range of generated social networks [110]. Clauset and Moore presented a general technique for inferring hierarchical structure from network data [111]. They showed how knowledge of the hierachical structure can be used to predict edge existence with high accuracy. Zhao et al approached edge prediction for a partially observed network by ranking the probabilty of an edge existing based on observed edges and node covariates [112]. The authors relied on an assumption that if one pair of nodes is similar to another pair of nodes then the probability of a edge existing hierarchical structure between the second pair is directly proportional to their similarity. Then they predict the probability of the missing edge using the similarity matrix based on a node similarity measure. Fire et al used network topology features to predict edge existence [113]. The additional contribution was the selected features were scalable for large social

networks. Many other methods have been used including using common neighbors [114], deep learning [115] and graph neural nets [116].

Despite extensive research on prediction of edge existence, much less work has been done on weighted graphs. Zhao et al used similarity measures based on reliable routes that incorporated the weights of the paths between two non-adjacent nodes of the network to predict the existence of edges and their weights using logistic regression [117]. The approach performed a transformation of the weights to normalize them between 0 and 1 and then performed logistic regression to estimate the log odds ratio. The work carried the assumption that the product of the weights along a path from one node to the next is a reliable estimator for weights (i.e. that nodes are more likely to have a edge if they share similar paths. This may be true for protein-protein interactions (as seen in the paper) but not necessarily true for social networks. Hou and Holder used deep learning with similarity of the nodes based on their common neighbors [118]. Though a promising approach that provided good accuracy, interpretability of the data is lost with this estimation method. Sa and Prudencio focused on a supervised learning approach for edge prediction based on topological features of weighted social networks [119]. We intend to expand on this work by using weighted network characteristics but instead of edge prediction, we will focus on weight prediction.

For this chapter, the problem statement is to develop a supervised estimation technique for edge weight prediction of a partially observed social network. We leverage a range of topological characteristics of the graph to inform and train the model. Additional benefits for the approach is the flexibility to incorporate metadata, something that has only been done for prediction of edge existence.

5.2 Problem Set-up and Definition

We consider the problem of weight estimation of an existing edge. We define the graph, G = (*V*,*E*,*W*), V is a vector of all the nodes of the network and $|V|$ is the cardinality, or the number of nodes in the network. Likewise, $|E|$ is the number of edges and E_{vxv} is a matrix of the edges where $E_{i,j} = 1$ indicates an edge exists from node i to node j. W_{vxv} is the matrix of the weights of the edges. Let $w_{i,j}$ be the entry of the *i*th row and *j*th column of the *W* matrix and represents the weight of the edge between nodes *i* and *j*. The setting for this study consists of having observed all the edges of the network and partially observed the weights. Hence, *W* will contain some missing values.

5.2.1 Sample Problem

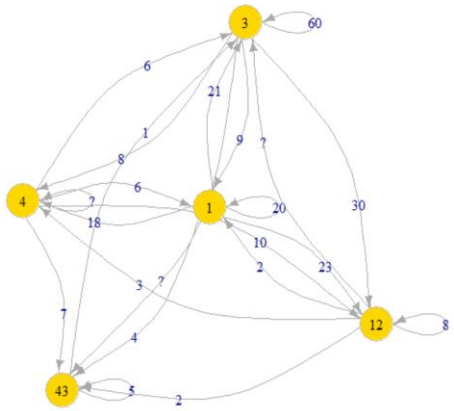


Figure 19: Sample of Faculty Network

We present a simplified example made from a small sample of a faculty hiring network for business schools. In this directed network, each node represents an institution

and the weight of the directed edge ($w_{i,j}$) is the total number of individuals who received their PhD from institution *i* and was hired as a faculty member at institution *j*.

Table 16: Adjacency Matrix of Sample Faculty Network

Nodes	***1***	***3***	***4***	***12***	***43***
1	20	21	18	23	4
3	9	60	8	30	**?**
4	6	6	**?**	10	7
12	2	**?**	3	8	2
43	0	1	0	0	5

Table 17: Metadata of Sample Network

Node ID	**USN2012**	**Region**	**Institution**
1	1	West	Stanford
3	2	Northeast	Harvard
4	7	West	UC Berkeley
12	3	Northeast	University of Pennsylvania
43	34	Northeast	Boston University

In this setting, we see there is high variation among the values of the edge weight. For example, there are lower frequencies, such as the edge $e_{12,4}$, and higher frequencies, such as the edge $e_{3,12}$. Self-connections are permitted, and it appears that most schools prefer hiring alumni from their own institution. There is metadata available for the nodes and it is listed in Table 17.

5.2.2 Data Description

The faculty hiring dataset is found in the Index of Complex Networks (ICON) [120]. It contains information of faculty hiring for years 2011-2013 and including approximately 19,000 tenure-track or tenured faculty from 461 North American departments in business (206 nodes), computer science (113 nodes), and history (145 nodes). Nodes are academic institutions and a directed weighted edge from node *i* to node *j* represents that the number of graduates from institution *i* who are now faculty members at institution *j*. The attributes and their descriptions are listed below. We used a subset of the observed attributes that were relevant to weight prediction. The institutions that were not included in the 461 institutions but granted PhDs to current faculty members within the sample, were aggregated into the observations of "All others."

The data provides information for each specific hire. Weight of the edge is determined by aggregating all the hires from a PhD-issuing university to the hiring institution. A network was created for each discipline (Figure 20). The business network has the most nodes and is also the most dense (i.e. highest average node degree). Four attributes are provided in the data set for each node. US News and World Report ranking is a factor variable that is the ranking of the specific department from the US News college ranking for the year 2012. Canadian schools were included within the computer science departments and were given a NA for this variable. Geographic region is a factor variable with values Northeast, Midwest, South and West for the US as well as Canada. The factor level "Earth" is used for the "All others" observation. Percentage of male is a continuous variable between 0 and 1 and is the percentage of hires from that PhD-issuing node (source node) that were male. This is a derived feature since the data only provided the gender for

each hire. Percentage of assistant professors is the percentage of the hires for the hiring institution that are the assistant professors. This also is a derived feature since the data only provides the current rank for each hire.

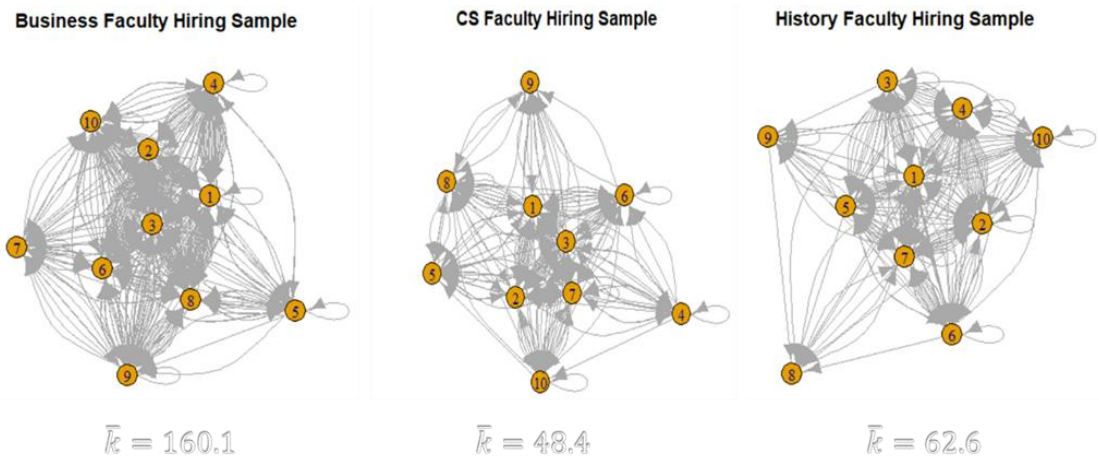


Figure 20: Sample Networks with the Mean Number of Edges per a Node for Faculty Hiring Data

5.3 Statistical Modeling

5.3.1 Deriving and Incorporating Covariates

We introduce $S_{i,j}$, the calculated similarity measure of the two nodes *i* and *j*, where the S_{vxv} matrix contains the similarity measures for all the nodes in the network. The similarity measure quantifies the closeness of two vectors s_i and s_j; the vector comprised of the covariate values (metadata) for nodes *i* and *j* respectively. To calculate the similarity measure, $S_{i,j}$, we have to consider two scenarios. The first, is when all the covariates are

numerical values that have been standardized, then we can calculate $S_{i,j}$ using any version of a p-norm, such as the exponential decay kernel [112]. That is, $S_{i,j} = exp\left\{\frac{\|s_i - s_j\|^2}{\sigma^2}\right\}$.

However, when the data is not continuous, we cannot use functions based on the p-norm. This is because if a variable is categorial with multiple categories, the categories are not necessarily ordered. Therefore, the magnitude of the difference is arbitrary based on the number assigned to each category. For example, if a variable is school subject and Business is assigned the number 1, History the number 2 and Computer Science the number 3. Business is not any more similar to History than Computer Science, but the number assignment would indicate that Business is 2x more similar to History than Computer Science. It is possible to create dummy variables for each categorical variable, if the number of categories is not too great. This is a helpful approach when the data is mixed with continuous and categorical data. Our example is a case where this could work because USN is an ordered categorical variable and Region does not contain too many categories. However, this is not the case for all situations, so in this case we propose calculating $S_{i,j}$ as an average of the similarity for each covariate. We let $S_{i,j}(r)$ represent the closeness of node *i* to node *j* for covariate *r*. We then calculate $S_{i,j}(r)$ across all *R* categorical attributes and take the mean. Numerous metrics exist to calculate $S_{i,j}(r)$, but we suggest Eskin's work [121] due to its simplicity and its consistent performance under a range of data characteristics [122]. The Eskin similarity measure assigns weights to mismatches that is proportional to the number of categories within each attribute. A mismatch that has more categories gets weighted heavier than those attributes with less categories. Eskin's limitation is that it does not perform well if the data has a large number of attributes. Given

metadata is often not available for networks and if available contains very few attributes (i.e. covariates), we do not believe this limitation will affect the metric's appropriateness for our use. See Boriah paper for a summary of similarity measures for categorical data [122]. We let d_r represent the number of categories within attribute *r*, and $s_i(r)$ is the *r*th component of vector s_i . For example, if *r* were an indication variable for female gender and the *i*th node was a female, then d_r = 2 and $s_i(r)$ = 1. Using Eskin, we calculate $S_{i,j}(r)$ and S_{ij} as the following:

$$S_{i,j}(r) = \begin{cases} 1 \ \ if \ s_i(r) = s_j(r) \\ \dfrac{d_r^2}{d_r^2 + 2} \ \ if \ s_i(r) \neq s_j(r) \end{cases}$$

$$S_{i,j} = \sum_{r=1}^{R} \frac{1}{R} \, S_{i,j}(r)$$

As aforementioned, metadata is often not used for the edge prediction and weight prediction because the information is rarely available. More often, topological features of the network are used to create the predictors. This information is valuable because it incorporates relational information of nodes since the metrics are derived for the edge instead of the node. Given that we are interested in an attribute of the edge (i.e. weight), this means that unlike the node measures, these metrics can be directly entered into the algorithm. We specific the features derived from the topology of the network as x_m , where $m \in \{1,2, \dots p\}$ is the index for the topological feature we include in the model and p is the number of topological features considered. Note, that X_{pxE} is a matrix in which each column contains the values of the topological features an edge.

5.3.2 Description of Considered Topological Features

In this section, we discuss the definitions of the topological features used as predictors for estimation. Both local and macro metrics are included in the study. Local metrics provide local structural information around the nodes of an edge, such as common neighbors. Macro metrics are based on information of the network a few degrees away from the nodes connected by an edge. We introduce the features and discuss their derivation when appropriate. We specify each feature with italics to indicate that it is a variable.

Source Node ID and *Target Node ID* are factor variables that specify the identification numbers of the two nodes that an edge connects. A directed edge begins at the source node and ends at the target node. If the network is undirected then the source and target node ids can be interchangeable. However, we distinguish between the source and target node for directed graphs so we can account for reciprocity. *Reciprocity* considers the edge from node *i* to node *j* and the existence of its inverse (edge from node *j* to node *i*). *Reciprocity* is featured as an indicator variable that specifies if the inverse of the edge exists and *Reciprocity Weight* is the weight of the inverse edge. If the inverse edge does not exist, the reciprocity weight is designated as 0. *Average Source Outcoming Weight* is the average weight of the outgoing edges from the source node and *Average Target Incoming Weight* is the average weight of the incoming edges to the target node. *Source Weight* is also included and is the total sum of the outgoing weights of source node.

Three common neighbor similarity metrics are calculated: inverse log-weighted [123], Jaccard [113], and Dice [124] . To find the common neighbors of two connected

nodes *i* and *j*, you first identify the set of node *i*'s neighbors (i.e. nodes to which *i* is connected) and the set of node *j*'s neighbors. Common neighbors is the intersection of both of those sets. Each of the common neighbor similarity measures are based on the number of common neighbors but not the weight of the edges between them. Hence, for each of these features, we also derived a weighted version to incorporate the strength of the connections to the common neighbors. A node's degree is the number of edges connected to it. The inverse log similarity of two nodes is the number of common neighbors multiplied by the inverse logarithm of their degrees. For the weighted version, instead of the multiplying by the inverse logarithm of their degrees, we would multiply by the inverse logarithm of the edge weights of the common neighbors. The Jaccard similarity is calculated by the unique number of common neighbors shared by both nodes divided by the total degree of the source and target node. The Dice similarity metric is calculated as twice the number of common neighbors divided by the total degree of the source and target node. For Jaccard and Dice metrics, we replace the number of common neighbors with the weights of the edges to the common neighbors.

Community Membership is an indicator variable which specifies if the source and target node belong to the same community. Community assignment was determined via a fast-greedy algorithm based on modularity optimization as in Clauset work [125]. *Edge Betweenness* is the feature that is the number of shortest paths that go through the edge. In addition, we include several directed motifs. A directed motif is a directed triangle in which three nodes (u, v, w) are connected in a certain pattern. If a directed motif exists for the edge of interest, the edge of interest is one of the edges that connects the three nodes in the specified pattern. The value for the directed motif feature is the average weight of the

other two edges included in the directed motif. We first consider *Direct Motif* which is the variable for the pattern u-> v -> w-> u. For example, in Figure 21, let u =1, v=2, and w = 12. The edge $e_{1,2}$ would have the Direct Motif value of 14.5 from the average of the edge, $e_{2,12}$, from node 2 to node 12 and the edge, $e_{12,1}$, from node 12 to node 1. If there are multiple such triangles, we can take the average of across all triangles.

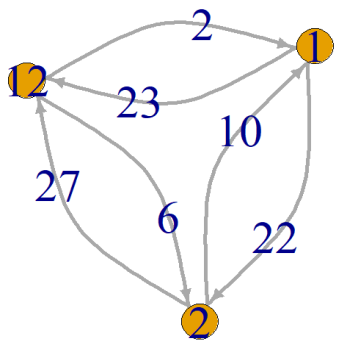


Figure 21: Sample Network with Directed Motifs

Feed Forward W Motif has the pattern u -> v -> w , u -> w. From Figure 21, we see the Feed Forward W Motif variable for the edge $e_{1,2}$ would be the average of edges $e_{1,12}$ and $e_{2,12}$ which is 25 (the mean of 23 and 27). Feed Forward V Motif has the pattern w->u->v, w->v and the value for the edge $e_{1,2}$ in Figure 21 would be the average of edges $e_{12,1}$ and $e_{12,2}$ which is 4 (mean of 2 and 6).

5.4 Proposed Estimators

In this section, we discuss the three different estimators that we compare in the study. We introduce the estimator and then discuss how we apply it for the purpose of predicting edge weights for a social network.

5.4.1 Poisson Regression with Covariates

The baseline estimator provides the least flexibility for estimation by constraining it to the functional form of a single Poisson distribution. Poisson regression is appropriate since we are only looking at social networks. Weights in social networks are a discrete variable based on some metric of strength that is used (i.e. number of interactions, counts, etc). For ease of notation, we collapse the indexes *i,j* to g where $g \in \{1, \dots |E|\}$ represents the edge number. Let $w_g \sim Poisson(\lambda_g)$. Hence, we model the weight as the following:

$$P(\mathrm{w}_g) = \frac{\lambda_g{}^{\mathrm{w_g}}}{\mathrm{w}_g!} e^{-\lambda_g}$$

We use the canonical link function for the Poisson Regression. The similarity measure and the topological metrics are predictors for the weight. β_1 is the coefficient of similarity measure, S_g, which is derived using one of the methods discussed in 5.3.1. This includes the node specific information (metadata). For the faculty hiring example, this includes: US News Ranking, Region, Percentage Male, Percentage Assistant Professors. Since the data is mixed (categorical and continuous), the Eskin similarity measure is used to calculate S_g. β_m is the coefficient for the topological metric, $x_{m,g}$ for edge e_g (described

in 5.3.2) and $m \in \{1,2,\ldots p\}$ is the index for the topological feature we include in the model. For our example, $p = 14$, and includes Source ID, Target ID, Community Membership, Edge Betweenness, Similarity Inverse log-weighted, Similarity Jaccard, Similarity Dice, Direct Motif, Feed Forward V Motif, Feed Forward W Motif, Avg Source Outcoming Weight, Average Target Incoming Weight, Reciprocity, and Reciprocity Weight. This yields the following regression:

$$log(w_g) = \beta_0 + \beta_1 \cdot S_g + \sum_{m=1}^{p} \beta_m x_{m,g}$$

5.4.2 Comparative Estimator: Finite Mixture Model

We suspect that in many cases there is more than one Poisson distribution from which the weights of the social network are drawn. For example, in an online social network we expect the social influencers (i.e. hubs) to have higher weights than the average user. Therefore, we want to have an estimator that can account for multiple subpopulations each with its own individual rate (i.e. one for the hubs and one for the non-hubs). To account for this, we introduce the finite mixture model [126, 127], where the edge weight, w_g, can be modeled as a finite mixture of Poisson distributions. Each Poisson distribution is the distribution for one of the subpopulations of the data. We focus on the generalized case of a finite mixture model in which the number of subpopulations, called components, is known a priori. We let *C* represent the total number of components. Hence, the edge e_g, can belong to one of C number of components and be estimated via Poisson regression for that component. The indicator variable Z contains the label of the component membership. The label vector $Z_g = (z_{g1}, z_{g2}, \ldots z_{gC})$ consists of the component indicators for edge e_g. Let

$$z_{gk} = \begin{cases} 1 & if\ edge\ g\ belongs\ to\ component\ k \\ 0 & o/w \end{cases}$$

Each component, *k*, has its own Poisson distribution of $f_\Lambda(w_g, \lambda_k)$ and its own proportion (i.e. mixing weight) π_k. The model for the edge weight can then be written as:

$$f(w_g, \Lambda) = \sum_{k=1}^{C} \pi_k f_\Lambda(w_g, \lambda_k)$$

The $\sum_{k=1}^{C} \pi_k = 1$, $\Lambda = \{\lambda_1, \ldots, \lambda_C\}$ is the vector of the parameters (i.e. rates) for all C clusters and $\Theta = \{\pi_1, \ldots, \pi_C\}$ is the vector of mixing weights. We want to solve for the maximum likelihood estimate (MLE) of Λ and Θ, given the data (*W*). We show how the EM algorithm can be used to solve the MLE. First, we identify the likelihood for the occurrence of $(w_g, z_{g1}, \ldots, z_{gC})^T$ as the joint distribution and using the Law of Total Probability we get

$$P(W_g = w_g, Z_{g1} = z_{g1}, \ldots, Z_{gC} = z_{gC})$$

$$= P(W_g = w_g \mid Z_{g1} = z_{g1}, \ldots, Z_{gC} = z_{gC})\, P(Z_{g1} = z_{g1}, \ldots, Z_{gC} = z_{gC})$$

$$= \prod_{k=1}^{C} \pi_g^{z_{gk}} f_\Lambda(w_g, \lambda_k)^{z_{gk}}$$

Leading to the likelihood and log-likelihood being the following, respectively:

$$L(\Lambda, \Theta) = \prod_{g=1}^{E} \prod_{k=1}^{C} \pi_g^{z_{gk}} f_\Lambda(w_g, \lambda_k)^{z_{gk}}$$

$$\log L(\Lambda, \Theta) = \sum_{g=1}^{E}\sum_{k=1}^{C} z_{gk} \log \pi_g + \sum_{g=1}^{E}\sum_{k=1}^{C} z_{gk} \log f_\Lambda(w_g, \lambda_k) \quad (3)$$

Next, we can implement the expectation step (E-step); the first step of the EM algorithm, by replacing the unobserved data, Z_{gk}, with the conditional expectation given the observed weights, W

$$P(Z_{gk} = 1|W) = \frac{P(W|Z_{gk} = 1)P(Z_{gk} = 1)}{\sum_k P(W|Z_{gk} = 1)P(Z_{gk} = 1)}$$

$$= \frac{\pi_k f_\Lambda(w_g, \lambda_k)}{\sum_k \pi_k f_\Lambda(w_g, \lambda_k)} = \tilde{z}_{gk}$$

Plugging in $\tilde{z}_{gk}$, the expected value of z_{gk}, for z_{gk} in Equation (3) , the update rule for the MLE becomes:

$$Q(\Lambda^{(l)}, \Theta^{(l)}) = \sum_{g=1}^{E}\sum_{k=1}^{C} \tilde{z}_{gk} \log \pi_k + \sum_{g=1}^{E}\sum_{k=1}^{C} \tilde{z}_{gk} \log f_\Lambda(w_g, \lambda_k)$$

The maximization step (M-step) is the second step in the EM algorithm. We update the current parameters with the expected value using the complete data likelihood. The mixing weights, Θ , can be found via the mean of the expected value of the conditional expectation, where $|E|$ is the number of edges. You can see here that the proportion is equal to the mean probability across all the edges belonging to component *k*.

$$\pi_k^{(l+1)} = \frac{\sum_g \tilde{z}_{gk}}{|E|}$$

If the component parameters (i.e. rate of the component) are unknown as well, we can estimate the rate of component *k* by taking the mean of the conditional expectations, as seen here:

$$\lambda_k^{(l+1)} = \frac{\sum_e \tilde{z}_{ek} w_e}{\sum_e \tilde{z}_{ek}}$$

5.4.2.1 Incorporating Topological Information via Covariates

We have determined how to solve for the parameters and the mixing weights of the components. We now move to the next step of incorporating the derived features to use as predictors for the finite mixture model. We start with the same log-likelihood model as before, except the probability distribution of *w* is now also dependent on the features (X_g, S_g) :

$$\log L(\Lambda, \Theta) = \sum_{g=1}^{E} \sum_{k=1}^{C} z_{gk} \log \pi_g + \sum_{g=1}^{E} \sum_{k=1}^{C} z_{gk} \log f_\Lambda(w_g, \lambda_k, X_g, S_g)$$

The EM equations are similar from the previous scenario, except now there is this added dependency. Recall, $\tilde{z}_{gk}$, is the probability that an observed edge g belongs to component *k* based on the observations, *W*. Now $\tilde{z}_{gk}$ will consider the similarity measure,

S_g, and all of the edge metrics, X_g . Z can be estimated using a logistic regression and solving for the coefficients via MLE.

$$P(Z_{gk} = 1 | X_g, W, S_g) = \frac{\exp\left(B_{0k} + \beta_1 \cdot S_g + \sum_{m=1}^{p} B_{mk} x_{mg}\right)}{\sum_{k=1}^{C} \exp\left(B_{0k} + \beta_1 \cdot S_g + \sum_{m=1}^{p} B_{mk} x_{mg}\right)} = \tilde{z}_{gk}$$

The weight is estimated using our initial formula:

$$f(\mathrm{w}_g, \lambda_k | X_g, S_g) = \sum_{k=1}^{C} P(Z_{gk} = 1 | X_g, W, S_g) f_\Lambda(w_g, \lambda_k, X_g, S_g)$$

Additional discussion on how to incorporate covariates into mixture models can be found in work by Gudicha and Vermunt [128].

5.4.3 Comparative Estimator: Ensemble Machine Learning

For the third estimator, we consider super learner [129]. Super learner is an ensemble machine learning algorithm used for prediction. It implements a library of prediction algorithm. Using cross validation, each algorithm is trained and gives a prediction for each observation. Then the predictions of each algorithm are used as the predictors in a linear regression with the outcome. Hence, the final prediction is a weighted average of each algorithm's prediction. We incorporate five algorithms into the library: generalized linear model (SL.glm), glm with an elastic net penalty (SL.glmnet), random forest (SL.randomForest), neural net (SL.nnet), stepwise regression (SL.step) .

5.5 Evaluation of Faculty Hiring Social Network

In this section, we discuss the evaluation process and the results from the three estimators for the faculty hiring example. We use precision error to assess the performance of each estimator:

$$Precision\ Error = \frac{\sum_{g=1}^{E}\left(w_g - \widehat{w}_g\right)^2}{\sum_{g=1}^{E}\left(w_g - \bar{w}_g\right)^2}$$

We first create the training and test sets. Random sampling is not an effective way to sample networks because the random selection can result in a disconnected sample graph and essential network features lost. Therefore, we sample the network using shortest paths; a method proven to work well for social networks [130]. The process for shortest path sampling is as follows.

(1) Two nodes are randomly selected in the network

(2) The shortest path between those two nodes are calculated using Dijkstra's algorithm and all the nodes on the shortest path are logged. For this example, the distance between two nodes is the inverse of their edge weight, since we wanted the edges in the sample network to be those with the highest weights.

(3) Repeat steps (1) and (2) *n* times (we use n=2*|*E*|)

(4) The nodes are ranked in descending order based on how many times each appeared in a shortest path

(5) The top 60% of the nodes are kept and their corresponding edges are used to create the core sample network. Another set of nodes from the bottom 40% are selected at random to bring the total to 80% of nodes in the sample network.

(6) Derive all the features based on the sample network and use it to train the estimators

(7) Estimate all the weights in the sample network

(8) Repeat steps (5) to (7) for 50 runs and then take the mean precision across all the runs

The results of the analysis by discipline are presented in Table 18. We see that increasing the number of components from the glm (k=1) to the 3-component finite mixture model improves the performance; demonstrating our assumption that there are multiple subpopulations within the network. From Figure 20, we know that the network based on the business hires has the highest mean degree. The results suggest that more dense networks produce lower precision error for the regression estimators. We also see that super learner has a major improvement in performance. This lends us to believe the ensemble machine learning method is the best estimator for the edge weights.

Table 18: Precision Measure by Faculty Discipline for Varying Number of Sub-Components (k) of a Mixed Poisson Estimation for the Training Set

	Precision Measure			
Faculty Discipline	k=1	k=2	k=3	SL
Business	0.37	0.21	0.15	0.06
Computer Science	0.45	0.28	0.27	0.05
History	0.44	0.40	0.26	0.04

5.6 Conclusions and Health Care Application

The motivation of this work is to predict the edge weights for social networks. Within health care, social networks are often used to measure the impact of social influence and the strength of the ties to one's community on health care outcomes and decisions. Hence, knowing the strengths of these relationships can be particularly useful. Unlike in our work, previous research in edge weight prediction focused on metrics originally designed for unweighted networks. The metrics were not adjusted to leverage the additional information provided by the strength of the relationship between the two nodes. In addition, to our knowledge, metadata has not been used in previous work to assess weight prediction in social networks. Given that metadata allows for some additional intuitive understanding of how node characteristics affect the strength of a relationship, this is a desired feature for health care applications. Our initial results are promising and showed that the finite mixture model does have higher performance with lower precision error. The implementation of super learner provides significant gains in performance, validating this approach for weight prediction. After demonstrating the value of this approach, we propose to implement the work on other social networks to see how different characteristics of the network can impact the precision of the estimator, such as mean degree. In addition, since our work only included a subset of potential topological features, more features can be added to improve estimation. Furthermore, we propose applying the work of this study to a health care network.

As interest in opioid research grows in the US, social networks from claims data is being used to identify networks with high-volume opioid prescribers [131]. Prescription claims data are also used in social network analysis to identify patients who are doctor

shopping [132]. The network specified opioid prescription recipients as nodes and an edge indicated that the two recipients shared a physician. The weight of the edge was the number of shared physicians between recipients. The research leveraged the knowledge that doctor shopping was a social process and provided a method to identify "pill mills". Expanding on this work, we suggest applying our approach to a network based on Medicaid claims data. Two networks could be created and assessed, a patient-sharing network among clinicians of opioid patients and a physician-sharing network among opioid users. In the patient-sharing network, the nodes are clinicians, determined by the NPI, and an edge indicates that the two clinicians share patients. The weight would be the number of patients the two clinicians share. For the physician-sharing network, it would be the opposite. The nodes would be patients and an edge would indicate that two patients share a physician. The weight of the edge would be the number of physicians shared between the two patients. With a near zero precision error, our estimator could perform quality assurance on the Medicaid network and predict which relationships will increase or decrease in strength in the next few time periods. Being able to predict changes in the relationships of the network can inform near-term effects such as health outcomes and help target high- risk individuals.

CHAPTER 6. CONCLUSION

In this thesis, we present a series of health care problems that are solved with techniques from analytics and machine learning. We focus our work on data-driven solutions from large data sets and show different approaches for addressing the limitations that arise with these data sets.

We begin with chapter 2, where we assess the impact of the CDC's contraceptive recommendations in the MEC. Since almost half of all US pregnancies are unintended [133], the study had major implications for women within the US. Our analysis included over 12 million reproductive-aged women enrolled in Medicaid across 14 states and over 4 years resulting in over 4 billion administrative claims. We identified women with chronic medical conditions by aggregating their claims across a two-year period. We also identified women who had undergone bariatric surgery or a solid organ transplant within the two-year period. The analysis informed an important question regarding the efficacy of the CDC's massive dissemination strategy for the new policy. We found there was an increase in the overall use of contraceptives for women with the identified chronic health conditions. However, not all conditions showed an increase in the use of contraceptives with the highest efficacy. A major takeaway is that since many women do not receive their routine gynecological care from an OB/gyn specialist. Therefore, it is important to also focus education efforts on the physician specialists that treat women with these chronic conditions. After assessing contraception and unintended pregnancies, we focused on adolescent pregnancy – the second highest age group for unintended pregnancy [133].

In chapter 3, we assess the health and wellness outcomes of infants born to adolescent mothers within their first year of life. Our study focused on Medicaid claims from over 65,000 infants across 42 states. We used the casual inference method of matching to isolate the causal effect of the age of the mother. We developed a sequential matching process based on the mother's demographics which can be used for future studies based on Medicaid claims data. An additional contribution is no previous study has included such a large cohort of infants enrolled in Medicaid. Nor has any study included such a broad range of outcomes within the first year of the infant's life. Our study found that for many of the outcomes the differences between the infants born to adolescent mothers versus adult mothers are not as large as previously believed. This indicates that there have been some changes that have improved situations for adolescent mothers and supports the research that stigma for adolescent pregnancy is unwarranted [134]. However, we do find that infants born to adolescent mothers had a higher number of emergency department visits (over 30% more visits) for infants with comparable health. Second, all infants were getting less wellness visits than recommended. We found a mean of 4.4 visits per a year instead of the recommended 7. This is a finding relevant to all mothers and suggests clinicians may not be informing mothers of the importance of preventive care for their infants.

In the first two studies, we established how advanced analytics and data pipelines on large data sets of administrative claims can be used to assess a health care policy and perform exact matching to determine causality. Next, we shifted to using machine learning to estimate a causal treatment effect of a health care policy. We use a motivating example with DID study setting for perinatal costs. The work extends existing machine learning methods to target the ATT. We demonstrate the performance of super learner through a

simulation in which compare its performance to the traditional outcome regression algorithm. The project applies the new approach to estimate the effects of episode-based payments on perinatal spending. We evaluate the algorithms under three effect sizes and two sample sizes. Our results show a decrease in bias when using super learner on the smaller effect sizes. Super learner also had a lower standard error and larger gains over the outcome regression for smaller sample sizes. This demonstrates the value of using machine learning, but the lack of improvement over all three effect sizes suggests a double robust technique may offer better results.

In the final project, presented in chapter 5, we apply machine learning to edge weight estimation for social network analysis. Network analysis is used to visualize and assess dependent relationships and within healthcare, social networks can be used to quantify the impact of social influence on healthcare interventions. Weight estimation can be a data quality tool to check if the weights provided in a network are correct. In this study, we derive 14 network metrics to include in the prediction of the edge weight. We alter current metrics made for unweighted graphs to account for the strength of the relationships between edges. We also incorporate metadata (i.e. covariates) through a similarity metric to account for node attributes. This technique leverages all available information for the weight prediction. We implement three different estimators and compare the results via precision error. The baseline estimator was Poisson regression and we used a Poisson finite mixture model and super learner as the comparative estimators. Super learner resulted in the lowest precision error of the three methods for our example of a faculty hiring network. Suggested future work will include applying these estimators to other data sets with different characteristics to further validate the technique. The final goal is to use this

method for a social network of Medicaid enrollees with substance dependence. We will analyze the patient-sharing network of their clinicians to evaluate health outcomes and how physician relationships may affect quality of care. We will also create and analyze a clinician-sharing network to try to identify abuses within the Medicaid system and what naturally occurring groups occur within patients.

REFERENCES

1. Reinsel, D., J. Ganrtz, and J. Rydning, *The Digitization of the World: From Edge to Core*. 2018, IDC

2. Foundation, K.F., *Total Monthly Medicaid and CHIP Enrollment*. 2020.

3. Carroll, C., et al., *Effects of episode-based payment on health care spending and utilization: Evidence from perinatal care in Arkansas.* Journal of Health Economics, 2018. **61**: p. 47-62.

4. Centers for Disease, C. and Prevention, *U S. Medical Eligibility Criteria for Contraceptive Use, 2010.* MMWR Recomm Rep, 2010. **59**(RR-4): p. 1-86.

5. Prevention, C.f.D.C.a. *US Medical Eligibility Criteria (US MEC) for Contraceptive Use*. 2016 [cited 2018s; Available from: https://www.cdc.gov/reproductivehealth/contraception/mmwr/mec/summary.html.

6. Russo, J.A., B.A. Chen, and M.D. Creinin, *Primary care physician familiarity with U.S. medical eligibility for contraceptive use.* Fam Med, 2015. **47**(1): p. 15-21.

7. Dehlendorf, C., et al., *Health care providers' knowledge about contraceptive evidence: a barrier to quality family planning care?* Contraception, 2010. **81**(4): p. 292-8.

8. Lindley, K.J., et al., *Contraceptive Use and Unintended Pregnancy in Women With Congenital Heart Disease.* Obstet Gynecol, 2015. **126**(2): p. 363-9.

9. Sun, M., et al., *Trends in contraceptive use among women with human immunodeficiency virus.* Obstet Gynecol, 2012. **120**(4): p. 783-90.

10. Massad, L.S., et al., *Contraceptive use among U.S. women with HIV.* J Womens Health (Larchmt), 2007. **16**(5): p. 657-66.

11. Finer, L.B. and M.R. Zolna, *Declines in Unintended Pregnancy in the United States, 2008-2011.* N Engl J Med, 2016. **374**(9): p. 843-52.

12. Foundation, K.F., *The Role of Medicaid for Adults with Chronic Illnesses*. 2013.

13. Institute, T.G. *Dramatic Gains in Insurance Coverage for Women of Reproductive Age Are Now in Jeopardy*. 2018.

14. Markus, A.R., et al., *Medicaid Covered Births, 2008 Through 2010, in the Context of the Implementation of Health Reform.* Women's Health Issues. **23**(5): p. e273-e280.

15. Daugherty, J., Ph.D., and Martinez, Gladys , Ph.D., *Birth Expectations of U.S. Women Aged 15–44.* NCHS Data Brie, 2016. **260**.

16. Reeves, S., et al., *Identifying sickle cell disease cases using administrative claims.* Acad Pediatr, 2014. **14**(5 Suppl): p. S61-7.

17. Westfall, J.M., MD, MPH; Mold, James , MD, MPH; Fagnan, Lyle , MD, *Practice-Based Research—"Blue Highways" on the NIH Roadmap.* JAMA, 2007. **297**(4): p. 403-406.

18. Kavanaugh, M.L.J., Jenna *Contraceptive method use in the United States: trends and characteristics between 2008, 2012 and 2014.* Contraception, 2018. **97**(1): p. 14-21.

19. Champaloux, S.W., et al., *Contraceptive Use Among Women With Medical Conditions in a Nationwide Privately Insured Population.* Obstet Gynecol, 2015. **126**(6): p. 1151-9.

20. Soma-Pillay, P., et al., *Physiological changes in pregnancy.* Cardiovasc J Afr, 2016. **27**(2): p. 89-94.

21. Collaborators, G.M.M., *Global, regional, and national levels of maternal mortality, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015.* The Lancet, 2015. **388**: p. 1775-1812.

22. Colorado Maternal Mortality Review Committee, D.M.M.R., Georgia Maternal Mortality, et al., *Building U.S. Capacity to Review and Prevent Maternal Deaths*. 2018.

23. Johnson, K., et al., *Recommendations to improve preconception health and health care--United States. A report of the CDC/ATSDR Preconception Care Work Group and the Select Panel on Preconception Care.* MMWR Recomm Rep, 2006. **55**(Rr-6): p. 1-23.

24. Elkayam, U., *Risk of Subsequent Pregnancy in Women With a History of Peripartum Cardiomyopathy.* Journal of the American College of Cardiology, 2014. **64**(15): p. 1629-1636.

25. Gynecologists, A.C.o.O.a., *Committee Opinion No. 654: Reproductive Life Planning to Reduce Unintended Pregnancy.* Obstet Gynecol, 2016. **127**(2): p. e66-9.

26. Women, C.o.H.C.f.U., *Access to Contraception*, in *American Journal of Obstetrics & Gynecology*. 2015. p. 250-5.

27. Bearak, J.M.P.J., Rachel K., PhD, *Did Contraceptive Use Patterns Change after the Affordable Care Act? A Descriptive Analysis.* Women's Health Issues, 2017. **27**(3): p. 316-321.

28. Pace LE1, D.S., Keating NL., *Early Impact of the Affordable Care Act on Uptake of Long-acting Reversible Contraceptive Methods.* PubMed, 2016. **54**(9): p. 811-7.

29. Services, U.D.o.H.a.H. *Title X Family Planning*. 2018; Available from: https://www.hhs.gov/opa/title-x-family-planning/about-title-x-grants/index.html.

30. *Evidence Summary: Prevent Unintended Pregnancy*. 2018; Available from: https://www.cdc.gov/sixeighteen/pregnancy/index.htm.

31. ACOG, *Understanding and using the U.S. Selected Practice Recommendations for Contraceptive Use, 2013.* Obstet Gynecol, 2013. **122**(5): p. 1132-3.

32. AAFG. *Clinical Practice Guideline: United States Medical Eligibility Criteria for Contraceptive Use*. 2016; Available from: https://www.aafp.org/patient-care/clinical-recommendations/all/contraceptive-use.html.

33. Ott, M.A. and G.S. Sucato, *Contraception for Adolescents.* Pediatrics, 2014. **134**(4): p. e1257.

34. Kearney, M.S. and P.B. Levine, *Why is the teen birth rate in the United States so high and why does it matter?* J Econ Perspect, 2012. **26**(2): p. 141-66.

35. Penman-Aguilar, A., et al., *Socioeconomic disadvantage as a social determinant of teen childbearing in the U.S.* Public Health Rep, 2013. **128 Suppl 1**: p. 5-22.

36. Perper, K.M.P., Kristen; Manlove, Jennifer PhD, *Diploma Attainment Among Teen Mothers.* Child Trends, 2010. **2010-01**.

37. Saul D. Hoffman, R.A.M., *Kids Having Kids; Economic Costs and Social Consequences of Teen Pregnancy.* Urban Institute, 2008.

38. Gortzak-Uzan, L., et al., *Teenage pregnancy: risk factors for adverse perinatal outcome.* J Matern Fetal Med, 2001. **10**(6): p. 393-7.

39. Chen, X.-K., et al., *Teenage pregnancy and adverse birth outcomes: a large population based retrospective cohort study.* International Journal of Epidemiology, 2007. **36**(2): p. 368-373.

40. Socolov, D.-G., et al., *Pregnancy during Adolescence and Associated Risks: An 8-Year Hospital-Based Cohort Study (2007-2014) in Romania, the Country with the Highest Rate of Teenage Pregnancy in Europe.* BioMed research international, 2017. **2017**: p. 9205016-9205016.

41. Gillman, M.W., *The first months of life: a critical period for development of obesity.* The American journal of clinical nutrition, 2008. **87**(6): p. 1587-1589.

42. Jornayvaz, F.R., et al., *Low birth weight leads to obesity, diabetes and increased leptin levels in adults: the CoLaus study.* Cardiovascular Diabetology, 2016. **15**: p. 73.

43. Haas, G.-M., E. Liepold, and P. Schwandt, *Low Birth Weight as a Predictor of Cardiovascular Risk Factors in Childhood and Adolescence? The PEP Family Heart Study.* International Journal of Preventive Medicine, 2015. **6**: p. 121.

44. Committee on Early Childhood, A.a.D.C., *Developmental Issues for Young Children in Foster Care.* Pediatrics, 2000. **106**(5): p. 1145-1150.

45. Spratt, E.G., et al., *The Effects of Early Neglect on Cognitive, Language, and Behavioral Functioning in Childhood.* Psychology (Irvine, Calif.), 2012. **3**(2): p. 175-182.

46. Sedgh, G., et al., *Adolescent Pregnancy, Birth, and Abortion Rates Across Countries: Levels and Recent Trends.* The Journal of adolescent health : official publication of the Society for Adolescent Medicine, 2015. **56**(2): p. 223-230.

47. Services, C.f.M.a.M. *Medicaid Eligibility*. 2018; Available from: https://www.medicaid.gov/medicaid/eligibility/index.html.

48. CMS, *2011 edition of the selected tables of the Data Compendium* 2011.

49. Society, C.P., *Age limits and adolescents.* Paediatrics & child health, 2003. **8**(9): p. 577-578.

50. Palmsten, K., et al., *Harnessing the Medicaid Analytic eXtract (MAX) to Evaluate Medications in Pregnancy: Design Considerations.* PLoS One, 2013. **8**(6): p. e67405.

51. *Medicaid Analytic Extracts (MAX) Documentation*, R.D.A. Center, Editor. 2018: Minneapolis, MN.

52. Blondel, B., et al., *The Impact of the Increasing Number of Multiple Births on the Rates of Preterm Birth and Low Birthweight: An International Study.* American Journal of Public Health, 2002. **92**(8): p. 1323-1330.

53. Zhang, S., et al., *Racial disparities in economic and clinical outcomes of pregnancy among Medicaid recipients.* Maternal and child health journal, 2013. **17**(8): p. 1518-1525.

54. Hughes, J., et al., *Clinical Risk Groups (CRGs): a classification system for risk-adjusted capitation-based payment and health care management*. Vol. 42. 2004. 81-90.

55. Parker, T., *Rural-Urban Continuum Codes*, E.R. Service, Editor. 2013.

56. Behnke, M. and V.C. Smith, *Prenatal Substance Abuse: Short- and Long-term Effects on the Exposed Fetus.* Pediatrics, 2013.

57. CDC, *Infant Mortality*. 2016.

58. Wier, L.M.M.P.H.Y., Hao, Ph.D., Owens, Pamela L. , Ph.D.; Washington, Raynard , Ph.D., M.P.H., *Overview of Children in the Emergency Department, 2010.* HEALTHCARE COST AND UTILIZATION PROJECT, 2013. **June 2013**.

59. Permanente, K. *Well-Child Visits and Immunizations*. 2014; Available from: https://wa.kaiserpermanente.org/healthAndWellness?item=%2Fcommon%2FhealthAndWellness%2Fchildren%2FchildVisits%2Foverview.html.

60. Larson, C.P., *Poverty during pregnancy: Its effects on child health outcomes.* Paediatrics & Child Health, 2007. **12**(8): p. 673-677.

61. Gavin, N.I., M.B. Benedict, and E.K. Adams, *Health service use and outcomes among disabled Medicaid pregnant women.* Women's Health Issues, 2006. **16**(6): p. 313-322.

62. Roth, J., et al., *The risk of teen mothers having low birth weight babies: implications of recent medical research for school health personnel.* J Sch Health, 1998. **68**(7): p. 271-5.

63. Olausson, P.O., S. Cnattingius, and B. Haglund, *Teenage pregnancies and risk of late fetal death and infant mortality.* BJOG: An International Journal of Obstetrics and Gynaecology, 1999. **106**(2): p. 116-121.

64. Foundation, K.F. *Medicaid Enrollment by Age*. 2013 [cited 2019; Available from: https://www.kff.org/medicaid/state-indicator/medicaid-enrollment-by-age/?dataView=1&currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D.

65. Pediatrics, B.F.A.A.o., *Recommendations for Preventive Pediatric Health Care*. 2017.

66. Smith, K.D., et al., *Improving the rate and quality of medicaid well child care exams in primary care practices.* J Okla State Med Assoc, 2010. **103**(7): p. 248-53.

67. Paul, D.A., et al., *Hospital Admission and Emergency Department Utilization in an Infant Medicaid Population.* Hospital Pediatrics, 2016. **6**(10): p. 587.

68. Pittard, W.B., 3rd, J.N. Laditka, and S.B. Laditka, *Associations between maternal age and infant health outcomes among Medicaid-insured infants in South Carolina: mediating effects of socioeconomic factors.* Pediatrics, 2008. **122**(1): p. e100-6.

69. Pomerantz, W.J., et al., *Characteristics of nonurgent emergency department use in the first 3 months of life.* Pediatr Emerg Care, 2002. **18**(6): p. 403-8.

70. Morrison, A.K., et al., *The Relationship Between Parent Health Literacy and Pediatric Emergency Department Utilization: A Systematic Review.* Academic pediatrics, 2013. **13**(5): p. 10.1016/j.acap.2013.03.001.

71. Development, N.I.o.C.H.a.H. *What are the risk factors for preterm labor and birth?* 2017 [cited 2018 8DEC2018]; Available from: https://www.nichd.nih.gov/health/topics/preterm/conditioninfo/who_risk.

72. Ickovics, J.R., et al., *Cluster Randomized Controlled Trial of Group Prenatal Care: Perinatal Outcomes Among Adolescents in New York City Health Centers.* American journal of public health, 2016. **106**(2): p. 359-365.

73. Grady, M.A. and K.C. Bloom, *Pregnancy outcomes of adolescents enrolled in a CenteringPregnancy program.* J Midwifery Womens Health, 2004. **49**(5): p. 412-20.

74. Raghavan, R., D.S. Brown, and B.T. Allaire, *Can Medicaid Claims Validly Ascertain Foster Care Status?* Child Maltreat, 2017. **22**(3): p. 227-235.

75. Angrist, J. and J. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*. 2008, Princeton, NJ: Princeton University Press;.

76. Wing, C., K. Simon, and R.A. Bello-Gomez, *Designing Difference in Difference Studies: Best Practices for Public Health Policy Research.* Annual Review of Public Health, 2018. **39**(1): p. 453-469.

77. Abrevaya, J., Y.-C. Hsu, and R.P. Lieli, *Estimating Conditional Average Treatment Effects.* Journal of Business & Economic Statistics, 2015. **33**(4): p. 485-505.

78. Callaway, B. and P. Sant'Anna, *Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment.* IDEAS Working Paper Series from RePEc, 2018.

79. Zimmert, M. and M. Lechner, *Nonparametric estimation of causal heterogeneity under high-dimensional confounding*. 2019.

80. Lunceford, J. and M. Davidian, *Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study.* Statistics in medicine, 2004. **23**: p. 2937-60.

81. Han, B., H. Yu, and M.W. Friedberg, *Evaluating the Impact of Parent-Reported Medical Home Status on Children's Health Care Utilization, Expenditures, and Quality: A Difference-in-Differences Analysis with Causal Inference Methods.* Health Services Research, 2017. **52**(2): p. 786-806.

82. Sant'Anna , P.H.C.Z., Jun B. , *Doubly Robust Difference-in-Differences Estimators.* arXiv, 2018.

83. Li, F.L., Fan *Double-Robust Estimation in Difference-in-Differences with an Application to Traffic Safety Evaluation.* Observational Studies, 2019. **5**.

84. Lu, C.N., Xinkun; Wager, Stefan *Robust Nonparametric Difference-in-Differences Estimation.* arXiv pre-print server, 2019.

85. Robins, J. and A.G. Rotnitzky, *Semiparametric Efficiency in Multivariate Regression Models with Missing Data.* Journal of The American Statistical Association - J AMER STATIST ASSN, 1995. **90**: p. 122-129.

86. Casale, A.S., et al., *"ProvenCareSM": a provider-driven pay-for-performance program for acute episodic cardiac surgical care.* Ann Surg, 2007. **246**(4): p. 613-21; discussion 621-3.

87. Amico, P., et al., *Episode-Based Payment for the Medicare Outpatient Therapy Benefit.* Archives of Physical Medicine and Rehabilitation, 2016. **97**(8): p. 1323-1328.

88. Agarwal, R., et al., *The Impact Of Bundled Payment On Health Care Spending, Utilization, And Quality: A Systematic Review.* Health Affairs, 2020. **39**(1): p. 50-57.

89. Wang, A., R.A. Nianogo, and O.A. Arah, *G-computation of average treatment effects on the treated and the untreated.* BMC Medical Research Methodology, 2017. **17**(1).

90. Schuler, M.S. and S. Rose, *Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies.* Am J Epidemiol, 2017. **185**(1): p. 65-73.

91. Weber, A.M., M.J. Van Der Laan, and M.L. Petersen, *Assumption Trade-Offs When Choosing Identification Strategies for Pre-Post Treatment Effect Estimation: An Illustration of a Community-Based Intervention in Madagascar.* Journal of Causal Inference, 2015. **3**(1).

92. van der Laan, M.J., M. Petersen, and W. Zheng, *Estimating the Effect of a Community-Based Intervention with Two Communities.* J Causal Inference, 2013. **1**(1): p. 83-106.

93. Rose, S. and S.L. Normand, *Double robust estimation for multiple unordered treatments and clustered observations: Evaluating drug‐eluting coronary artery stents.* Biometrics, 2018.

94. Brown, D.M., et al., *Occupational Exposure to PM2.5 and Incidence of Ischemic Heart Disease: Longitudinal Targeted Minimum Loss-based Estimation.* Epidemiology (Cambridge, Mass.), 2015. **26**(6): p. 806-814.

95. Blok, D.J., et al., *The role of smoking in social networks on smoking cessation and relapse among adults: A longitudinal study.* Preventive Medicine, 2017. **99**: p. 105-110.

96. Roberts, M.E., et al., *Adolescent social networks: general and smoking-specific characteristics associated with smoking.* Journal of studies on alcohol and drugs, 2015. **76**(2): p. 247-255.

97. Schaefer, D.R., J. Adams, and S.A. Haas, *Social networks and smoking: exploring the effects of peer influence and smoker popularity through simulations.* Health education & behavior : the official publication of the Society for Public Health Education, 2013. **40**(1 Suppl): p. 24S-32S.

98. Powell, K., et al., *The role of social networks in the development of overweight and obesity among adults: a scoping review.* BMC Public Health, 2015. **15**(1): p. 996.

99. Serrano Fuentes, N., A. Rogers, and M.C. Portillo, *Social network influences and the adoption of obesity-related behaviours in adults: a critical interpretative synthesis review.* BMC Public Health, 2019. **19**(1): p. 1178.

100. Brinkhues, S., et al., *Social Network Characteristics Are Associated With Type 2 Diabetes Complications: The Maastricht Study.* Diabetes Care, 2018. **41**(8): p. 1654-1662.

101. Heaney, C.A. and B.A. Israel, *Social networks and social support.* Health behavior and health education: Theory, research, and practice, 2008. **4**: p. 189-210.

102. Neaigus, A., et al., *Using dyadic data for a network analysis of HIV infection and risk behaviors among injecting drug users*, in *Social Networks, Drug Abuse, and HIV Transmission*, R. Needle, et al., Editors. 1995, United States Department of Health and Human Services, National Institute on Drug Abuse: Rockville, MD. p. 20-37.

103. Stattner, E. and N. Vidot. *Social network analysis in epidemiology: Current trends and perspectives*. in *2011 FIFTH INTERNATIONAL CONFERENCE ON RESEARCH CHALLENGES IN INFORMATION SCIENCE*. 2011.

104. Appel, A.P., et al., *A Social Network Analysis Framework for Modeling Health Insurance Claims Data.* 2018.

105. Herrin, J., et al., *An empiric approach to identifying physician peer groups from claims data: An example from breast cancer care*, in *Health Services Research*. 2019. p. 44+.

106. Landon, B.E., et al., *Using administrative data to identify naturally occurring networks of physicians.* Medical care, 2013. **51**(8): p. 715-721.

107. Landon, B.E., et al., *Variation in Patient-Sharing Networks of Physicians Across the United States.* JAMA, 2012. **308**(3): p. 265-273.

108. Leonard, C.E., et al., *The quality of Medicaid and Medicare data obtained from CMS and its contractors: implications for pharmacoepidemiology.* BMC health services research, 2017. **17**(1): p. 304-304.

109. Devoe, J.E., et al., *Electronic health records vs Medicaid claims: completeness of diabetes preventive care data in community health centers.* Annals of family medicine, 2011. **9**(4): p. 351-358.

110. Wasserman, S. and P. Pattison, *Logit Models and Logistic Regressions for Social Networks.* Psychometrika, 1996. **61**(3): p. 401-425.

111. Clauset, A., C. Moore, and M.E.J. Newman, *Hierarchical structure and the prediction of missing links in networks.* Nature, 2008. **453**: p. 98.

112. Zhao, Y.W., Yun-Jhoung; Levina, Elizaveta; Zhu, Ji, *Link Prediction for Partially Observed Networks.* JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS, 2017. **26**: p. 725-733.

113. Fire, M., et al. *Link Prediction in Social Networks Using Computationally Efficient Topological Features*. in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 2011.

114. Ahmad, I., et al., *Missing Link Prediction using Common Neighbor and Centrality based Parameterized Algorithm.* Scientific Reports, 2020. **10**(1): p. 364.

115. Wang, X.-W., Y. Chen, and Y.-Y. Liu, *Link Prediction through Deep Learning.* bioRxiv, 2018: p. 247577.

116. Crichton, G., et al., *Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches.* BMC Bioinformatics, 2018. **19**(1): p. 176.

117. Zhao, J., et al., *Prediction of Links and Weights in Networks by Reliable Routes.* 2015. **5**: p. 12261.

118. Hou, Y. and L.B. Holder. *Deep learning approach to link weight prediction*. in *2017 International Joint Conference on Neural Networks (IJCNN)*. 2017. Anchorage, AK, USA: IEEE.

119. Sa, H. and R. Prudêncio, *Supervised link prediction in weighted networks*. 2011. 2281-2288.

120. Aaron Clauset, E.T., Matthias Sainz. *The Colorado Index of Complex Networks*. 2016 [cited 2019; Available from: https://icon.colorado.edu/.

121. Eskin, E., et al., *A Geometric Framework for Unsupervised Anomaly Detection*, in *Applications of Data Mining in Computer Security*, D. Barbará and S. Jajodia, Editors. 2002, Springer US: Boston, MA. p. 77-101.

122. Boriah, S., V. Chandola, and V. Kumar. *Similarity Measures for Categorical Data: A Comparative Evaluation*. in *Society for Industrial and Applied Mathematics - 8th SIAM International Conference on Data Mining 2008, Proceedings in Applied Mathematics 130*. 2008. Atlanta, GA.

123. Adamic, L. and E. Adar, *Friends and Neighbors on the Web.* Social Networks, 2003. **25**: p. 211-230.

124. Dice, L.R., *Measures of the Amount of Ecologic Association Between Species.* Ecology, 1945. **26**(3): p. 297-302.

125. Clauset, A., M.E.J. Newman, and C. Moore, *Finding community structure in very large networks.* Physical Review E, 2004. **70**(6).

126. Schlattmann, P., *Medical Applications of Finite Mixture Models*. 2011.

127. McLachlan, G.J., S.X. Lee, and S.I. Rathnayake, *Finite Mixture Models.* Annual Review of Statistics and Its Application, 2019. **6**(1): p. 355-378.

128. Gudicha, D.W. and J.K. Vermunt. *Mixture Model Clustering with Covariates Using Adjusted Three-Step Approaches*. in *Algorithms from and for Nature and Life*. 2013.

129. Polley, E.C., M.J. van der Laan, and S. Rose, *Super Learning*, in *Targeted Learning: Causal Inference for Observational and Experimental Data*. 2011.

130. Rezvanian, A. and M.R. Meybodi, *Sampling social networks using shortest paths.* Physica A: Statistical Mechanics and its Applications, 2015. **424**: p. 254-268.

131. Hinami, K., et al., *Prescribing Associated with High-Risk Opioid Exposures Among Non-cancer Chronic Users of Opioid Analgesics: a Social Network Analysis.* J Gen Intern Med, 2019. **34**(11): p. 2443-2450.

132. Perry, B.L., et al., *Co-prescription network reveals social dynamics of opioid doctor shopping.(Research Article)(Report).* PLoS ONE, 2019. **14**(10): p. e0223849.

133. Finer, L.B. and M.R. Zolna, *Declines in Unintended Pregnancy in the United States, 2008–2011.* New England Journal of Medicine, 2016. **374**(9): p. 843-852.

134. Ellis-Sloan, K., *Teenage Mothers, Stigma and Their 'Presentations of Self'.* Sociological Research Online, 2014. **19**(1): p. 9.