INTERGENERATIONAL EFFECTS OF MATERNAL HEALTH ON PREGNANCY AND NEONATAL OUTCOMES IN NOVA SCOTIAN CHILDREN

by

Mary M. Brown

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
July 2022

For Mike

Table of Contents

List of Tables

List of Figures

Abstract

Maternal pre-pregnancy body mass index (BMI) is associated with first-generation health outcomes. The literature suggests an increased risk of low birthweight in infants born to mothers who are underweight, while infants born to mothers with obesity have increased risk of high birthweight, and of becoming obese themselves. Moderate associations between grandparental factors and child birthweight have been reported, but several studies have limitations affecting validity and precision and seldom examined mediation by first-generation factors. Two objectives of this research were to 1) examine the association between grandmaternal (G0) pre-pregnancy BMI and child (G2) birthweight, with investigation of mediation by maternal (G1) pre-pregnancy BMI, and 2) develop a prediction model for G2 fetal growth abnormalities using G0 risk factors, G1 birth characteristics, and G1 pregnancy characteristics in nulliparous G0s and G1s. These objectives were addressed using a subset of the Nova Scotia Atlee Perinatal Database (NSAPD) created by linking women's birth information with their pregnancy information in adulthood. The clustering structure of the NSAPD, where delivery-level data is nested within women, creates challenges when imputing missing data. The third objective was to assess the use of a recently proposed tree-based method, mixed-effects random forest (MERF), which incorporates clustering in the prediction procedure to impute BMI. This study found imputation using MERF was moderately biased when BMI was missing at random but severely biased when missing not at random, and imputation using standard random forest was least biased and most efficient. In analyses of 20822 G1-G2 dyads, estimates of the total effect of G0 pre-pregnancy BMI on G2 birthweight z-score and the mediator-specific effect via G1 pre-pregnancy BMI, assuming G0s had a BMI of 22 kg/m^2 as compared to the 'natural course' scenario, were small. G0 factors and G1 birth characteristics, together with G1 characteristics, modestly improved the prediction of fetal growth abnormalities as compared to models based solely on G1 characteristics in a sample of 9068 G2s. Key predictors included G1 gestational weight gain, pre-pregnancy BMI and birthweight z-score. These findings suggest negligible intergenerational effects of G0 pre-pregnancy BMI on G2 birthweight, but moderate predictive ability of G1 size at birth.

List of Abbreviations and Symbols Used

$Y_i(a)$ Potential outcome for subject i that would have been observed under exposure level a.

$Y_i(a, m)$ Potential outcome for subject i that would have been observed under exposure level a and mediator level m.

$\boldsymbol{R}$ $n \times p$ logical matrix where $r_{ij} = 1$ if y_{ij} is observed and 0 otherwise.

$\boldsymbol{Y}$ $n \times p$ matrix of sample data.

$\boldsymbol{Y}_{mis}$ Missing data of $\boldsymbol{Y}$.

$\boldsymbol{Y}_{obs}$ Observed data of $\boldsymbol{Y}$.

y_{ij} jth data value for the ith unit for $i \in (1, n)$ and $j \in (1, p)$.

AGA Appropriate for gestational age.

AGREMA A Guideline for Reporting Mediation Analyses.

AUC-PR Area under the precision-recall curve.

AUC-ROC Area under the receiver operating characteristic curve.

BMI Body mass index.

CART Classification and regression tree.

CCA Complete-case analysis.

CI Confidence interval.

CV Cross-validation.

DAG Directed acyclic graph.

DNA Deoxyribonucleic acid.

DOHaD Developmental Origins of Health and Disease.

FN False negative.

FP False positive.

G0 Identifier for the grandmaternal generation.

G1 Identifier for the maternal generation.

G2 Identifier for the child generation.

GAM Generalized additive model.

GDM Gestational diabetes mellitus.

GEE Generalized estimating equations.

GLL Generalized log-likelihood.

HBW High birthweight.

LASSO Least absolute shrinkage and selection operator.

LBW Low birthweight.

LGA Large for gestational age.

LM Normal linear model.

LMER Linear mixed-effects model.

MAR Missing at random.

MCAR Missing completely at random.

MERF Mixed-effects random forest.

MERT Mixed-effects regression tree.

MICE Multivariate imputation using chained equations.

MNAR Missing not at random.

MVNI Multivariate normal imputation.

NDE Natural direct effect.

NIE Natural indirect effect.

NPSEM Nonparametric structural equation model.

NPV Negative predictive value.

NSAPD Nova Scotia Atlee Perinatal Database.

OR Odds ratio.

PMM Predictive mean matching.

PMSE Predictive mean squared error.

PPV Positive predictive value.

PR Precision-recall.

QAIPPE Quintile of Annual Income Per Person Equivalent.

RF Random forest.

ROC Receiver operating characteristic.

SD Standard deviation.

SE Standard error.

SEM Structural equation modeling.

SES Socioeconomic status.

SGA Small for gestational age.

SMD Standardized mean difference.

SVM Support vector machine.

TE Total (causal) effect.

TMLE Targeted Maximum Likelihood Estimation.

TN True negative.

TP True positive.

TRIPOD Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

WHO World Health Organization.

WHR Waist-to-hip ratio.

XGBoost Extreme Gradient Boosting.

Acknowledgements

I would like to first and foremost thank and acknowledge my supervisory team. Dr. Christy Woolcott, Dr. Bruce Smith, Dr. Stefan Kuhle, Dr. Jennifer Payne, and Dr. Victoria Allen, I can easily say that I would not have completed this thesis without your unwavering support, kindness, motivation, and confidence in me. Thank you for your wisdom, patience, and good humour that pushed me to work harder on my good days, and helped me to persevere through my hard days (and trust me, there were a lot of those). Thank you for allowing me to grow and learn, letting me ask both thoughtful and not-so thoughtful questions, and empowering me to figure things out on my own. Thank you.

I extend my thanks to the Reproductive Care Program of Nova Scotia and Dr. Alexander (Alec) Allen (1933-2018) whose extraordinary vision to establish a population-based perinatal database in the 1980s has made the work in this thesis possible. And of course, a special thanks to John Fahey for always making me laugh and reminding me to be silly at least once a day.

I am grateful for the generous support from the Nova Scotia Health Research Foundation (Scotia Scholars Award), the province of Nova Scotia (Nova Scotia Research and Innovation Trust), and Dalhousie University.

I wish to thank my parents, my husband Alex, and many friends for their patience, support, inspiration, and laughs over the last five years. Thank you for keeping me afloat, and reminding me to eat, sleep, and leave the house sometimes. I owe it to my goofy pup Guinness for always giving me a reason to smile, and for making tough days and pandemic life bearable. Lastly, I am eternally grateful for the love and support from my late grandfather Melbourne "Mike" Brown (1927-2022), to whom this dissertation is dedicated. Thank you for being an exceptional role model, imparting me with a love of reading and learning, and showing me first-hand the importance of living life to the fullest.

Chapter 1

Introduction

The rising trend in overweight and obesity observed in the Canadian population poses risks for women before, during, and after pregnancy, as well as for their offspring. In 2015, 22% of Canadian women aged 18 to 34 were affected by overweight and 19% by obesity[1]. Women entering pregnancy with overweight or obesity (i.e., pre-pregnancy overweight or obesity) are at increased risk of adverse maternal, fetal, and neonatal outcomes[2], and require specialist care, and additional healthcare services, resulting in higher maternity costs for these women[3]. These issues are especially concerning in Nova Scotia, one of Canada's four Atlantic provinces, where more than half of the women entering pregnancy were either overweight or obese in 2019[4]. The prevalence of underweight among Canadian women aged 18 to 34 increased from 4.8% in 2005 to 6.6% in 2015[1], but the prevalence of pre-pregnancy underweight in Nova Scotia has remained relatively constant at approximately 4% between 2015 and 2019[4].

Elevated and low body mass index (BMI) before becoming pregnant is associated with a higher risk of adverse pregnancy and obstetrical outcomes[2, 5, 6]. Infants born to underweight mothers are at increased risk of low birthweight (LBW) and being born small for gestational age (SGA), while infants born to mothers with overweight or obesity have increased risk of high birthweight (HBW), being born large for gestational age (LGA)[2, 5], and of becoming overweight and obese in childhood and adolescence[6]. The observed associations between maternal pre-pregnancy BMI and offspring BMI, and the heritability of weight via genetic and epigenetic mechanisms[7] have suggested the possibility of an effect of grandmaternal pre-pregnancy BMI on child birthweight.

Multigenerational studies on the effects of in utero exposures on second-generation outcomes[8] have found small to moderate associations between grandparental risk

factors and child birthweight, including grandparental birthweight[9, 10], BMI[11–13], smoking in pregnancy[14–19], socioeconomic status[20–23], and diabetes[24, 25], with most studies focusing on the maternal line. The results of two studies that examined the association between maternal grandmother pre-pregnancy BMI and child birthweight suggested no large differences in child birthweight with each unit (kg/m^2) increase[12, 13]. Both studies, however, were limited by small sample sizes, and one study inappropriately adjusted for mediators of the association[12], thus potentially biasing the total effect estimate.

Maternal factors are likely to lie on the causal pathway between grandmaternal body weight and child birthweight, but few studies have conducted mediation analyses to investigate the role of maternal characteristics in these associations. Moreover, only traditional methods of mediation analysis have been used, which are known to be limited in many settings[26]. More modern counterfactual-based approaches are more flexible than traditional methods, and allow for more intuitive interpretations of the estimates by considering hypothetical conditions on the exposure. However, causal interpretation of estimates obtained using counterfactual-based approaches require strong and untestable assumptions, which are likely to be violated when assessing the effects of exposures such as body weight and BMI[27]. Nonetheless, pre-pregnancy BMI remains an important risk factor in pregnancy; examining its relationship with second-generation offspring birthweight and assessing the mediating effect of maternal pre-pregnancy BMI is valuable for guiding future research on specific interventions that could modify BMI.

As previously mentioned, the results of some studies have suggested an association between grandparental risk factors and child birthweight, but it remains unclear whether intergenerational information can aid in identifying pregnancies at greatest risk for fetuses with restricted or excessive growth. Accurate identification of these pregnancies has important implications for preconception counselling, antenatal assessment and intrapartum care. The risk of abnormal fetal growth is increased in infants born to mothers at the extremes of the BMI distribution, which, in turn, is associated with adverse health outcomes in infants[28]. Prediction models for fetal growth abnormalities have been developed using routinely collected data readily available in an antenatal setting but predictive performance remains relatively poor,

especially among women in their first pregnancy. Despite the well-established associations between maternal and offspring size-at-birth[29], a prediction model for fetal growth abnormalities based on maternal birth characteristics and grandmaternal risk factors, together with maternal pregnancy-related information, has yet to be developed and validated.

Two contributions of this thesis are concerned with the relationship between grandmaternal pregnancy-related characteristics and fetal growth. One with the goal of estimating the association between grandmaternal pre-pregnancy BMI and infant birthweight, with examination of mediation by maternal pre-pegnancy BMI, and the other with the goal of developing and validating a prediction model for fetal growth abnormalities using grandmaternal risk factors, maternal birth characteristics, and maternal pregnancy characteristics. These objectives were addressed using the 3G Multigenerational Cohort of Nova Scotian women and their offspring, a subsample of the Nova Scotia Atlee Perinatal Database (NSAPD), created by linking women's birth information with their pregnancy information in adulthood[30].

Like other perinatal databases, the NSAPD has a unique clustering structure where delivery records to the same woman are linked, resulting in a hierarchical structure of delivery-level data nested within women. Notable design aspects of this database are that women have differing numbers of deliveries (i.e., unequal cluster sizes), the time between deliveries varies (i.e., unequal spacing of measurements), and many women have only one delivery (i.e., large proportion of singleton clusters). These databases are also prone to missingness, particularly in maternal pre-pregnancy weight and height (information required to calculate BMI), and appropriately handling missingness is complicated by this complex clustering structure. Several studies have evaluated and compared multilevel imputation methods under varying conditions[31–33], but it remains unclear which method is best in data with small and unbalanced clusters.

Imputation using data-adaptive methods has been proposed as an alternative to parametric imputation in the case of independent observations[34–38]. These methods can capture complex relationships in the data without the need to explicitly specify the imputation model, and have been shown to perform comparably or better than parametric-based imputation[34, 36–39]. Tree-based algorithms are among the

most popular machine learning algorithms for prediction, some of which have been extended to accommodate clustered data for continuous outcomes[40–42]. Due to the complex structure of the NSPAD and considering weight measurements over time are likely to be correlated, imputation methods using multilevel tree-based algorithms may outperform other imputation strategies. Therefore, the final contribution of this thesis is the evaluation of a recently proposed tree-based algorithm that can accommodate clustering as a method for imputing pre-pregnancy BMI using real-life data drawn from the NSAPD.

This dissertation is presented in the form of a series of related manuscripts. Due to the manuscript-based nature of this thesis, some repetition occurs between chapters. Chapter Two states the objectives of this research. Chapter Three reviews the literature relevant to maternal pre-pregnancy BMI and associated maternal and neonatal outcomes, followed by the necessary background information required to address each objective. Chapters Four through Six are the individual manuscripts that have been or are in the process of being submitted for publication. Lastly, Chapter Seven provides a discussion of the overall findings, strengths and limitations, and highlights possible implications of the findings to policy and decision-making settings.

In some chapters of this dissertation, the grandmaternal generation is referred to as G0, the maternal generation as G1, and the second-generation offspring as G2. The decision to use this notation was based on the context and the target academic journal.

Chapter 2

Objectives

The overarching goal of this research was to examine the relationship of grandmaternal (G0) pre-pregnancy BMI and other pregnancy-related factors with second-generation (G2) offspring birthweight using the 3G Multigenerational Cohort. Specifically, the three primary objectives were to:

1. Investigate the performance of a recently proposed tree-based algorithm that accommodates clustering as a method for imputing pre-pregnancy BMI values in the NSAPD.

2. Examine the association between G0 pre-pregnancy BMI and G2 birthweight, with an investigation of mediation by maternal (G1) pre-pregnancy BMI.

3. Develop and validate a prediction model for G2 fetal growth abnormalities using G0 pregnancy-related factors and G1 birth characteristics together with G1 pregnancy-related factors.

Chapter 3

Literature Review

This chapter provides a definition of pre-pregnancy BMI and describes associated maternal and neonatal outcomes. Secondly, a literature review of the current evidence for the transmission of weight, including an overview of the epigenetic evidence by which grandmaternal pre-pregnancy BMI may influence child birthweight, is presented. Thirdly, methodologies for multiply imputing maternal pre-pregnancy BMI (used in Chapter Four), and for estimating total and mediation effects (used in Chapter Five), are discussed. Lastly, the limitations of current prediction models for fetal growth abnormalities are explored, with this last section providing an overview of the methodology used in Chapter Six.

3.1 Maternal pre-pregnancy BMI

BMI, a measure of weight adjusted for height, is often used as a surrogate measure of body adiposity. The most widely used BMI classification, followed by Canadian and World Health Organization (WHO) guidelines, are as follows: BMI < 18.5 kg/m^2 as underweight, BMI between 18.5 and 24.9 kg/m^2 as normal weight, BMI between 25.0 and 29.9 kg/m^2 as overweight, and BMI ≥ 30 kg/m^2 as obese[43]. Overweight is also sometimes defined as BMI ≥ 25 kg/m^2. Maternal pre-pregnancy BMI is the BMI of a woman immediately prior to becoming pregnant. Maternal pre-pregnancy BMI is often represented by a measurement done in early pregnancy at the first prenatal visit. Given that weight gain during the first trimester is, on average, slow[44], measurements done in early pregnancy are valid.

Pre-pregnancy BMI is a potentially modifiable risk factor for many pregnancy, obstetrical and neonatal outcomes: pregnancy loss[45], maternal thromboembolism[46], problems in labour such as shoulder dystocia[47], need for Caesarean delivery[2], preterm birth[2], and abnormal fetal growth[28]. Maternal BMI and some of the neonatal outcomes associated with it has been linked to obesity, diabetes, and cardiovascular

disease in the offspring in adulthood[28, 48, 49].

3.2 Evidence for the association between grandmaternal and child body weight, and for mediation by maternal body weight

Birthweight and weight in childhood are associated with adverse cardiovascular and metabolic outcomes. While the association between maternal and offspring BMI is established, the evidence for an association between grandparental and child body weight is limited in both amount and quality. In the following section, epidemiological studies investigating the total effect of grandmaternal body weight measures on offspring body weight measures and those estimating the mediated effect by maternal body weight measures are discussed. Next, epigenetic evidence that could be a biological mechanism for the potential effect of grandmaternal body weight measures on child body weight measures is presented.

3.2.1 Association between grandmaternal and child body weight

The association between maternal pre-pregnancy BMI and the short- and long-term health of the first-generation offspring has been established. A meta-analysis reported an increased odds of being born SGA (OR [odds ratio] 1.55, 95% confidence interval [CI] [1.49, 1.62]) and LBW (OR 1.51, 95% CI [1.48, 1.54]) in underweight mothers, and an increased odds of being born LGA (OR 2.36, 95% CI [2.17, 2.56]) and HBW (OR 2.28, 95% CI [2.15, 2.41]) in mothers with obesity, relative to normal weight mothers. Beyond the first-generation offspring, a meta-analysis of seven studies totalling 23 033 participants reported an increased odds of overweight or obesity in children with overweight or obese grandparents compared to those with normal weight grandparents (OR 1.79, 95% CI [1.01, 2.57])[50]. Intergenerational associations between grandparental body weight measurements and first- and second-generation offspring body weight measurements have been shown to be stronger along the maternal line than along the paternal line[11, 12], thus suggesting that maternal genetics and the intrauterine environment play a key role in explaining the observed associations.

The mechanisms underlying the intergenerational transmission of birthweight remain unclear but are thought to include a combination of genetics, epigenetics, and

prenatal environmental factors. Prenatal environmental factors that are predictors of birthweight and fetal growth include diabetes[24] and smoking[18]. While studies have demonstrated multigenerational patterns of birthweight[12, 51, 52], prospective studies examining the persistence of the effects of in utero exposures beyond the first-generation in human populations are scarce.

Evidence from studies investigating grandparental body weight-related measures (e.g., birthweight, BMI, waist circumference) and offspring body-weight related measures (e.g., birthweight, BMI in childhood) is mixed with most reporting weak or null associations (Table A.1). Overall, most studies have used a cross-sectional design, but vary in terms of sample size, time frame for ascertainment of exposure and outcome, adjustment variables, and statistical analysis. Studies with cross-sectional designs are limited because the exposure and outcome are measured at the same time, making it impossible to establish a temporal link between the exposure and outcome, and infer causal relationships.

The intergenerational transmission of size-at-birth was investigated by some studies. An unadjusted correlation of -0.41 (p=0.37) between maternal grandmother birthweight and child birthweight was reported in a small study of 34 female students from Tokyo[53]. In a larger sample of 6169 second-generation offspring from the Uppsala Birth Cohort Multigenerational Study, a small correlation of 0.124 (95% CI [0.095, 0.153]) was found between maternal grandmother birthweight and child birthweight, with this correlation being the largest amongst all grandparents (i.e., paternal grandmother, and grandfathers). No information on the parental generation was available to the authors of this study, so investigation of mediation by parental characteristics was impossible. Moreover, participants with missing information on important confounders, such as grandparental smoking, were dropped in the adjusted analyses, leading to a large reduction in sample size and potential selection bias by conducting complete case analyses.

Another study used prospectively collected data in the Aberdeen Maternity Neonatal Databank to examine the transmission of both birthweight and fetal growth across three generations[10]. After adjusting for sociodemographic and prenatal covariates in all three generations, the estimated association between grandmaternal birthweight z-score and child birthweight z-score was 0.17 standard deviation (SD)

units (95% CI [0.12, 0.23]). This estimate decreased to 0.12 SD units (95% CI [0.07, 0.18]) after further adjusting for maternal birthweight, height, and BMI. Both adjusted models included covariates that are likely to lie on the causal pathway between grandmaternal birthweight and child birthweight, such as maternal birthweight, therefore biasing the estimate of the total association. Smoking information, a likely confounder in the analyses, was only available after 1965 and, was therefore missing for approximately 90% and 30% of the pregnancies of the great-grandmothers and grandmothers, respectively. Lastly, the authors dummy-coded participants with missing data into a separate category. This popular method to accommodate missing data is known to produce biased estimates[54].

Beyond birthweight, a few studies have specifically examined grandmaternal pre-pregnancy weight or BMI and offspring birthweight. A study of Maltese women born in 1987 and who, as adults, delivered at the same hospital, found that infants born to mothers with in utero exposure to overweight or obesity were, on average, 280 g (95% CI [149, 411]) heavier at birth than infants born to mothers without in utero exposure to overweight or obesity[55]. However, an early study from 1992 found no association between maternal grandmother pre-pregnancy weight and child birthweight using data from the National Child Development Study[51]. Similarly, later studies using the Bogalusa Heart Study and the Isle of Wight birth cohort did not find large differences in offspring birthweight with each unit (kg/m^2) increase in BMI (-12 g, 95% CI [-31.6, 8.0][12], and 8 g [p=0.32][13]). All studies were limited by small sample sizes, two studies were unable to control for all relevant confounders[13, 55], and two studies inappropriately adjusted for mediators of the association[12, 51], thus potentially biasing the total effect estimate.

The influence of grandmaternal pre-pregnancy BMI on second-generation birthweight has not been thoroughly investigated using prospectively collected data. As most studies of this association had a cross-sectional design, the temporal ordering of exposure and outcome was impossible, which limits the interpretation of the associations found. The current available evidence is also limited by small sample sizes, the use of self-reported information, and inappropriate treatment of missing data. Most importantly, recent studies were unable to control for relevant confounders, such as grandparental smoking, and improperly adjusted for mediating factors.

3.2.2 Maternal weight as a potential mediator

The evidence for an intergenerational effect of weight is mixed and warrants further investigation. Maternal weight measures, such as birthweight and pre-pregnancy weight, are likely to fall on the pathway between grandmaternal body weight at various time points and child birthweight, but the strength of their mediating effect remains unclear. Investigating the effect of grandmaternal weight on child birthweight that operates via maternal weight can help clarify the underlying causal mechanisms involved.

Mediation analyses are used to assess the degree to which the relationship between an exposure and outcome is mediated by another variable (i.e., a mediator). Consider the simple example of mediation shown in Figure 3.1. The direct effect measures the extent to which the outcome changes in response to changing the exposure while holding the mediator fixed and is represented by the pathway "Exposure $\rightarrow$ Outcome". The indirect effect measures the extent to which the outcome changes as a result of changing the mediator while holding the exposure fixed and is represented by the pathway "Exposure $\rightarrow$ Mediator $\rightarrow$ Outcome". Causal language (i.e., total effect, direct effect, and indirect effect) was used in the subsequent discussion to indicate which effect types are being targeted in the analyses, but these are in fact statistical associations that require strong and carefully assessed assumptions to be interpreted causally. The details of the statistical approaches to estimate direct and indirect effects and the assumptions required to interpret these as causal effects are given in Section 3.4.

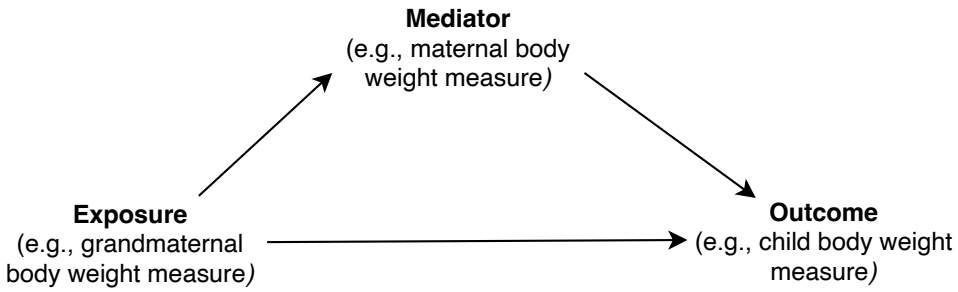


Figure 3.1: An example of a simple mediation analysis

Two studies investigating the intergenerational effect of grandmaternal body

weight reported on mediation by maternal body weight[10, 13]. Lahti-Pulkkinen et al.[10] used prospectively collected data to investigate the relationship between maternal grandmother birthweight z-score and child birthweight z-score. Using a partially adjusted model, the total effect was estimated to be 0.17 SD units (95% CI [0.12, 0.23]), or, for each increase in 1 SD unit in grandmaternal birthweight z-score, child birthweight z-score is expected to increase, on average, by 0.17 SD units. Using maternal birthweight z-score as a mediator, the indirect effect was estimated to be 0.06 SD units (95% CI [0.04, 0.06]). Since only birthweight variables were considered in the mediation analysis, the validity of the results is threatened by the potential confounding of the exposure-outcome, exposure-mediator, and mediator-outcome relationships. Failure to control for such variables leaves the possibility that these associations explain the significant direct and indirect effects observed. Furthermore, the authors did not account for possible exposure-induced mediator outcome confounders (i.e., intermediate confounders), such as grandmaternal diabetes, did not investigate the possibility of exposure-mediator interaction, and did not assess the possibility of nonlinear associations (e.g., between grandmaternal and child birthweight z-score), which may have resulted in invalid inference[56, 57].

In another study, the total effect of maternal grandmother BMI on child birthweight was estimated to be 8 g per kg/m^2 increase[13]. Using structural equation modeling (SEM), the estimated coefficient of the pathway via maternal body weight measures (i.e., indirect effect through maternal birthweight and BMI at age 18) was 6.6 g per kg/m^2 (p=0.04), and the coefficient of the pathway not via maternal body weight measures (i.e., direct effect) was 1.3 g per kg/m^2. This study was limited in its ability to control for relevant confounders (e.g., grandmother's age at the time of the mother's delivery), and is it unclear whether possible intermediate confounders (e.g., grandmaternal gestational diabetes mellitus [GDM] and hypertension) were addressed. Stepwise variable selection was used that, when combined with a small sample size, likely resulted in the exclusion of several important pathways (e.g., maternal birthweight to BMI at age 18) and may have introduced residual confounding bias. Lastly, SEM approaches, like the traditional regression-based approaches to mediation, assume all relationships are linear (or log-linear in the case of a binary outcome), which may be unrealistic when modeling associations with BMI.

To date, the investigation of the role of maternal pre-pregnancy weight in the association between grandmaternal pre-pregnancy weight and child birthweight, has been insufficient. Studies investigating mediation by maternal characteristics have relied on SEM approaches, which are only valid under strong, and often unrealistic, assumptions. No studies of this association have used newer counterfactual-based approaches to mediation that can accommodate intermediate confounding and allow for more intuitive interpretations of the direct and indirect effects by considering hypothetical conditions on the exposure.

3.2.3 Epigenetic mechanisms of maternal pre-pregnancy BMI

There is growing support for the Developmental Origins of Health and Disease (DOHaD) hypothesis that is concerned with exposures to environmental factors during critical periods of development and their effect on the short- and long-term health of the offspring. Several epidemiological studies have demonstrated a relationship between the early nutritional environment of a growing fetus and the development of obesity in adulthood[58]. For example, a study investigated the effects of in utero undernutrition on babies exposed to the Dutch Famine in 1944-45 who would later be faced with an abundance of food[59]. Offspring born to women exposed in their first two trimesters had a lower birthweight, but a higher incidence of obesity later in life compared to the general population.

The complex processes that mediate how these early life exposures impact later life are thought to be partly epigenetic in nature. In *Obesity Before Birth: Maternal and Prenatal Influences on the Offspring*, epigenetics is described as heritable but reversible changes in gene expression that do not directly alter the DNA sequence itself[60]. Epigenetic mechanisms include DNA methylation and post-translational histone modifications. DNA methylation occurs when a methyl group is added at a cystosine base that precedes a guanine in the DNA strand, known as the CpG dinucleotides. Histones are the proteins located inside the nucleus of the cell that package and order the DNA and are referred to collectively as chromatin (DNA wrapped around histones). Types of histone modifications that regulate the chromatin structure and, as a result, gene activity, include acetylation, methylation, phosphorylation, and ubiquitination.

During the formation and development of the embryo, the DNA is first hypomethylated, and later, DNA methylation increases, which leads to cell differentiation and organogenesis[61]. Highly methylated regions of the DNA are inaccessible to the enzymes responsible for transcription and gene expression, thus resulting in gene silencing. Genomic imprinting is a well-studied epigenetic phenomenon due to DNA methylation whereby the phenotype is modified depending on the parental sex contributing the allele[60]. For most autosomal genes, expression occurs from both alleles (one from each parent) simultaneously, but when genomic imprinting is present, gene expression occurs from only one allele. Post-translational modifications of the histones typically include methylation or demethylation and acetylation or deacetylation of the lysine residues on the histone tails. Histone methylation can result in either an increase or decrease in transcription whereas changes in acetylation suppress gene expression.

Animal models have provided the greatest evidence for the role of epigenetic changes in the development of obesity in the offspring. For example, consider the animal model of the *agouti* mouse exhibiting the *agouti* viable yellow mutation. The *agouti* gene is normally methylated, which results in a thin phenotype and a brown coat colour. However, demethylation of the *agouti* gene promotes gene expression, thus resulting in a coat color change to yellow and inducing obesity[61]. Relative to human models, animal models allow multiple generations of offspring to be studied quickly and, most importantly, allow researchers to assess the intergenerational transmission of obesity risk.

Several small human studies have investigated the association between maternal pre-pregnancy BMI and epigenetic markers in the mother, placenta, and her offspring, but the evidence is inconclusive[62]. Studies of the association between maternal underweight and offspring health, or epigenetic signatures related to maternal underweight, are scarce[63]. This may be due to underweight being more infrequent than overweight and obesity. One possible theory related to in utero exposure to maternal underweight is the "thrifty phenotype hypothesis." This hypothesis posits that when the intrauterine environment is nutritionally poor, the fetus adapts by increasing metabolic efficiency to increase the chance of short-term survival in a post-natal nutritionally poor environment. However, when the fetus is

in fact exposed to a normal nutrient exposure post-natally, those adaptations can be harmful, leading to alterations in energy balance, which increases predisposition to metabolic disorders in adulthood[64, 65].

Some evidence suggests that the epigenetic mechanisms involved in maternal programming of obesity can be passed across multiple generations. It is important to clarify the difference between intergenerational and transgenerational effects. Intergenerational effects include effects on the developing embryo and its germline (i.e., first- and second-generations) as opposed to transgenerational effects, which are those that persist in generations that were not exposed to the initial insult (i.e., third generation)[66]. Intergenerational effects from mother to offspring have been well studied using both human and animal models, but studies of the effects that persist into the second generation and beyond are limited. Although research in this area is still growing, most of the current evidence is derived using rodents. For example, a study found that pregnant mice continuously fed high-fat diets led to an increase in birthweight and adiposity across three generations with the greatest effect observed in the grand-offspring generation[67]. Alternatively, another study reported that offspring and grand-offspring of a rat fed a high-fat diet (with the intermediate generation fed only a chow diet) had increased body weight compared to controls with the effect less pronounced in the grand-offspring[68].

Primary mechanisms that mediate parental programming effects are thought to be the epigenetic state of sperm and oocyte[69]. Any in utero exposure not only influences the growing fetus, but also the germline of the developing embryo. This can lead to changes in the epigenome, thus resulting in a phenotypic change in the offspring that may eventually develop from these gametes. As mentioned above, the developing embryo undergoes a state of hypomethylation followed by hypermethylation that leads cells to differentiate into their respective tissues. This can be viewed as a sort of epigenetic erasure that could eliminate any epigenetic changes that had been passed along. It has been recently shown, however, that this reprogramming is not complete and some genomic sequences and associated epigenetic marks are resistant to reprogramming[69].

3.3 Imputing maternal pre-pregnancy weight information

The NSAPD, like other birth registers and perinatal databases, has a unique clustering structure where delivery records to the same individuals can be linked, resulting in a hierarchical structure of delivery-level data nested within individuals. These databases, however, are prone to missingness, particularly in maternal pre-pregnancy weight, which creates challenges when examining the effects of maternal pre-pregnancy BMI. Missing data are often addressed using multiple imputation, which accounts for the uncertainty of the missing values by creating multiple complete datasets.

Multivariate imputation by chained equations (MICE) is a population imputation procedure that enables imputation using parametric-based approaches (e.g., linear regression) or nonparametric-based approaches, such as machine learning techniques (e.g., random forest). However, in the case of clustered data, little work has been done to extend these techniques by, for example, incorporating random effects. Random effect terms model the extent to which average trends in the data vary across levels of a group factor (e.g., women in perinatal databases) and are included in mixed-effects models to account for the fact that differences may exist between the behaviour of the cluster and the average effect. In the subsequent section, a brief introduction to the methodological aspects of multiple imputation is given, followed by an overview of the machine learning imputation technique based on the recently proposed mixed-effects random forest algorithm[41] used in Chapter Four.

3.3.1 Introduction

Choosing the best method to handle multivariate missing data is an obstacle often faced in observational studies. One issue with simple methods for dealing with missing data, including listwise deletion, mean imputation, and the indicator method, is that they may produce standard errors that are either too large (deletion methods) or too small (single imputation methods)[54]. As a solution to the problem of standard errors that are too small, Rubin[70] proposed multiple imputation, a mechanism for dealing with the inherent uncertainty of the imputed data values themselves.

Multiple imputation techniques comprise three major steps: imputation, analysis,

and pooling. In the imputation stage, the analyst creates $m \geq 2$ complete data sets using a given imputation technique. The imputed data sets are identical for observed data values but may differ in the imputed data values. Secondly, each of these data sets is analyzed and thirdly, the m results are pooled into a final point estimate and corresponding standard error using a specific set of rules known as Rubin's rules. This estimate of the standard error combines the conventional sampling variance (within-imputation variance) and the extra variance caused by the missing data (between-imputation variance).

One popular multiple imputation method for handling multivariate missing data is MICE. MICE is an iterative procedure that cycles through incomplete variables one at a time, drawing predictions for that variable from a series of univariate conditional regression models. Continuous variables are typically imputed in MICE using a linear model that does not include interaction terms. This model assumes continuous variables are normally distributed and that no nonlinear relationships exist between either the outcome and the predictors or between the predictors themselves. It may be difficult to appropriately accommodate for all nonlinear relationships and omission of these terms may bias results[71].

Random forest-based imputation has been proposed as an alternative to other parametric- and nonparametric-based imputation strategies within the MICE framework[37]. Random forest combines the results of multiple decision trees constructed using bootstrap samples of the data. Multiple trees are used rather than a single decision tree in order to reduce overfitting and increase predictive accuracy. Random forest does not rely on distributional assumptions and can accommodate nonlinear relationships and interactions without explicit specification in the imputation model. Random forest-based MICE has been found to produce more efficient estimates than parametric-based MICE, and was especially advantageous when the data set contained nonlinear relationships[39].

One limitation of many imputation techniques, including random forest-based MICE, is their inability to accommodate clustering in the imputation procedure. Analyses are further complicated by unbalanced and small cluster sizes, and it remains unclear which MICE-based multilevel imputation technique is best in these cases[72]. Several modifications to tree-based algorithms to accommodate clustered

data for continuous outcomes have been proposed[40–42]. Hajjem et al.[40] incorporated mixed-effects in the regression tree algorithm using an iterative procedure similar to the expectation-maximization algorithm for linear mixed-effects models (LMER). Then, Hajjem et al.[41] extended the mixed-effects regression tree method to the random forest setting for predicting a continuous outcome and found this method to have a smaller predictive mean squared error (PMSE), on average, than random forest with the largest gain in performance observed in settings with large random effects. However, the mixed-effects random forest (MERF) algorithm has not been implemented or evaluated in an imputation setting.

3.3.2 General overview of multiple imputation

Notation

Let $\boldsymbol{Y}$ denote the $n \times p$ matrix containing the sample data on p variables and n units where y_{ij} represents the jth data value for the ith unit for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. Define the response indicator $\boldsymbol{R}$ to be a $n \times p$ logical matrix where $r_{ij} = 1$ if y_{ij} is observed and is zero if y_{ij} is missing. The observed data (where $r_{ij} = 1$) and missing data (where $r_{ij} = 0$) are denoted by $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{mis}$, respectively. Considered together, $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ contains the complete data values, where the values of $\boldsymbol{Y}_{mis}$ are unknown to the analyst.

Missing data mechanisms and ignorability

The missing data model, whose parameters are denoted by $\boldsymbol{\phi}$, describes the relationship between $\boldsymbol{R}$ and $\boldsymbol{Y}$ and can be written as $P(\boldsymbol{R} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\phi})$. The data are said to be missing completely at random (MCAR) if the probability of being missing depends only on some parameters $\boldsymbol{\phi}$. Thus, in this case, the overall probability of being missing is

$$P(\boldsymbol{R} = 0 \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\phi}) = P(\boldsymbol{R} = 0; \boldsymbol{\phi}). \tag{3.1}$$

If the probability of missingness depends only on the observed data, the missingness mechanism is said to be missing at random (MAR), and the overall probability of

being missing is

$$P(\boldsymbol{R} = 0 \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\phi}) = P(\boldsymbol{R} = 0 \mid \boldsymbol{Y}_{obs}; \boldsymbol{\phi}). \tag{3.2}$$

When the probability of missingness depends on unobserved information, whether it be a variable that was not collected or $\boldsymbol{Y}_{mis}$ itself, the data are said to be missing not at random (MNAR) and the left hand side of (3.1) does not simplify. It is important to note that from the data alone, a MAR mechanism is indistinguishable from a MNAR mechanism.

A missing data mechanism can be classified as ignorable if i) the data are MAR (or MCAR), and ii) the parameters of interest, say $\boldsymbol{\theta}$, and the parameters governing the missing data model, $\boldsymbol{\phi}$, are independent (i.e., $P(\boldsymbol{\theta}, \boldsymbol{\phi}) = P(\boldsymbol{\theta})P(\boldsymbol{\phi})$). The implication of an ignorable missing data mechanism is that the analyst can accurately estimate $\boldsymbol{\theta}$ without knowing $\boldsymbol{\phi}$. The assumption of ignorability is required when constructing imputation models, and if ignorability holds, the posterior distribution of the missing data, from which imputations are drawn, does not depend on $\boldsymbol{R}$, thus

$$P(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}, \boldsymbol{R}) = P(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}).$$

Multiple imputation may be used when data are MNAR, but requires the analyst to explicitly model the missingness mechanism.

Drawing imputations and inference

Suppose the scientific quantity of interest is Q, where Q can be expressed as a function of the population data. For example, Q may be a population mean or regression coefficient. Let $\hat{Q}$ be the estimator of Q with sample variance U estimated by $\hat{U}$. Denote the collection of m imputations for $\boldsymbol{Y}_{mis}$ to be $\left\{\boldsymbol{Y}_{mis}^{(1)}, \boldsymbol{Y}_{mis}^{(2)}, \ldots, \boldsymbol{Y}_{mis}^{(m)}\right\}$ and define $\hat{Q}^m = \hat{Q}^{(m)}\left(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(m)}\right)$ and $\hat{U}^{(m)} = \hat{U}\left(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(m)}\right)$ to be estimates of Q and U computed using the mth complete data set $\left(\boldsymbol{Y}^{(m)} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(m)})\right)$.

The goal of multiple imputation is to obtain an estimate of Q that is both unbiased and at least confidence valid[73]. Unbiased means that the average $\hat{Q}$ over all possible

samples from the population is equal to Q, or

$$E[\hat{Q}] = Q. \tag{3.3}$$

At least confidence valid means that a nominal $100 \times (1-\alpha)\%$ CI has actual coverage of at least $100 \times (1-\alpha)\%$. Q is said to be at least confidence valid if the average of $\hat{U}$ over all possible samples from the population is greater or equal to the variance of $\hat{Q}$, or

$$E[\hat{U}] \geq Var[\hat{Q}]. \tag{3.4}$$

Bayesian methods are the motivation behind multiple imputation techniques. When some data are missing, the distribution of Q needs to be summarized under varying $\boldsymbol{Y}_{mis}$. The possible values of Q given what is known about $\boldsymbol{Y}_{obs}$ is captured in the posterior distribution of Q, $P(Q \mid \boldsymbol{Y}_{obs})$. This distribution is difficult to estimate directly and so it is decomposed and rewritten as a function of two simpler posteriors,

$$P(Q \mid \boldsymbol{Y}_{obs}) = \int P(Q \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) P(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}) d\boldsymbol{Y}_{mis}. \tag{3.5}$$

Repeated draws are taken from the posterior predictive distribution of the missing data given the observed data, $P(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs})$. For a given draw of $\boldsymbol{Y}_{mis}$ from this distribution, say $\dot{\boldsymbol{Y}}_{mis}$, $P(Q \mid \boldsymbol{Y}_{obs}, \dot{\boldsymbol{Y}}_{mis})$ is used to calculate Q from the complete data $(\boldsymbol{Y}_{obs}, \dot{\boldsymbol{Y}}_{mis})$. This process is repeated a number of times with new draws for $\dot{\boldsymbol{Y}}_{mis}$. Thus, the actual posterior distribution of Q is equal to the complete-data posterior distribution of Q averaged over the posterior predictive distribution of $\boldsymbol{Y}_{mis}$, or in other words, over the repeated imputations.

From (3.5), the posterior mean of $P(Q \mid \boldsymbol{Y}_{obs})$ is equal to

$$E[Q \mid \boldsymbol{Y}_{obs}] = E(E[Q \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}] \mid \boldsymbol{Y}_{obs}), \tag{3.6}$$

which is the average of the posterior means of Q over the repeated imputations. The

posterior variance of $P(Q \mid \boldsymbol{Y}_{obs})$ is equal to

$$Var[Q \mid \boldsymbol{Y}_{obs}] = E[Var(Q \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) \mid \boldsymbol{Y}_{obs}] + Var[E(Q \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) \mid \boldsymbol{Y}_{obs}], \quad (3.7)$$

which is the sum of two variance components. The first is the average of the repeated complete-data posterior variances of Q, or the within-variance, and the second is the variance between the complete-data posterior means of Q, or the between-variance. Assuming an infinitely large number of imputations m, denote the estimated within- and between-variance components as $\bar{U}_\infty$ and B_∞, respectively.

Equations (3.6) and (3.7) suggest the following method to combine the results over the m imputations. An overall estimate of Q is obtained by taking an average of the $\hat{Q}^{(m)}$ estimates,

$$\bar{Q} = \sum_{m=1}^{M} \frac{\hat{Q}^{(m)}}{m}. \quad (3.8)$$

Computing a final estimate of the variance of Q requires first estimating both the within-variance (averaging the complete-data variances $\hat{U}^{(m)}$) and the between-variance (standard unbiased estimate of the variance between the m complete data estimates):

$$\bar{U} = \sum_{m=1}^{M} \frac{\hat{U}^{(m)}}{m} \quad (3.9)$$

$$B = \frac{1}{m-1} \sum_{m=1}^{M} (\hat{Q}^{(m)} - \bar{Q})^2. \quad (3.10)$$

Thus, the variance of $\bar{Q}$ is equal to

$$T = \bar{U} + (1 + m^{-1})B, \quad (3.11)$$

where the inflation factor $(1 + m^{-1})$ is used to account for the additional variance due to taking a finite number of imputations.

Specifying the imputation model

Often the most challenging step in multiple imputation is correctly specifying the imputation model (the model used to generate imputed values). At this step in the multiple imputation procedure, the analyst must decide on the functional form of the model (discussed in Section 3.3.3), which variables are to be included in the model, how possible nonlinear relationships will be handled, and how binary, categorical, derived, and non-normal continuous variables will be imputed. Although some general guidelines address each of these points[74], no definite rules exist, and it may be difficult for analysts to avoid the various pitfalls associated with the use of multiple imputation methods[72].

Based on current recommendations, the proposed imputation model should contain at least all variables in the analysis model (the final model fitted to each imputed dataset) including the outcome and any interaction terms that are of interest. Incorporating additional, or auxiliary variables, in the imputation model that are predictive of the incomplete variables and of the missingness mechanism is also beneficial. This may reduce bias by making the MAR assumption more plausible (since MAR is indistinguishable from MNAR in practice) and may improve the quality of the imputations, thus resulting in a gain in precision of the final estimates[72, 75].

With respect to imputing non-normal continuous variables and variables with a nonlinear relationship with the outcome, predictive mean matching (PMM) has been shown to be a useful nonparametric method[71, 72, 76–78]. Using a given imputation model and variable X with missingness, PMM first calculates a predicted value for both observed and unobserved values of X. For each missing $X = x$ value, a set of k candidate donors with the closest predicted value to that of the missing value are selected. One of these donor values is randomly chosen and the observed value of the donor is taken as the imputed value for the missing entry. It is assumed that the distribution of the missing value is the same as the observed data of the selected donors and so PMM is an attractive approach for imputing all types of variables, especially skewed continuous or semi-continuous variables since it ensures the imputed values will fall within the range of the observed data[79].

Imputing derived variables such as interaction terms, higher-order terms (e.g., squared or cubed variables) and variables that are functions of others, such as BMI

and waist-to-hip ratio (WHR), poses a challenge. It is unclear if derived variables should be imputed directly (often known as active imputation) or calculated from the imputed values of its individual parts (passive imputation). The advantage of passive imputation is that the derived variable respects the interrelationships of its individual parts, whereas active imputation does not and it may also produce implausible values for the derived variable[54]. However, active imputation is the easiest to implement and ensures the imputation model is compatible with the analysis model, meaning that there exists a joint model for which the imputation model and the analysis model are conditionals[72]. As opposed to incompatible imputation models, compatible models have been shown to be robust to model misspecification[71, 80].

The literature concerning the optimal method for imputing incomplete variables that are ratios is limited, with some authors suggesting the use of active imputation[81] while others suggest using passive imputation[82]. When imputing BMI in practice, however, one group of researchers found virtually no difference between active and passive imputation when height and weight were MCAR and active imputation was only favored slightly when data were MAR[83]. Little difference between the two approaches when imputing BMI in a real dataset was also found another study[81]. However, this study found significant differences in the two methods when imputing cholesterol ratio and attributed the observed differences to the large coefficient of variation (i.e., relative variability) of the denominator in the cholesterol ratio. Based on the results of a simulation study, the authors advised the use of active and passive imputation after log transformation, especially when the coefficient of variation of the denominator is greater than 0.1, and also found that using passive imputation with PMM yielded less biased results than those resulting from the joint normal approach[81].

3.3.3 Multivariate imputation by chained equations (MICE)

Two current widely available methods to handle multivariate missing data and to model relationships between the missing and observed data are multiple imputation based on the multivariate normal distribution (MVNI)[84] and by chained equations[85]. MVNI is a joint model approach that assumes all variables in the imputation model jointly follow a multivariate normal distribution. The assumption of

joint normality is not always feasible, especially when the imputation model contains both binary and categorical variables. However, the joint normal approach has been shown to be robust to model misspecification and to perform well under conditions in which data are not normal[79, 84].

An alternative to the joint model approach is to use a sequential modeling approach where a conditional regression model is specified for each variable with missingness. These conditional models reflect the distribution of the variable (e.g., logistic regression for a binary variable). This method is more flexible as it does not rely on the assumption of multivariate normality and is attractive when imputing binary and categorical variables. Imputations are generated by an iterative procedure where each conditional regression model is estimated in turn, using only observed data for that variable and imputed values for the other variables at that iteration.

Rather than specifying the joint model directly as in the MVNI approach, this step is bypassed in MICE. The imputation model is defined by $P(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}, \boldsymbol{R})$ and describes how synthetic values for $\boldsymbol{Y}_{mis}$ are generated. In the MICE approach, data are imputed on a variable-by-variable basis by specifying an imputation model for each variable subject to missingness. In other words, this method attempts to define the joint density $P(\boldsymbol{Y}, \boldsymbol{R} \mid \boldsymbol{\theta})$ by specifying a conditional density $P(Y_j \mid \boldsymbol{Y}_{-j}, \boldsymbol{R}, \boldsymbol{\theta}_j)$ for each Y_j, where Y_j is one incomplete variable and $\boldsymbol{Y}_{-j}$ is the collection of variables in $\boldsymbol{Y}$ except Y_j. This density is used to impute $Y_{j,mis}$ given $\boldsymbol{Y}_{-j}$ and $\boldsymbol{R}$. Initially, starting with simple guesses for the missing values (e.g., mean imputation or a random sample from the observed values), MICE operates by iterating over all conditionally specified imputation models until apparent convergence is reached (about 10-20 iterations). MICE is implemented in R using the *mice* package[82].

Although MICE has the advantage of being a more flexible approach than MVNI, it has several limitations. Justification of the MICE procedure has been largely based on empirical studies rather than on theoretical arguments[72]. For example, one may specify a series of conditional distributions for which no multivariate density exists and therefore the concern of incompatible conditionals may arise. In other words, two conditional models are compatible if there exists some joint distribution that has these two models as its conditional densities. Compatibility is a theoretical requirement of the MICE algorithm, but little evidence of the influence of incompatibility

in practice has been reported[85]. It is also advised that analysts assess convergence of the algorithm and the chosen imputation models, but since the range of available imputation diagnostic tools is limited, this may be a challenging task[74].

3.3.4 Random forest-based MICE

First introduced by Breiman et al. in 1984[86], the classification and regression tree (CART) model is commonly used for prediction and classification. CART algorithms involve recursively splitting the predictor space into smaller regions and using these regions (or nodes) to predict the response for a new observation. For example, a regression tree is typically used to model a continuous outcome and predicts the response of a new observation by taking the mean of the training observations in the node to which the new observation resides. Although CART models are widely used for their ease of interpretability and intuitive representation, these methods can be limited in their predictive accuracy compared to other regression and classification approaches. Decision trees can also be non-robust, meaning that small changes in the data set used to build the tree can have a large effect on the final estimated structure of the tree.

To address the shortcomings of decision trees, Breiman et al.[87] introduced the ensemble method called bagging, or bootstrap aggregation. This bagging method combines the results from a collection of decision trees trained on bootstrap samples of the data to reduce the inherent high variance of decision trees. One limitation of bagged trees is that they could have similar structures if a very strong predictor exists in the data set, since most or all trees will use this predictor in the top split. If all bagged trees have a similar structure, their predictions will be highly correlated and little improvement is made over the use of a single tree. Building on the bagging method, Breiman et al.[88] proposed the random forest technique, which adds an additional layer of randomness that decorrelates the trees. Random forest addresses the limitations of bagged trees by considering only a random subset of the predictors at each split. In both the bagging and random forest algorithms, a fixed number of trees are constructed, each using a different bootstrap sample of the data. However, the methods differ in that for bagged trees, each node is split using the best split among all predictors, whereas in random forest, each node is split using the best

among a subset of predictors randomly chosen at the node.

For a continuous variable with missing values, a possible strategy in MICE is to impute using a linear model. It has been shown that omitting nonlinear terms from the imputation model can lead to biased results[71, 80, 89]. Although these interaction terms can be added to the imputation model, it may be difficult to appropriately accommodate all underlying interactions in the data set. An imputation technique proposed by Shah et al.[37] aims to overcome these issues by imputing using random forest. This method does not rely on distributional assumptions and can accommodate nonlinear relations and interactions without explicit specification in the imputation model.

Within the MICE framework, random forest imputation proposed by Shah et al.[37] was derived from the "mice.impute.norm.boot" function in the *mice* R package[82]. First, "mice.impute.norm.boot" fits a linear regression model to a bootstrap sample of those with observed values of the variable to be imputed, which accommodates sampling variation in estimating population parameters. Building on this function, the random forest algorithm ("mice.impute.rf.cont") involves an additional level of bootstrap sampling, where each tree in the forest is constructed using another bootstrap sample. Then, observations with missing values are imputed by taking random draws from independent normal distributions with conditional means predicted using the random forest. The variances of these distributions are taken to be the out-of-bag mean square error, or the mean of the squared differences between the observed value and the prediction using the trees for which that observation was not included in the bootstrap sample.

Note that the random forest-based imputation method described above is different from that based on Breiman's random forest algorithm developed by Doove et al.[36] ("mice.impute.rf"). In "mice.impute.rf", missing values are imputed by taking the observed value of one randomly selected donor from the set of observations in the terminal nodes of the trees used to build the random forest in which the observation with a missing value resides.

Using simulation, Shah et al.[37] found that random forest-based MICE produced more efficient estimates and therefore narrower CIs compared to parametric-based MICE. In simulated data sets with interactions between predictor variables, estimates

of statistical parameters were less biased under random forest-based MICE than under imputation with a main-effects linear regression model. One limitation of this method is that it can be biased in some situations when imputing continuous variables, since random forest predictions at the extremes of their range are biased towards less extreme values. In an additional simulation study, the authors found that random forest-based MICE led to bias when the distribution of missing values was very different from that of observed values[37]. However, in these situations, any imputation method may produce poor results.

3.3.5 Mixed-effects random forest-based MICE

Like many imputation techniques, random forest-based MICE does not accommodate clustering in the imputation procedure. This creates challenges when imputing missing values in datasets where observations are not independent. For example, consider the NSAPD where delivery-level data are nested within women. Since deliveries from the same woman are more similar than deliveries from different women, the variance of the parameter estimates might be underestimated, and the parameter estimates themselves may also be biased. Imputation that ignores clustering may underestimate standard errors even if the analysis model allows for clustering, but imputation techniques that allow for clustering through fixed effects (including dummy variables representing cluster membership in the imputation model) may overestimate standard errors[32, 90].

The *mice*[82] and *miceadds*[91] R packages offer several multilevel imputation techniques for imputing continuous variables. Many of these methods have fairly similar algorithms and are intended for normally distributed variables, but differ in their assumption about the error variance across clusters (heteroscedasticity versus homoscedasticity). In practice, complications may arise when the data contain unbalanced and small cluster sizes, and although several studies have compared multilevel imputation methods[31–33], it remains unclear which method is best in these cases. Additionally, studies examining the robustness to cluster size have not thoroughly investigated multilevel imputation techniques on data sets that have many small clusters with sizes as small as 1 or 2 observations.

As previously mentioned, random forest-based MICE accommodates nonlinearities among predictors and has been found to produce more efficient estimates than parametric-based MICE[37]. Several modifications to tree-based algorithms to accommodate clustered data for continuous outcomes have been proposed[40–42]. Hajjem et al.[40] incorporated mixed-effects in the regression tree algorithm (mixed-effects regression tree [MERT]) using an iterative procedure similar to the expectation-maximization algorithm for LMER. Then, Hajjem et al.[41] extended the mixed-effects regression tree method to the random forest setting.

Hajjem et al.[41] proposed replacing the regression tree within each iteration of the MERT algorithm with a forest of regression trees, and called this a MERF. A MERF of regression trees is defined as follows

$$
\begin{aligned}
&y_i = f(X_i) + Z_i b_i + \epsilon_i, \\
&b_i \sim N(0, D), \qquad \epsilon_i \sim N(0, R_i), \qquad i = 1, \dots n,
\end{aligned}
$$

where $y_i = [y_{i1}, \dots, y_{in_i}]^T$ is the $n_i \times 1$ vector of responses for the n_i observations in cluster i, $X_i = [x_{i1}, \dots, x_{in_i}]^T$ is the $n_i \times p$ matrix of fixed-effects covariates, $Z_i = [z_{i1}, \dots, z_{in_i}]^T$ is the $n_i \times q$ matrix of random-effects covariates, $b_i = [b_{i1}, \dots, b_{iq}]^T$ is the $q \times 1$ unknown vector of random effects for cluster i, and $\epsilon_i = [\epsilon_{i1}, \dots, \epsilon_{in_i}]^T$ is the $n_i \times 1$ vector of errors. The unknown function, $f(X_i)$, is estimated using a standard forest of regression trees, and the random part, $Z_i b_i$, is assumed to be linear. Lastly, the total number of observations is $N = \sum_{i=1}^{n} n_i$, and D and R_i are the covariance matrices of b_i and ϵ_i, respectively. In the random intercept case where $q = 1$, the MERF model simplifies to the following

$$
\begin{aligned}
&y_i = f(X_i) + b_i + \epsilon_i, \\
&b_i \sim N(0, D), \qquad \epsilon_i \sim N(0, R_i), \qquad i = 1, \dots n,
\end{aligned}
$$

where b_i is the random intercept for the ith cluster.

Assumptions of the model are that b_i and ϵ_i are independent and normally distributed, and that the between-cluster observations are independent. The covariance matrix of the vector of observations y_i in cluster i is therefore defined as $V_i = Cov(y_i) = Z_i D Z_i^T + R_i$, and $V = Cov(y) = \text{diag}(V_1, \dots, V_n)$, where

$y = [y_1^T, \ldots, y_n^T]^T$. This model further assumes a compound symmetry covariance structure, or that $R_i = \sigma^2 I_{n_i}$ for $i = 1, \ldots, n$, is diagonal.

The MERF algorithm (Algorithm 1) is similar to the expectation-maximization algorithm for LMER. For simplicity, consider the random-intercept case (i.e., $q = 1$). The algorithm initializes by setting $\hat{b}_i$, $\hat{\sigma}^2$, and $\hat{D}$ to starting values (e.g., $\hat{b}_i = 0$, $\hat{\sigma}^2 = 1$, and $\hat{D} = 0.01$). In Step 1, the fixed part of the response variable, y_i^*, is calculated by removing the current available value of the random part (e.g., $\hat{b}_i = 0$) from y_i. Regression trees are then built from bootstrap samples, taken with replacement, from the training set (y_{ij}^*, x_{ij}). The out-of-bag prediction for each observation j in cluster i is obtained by taking the mean of the subset of trees built using the bootstrap samples not containing this observation. Then, $\hat{b}_i$ is computed using the updated estimate of the random part. In Step 2, the variance components $\hat{\sigma}^2$ and $\hat{D}$ are updated based on updated estimates of the residuals. Steps 1 and 2 are repeated until the generalized log-likelihood (GLL) converges, or changes by a very small amount (ϵ) between iterations. The GLL is a measure of the loss, and as model fit improves, the GLL will decrease.

After fitting the MERF, the predicted response of a new observation j that belongs to cluster i is calculated in one of two ways. If cluster i was used to build the model, the predicted value is calculated using its corresponding population-averaged random forest prediction, $\hat{f}(x_{ij})$, and the predicted random part corresponding to its cluster, $Z_i\hat{b}_i$. If cluster i was not used to build the model, the predicted value is calculated using solely its corresponding population-averaged random forest prediction, $\hat{f}(x_{ij})$.

The performance of MERF in the setting of predicting a continuous outcome was compared to 1) LMER, 2) standard regression tree, 3) random forest, and 4) MERT in a simulation study[41]. Overall, MERF had the smallest PMSE, on average, than the four alternative models with the most pronounced improvements made over models without random effects (i.e., random forest, regression tree, and linear model) in settings with large random effects, and models with random effects (i.e., linear-mixed effects model and MERT) in settings with small random effects. As opposed to random forest, MERF uses the cluster random effect in the final prediction and so the more important the random effect, the greater gain in predictive performance.

Algorithm 1: Mixed-effects random forest

Step 0. Set $r = 0$. Let $b_{i(0)} = \overrightarrow{\mathbf{0}}_q$, $\hat{\sigma}^2_{(0)} = 1$, and $\hat{D}_{(0)} = 100^{-1} I_q$.

while $GLL \geq \epsilon$ **do**

Step 1. Set $r = r + 1$.

i) Update $y^*_{i(r)} = y_i - Z_i \hat{b}_{i(r-1)}$, $i = 1, \ldots, n$.

ii) Using $y^*_{ij(r)}$ and x_{ij} for $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, as the full set of training responses and covariates, build *ntree* regression trees using the random forest algorithm, where each tree is built using a bootstrap sample drawn with replacement from $(y^*_{ij(r)}, x_{ij})$.

iii) Obtain an estimate $\hat{f}(x_{ij})_{(r)}$ of $f(x_{ij})$ by taking the mean prediction from the subset of trees that are built with the bootstrap samples not containing observation j in cluster i, or the out-of-bag prediction

$$\hat{f}(X_i)_{(r)} = [\hat{f}(x_{i1})_{(r)}, \ldots, \hat{f}(x_{in_i})_{(r)}]^T.$$

iv) Update $\hat{b}_{i(r)}$ using

$$\hat{b}_{i(r)} = \hat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} (y_i - \hat{f}(X_i)_{(r)})$$

where $\hat{V}_{i(r-1)} = Z_i \hat{D}_{(r-1)} Z_i^T + \hat{\sigma}^2_{(r-1)} I_{n_i}$ for $i = 1, \ldots, n$.

Step 2. Update $\hat{\sigma}^2_{(r)}$ and $\hat{D}_{(r)}$ using

$$\hat{\sigma}^2_{(r)} = N^{-1} \sum_{i=1}^{n} \{\hat{\epsilon}^T_{i(r)} \hat{\epsilon}_{i(r)} + \hat{\sigma}^2_{(r-1)} [n_i - \hat{\sigma}^2_{(r-1)} \text{trace}(\hat{V}_{i(r-1)})]\}$$

$$\hat{D}_{(r)} = n^{-1} \sum_{i=1}^{n} \{\hat{b}_{i(r)} \hat{b}^T_{i(r)} + [\hat{D}_{(r-1)} - \hat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} Z_i \hat{D}_{(r-1)}]\}$$

where $\hat{\epsilon}_{i(r)} = y_i - \hat{f}(X_i)_{(r)} - Z_i \hat{b}_{i(r)}$.

Step 3. Calculate the generalized log-likelihood (GLL) criterion

$$GLL(f, b_i \mid y) = \sum_{i=1}^{n} \{[y_i - f(X_i) - Z_i b_i]^T R_i^{-1} [y_i - f(X_i) - Z_i b_i] + b_i^T D^{-1} b_i + \log |D| + \log |R_i|\}.$$

end

However, when prediction is made for an observation in a new cluster that was not present in the training sample, MERF uses only the random forest prediction (since the random effect estimate is unavailable). For most of these observations, MERF generally performs better than random forest, but the improvement is modest. Using real box office revenue data, MERF exhibited the best predictive performance among all alternative models with a PMSE of 0.47 compared to 0.60 and 0.53 for random forest and MERT, respectively[41].

The best strategy for imputing skewed variables, particularly in the context of missing correlated BMI values in studies of pregnancy-related outcomes remains unclear. One alternative to parametric-based MICE is to impute continuous variables using random forest, a machine learning technique that does not rely on distributional assumptions and can naturally accommodate nonlinearities in the data. Building on the random forest algorithm, MERFs adjust for clustering in the data by calculating a random effect component, and have yet to be evaluated as an imputation technique.

3.4 Estimating total and mediation effects in the analysis of grandmaternal BMI and child birthweight

3.4.1 Introduction

Mediation analysis is becoming increasingly popular in the field of epidemiology. The goal of mediation analyses is to quantify specific causal pathways described by one or more variables that are assumed to be affected by the exposure and also affect the outcome of interest. Consider again the simple example of mediation shown in Figure 3.2. The direct effect is represented by the pathway *c'* and, using traditional approaches to mediation analysis, can be estimated by the exposure coefficient from the regression of the outcome on the exposure and mediator (e.g., a correctly specified linear regression model). The indirect effect can be estimated either as the product of the a and b pathways (product method), or as the difference between the total effect and the direct effect (difference method). The a and b pathways can be estimated by the exposure coefficient from the regression of the mediator on the exposure, and by the mediator coefficient from the regression of the outcome on the exposure and mediator. The total effect is estimated by the exposure coefficient from the regression

of the outcome on the exposure.

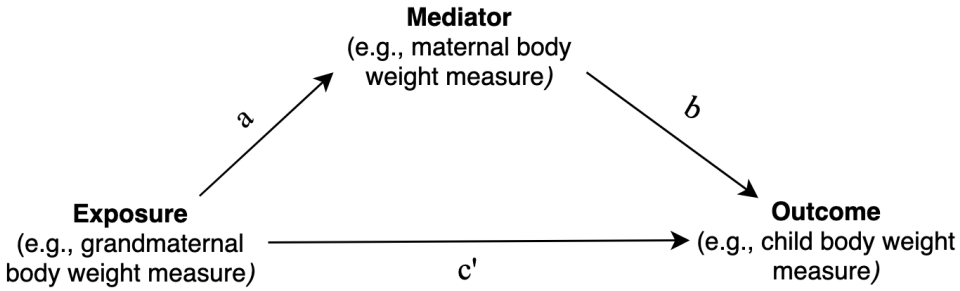


Figure 3.2: An example of a simple mediation analysis with indicated pathways

These traditional approaches to mediation analysis proposed by Baron and Kenny[92] are prone to bias arising both from incorrect statistical analysis and suboptimal study design[26]. The traditional approaches also have several limitations concerning their applicability in models with interactions or nonlinearities[93, 94], which are overcome by using a counterfactual-based approach to mediation analysis.

Causal mediation analysis is based within the counterfactual framework[93, 94] and defines causal effects as contrasts of potential outcomes. A potential outcome is an individual's outcome value that would have been observed had their exposure been set to a specified value. For example, assuming a binary exposure, an individual's potential outcomes under exposure values 0 and 1 are defined as $Y_i(0)$ and $Y_i(1)$, respectively, and the individual-level causal effect of the exposure on the outcome is estimated by $Y_i(1) - Y_i(0)$. In the context of mediation, potential outcomes depend on both exposure and mediator values. Causal mediation analysis differentiates between causal effect definitions and causal effect estimation[95], a strength of these definitions being that they are nonparametric and can be applied to any type of model - for example, mediation models with nonlinear and interaction terms, and models with non-continuous mediator and outcome variables. Causal effects are defined at the individual level, but since only one outcome value is observed for each individual while the other is from an unobserved scenario, population-average causal effects are estimated.

In Chapter Five, parametric g-computation, a counterfactual-based approach, was used to estimate the total effect of grandmaternal pre-pregnancy BMI on child

birthweight z-score and path specific-effects (e.g., indirect effect via maternal pre-pregnancy BMI). In the following section, directed acyclic graphs (DAGs) will be briefly introduced, as they will be used throughout this section and thesis. Next, the traditional and counter-factual based approach to mediation analysis are presented. Lastly, an overview of parametric g-computation in the context of mediation is given.

3.4.2 Directed acyclic graphs (DAGs)

Causal diagrams or DAGs are used to graphically represent hypothesized relationships between variables and to identify sources of potential bias. In observational studies, DAGs are particularly useful for representing sources of confounding and selection bias. In the specific context of the research presented in Chapter Five, they are useful for representing hypothesized causal relationships with mediators. Formal rules are used to develop these graphs and to guide appropriate statistical analyses.

Consider the following example illustrated in Figure 3.3 depicting the hypothesized relationships among variables in estimating the effect of maternal pre-pregnancy BMI on age- and sex-specific birthweight z-score.

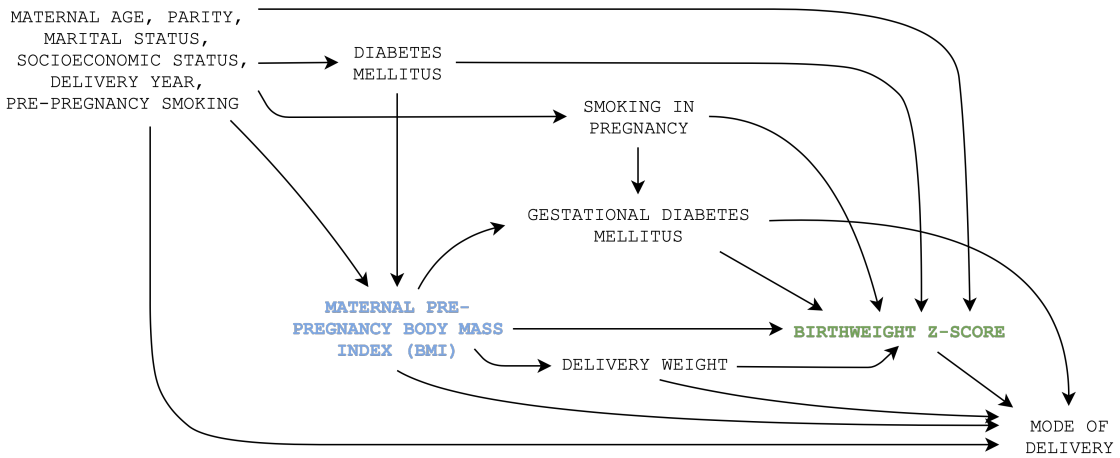


Figure 3.3: Directed acyclic graph depicting the hypothesized relationships between maternal pre-pregnancy body mass index (exposure) and birthweight z-score

The notation for describing DAGs is given in Greenland et al.[96]. A line connecting two variables is called an arc or edge, and a causal association is indicated by a single-headed arrow from cause to effect. Ancestors or causes of a variable A are

variables that are on a directed path leading to A, whereas the descendants of A are variables that lie on a directed path leading away from A. For example, maternal pre-pregnancy BMI is a descendant of maternal age, and an ancestor of delivery weight. A collider is a variable that has two arrows pointing towards it (e.g., birthweight z-score is a collider on the pathway *smoking in pregnancy* $\rightarrow$ *birthweight z-score* $\leftarrow$ *delivery weight*).

Causal diagrams are a simple way to encode background knowledge and assumptions about the variables under study. They are also used to determine whether a set of measured variables is sufficient for analyzing the association under investigation. Traditionally, a confounder was defined as a variable that is associated with both the exposure and the outcome and is not itself affected by the exposure[97]. However, developments in causal inference have demonstrated this definition to be inadequate and more recent definitions put more emphasis on "confounding" rather than "confounder."

The definition of "confounder" that is used in this dissertation is that suggested by VanderWeele and Shpitser[98] and used in recent epidemiology textbooks[99]. That is, pre-exposure covariate C is considered a confounder for the effect of exposure A on outcome Y if there exists a set of covariates X such that the effect of A on Y is unconfounded conditional on (X, C) but for no proper subset of (X, C) is the effect of A on Y unconfounded given the subset. Or, equivalently, a "confounder" is a member of a minimally sufficient adjustment set for which the effect of A on Y is unconfounded. For example, in Figure 3.3, the variable *pre-existing diabetes* would be included in the minimally sufficient adjustment set needed to control for confounding of the exposure-outcome relationship. The variables in this set should be adjusted for to get a valid estimate of the total effect, whereas variables on the causal pathway between the exposure and the outcome (e.g., delivery weight), known as mediators, should not be adjusted for.

To examine sources of bias using a DAG, we need to identify backdoor paths, or noncausal pathways from exposure to outcome that would remain if any arrows pointing away from the exposure were removed. In Figure 3.3, an open pathway from maternal BMI to birthweight z-score is the pathway *maternal BMI* $\leftarrow$ *pre-existing diabetes* $\rightarrow$ *birthweight z-score*. If a backdoor path is blocked by a collider, then this

path is ignored. Once all unblocked backdoor paths have been identified, the set of covariates that would block all existing open backdoor paths is the adjustment set required to control for confounding. It is important to note that there may be more than one possible adjustment set that can control for confounding.

3.4.3 Traditional approaches to mediation analysis

The "product method" and "difference method" are two traditional approaches to mediation analysis first proposed by Baron and Kenny[92]. Consider the simple causal diagram in Figure 3.4 where A, M, and Y represent the exposure, mediator and outcome variables, respectively, and let C be additional covariates (not shown).

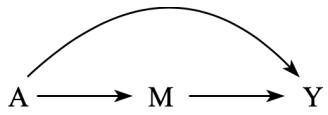


Figure 3.4: Directed acyclic graph with exposure A, mediator M, and outcome Y

For the case where both A and Y are continuous, the difference method begins by fitting the following two regression models:

$$E[Y \mid A = a, C = c] = \beta_0 + \beta_1 a + \beta_2' c, \tag{3.12}$$

$$E[Y \mid A = a, M = m, C = c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3' c. \tag{3.13}$$

The total effect of the exposure on the outcome is the estimate of β_1. The difference between the estimate of the exposure in the model with the mediator and the one without, $\beta_1 - \theta_1$, is interpreted as the mediation or indirect effect (the pathway from A to Y through M). The direct effect of A on Y (the pathway from A to Y not through M) is the estimate of θ_1 as this is the effect of the exposure on the outcome that remains after controlling for M.

The "product method" fits (3.13) and a model for the mediator,

$$E[M \mid A = a, C = c] = \phi_0 + \phi_1 a + \phi_2' c. \tag{3.14}$$

The direct effect is again the estimate of θ_1 as in the difference method. The indirect

effect of A on Y is now taken as the product of the exposure coefficient in the mediator model and the mediator coefficient in the outcome model, i.e., $\theta_2\phi_1$. An estimate of the total effect is the sum of the indirect and direct effects, $\theta_1 + \theta_2\phi_1$, or alternatively the exposure coefficient in the outcome model, β_1. The difference method and product method will yield the same results when Y and M are continuous and models are fitted using ordinary least squares regression.

Extra caution must be taken when estimating direct and indirect effects with binary outcomes. Valeri and Vanderweele[56] proposed regression-based estimators of these effects on the OR scale by invoking approximations based on the rare outcome assumption. The approximate natural effects OR estimator derived from correctly specified logistic regression models are unbiased in the case of a rare binary outcome. Exact regression-based estimators of the natural direct and indirect effects have been proposed[100–103]. For example, those described by Samoilenko and Lefebvre[103] were found to be unbiased, regardless of the effect scale and the prevalence of the outcome.

3.4.4 Causal mediation analysis

For the subsequent section, let A, M, and Y be continuous exposure, mediator, and outcome variables, respectively, with assumed relationships as shown in Figure 3.4. Robins and Greenland[94] and Pearl[93] define total and natural (in)direct effects using the counterfactuals $M(a)$, $Y(a)$, and $Y(a, M(a'))$, where $M(a)$ is the value of M that would have been observed if A were set to a, $Y(a)$ is the value of Y that would have been observed if A were set to a, and $Y(a, M(a'))$ is the value of Y that would have been observed if A were set to a and M to $M(a')$.

For two different values of continuous exposure, say $A = a'$ and $A = a$, the average or total (causal) effect (TE) of the exposure on the outcome for a vs. a' is $E[Y(a) - Y(a')]$. More generally, the conditional causal effect of the exposure on the outcome for a vs. a', given pre-exposure covariates C, is $E[Y(a) - Y(a') \mid C]$. The natural direct effect (NDE) and natural indirect effect (NIE) are defined as follows:

$$NDE = E[Y(a, M(a')) - Y(a', M(a'))] \tag{3.15}$$

$$NIE = E[Y(a, M(a)) - Y(a, M(a'))] \tag{3.16}$$

The NDE estimates how much the outcome is expected to change, on average, if the exposure was set at level a versus a' and for each individual the value of the mediator was fixed to the value that would be observed if $A = a'$. The NIE estimates how much the outcome is expected to change, on average, if the exposure was fixed at level a but the mediator was changed from the value that it would take under a' to the value it would take under a. Thus, the NDE captures the effect of the exposure on the outcome that would remain if the pathway from the exposure to the mediator was removed, and the NIE captures the effect of the exposure on the outcome that operates by changing the mediator.

Note that the sum of the NIE and NDE is the TE of A on Y for a change in exposure from a' to a:

$$\begin{aligned} Y(a) - Y(a') &= E[Y\,(a, M(a))] - E[Y(a', M(a'))] \\ &= (E[Y(a, M(a))] - E[Y(a, M(a'))]) + \\ &\qquad (E[Y(a, M(a'))] - E[Y(a', M(a'))]) \\ &= \text{NIE} + \text{NDE} \end{aligned}$$

Counterfactual-based definitions of direct and indirect effects can be estimated from the regression models in Section 3.4.3 given that they are correctly specified and certain no-confounding assumptions hold. Similarly as above, suppose a continuous outcome Y, a continuous mediator M, and the mediator model in (3.14). Define the following outcome model that now allows for exposure-mediator interaction in the linear model for the outcome as

$$E[Y \mid A = a, M = m, C = c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c. \tag{3.17}$$

For a change in exposure from level a' to a, (3.14) and (3.17) allow the natural direct effect (NDE) and the natural indirect effect (NIE) to be defined as follows:

$$NDE = (\theta_1 + \theta_3\phi_0 + \theta_3\phi_1 a' + \theta_3\phi_2' c)(a - a') \tag{3.18}$$

$$NIE = (\theta_2\phi_1 + \theta_3\phi_1 a)(a - a') \tag{3.19}$$

These expressions extend those proposed by Baron and Kenny[92] to allow for interaction between the exposure and the mediator. In the absence of interaction, $\theta_3 = 0$ and the NDE is equal to the direct effect estimate θ_1 obtained using the traditional methods multiplied by $(a - a')$. The NIE is equal to that of the product method $\theta_2\phi_1$ multiplied by $(a - a')$.

3.4.5 Assumptions required to identify total causal effects and natural effects

Causal effects are defined as contrasts of potential outcomes. To be able to identify $E[Y(a)]$ from the observed data, several assumptions must be made. The first is called the consistency assumption and asserts that amongst individuals with $A = a$, the observed outcome Y is equal to the potential outcome $Y(a)$. Under this assumption, one potential outcome is observed for each individual (i.e., the one corresponding to the observed exposure level). The second assumption is that of conditional exchangeability and states that individuals with different observed exposure values A but the same pre-exposure covariate values C are comparable such that $Y(a) \perp\!\!\!\perp A \mid C$. This assumption is also referred to as the no unmeasured confounding assumption. Both of these assumptions cannot be tested from the observed data alone, but are sufficient for identifying the conditional causal effect from the data as

$$E[Y(a) - Y(a') \mid C] = E[Y(a) \mid A = a, C] - E[Y(a') \mid A = a', C] \tag{3.20}$$

$$= E[Y \mid A = a, C] - E[Y \mid A = a', C]. \tag{3.21}$$

The TE can be obtained by averaging over the distribution of C. Conditional exchangeability enables replacement of $E[Y(a) \mid C)$ with $E[Y(a) \mid A = a, C)$ and consistency enables replacement of $E[Y(a) \mid A = a, C]$ with $E[Y \mid A = a, C]$. The latter replacement is only legitimate if $E[Y \mid A = a, C]$ is well defined for all values of C. This is the assumption of positivity, and is formally written as $P(A = a \mid C) > 0$.

Similar assumptions are needed to identify $E[Y(a, M(a')]$ from which definitions of the NDE and NIE are constructed. The potential outcome when $a \neq a'$ is called the cross-world potential outcome. This potential outcome differs from $Y(a, M(a))$

in that it is unobservable and belongs in a completely counterfactual world where M is set to the value that would have been observed if all individuals had a different exposure value. For the consistency assumption, it is assumed i) consistency of potential outcomes, $Y = Y(a, m)$ for all values of a and m, ii) consistency of the potential mediator, $M = M(a')$ if $A = a'$, and iii) consistency of the cross-world potential outcome, $Y(a, M(a')) = Y(a, m)$ if $M(a') = m$.

Related to conditional independence, four conditional exchangeability assumptions must met. That is, i) $Y(a, m) \perp\!\!\!\perp A \mid C$ for all values of a and m, ii) $Y(a, m) \perp\!\!\!\perp M \mid A, C$ for all values of a and m, iii) $M(a) \perp\!\!\!\perp A \mid C$ for all values of a, and iv) $Y(a, m) \perp\!\!\!\perp M(a') \mid C$ for all values of a, a', and m. If the data are assumed to be generated from a nonparametric structural equation model (NPSEM)[104], then the condition iv) simplifies to no mediator-outcome confounders affected by the exposure. Lastly, positivity of the exposure conditions and mediator values must met, meaning that there is a non-zero probability that $A = a$ and $A = a'$ conditional on C and a non-zero probability that M is equal to each of its potential values conditional on A and C.

Consider Figure 3.5 where, as before, A, M, and Y are the exposure, mediator, and outcome of interest, and C represents confounders of the exposure-outcome, exposure-mediator, and mediator-outcome relationships. The first three conditional exchangeability conditions would be met by controlling for C. The final condition for the identification of natural effects is slightly more difficult to interpret, and under the NPSEM assumption, will hold if there are no variables that are causes of the exposure and that confound the mediator-outcome relationship (represented by L in Figure 3.5)[93]. Such variables L will be referred to as intermediate confounders.

3.4.6 Mediator-specific effects in the presence of intermediate confounding

In many contexts, the condition iv) of the conditional exchangability assumption required to identify natural (in)direct effects will be violated. That is, variables that confound the mediator-outcome relationship and are affected by the exposure are likely to exist (L in Figure 3.5). In the presence of such intermediate confounders, the NDE and NIE are nonidentifiable. In this case, an alternative is to treat L and

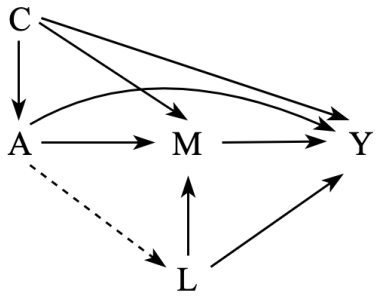


Figure 3.5: Directed acyclic graph with exposure A, mediator M, and outcome Y, where C represents confounders of the exposure-outcome, exposure-mediator, and mediator-outcome relationships, and L is a mediator-outcome confounder that is affected by exposure

M both as mediators and target path-specific effects[105].

Let A, M, and Y be continuous exposure, mediator, and outcome variables, respectively, and L be a continuous intermediate confounder (note that L is assumed to affect M but not vice versa). Furthermore, let the counterfactuals $L(a)$, $M(a, L(a'))$, $Y(a, L(a'), M(a'', L(a''')))$ for $a \neq a' \neq a'' \neq a'''$ be defined according to extensions of those given in Section 3.4.4.

Daniel et al.[105] provide a natural extension of the NDE definition given in 3.15 to the case of two mediators is

$$NDE_{000} = E[Y(a, L(a'), M(a', L(a'))) - Y(a', L(a'), M(a', L(a')))],$$

and is the direct effect through neither L nor M. Indirect effects can be defined through L alone, M alone, or through both L and M, so that their sum, with NDE_{000} is equal to the TE. The natural indirect effect through L alone is

$$NIE_{100} = E[Y(a, L(a), M(a', L(a'))) - Y(a, L(a'), M(a', L(a')))].$$

Similarly, the indirect effect through M alone is

$$NIE_{110} = E[Y(a, L(a), M(a, L(a'))) - Y(a, L(a), M(a', L(a')))],$$

and that through both L and M is

$$NIE_{111} = E[Y(a, L(a), M(a, L(a))) - Y(a, L(a), M(a, L(a')))].$$

These definitions are generalizations of the two mediator setting shown by Daniel et al.[105] to the setting of an intermediate confounder L (i.e., M_1) and a primary mediator of interest (i.e., M_2). Definitions of these effects have also been defined in previous literature[106, 107]. Note that the TE decomposes into the sum of the NDE and NIEs defined above.

Although a four-way decomposition is possible, this increases the complexity of the estimation procedure. An alternative option is to combine the effects through some pathways and only estimate those of primary interest. In this case, only the indirect effect via M is of interest and any effect that operates via L is not, and so a coarser decomposition would be that into three effects: the direct effect, and two mediator-specific effects. That is, the direct effect through neither L nor M, the indirect effect though M alone, and the sum of the indirect effects through L (i.e., the pathways $A \to L \to Y$ and $A \to L \to M \to Y$). The counterfactual definitions of these effects are given in Table 3.1, where "MS^2" indicates that the mediator-specific effect via the second mediator (i.e., M) is being targeted[105].

Table 3.1: Counterfactual definition and description of the total effect and of the components of its 3-way decomposition where A is the exposure, L is the intermediate confounder, M is the mediator, and Y is the outcome

Effect	**Counterfactual definition**
TE	$E\left[Y\left(a, L(a), M(a, L(a))\right) - Y\left(a', L(a'), M(a', L(a'))\right)\right]$
MS^2-NDE-00: $A \to Y$	$E\left[Y\left(a, L(a'), M(a', L(a'))\right) - Y\left(a', L(a'), M(a', L(a'))\right)\right]$
MS^2-NIE-10: $A \to M \to Y$	$E\left[Y\left(a, L(a'), M(a, L(a'))\right) - Y\left(a, L(a'), M(a', L(a'))\right)\right]$
MS^2-NIE-11: $A \to L \to Y$ + $A \to L \to M \to Y$	$E\left[Y\left(a, L(a), M(a, L(a))\right) - Y\left(a, L(a'), M(a, L(a'))\right)\right]$

The identification assumptions for the effects given in Table 3.1 are similar to those described previously for the case of one mediator. The TE is identifiable under the conditions given in Section 3.4.5 and extensions of the assumptions for natural effects to the case of two causally ordered mediators under the assumption that the

data are generated from a NPSEM are: consistency of i) (A, L, M) on Y, ii) A on L and iii) (A, L) on M; no unmeasured confounding of the exposure-outcome relationship, $Y(a, l, m) \perp\!\!\!\perp A \mid C$ for all c, a, l and m; no unmeasured confounding of the mediator-outcome relationships, $Y(a, l, m) \perp\!\!\!\perp L \mid C, A$ and $Y(a, l, m) \perp\!\!\!\perp M \mid C, A, L$ for all c, a, l and m; no unmeasured confounding of the exposure-mediator or mediator-mediator relationships, $L(a) \perp\!\!\!\perp A \mid C$, $M(a, l) \perp\!\!\!\perp A \mid C$, and $M(a, l) \perp\!\!\!\perp L \mid C, A$ for all c, a, l and m; and no mediator-outcome confounder affected by the exposure (excluding L).

Each half of each of the mediator-specific natural effects in Table 3.1 is of the form

$$E[Y(a, L(a'), M(a'', L(a')))]$$

and, under the assumptions above, can be nonparametrically identified from the observed data by

$$\int_c \int_l \int_m E[Y \mid A = a, L = l, M = m, C = c] \\ f_M(m \mid A = a'', L = l, C = c)\, f_L(l \mid A = a', C = c) f_C(c)\, dm\, dl\, dc. \quad (3.22)$$

Estimating mediator-specific effects using g-computation

G-computation was proposed by Robins[108] as a method for estimating causal effects in the presence of time-varying exposures and confounders. Recently, an extension of the g-computation algorithm that incorporates the mediation formula[109] has been used in mediation analysis. The steps of this method involve specifying regression models for each density and expectation in the identifying equations, estimating their parameters from the observed data, and then evaluating the integral analytically[110]. In many cases, the g-computation formula is too difficult to evaluate analytically, and so the integration can be approximated through Monte Carlo simulation[110, 111].

Estimation of the mediator-specific natural effects (Table 3.1) in the mediation analysis presented in Chapter Five was performed via parametric g-computation using Monte Carlo simulation. The advantage of this approach is its flexibility, as any combination of types of outcomes, mediators, and intermediate confounders can

be modeled. This approach, however, operates under the assumption of correct model specification. Since this mediation analysis was complicated by the presence of intermediate confounding, the approach to parameter identifiability for the case of causally ordered mediators given in Daniel et al.[105] and described in Section 3.4.6 was followed. This mediation analysis was complicated by four intermediate confounders, and since the indirect effects via pathways involving these variables were not of interest, they were treated in the analysis as a group of mediators. The identifiability assumptions previously described would therefore apply to this joint set of mediators.

Estimation by parametric g-computation via Monte Carlo simulation for the mediation analysis in Chapter Five is briefly discussed. First, flexible parametric models (e.g., generalized additive models [GAM] with smooth terms for continuous covariates) for i) each intermediate confounder given the exposure and baseline covariates (and possibly previously occurring intermediate confounders), ii) the mediator of interest given the exposure, intermediate confounders, and baseline covariates, and iii) the outcome given the exposure, mediator, intermediate confounders and baseline covariates were fitted using the observed data. The outcome model also included an interaction term between the exposure and mediator. Then, simulations were carried out forward in time based on the hypothesized causal structure. That is, using simulated baseline covariate values, and for $A = \{a, a'\}$, and $i = \{1, \ldots, n\}$, potential values of the intermediate confounders were simulated (e.g., $\boldsymbol{L}(a)$), followed by potential values of the mediator (e.g., $M(a, \boldsymbol{L}(a))$), and lastly, potential values of the outcome (e.g., $Y(a, \boldsymbol{L}(a), M(a, \boldsymbol{L}(a)))$). Contrasts of the i potential outcomes corresponding to the mediator-specific natural effects in Table 3.1 were averaged to estimate population average estimates. To reduce Monte Carlo error, simulations were performed on an enlarged sample of $n = 100,000$ simulated observations (estimation of model parameters is based the original sample). Standard errors were obtained using bootstrapping, whereby all steps of the procedure were repeated on random samples of the observed data drawn with replacement.

Although advanced methods for estimating total and mediation effects are becoming more readily available and accessible to analysts, a 2022 systematic review found that few observational studies have applied modern approaches of mediation

analysis such as those based within the counterfactual framework[112]. Furthermore, many studies of mediation either ignored underlying assumptions or were not explicit about the assumptions of the causal model, and did not report how missing data were accommodated in analyses[112]. The study presented in Chapter Five sought to address these shortcomings of previous observational studies in the analysis of grandmaternal BMI on child birthweight.

3.5 Predicting fetal growth abnormalities

As previously discussed, deviations from normal fetal growth are associated with adverse short- and long-term health outcomes in infants[28]. Infants with birthweight at the extremes of their respective size for gestational age distributions are more likely to accrue slightly higher health care costs in their first year of life compared to infants with normal fetal growth[113]. The ability to correctly identify women at higher risk of delivering both large and small for gestational age infants would therefore be beneficial to the mother, baby and the health care system. In the following section, an overview of current prediction models for SGA and LGA based on early-pregnancy factors is discussed. Then, the methodological aspects of prediction using Super Learner (an ensemble-based approach used in Chapter Six) and metrics for evaluating model performance are presented.

3.5.1 Introduction

Several models for predicting the risk of SGA and LGA have been developed using maternal characteristics, including sociodemographics, pregnancy risk factors, past pregnancy history, and clinical characteristics, but predictive performance remains relatively poor, especially among women in their first pregnancy (i.e., nulliparous women). Most of these models were developed using conventional regression-based methods, such as logistic regression, enabling them to be easily compared and validated on other data sets. For example, one study validated six prediction models for SGA and LGA using an independent cohort of 1311 nulliparous women and found discriminative performance, measured using the area under the receiver operating characteristic curve (AUC-ROC), to be between 0.50-0.66 for SGA and 0.58-0.67 for LGA[114]. The AUC-ROC measures how well the model can distinguish between two

classes (e.g., infants born SGA versus infants born non-SGA); an AUC-ROC of 0.5 indicates no discrimination and 1.0 indicates perfect discrimination. Current efforts to improve prediction models for early detection of SGA and LGA include adding ultrasound measurements, biochemical markers, and results of biophysical tests, but only modest improvements have been reported[115–131] and measurement of some predictors may be costly, time-consuming, and inconvenient for pregnant women. A summary of prediction models for SGA and LGA based on early-pregnancy factors is given in Table A.2.

The discriminative ability of prediction models based solely on maternal characteristics may be improved by including grandmaternal pregnancy-related factors and maternal birth characteristics. The literature for the evidence of an association between grandmaternal body weight measures and offspring body weight measures was discussed in Section 3.2.1, but other multigenerational studies of the effects of in utero exposures on second-generation outcomes[8] have also found small to moderate associations between grandparental risk factors and child birthweight[14–25]. Although many have reported associations between maternal size-at-birth and offspring size-at-birth[29], maternal birth characteristics and grandmaternal risk factors have not been explicitly examined as candidate predictors to improve the prediction of SGA and LGA.

Prediction models based on conventional regression methods have been widely used because they are easily interpreted, compared, and implemented in standard statistical software. However, they are limited in that they rely on strong assumptions, such as the type of error distribution, and in their ability to incorporate complex interactions of the predictors. Machine learning techniques such as random forest, elastic net, and support vector machines (SVM), offer solutions to the limitations of conventional regression-based models. Essentially, these models learn from the data to make predictions on new data and offer advantages in that they do not require explicit specification of a model and can handle a very large number of predictors. Since grandmaternal risk factors may act as potential effect modifiers of the maternal-offspring associations, and relations between predictors and fetal growth may be nonlinear, predictions may benefit from more data-adaptive techniques like machine learning algorithms.

One limitation of using many machine learning techniques in a single prediction setting is that the performance of a particular algorithm will depend on the true data-generating process. In practice, it is nearly impossible to determine which algorithm will perform best in the data under study[132]. To solve this problem, van der Laan and others proposed the Super Learner algorithm, a method based on previous work by Wolpert[133] and Breiman[134] in the 1990s. The Super Learner algorithm uses cross validation (CV) to build a new prediction algorithm, created as the optimal weighted combination of predictions from a library of candidate learners (e.g., machine learning algorithms, logistic regression). The resulting Super Learner model has been shown to perform as well as, or better than, the best algorithm in the ensemble in large samples[135].

Machine learning techniques have been used to predict various clinical conditions and their performance has been compared to traditional logistic regression[136–141]. Overall, the literature is conflicting, but there is a consensus that performance of these methods depends on the data. For example, one study added variables to an established cardiovascular risk prediction model and found machine learning methods performed similarly to logistic regression[140]. On the contrary, another study of the prediction of all-cause mortality using fitness data found that machine learning methods typically outperformed logistic regression, but the performance varied between machine learning methods[138].

Machine learning techniques are gaining popularity as alternative approaches to classification and prediction problems in clinical medicine[142, 143]. Although these techniques have better performance in some situations, the benefit over logistic regression remains unclear, and recent research has suggested that some machine learning methods require more data points than logistic regression[144]. A 2019 systematic review comparing logistic regression to machine learning methods found that, on average, machine learning techniques performed similarly when the studies had low risk of bias[143]. However, only 32% of the 71 studies were deemed to have low risk of bias. One key finding of this review was that reporting of methodology and findings was often incomplete and unclear. For example, information on handling of missing data was lacking or unclear in 32 studies, and model calibration was often ignored. Therefore, further investigation is required to determine if machine learning

techniques lead to better clinical prediction models compared to traditional logistic regression, and reporting of such comparisons needs improvement.

The final recommendations of the aforementioned systematic review[143] and the guidelines described by the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement[145] were used to guide the development and validation of Super Learner models for predicting SGA and LGA in Chapter Six. This includes providing sufficient detail to maximize transparency and reproducibility and using calibration curves to investigate model calibration.

3.5.2 Prediction using Super Learner

The Super Learner algorithm is a technique that uses V-fold CV to construct an optimal weighted average of a set of candidate algorithms with weights estimated according to a user-specified loss function[132]. In general, CV is a tool for evaluating how well an algorithm performs in a sample from the same target population from which the data to fit the model was derived. CV works by first dividing the sample into equal subsets (e.g., $V = 10$ subsets for 10-fold CV), and then, one at a time, a subset is set aside (test sample) and the algorithm is fitted to the remaining $V - 1$ subsets (training sample). Then, in the case of a binary outcome Y, the probability of $Y = 1$ is predicted for each observation in the test sample using the model fitted to the respective training sample. When the outcome is binary, the CV procedure should maintain the prevalence of the outcome across training samples. This can be done using stratified V-fold CV where observations are randomized to folds within strata of the outcome.

In the Super Learner algorithm, the V-fold CV procedure described above is repeated for each of the algorithms in the library of candidate learners. For each learner, the risk is estimated according to a user-specified loss function (e.g., mean squared-error loss for continuous outcomes, or equivalently, the Brier score for binary outcomes defined as $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2/n$). The risk estimates from each of the V test samples are averaged resulting in one cross-validated risk estimate for each learner.

At this point, selecting the algorithm with the smallest cross-validated risk would result in a discrete Super Learner. However, improved performance can be attained

by creating an optimal weighted average of the learners. To do this, non-negative least squares is used to regress the observed outcome Y on the predicted probabilities $\hat{Y}$, with the constraint that all estimated regression coefficients are non-negative. In other words, non-negative least squares is used to solve the equation below for $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ where k is the number of learners

$$E[Y \mid \hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k] = \alpha_1 \hat{Y}_1 + \alpha_2 \hat{Y}_2 + \cdots + \alpha_k \hat{Y}_k \tag{3.23}$$

such that $\alpha_1, \alpha_2, \ldots, \alpha_k \geq 0$. Then, the estimated α_i for $i \in \{1, \ldots, k\}$ values are reweighted so that they sum to 1. This is known as a convex combination of weights and provides greater stability for the final Super Learner prediction[135]. Lastly, the above weights are used to generate the Super Learner that can then be used to predict the outcome in new data. Super Learner predictions are obtained by first predicting the outcome using the learners, and then using the weights to calculate a final weighted prediction. To avoid over-fitting, it is recommended that the performance of the Super Learner itself be evaluated using V-fold CV. To estimate the cross-validated risk of the ensemble itself, nested CV is used, where the Super Learner is trained on $k-1$ subsets of the data and validated using the holdout sample.

Choosing a library of candidate learners

The Super Learner performs asymptotically as well as the best learner used in the ensemble[135]. Incorporating a rich collection of algorithms can only improve Super Learner predictions but increases the computational burden of the algorithm. Another issue with implementing the Super Learner is choosing which learners to include in the library of algorithms. It is suggested that the library include a wide variety of learners, such as random forest, elastic net, and SVM, but the set should reflect what is computationally feasible. For the study in Chapter Six, the library of candidate learners consisted of generalized linear models [logistic regression (main and interaction term models), GAM] and a diverse set of machine learning algorithms [elastic net[146], random forest[88], tree-based extreme gradient boosting (XGBoost)[147], and kernel-based SVM[148]. Each of these learners will be briefly discussed in the

context of predicting a binary outcome.

Consider the training data set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ where each observation i has a p dimensional vector of observed predictors $\boldsymbol{x}_i$ and a binary outcome y_i. The observed binary response of y_i of the random variable Y_i can take one of two values, 0 or 1 (where $y_i = 1$ indicates observation i exhibits the outcome of interest), with corresponding probabilities π_i and $1 - \pi_i$, respectively. The logistic regression model belongs to the family of models called generalized linear models and is defined as

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}, \tag{3.24}$$

where the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are estimated using numerical methods like iteratively re-weighted least squares. The resulting model is used to generate the predicted probability of the outcome in the test sample.

GAMs extend generalized linear models to include smooth functions of continuous predictor variables. That is, the model is defined as

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\theta} + \sum_j f_j(x_{ji}), \tag{3.25}$$

where $\boldsymbol{x}_i$ contains the linear model component with corresponding parameter vector $\boldsymbol{\theta}$, and the $f_j(x_{ji})$ are smooth functions of the covariates x_j. As opposed to one fixed coefficient β as in logistic regression, the function f can change over the range of the predictor. The smoothness of the function is controlled by the degrees of freedom ($deg.gam$) for the smoother[149].

Elastic net is a penalized regression method that shrinks coefficients towards zero[146]. Elastic net solves the problem

$$\min_{\beta_0,\beta} \frac{1}{n}\sum_i^n w_i l(y_i, \boldsymbol{x}_i^T\boldsymbol{\beta}) + \lambda\left[\frac{(1-\alpha)}{2}\sum_{j=1}^p \beta_j^2 + \alpha\sum_{j=1}^p |\beta_j|\right] \tag{3.26}$$

over a grid of values λ. In the equation above, $l(y_i, \boldsymbol{x}_i^T\boldsymbol{\beta})$ is the negative log-likelihood contribution for observation i, and the remaining term is a penalty term that shrinks the estimates of the coefficients. The value of α controls the weight given to the different components of the penalty term where if $\alpha = 1$, (3.26) is called ridge

regression and if $\alpha = 0$ is called least absolute shrinkage and selection operator (LASSO) regression. The tuning parameter λ controls the weight given to the penalty term. Elastic net is useful for reducing the number of predictors in the model and for handling groups of correlated predictors.

Random forest, described above but briefly again here, is an ensemble method that aggregates the predictions of a collection of decision trees. Classification trees are constructed using recursive binary splitting where the predictor space is partitioned into regions to maximize the discrimination between observations with $y_i = 1$ and those with $y_i = 0$. Beginning at the top of the tree, the predictor space is successively split into two new regions, which results in the formation of two new branches. The regions are split until a stopping criterion is met (e.g., a minimum *nodesize* of 5 observations). The random forest algorithm involves repeatedly constructing classification trees using bootstrap samples of size n drawn with replacement from the training sample until the desired number of trees have been made, denoted by *ntree*. To increase variability in the ensemble, a random subset of the predictors is selected at each potential split (*mtry*). The predicted probability of a new observation is found by averaging the outcome values in the terminal nodes in which the new observation resides.

Similar to random forest, tree-based gradient boosting is an ensemble method that aggregates predictions from a collection of decision trees. However, instead of building independent classification trees as in random forest, the decision trees are added sequentially, with each new tree focusing on the reducing the errors of the previous tree. This is done by repeatedly building a classification tree to a weighted version of the training data set with larger weights assigned to misclassified observations[147]. XGBoost is an improved version of gradient boosting designed for enhanced computational speed and model performance[147].

The SVM algorithm works by identifying a decision boundary (i.e., a multidimensional surface) that best separates the data into observations with $y_i = 1$ and those with $y_i = 0$, defined by obtaining the largest possible margin between the two classes. SVM then generates predictions based on this separation boundary. SVM can be specified with a linear hyperplane, but more often data are not linearly separable. In this case, SVMs use kernel functions (e.g., radial basis function kernel) to map

the data to higher dimensions where the decision boundary can be linear. Another feature of the algorithm is defining the boundary margin, or the buffer zone around the boundary. If the margin is small, the model will try to find a boundary that makes fewer errors in the training data, and if the margin is large, the model will tolerate more errors in the training data. This is controlled by the cost parameter C and the optimal value of C balances the trade-off between misclassification and simplicity of the model.

Hyperparameter tuning

The performance of machine learning methods and GAMs depends on the choice of hyperparameter values (e.g., the degrees of freedom for the smoothers in GAM). Hyperparameters adjust an algorithm's characteristics to different aspects of the data, and knowing which values work best a priori is nearly impossible. The traditional approach to selecting the best combination of hyperparameters has been to use a grid search coupled with CV. For user-specified values of each hyperparameter, the grid search method exhaustively considers all possible combinations and selects that which has the smallest cross-validated risk (e.g., Brier score).

Alternatively, the tuning strategy employed in Chapter Six uses the Super Learner algorithm to solve the problem of hyperparameter optimization by creating a weighted ensemble of each learner specified with different parameter settings. To do this, small grids of hyperparameter configurations were constructed for each base learner, and then the Super Learner algorithm (fitted using V-fold CV) was used to estimate the optimal weighted average of the various instances of the same learner with different settings based on the Brier score. This tuning strategy limits the number of algorithms included in the main Super Learner algorithm by allowing each base learner to be included only once as a tuned algorithm. Specific details on the grid of hyperparameter configurations over which each algorithm was tuned is given in Chapter Six.

3.5.3 Model performance and validation

Predictive models built using different algorithms or different variables are often compared and assessed in terms of accuracy. In the context of a binary outcome,

this often consists of assessing and comparing discriminative ability and checking if the models are well-calibrated. Discrimination refers to how well a model can discriminate, or accurately distinguish between, those with and those without the outcome, whereas calibration assesses the agreement between observed outcomes and predicted probabilities. Prior to discussing each of these metrics, definitions for standard measures of diagnostic accuracy are given.

Standard measures of diagnostic accuracy

For the following discussion, consider the confusion matrix for the case of a binary outcome with values "event" and "non-event" given in Table 3.2. Some of those with the event and without the event are correctly predicted as so (true positives (TP) and true negatives (TN)), while others are incorrectly classified as either having the event when in fact they do not (false positive (FP)), or not having the event when in fact they do (false negative (FP)).

Table 3.2: Confusion matrix for a binary classification problem with possible outcomes "event" and "non-event" where cells indicate the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

	Observed	
Predicted	Event	Non-event
Event	TP	FP
Non-event	FN	TN

Suppose the outcome of interest is whether a woman will deliver a LGA infant, where in this case the event is delivering LGA and the non-event is not delivering LGA. Further assume that a woman is classified by the model as delivering LGA if her predicted probability of so is greater than α. The sensitivity of the model is the probability that LGA is predicted correctly for all women who delivered LGA, or

$$\begin{aligned}\text{Sensitivity} &= \frac{\#\text{ who delivered LGA and predicted to deliver LGA}}{\#\text{ who delivered LGA}} \\ &= \frac{TP}{TP+FN}.\end{aligned}$$

This measure is sometimes referred to as a true positive (TP) rate. The specificity of the model, or the true negative (TN) rate, is defined as the probability that women

not delivering LGA are correctly predicted as so, or

$$\begin{aligned}\text{Specificity} &= \frac{\#\text{ who did not deliver LGA and predicted to not deliver LGA}}{\#\text{ who did not deliver LGA}} \\ &= \frac{TN}{FP+TN}.\end{aligned}$$

The false positive (FP) rate is defined as $1 - Specificity$. Assuming a fixed level of accuracy, there is typically a trade-off between sensitivity and specificity. By increasing the sensitivity of a model (i.e., decreasing α), specificity will decrease since more women will be predicted as delivering LGA.

Sensitivity and specificity are conditional measures, and only reflect the probability in the event and non-event groups. For example, the sensitivity is the accuracy rate only for the women who have the event, or delivered LGA. The positive and negative predictive values are analogues of the sensitivity and specificity that additionally account for the prevalence of the outcome in the population, denoted by p. Based on the sensitivity and specificity of particular model or test, the positive predictive value (PPV) represents the probability that a woman delivers LGA given that her predicted probability is greater than α. Conversely, the negative predictive value (NPV) represents the probability that a woman does not deliver LGA given that her predicted probability is less than α. Using the sensitivity and specificity, the PPV is calculated by

$$PPV = \frac{Sensitivity \times p}{(Sensitivity \times p) + \big((1 - Specificity) \times (1-p)\big)}, \tag{3.27}$$

and similarly the negative predictive value (NPV) is calculated by

$$NPV = \frac{Specificity \times (1-p)}{\big(p \times (1 - Sensitivity)\big) + \big(Specificity \times (1-p)\big)}. \tag{3.28}$$

As p increases, the PPV also increases but the NPV decreases. Conversely, as p decreases, the PPV decreases while the NPV increases. For example, if p increases towards 1, $(1-p)$ in the denominator of (3.27) approaches 0 while the term $(Sensitivity \times p)$ in both the numerator and denominator increase, thus increasing the PPV. Similarly, $(1-p)$ in the numerator and denominator of (3.28) approach 0

while $\big(p \times (1 - Sensitivity)\big)$ in the denominator increases, thus decreasing the NPV. PPV and NPV can also be calculated using Table 3.2 by

$$PPV = \frac{TP}{TP + FP}, \tag{3.29}$$

$$NPV = \frac{TN}{TN + FN}. \tag{3.30}$$

Discrimination

Accurate classifiers can effectively discriminate between those with and those without the outcome, or event. The most commonly used measure to assess how well the model classifies observations in a binary prediction problem is the concordance (c) statistic. When the outcome is binary, the c statistic is identical to the AUC-ROC. The c statistic is interpreted as the probability that in a randomly selected pair of observations, one with the outcome and one without, the observation with the outcome has a higher predicted probability.

Receiver operating characteristic (ROC) curves are a valuable tool to examine how the model distinguishes between those with and those without the outcome at various decision rules. The ROC curve plots the *sensitivity* (TP rate) against $1 - specificity$ (FP rate) across a continuum of thresholds corresponding to the probability of the outcome[150]. Plotting the TP rate against the FP rate for each candidate threshold is helpful for determining the threshold that appropriately maximizes the trade-off between sensitivity and specificity. A completely ineffective model would result in an AUC-ROC of 0.5 (comparable to a coin toss), whereas a perfect model that completely separates the two classes would have an AUC-ROC of 1. Superimposing ROC curves is useful to contrast two or more models with different predictor sets, or two different classifiers.

One limitation of ROC curve analyses is that they may be misleading in data with class imbalance (i.e., when a large difference in the number of observations with and without the outcome exist). When the prevalence of the outcome is low, the AUC-ROC may indicate overly optimistic performance because the FP rate is much smaller than the number of TNs, and the AUC-ROC will not change very much even as the number of FPs decreases, as indicative of a better classifier[151, 152]. A proposed

alternative metric is to evaluate a classifier using the area under the precision-recall curve (AUC-PR). The precision-recall (PR) curve is a plot of precision (i.e., PPV) versus recall (i.e., sensitivity or TP rate) across a continuum of thresholds. Since PPV does not incorporate the number of TNs, it is more sensitive to changes in the number of TPs. A completely ineffective model would result in an AUC-PR equal to the prevalence of the outcome, whereas a perfect model that completely separates the two classes would have an AUC-PR of 1.

Calibration

Calibration evaluates the agreement between the predicted risk from the model and the observed risk. A well-calibrated model is one in which the estimated class probabilities are reflective of the true underlying probability of the sample. In predictive models based on binary outcomes, the final output is the estimated probability that the outcome of interest will occur (e.g., a particular woman has a $p\%$ chance of delivering LGA). Calibration is assessed for each individual by checking how close this prediction is to the true underlying probability for that individual. Since it is not possible to determine these underlying probabilities, the probability of the outcome in a similar group of individuals is used as a proxy. For example, calibration is often assessed by comparing the mean observed and predicted probabilities in observations grouped based on deciles calculated on model predictions.

Calibration can help diagnose potential lack of fit and can be assessed graphically using calibration plots. Calibration plots are generated by dividing the sample into groups of equal size (e.g., deciles) based on predicted probabilities and plotting the midpoint of the predicted probability for each group on the x-axis against the true prevalence of the outcome in that group on the y-axis. A well calibrated model would have predictions on the line with zero-intercept and a slope of 1[150].

Chapter 4

Comparison of parametric and nonparametric imputation of missing correlated body mass index values: a simulation study applied to the context of perinatal epidemiology

The target journal for this manuscript is *Statistics in Medicine*.

MM Brown conceptualized and designed the study, performed the analysis, and wrote the initial manuscript draft. C Woolcott, B Smith, and S Kuhle contributed to the design of the study, and provided supervisory input regarding methodology and continuous feedback on the manuscript. J Payne and V Allen provided content expertise and guidance for implications and design of the study. All authors reviewed and revised the manuscript, and approved the final manuscript as presented in this thesis.

Comparison of parametric and nonparametric imputation of correlated missing body mass index values: a simulation study applied to the context of perinatal epidemiology

Mary M Brown[a,b,c], MSc
Stefan Kuhle[a,b], MD, PhD
Bruce Smith[c], PhD
Victoria M. Allen[a], MD, MSc
Jennifer Payne[d], PhD
Christy G. Woolcott[a,b], PhD

Affiliations:
[a]Dept of Obstetrics & Gynaecology
[b]Dept of Pediatrics
[c]Dept of Mathematics and Statistics
[d]Dept of Diagnostic Radiology
Dalhousie University, Halifax, NS, Canada

Corresponding author: Dr. Christy Woolcott, Perinatal Epidemiology Research Unit, IWK Health Centre, 5980 University Avenue, Halifax, NS B3K 6R8, Canada. Email: christy.woolcott@iwk.nshealth.ca

Funding sources: Mary M. Brown received a Nova Scotia Graduate Scholarship from the Nova Scotia Research and Innovation Trust and a Scotia Scholar Doctoral Award from the Nova Scotia Health Research Foundation.

Potential Conflicts of Interest: The authors have no conflicts of interest relevant to this article to disclose.

4.1 Abstract

Missing data present challenges in epidemiologic studies and are frequently addressed using multiple imputation. Appropriate conduct of multiple imputation procedures is further complicated by clustered data structures, like those of perinatal databases, which arise when the data can be separated into naturally occurring groups (e.g., deliveries to the same woman). To ensure valid inference, the imputation procedure needs to preserve all features of the data, and in the case of clustered data, the within-cluster correlation. Modifications to tree-based algorithms to accommodate clustered data have been proposed but have yet to be implemented or evaluated as imputation techniques. We compared imputation of maternal pre-pregnancy body mass index (BMI; weight divided by height squared) using mixed-effects random forest (MERF), an extension of the random forest algorithm to clustered data, to other widely used imputation techniques. We drew 100 samples of all deliveries to 2000 randomly selected women drawn from the 41809 eligible women with complete data in the Nova Scotia Atlee Perinatal Database. Maternal pre-pregnancy weight was simulated to be missing at random (MAR) and missing not at random (MNAR), and height to be missing completely at random (MCAR). Analyses of the association between pre-pregnancy BMI and one continuous and one binary outcome after imputation using MERF resulted in overestimation of the true parameter value by 8.3 to 15.8%. The bias was most pronounced in scenarios where weight was simulated to be MNAR. Parameter estimates were least biased using standard random forest-based imputation, which may be the best choice for imputing missingness in complex clustered data.

4.2 Introduction

Missing data present challenges in health research and are frequently addressed using multiple imputation. Multiple imputation is a general approach that accounts for the uncertainty of the imputations by creating multiple complete datasets where the missing values have been filled in with different plausible values based on a statistical model. The most common implementations of multiple imputation in statistical software packages assume data to be missing at random (MAR)[54, 84] where the probability of missingness depends only on observed data, or missing completely at random (MCAR), a special case of MAR where the probability of missingness depends neither on observed nor unobserved data. Missing not at random (MNAR) occurs if the probability of missingness depends on unobserved data. A test for whether data are MAR or MNAR does not exist, so assessing the likelihood of either mechanism relies on content expertise[153].

Clustered data structures are common in health research. These structures may arise when the data can be separated into naturally occurring groups or clusters (e.g., children in a school, patients in a hospital) or when repeated measurements are taken on the same individuals. Since observations within a cluster tend to be correlated, the assumption of independent observations required by statistical procedures is violated. To obtain valid inference in these settings, methods that can account for the non-independence of observations are required.

Depending on the target inference, clustering is often accommodated by mixed-effects models, estimation using generalized estimating equations (GEE), or standard regression models with clustered-robust standard errors. To ensure valid inference in the analyses of multiply imputed datasets, the imputation model needs to preserve all features of the data, including interactions, nonlinear relationships, and in the case of clustered data, the correlation among observations in the same cluster[154]. Although multilevel imputation techniques are available in standard software, imputation using data-adaptive methods has been proposed as an alternative to parametric imputation in the case of independent observations[34–38]. These methods can capture complex relationships in the data without the need to explicitly specify the imputation model, and have been shown to perform comparably or, in some cases, better than parametric-based imputation[34, 36–39].

Among these proposed data-adaptive imputation methods, tree-based algorithms such as classification trees, regression trees and random forest are commonly used. Several modifications to tree-based algorithms to accommodate clustered data for continuous outcomes have been proposed[40–42]. Hajjem et al. incorporated mixed-effects in the regression tree algorithm using an iterative procedure similar to the expectation-maximization algorithm for linear mixed-effects models (LMER)[40]. They later extended the mixed-effects regression tree method to the random forest algorithm and found that this method had a smaller predictive mean squared error, on average, than random forest with the largest gain in performance observed in settings with large random effects[41]. However, the mixed-effects random forest (MERF) algorithm has not been implemented or evaluated yet in an imputation setting.

Perinatal databases contain pregnancy and birth information for women and their offspring. These databases exhibit a unique clustering structure where the delivery records of the same woman can be linked, resulting in a hierarchical structure of delivery-level data nested within women. This type of design is complicated by women with differing numbers of deliveries (i.e., unequal cluster sizes), variable time between deliveries (i.e., unequal spacing of measurements), and many having only one delivery (i.e., large proportion of singleton clusters). These databases, however, often have missing values, particularly for maternal pre-pregnancy weight, a key variable in perinatal epidemiology. In the context of a complicated clustering structure and weight measurements over time likely being correlated, imputing pre-pregnancy weight using a MERF may perform better than alternative parametric and nonparametric imputation methods.

Therefore, the aim of this study was to describe the performance of MERF as a method for imputing maternal pre-pregnancy body mass index (BMI) values in a real-life perinatal dataset. Imputation using MERF was compared to both parametric and nonparametric imputation methods in the context of estimating population-averaged estimates of the association of maternal pre-pregnancy BMI with birthweight z-score and large for gestational age (LGA) birth, two common perinatal outcomes[155]. Since pre-pregnancy weight, the numerator of pre-pregnancy BMI, may be MNAR, comparisons were made in both MAR and MNAR scenarios.

4.3 Methods

4.3.1 Data source and study population

The Nova Scotia Atlee Perinatal Database (NSAPD) is a long-standing perinatal database containing detailed information on all births in Nova Scotia, Canada since 1987. It grows by approximately 8000 deliveries annually. All residents in Nova Scotia are assigned a provincial health card number that grants access to publicly funded medical and hospital services. The NSAPD records the provincial health card number of women and their offspring in the database, thus enabling internal linkage of all deliveries within a woman.

Detailed demographic and birth information on deliveries occurring between 2000 and 2019 was obtained from the NSAPD. Multiple gestations (i.e., twins, triplets, etc.) and pregnancies affected by major congenital anomalies were excluded. The exposure of interest was pre-pregnancy BMI (pre-pregnancy weight in kg over height in meters squared), and the outcomes were birthweight z-score and LGA birth. Birthweight for gestational age and sex z-scores were calculated relative to a Canadian reference population[156] and LGA birth was defined as birthweight >90th percentile for gestational age and sex relative to the same reference population. Confounding variables of the association between pre-pregnancy BMI and the outcomes included year of delivery, maternal age, parity, marital status, area-level income quintile (proxy measure for socioeconomic status), smoking in pregnancy, and pre-existing diabetes. Auxiliary variables (those that may be correlated with the missing variables or associated with their missingness) used in the imputation procedure included pre-pregnancy weight, delivery weight, gestational diabetes, and mode of delivery. Additional details on these variables can be found elsewhere[30, 157].

4.3.2 Inducing missingness

A dataset of women with complete information was created by excluding deliveries with implausible values of birthweight z-score (>5 in absolute value) and women with incomplete data in any of their first five recorded eligible deliveries. One-hundred samples of 2000 women randomly selected without replacement from the complete dataset were created. All deliveries from each of these women were included. For

each of the 100 samples, the missing data pattern shown in Table 4.1 was simulated assuming a MAR and a MNAR mechanism for pre-pregnancy weight, resulting in 100 MAR and 100 MNAR datasets.

Evidence from the literature and the original data (with missingness) were used to inform the missing data pattern and the factors affecting the missing mechanism. For example, logistic regression models were fitted in the original dataset to explore possible factors associated with missingness in pre-pregnancy weight. Under both mechanisms, the probability that a delivery record had a missing pre-pregnancy weight value was higher if the woman was older, not married or common-law, and reported smoking in her pregnancy, while for MNAR, the probability of missingness also increased by, on average, 0.015 for each 1-kg increase in weight. Pre-pregnancy height and marital status were simulated to be MCAR. Smoking in pregnancy and delivery weight were simulated to be MAR with the probability of missing smoking information being higher if the woman was of lower socioeconomic status and the probability of missing delivery weight being higher if the woman was not married or common-law and had a planned Caesarean section. Pre-pregnancy BMI was set to missing if either weight, height, or both were missing. The outcome variables had complete data.

The "ampute" function in the *mice*[82] R package was used to induce missingness. The procedure creates k subsets of the complete data where k is the number of missing data patterns. For MAR and MNAR, a weighted sum score is calculated as the outcome of a linear regression equation where the coefficients or weights are determined by the user. In general, larger weights will have higher sum scores and results in that particular covariate in the pattern having a higher influence on the overall score. Based on the weighted sum score, an observation obtains a certain probability of being set to missing. In the case of MCAR, all observations have the same probability of being set to missing. These probabilities are assigned using a continuous logistic distribution function where cases with high weighted sum scores have a higher probability of being missing than those with low weighted sum scores. In every subset, the user-specified proportion of observations is made incomplete according to the missing data pattern and all subsets are merged to create one dataset with complete and incomplete observations.

4.3.3 Imputation methods

For each of the samples, multiple imputation using chained equations (MICE) was used to generate 10 imputed datasets (10 iterations) in which pre-pregnancy BMI was actively imputed (i.e., imputed directly) using MERF and four other widely used imputation techniques (LM - normal linear model ["mice.impute.norm"], PMM - predictive mean matching ["mice.impute.pmm"], RF - random forest ["mice.impute.rf"] with default settings, and LMER - linear mixed-effects model with a random intercept ["mice.impute.2l.norm"]).

Since the MERF algorithm has not been used for imputation, more details of the algorithm and its use as an imputation method are outlined below. Imputation models for all techniques included the outcome, confounders, and auxiliary variables. Missing values of confounders and auxiliary variables were imputed using logistic regression (marital status and smoking in pregnancy) and predictive mean matching (delivery weight and pre-pregnancy weight). A smaller number of imputed datasets was used than what would generally be recommended[158] due to the computational burden of the MERF algorithm. Although MERF could have been used to impute all continuous variables included in the imputation model, predictive mean matching was used to reduce computation time and to focus on differences in the imputation of pre-pregnancy BMI.

MERF-based imputation

A MERF of regression trees is defined as

$$
\begin{aligned}
&y_i = f(X_i) + Z_i b_i + \epsilon_i, \\
&b_i \sim N(0, D), \qquad \epsilon_i \sim N(0, R_i), \qquad i = 1, \ldots n,
\end{aligned}
$$

where $y_i = [y_{i1}, \ldots, y_{in_i}]^T$ is the $n_i \times 1$ vector of responses for the n_i observations in cluster i, $X_i = [x_{i1}, \ldots, x_{in_i}]^T$ is the $n_i \times p$ matrix of fixed-effects covariates, $Z_i = [z_{i1}, \ldots, z_{in_i}]^T$ is the $n_i \times q$ matrix of random-effects covariates, $b_i = [b_{i1}, \ldots, b_{iq}]^T$ is the $q \times 1$ unknown vector of random effects for cluster i, $\epsilon_i = [\epsilon_{i1}, \ldots, \epsilon_{in_i}]^T$ is the $n_i \times 1$ vector of errors, and D and $R_i = \sigma^2 I_{n_i}$ are the covariance matrices of b_i and ϵ_i, respectively. The model assumes b_i and ϵ_i are independent and normally distributed,

and that between-cluster observations are independent. The covariance matrix of the vector of observations y_i in cluster i is defined as $V_i = Cov(y_i) = Z_i D Z_i^T + R_i$, and $V = Cov(y) = diag(V_1, \ldots, V_n)$, where $y = [y_1^T, \ldots, y_n^T]^T$[41].

The unknown function $f(X_i)$ is estimated using a forest of regression trees, and the random part $Z_i b_i$ is assumed to be linear. Estimation of the model parameters is similar to the expectation-maximization algorithm for LMER. More details of this algorithm can be found in Supplementary Methods 1 and in Hajjem et al[41]. The predicted response of a new observation j that belongs to cluster i is calculated using its corresponding population-averaged random forest prediction, $\hat{f}(x_{ij})$, and, if cluster i was used to build the model, additionally using the random part corresponding to its cluster $Z_i \hat{b}_i$.

The proposed imputation technique using MERF modifies the algorithm for random forest-based imputation developed by Shah et al.[37] ("mice.impute.rfcont") by replacing the random forest with a MERF. Briefly, the random forest imputation technique takes a bootstrap sample (with replacement) of those with observed values in the variable to be imputed and then constructs each regression tree in the forest using another bootstrap sample of the data. Observations with missing values are imputed by taking random draws from independent normal distributions with conditional means predicted using the random forest and variances estimated from the out-of-bag mean square error. Since the datasets with induced missingness consisted mostly of clusters with size 1 or 2 (average cluster size of 1.5), only a random intercept term was estimated.

The MERF imputation method uses the *randomForest*[159] R package and requires specification of the random forest parameters. For the simulation study, each forest was created using 300 regression trees, the number of predictors considered for splitting at each node was set to the square root of the number of predictors (rounded down to the nearest integer), and the minimum size of the terminal nodes was set to 5 observations. The MERF algorithm was forced to iterate a minimum of 100 times and was set to continue iterating while the absolute change in generalized-log likelihood criterion was greater than 0.0001 or a maximum of 150 iterations was reached.

4.3.4 Analysis

Pooled population-averaged estimates of the association between pre-pregnancy BMI and the outcomes of interest, adjusted for relevant confounders, were obtained for each set of multiply imputed datasets using GEE with an exchangeable working correlation structure. For each of the five imputation techniques, this resulted in pooled estimates of the marginal change in mean birthweight z-score and the log odds ratio of LGA birth for a 1-kg/m^2 increase in pre-pregnancy BMI. To assess the robustness of the results to the choice of analysis model, pooled estimates were additionally obtained by fitting generalized linear models (identity and logit link function) with clustered-robust standard errors. The true parameter estimates were obtained from the same models fitted to the complete sample. All analyses were performed in R (version 4.1.2)[160] and R Studio[161]. A flowchart of the simulation procedure is shown in Figure 4.1.

4.3.5 Performance measures

For each imputation technique, absolute mean bias (mean difference in the estimate and the true parameter value), percent relative bias (absolute mean bias divided by the true parameter value multiplied by 100), efficiency (expressed by the empirical standard error (SE) over all simulations), and coverage probability of 95% confidence intervals (proportion of 95% confidence intervals for the estimate that contain the true parameter value) were computed. A coverage probability between 0.906 and 0.994 was deemed appropriate as per the suggestion that coverage be within two SE of the confidence level, p, where $SE(p) = \sqrt{(p(1-p)/100)}$[162]. Results of the analysis of complete-cases (dropping missing observations) are provided for comparison.

4.4 Results

The original sample consisted of 166473 deliveries to 102864 women, and after removing delivery records with implausible values of birthweight z-score and women with incomplete data in any of their first five recorded deliveries, the final complete-data sample consisted of 64204 deliveries to 41809 women.

Across the 100 complete samples of 2000 randomly selected women, on average,

56.1%, 35.7%, 6.9%, 1.0%, and 0.3% had had one, two, three, four, and five deliveries, respectively, resulting in an average of 3070 deliveries and an average cluster size of 1.5 within each sample. Descriptive statistics on deliveries in the original dataset, eligible women with complete data in all recorded deliveries, and the samples with missingness in pre-pregnancy weight simulated to be MAR and MNAR (averaged across the 100 samples) are shown in Table 4.5. Boxplots of the distribution of pre-pregnancy height, weight, and BMI values in 20 of the 100 datasets before inducing missingness (complete) and after simulating pre-pregnancy weight to be MAR and MNAR are shown in Figure 4.2. In general, the distribution of pre-pregnancy height was similar in all samples, and the distribution of weight and BMI values in complete and MAR datasets were comparable. In MNAR samples, the distributions of pre-pregnancy weight and BMI were shifted to the left, had smaller medians, and contained less positive extreme values (as expected based on how missingness was simulated).

True parameter values for the associations of interest and estimates of bias (absolute and relative to the true parameter value [%]), efficiency, and coverage probability of the 95% confidence intervals in complete-case analyses and in analyses using each imputation method are shown in Table 4.3, and boxplots of the distribution of estimates are shown in Figure S3. In analyses of the continuous outcome, birthweight z-score, imputation using MERF resulted in less bias than imputation using a linear model, predictive mean matching, and a linear mixed-effects model with a random intercept in MAR data (|relative bias|, 8.3% vs. 11.7-15.8%), but similarly biased in MNAR data (14.1% vs. 10.9-12.1%). Under MAR, imputation using MERF was slightly more biased than random forest (8.3% vs. 5.3%) and substantially more in MNAR scenarios (14.1% vs. 1.6%). In analyses of the binary outcome, LGA birth, and with the exception of random forest, imputation using MERF in MAR data resulted in similar bias to all imputation methods considered (9.9% vs. 10.7-11.7%) and had comparable SEs (0.009 vs. 0.008-0.009). Imputation using MERF was more biased than imputation using random forest in both MAR and MNAR scenarios (9.9% vs. 2.8%, and 15.8% vs. 0.0%, respectively) and was less efficient (0.009-0.012 vs. 0.008-0.010). In all scenarios, MERF was positively biased, resulting in an overestimation of the true effect.

When the associations of interest were estimated using generalized linear models with clustered-robust standard errors (Table 4.4), MERF was positively biased in all scenarios (12.9-21.1%), with the largest bias occurring in analyses of MNAR samples. In all scenarios, imputation using a linear model and predictive mean matching were less biased than those when the analysis model was estimated using GEE (0.7-6.7% vs. 5.7-15.8%) while imputation using a linear mixed-effects model was resulted in similar bias (10.6-11.9% vs. 8.0-9.2%). Imputation using random forest resulted in similar or slightly more bias compared to the results of the GEE analyses (1.5-7.4% vs. 0.0-5.3%).

4.5 Discussion

The present study aimed to evaluate MICE-based imputation of clustered data using MERF; we used real-life data from a perinatal database with clusters of deliveries nested within women. Analyses of the association between pre-pregnancy BMI and one continuous and one binary outcome after imputing pre-pregnancy BMI using MERF resulted in overestimation of the true parameter value by 8.3 to 15.8%. When pregnancy weight was simulated to be MNAR, the bias was most pronounced and, as expected, coverage of the 95% confidence intervals was poor (0.840). With the exception of random forest, imputation using MERF performed better than that using a linear model, predictive mean matching, and linear-mixed effects model when pre-pregnancy weight was simulated to be MAR, and performed worse or similarly when pre-pregnancy weight was simulated to be MNAR. Imputation using random forest was minimally biased in all settings.

Random forest was generally the least biased, most efficient, and had the highest coverage probability of the imputation methods considered. Although imputation using MERF uses the random forest algorithm, differences between the two techniques may explain, at least in part, the differences observed in this study. First, the MERF algorithm is similar to the expectation-maximization algorithm for a linear mixed-effects model, so predictions from the random forest are updated using the current available estimate of the random part, which are then used in the subsequent iteration to build the next random forest. The final prediction from the MERF for observation j from cluster i is calculated in one of two ways. If cluster i was not used

to build the MERF, the predicted value is calculated using solely its corresponding population-averaged random forest prediction, and if cluster i was used to build the MERF, the predicted value is calculated using its population-averaged random forest prediction and the predicted random part corresponding to its cluster. The predictions from a MERF fitted to the complete-data set indicated that the random effect estimates were small, and so, as expected, predictions from the MERF and random forest were similar, so it is unlikely the algorithm differences impacted the results greatly.

Secondly, the proposed imputation technique using MERF modifies the imputation algorithm for random forest developed by Shah et al.[37] ("mice.impute.rfcont"), which generates imputations differently than the default random forest imputation algorithm in the mice package developed by Doove et al.[36] ("mice.impute.rf"). In "mice.impute.rfcont", observations with missing values are imputed by taking random draws from independent normal distributions with conditional means predicted using the random forest and variances estimated from the out-of-bag mean square error. In "mice.impute.rf", missing values are imputed by taking the observed value of one randomly selected donor from the set of observations in the terminal nodes of the random forest trees in which the observation with a missing value resides. One assumption of the "mice.impute.rfcont" method is the assumption that the residuals from the random forest regression are normally distributed with constant variance. Investigation of residual plots from MERFs fitted to complete-data samples indicated this assumption was likely not met and may have resulted in poor imputations.

Imputation using MERF performed best when the analysis model was estimated using GEE and pre-pregnancy weight was simulated to be MAR. Since parameters estimated using GEE are fitted by minimizing the weighted sum of squared residuals using the working covariance matrix as the weight, an imputation method that uses the correlation among observations to generate imputations, even if it is nonparametric, may perform better than imputation methods that ignore the correlation structure (e.g., normal linear model or predictive mean matching). However, when clustering is accommodated by using standard generalized linear models with clustered-robust standard errors, methods that ignore the clustering were less biased

with appropriate coverage than both MERF and imputation using a linear mixed-effects model.

In the present study, analyses of complete-cases resulted in minimal bias in all settings considered. Complete-case analysis will be valid under MAR and MNAR when the outcome is complete, and the complete cases can be thought of as a random sample of all cases within strata formed by the covariates used for adjustment[163]. For this study, the variables used to induce missingness under MAR were also included in the final analytical model since they were confounding variables. Under MNAR, the probability of missingness in pre-pregnancy weight was simulated to increase with weight values, and since weight is highly correlated with BMI (derived from pre-pregnancy weight and height), analyses of complete-cases is likely to be minimally biased. Although a complete-case analysis resulted in unbiased estimates of the associations considered in this study, because of the smaller sample size, the estimates had larger standard errors than those obtained with multiple imputation. The proportion of missing pre-pregnancy BMI can be quite large in perinatal databases, so complete-case analysis may be insufficiently powered. The loss of statistical efficiency might become quite problematic when examining rare outcomes or outcomes in specific subgroups of the population.

This study used data with a specific and unique clustering structure, that is, one that consists of nearly 50% singleton clusters and an average cluster size of 1.5 observations. In this data structure, MERF performed poorly as an imputation technique as it led to gross over-estimation of the true effect and was much more computationally intensive compared to random forest (approximately 70 times longer to generate one imputed dataset). The current study is the first to investigate random forest as an imputation method in a clustered data structure and we found it to be the best method for imputing pre-pregnancy BMI. The results of other studies comparing random forest-based imputation to parametric methods in datasets with independent observations suggest this method is superior for imputation in complex datasets (e.g., interactions and nonlinearities) when the parametric imputation model is likely to be misspecified[37, 39]. Together, random forest-based imputation may be useful for imputing complex epidemiologic datasets with either dependent or independent observations. However, further research is required, and although random forest

was found to be the best method in this study, there may be scenarios in which imputation using MERF would be better. For example, in data with larger cluster sizes that enable the estimation of more complex models (e.g., random slope), and larger random effects, and for imputing variables that do not exhibit the extreme positive skewedness typically observed in BMI distributions.

The main strengths of this study were that simulations were performed using real data and that the analysis was realistically complex. Resampling from a real dataset yields samples that more accurately reflect the true underlying distributions of the variables and the possible complex relationships they have with each other, which may be difficult to recreate in simulated samples. However, this study has several limitations. First, since samples were drawn from a real dataset and missingness was artificially simulated to reflect reality, it is unclear if the results of this study can be generalized to other data sets, such as those with different clustering structures, assumed missingness mechanisms, and proportion of missing data; to other outcome types, such as time-to-event outcomes; and to other associations with different effect sizes. Secondly, this study assumed that there was no woman-specific effect on the probability of a delivery having a missing pre-pregnancy BMI value. As it is plausible that there may exist within-women differences in the baseline probability of missingness, investigating the performance of MERF and other imputation methods in data where missingness in pre-pregnancy BMI is simulated using a random effects model is an important area for future research. Thirdly, imputation using MERF (and RF) relies on drawing bootstrap samples from deliveries with complete information, and thus does not necessarily preserve the clustering structure of the data. It is unclear whether imputation of missingness would improve by using a bootstrapping procedure that maintains the clustering structure and should be the subject of future research. Lastly, to avoid excessive computation time, only 100 simulations were performed, and 10 imputed datasets were used for the comparisons, which may have resulted in noisy estimates of between-imputation variability.

In summary, this study investigated the use of a novel tree-based imputation technique based on mixed-effects random forest and compared its performance to widely used parametric and nonparametric imputation methods, such as imputation using a linear model, a linear mixed-effects model, predictive mean matching, and

random forest. Imputation using MERF was moderately biased when pre-pregnancy weight was MAR but was severely biased when MNAR. We found random forest-based imputation to be the least biased and most efficient method for this specific type of clustered data structure and that this method performed better than imputation methods that both ignore and include clustering in the imputation procedure. Further research on the use of tree-based imputation methods in data with different clustering structures, assumed missingness mechanisms, and other outcome types is needed.

Table 4.1: Missing data pattern simulated in each of the 100 samples where a "1" indicates the variable was complete and a "0" for incomplete. The first column corresponds to the proportion of observations with the corresponding pattern. The final row is the overall proportion of missingness for that variable

Prop.	**DL year**	**Age**	**Parity**	**Marr-ied**	**SES**	**Smk**	**Prx-DM**	**DL weight**	**Mode of DL**	**GDM**	**BWZ/ LGA**	**PP weight**	**PP height**	**PP BMI**
0.45	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
0.1	1	1	1	1	1	1	1	1	1	1	1	0	1	0
0.1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
0.05	1	1	1	1	1	1	1	1	1	1	1	0	0	0
0.1	1	1	1	1	1	1	1	0	1	1	1	0	1	0
0.05	1	1	1	0	1	1	1	1	1	1	1	1	1	1
0.05	1	1	1	1	1	0	1	1	1	1	1	1	1	1
	0	0	0	0.05	0	0.05	0	0.2	0	0	0	0.25	0.15	0.35

Abbreviations: *BWZ* birthweight z-score; *DL* delivery; *GDM* gestational diabetes; *LGA* large for gestational age; *PP* pre-pregnancy; *Prx-DM* pre-existing diabetes; *SES* socioeconomic status; *Smk* smoking in pregnancy

Table 4.2: Descriptive statistics of the original dataset, the subsample of women with complete data on all their deliveries, and the 100 missing samples (averaged across samples) in which pre-pregnancy weight was simulated to be missing at random (MAR) and missing not at random (MNAR)

	Original (n=166473)		**Complete (n=64204)**	**Amputed samples (mean size of n=3070)**			
				MAR		**MNAR**	
	missing, %	mean (SD) or %	mean (SD) or %	missing, %	mean (SD) or %	missing, %	mean (SD) or %
Maternal pregnancy and delivery characteristics							
Pre-pregnancy height [cm]	22.3	164.3 (6.8)	164.3 (6.8)	15.1	164.3 (6.8)	15.1	164.3 (6.8)
Pre-pregnancy weight [kg]	16.5	70.7 (18.0),	70.7 (18.2),	24.7	70.6 (18.1),	24.5	68.9 (17.1),
mean (SD), median (IQR)		66.7 (58.1-79.4)	66.2 (58.1-79.4)		66.2 (58.1-79.4)		64.9 (56.7-77.1)
Pre-pregnancy BMI [kg/m^2]	32.7	26.2 (6.5),	26.2 (6.4),	34.7	26.1 (6.4),	34.7	25.4 (6.0),
mean (SD), median (IQR)		24.6 (21.6-29.4)	24.6 (21.6-29.3)		24.5 (21.5-29.2)		23.9 (21.2-28.2)
Age [years]	0.0	29.3 (5.6)	29.5 (5.6)	0.0	29.5 (5.6)	0.0	29.5 (5.6)
Married/common-law	6.7	73.1	74.7	5.0	74.6	5.0	74.6
Socioeconomic status	1.9			0.0		0.0	
Low		18.9	17.4		17.3		17.3
Middle		65.5	66.2		66.4		66.4
High		15.6	16.4		16.3		16.3
Parity	0.0			0.0		0.0	
0		61.8	65.1		65.1		65.1
1		28.7	28.6		28.6		28.6
2		7.3	5.3		5.3		5.3
3		1.7	0.8		0.8		0.8
4		0.5	0.2		0.2		0.2
Smoking in pregnancy	0.9	20.5	16.5	4.8	16.7	4.8	16.7
Pre-existing diabetes	0.0	0.9	1.0	0.0	1.0	0.0	1.0
Gestational diabetes	0.0	4.9	5.6	0.0	5.7	0.0	5.7
Delivery weight [kg], mean	18.4	85.6 (17.8),	85.7 (17.9),	19.5	85.6 (17.7),	19.5	85.0 (17.5),
(SD), median (IQR)		82.6 (73.0-95.3)	82.6 (73.0-95.3)		82.6 (73.0-94.8)		81.6 (72.6-94.3)
Vaginal delivery	0.0	73.7	71.7	0.0	71.8	0.0	71.8

Table 4.2: continued

	Original (n=166473)		**Complete (n=64204)**	**Amputed samples (mean size of n=3070)**			
				MAR		**MNAR**	
	missing, %	mean (SD) or %	mean (SD) or %	missing, %	mean (SD) or %	missing, %	mean (SD) or %
Neonatal characteristics							
Birthweight z-score	0.5	0.19 (1.0)	0.15 (1.0)	0.0	0.15 (1.0)	0.0	0.15 (1.0)
LGA birth	8.1	14.5	13.5	0.0	13.4	0.0	13.4

Abbreviations: *BMI* body mass index; *IQR* interquartile range (first quartile - third quartile); *LGA* large for gestational age; *MAR* missing at random; *MNAR* missing not at random; *SD* standard deviation

Table 4.3: Bias (absolute and relative to the true parameter value), empirical standard error (SE), and coverage probability of the 95% confidence intervals (CI) for the population-averaged estimates of the association of pre-pregnancy BMI and the outcomes of interest obtained using generalized estimating equations in analyses of complete-cases and imputed data

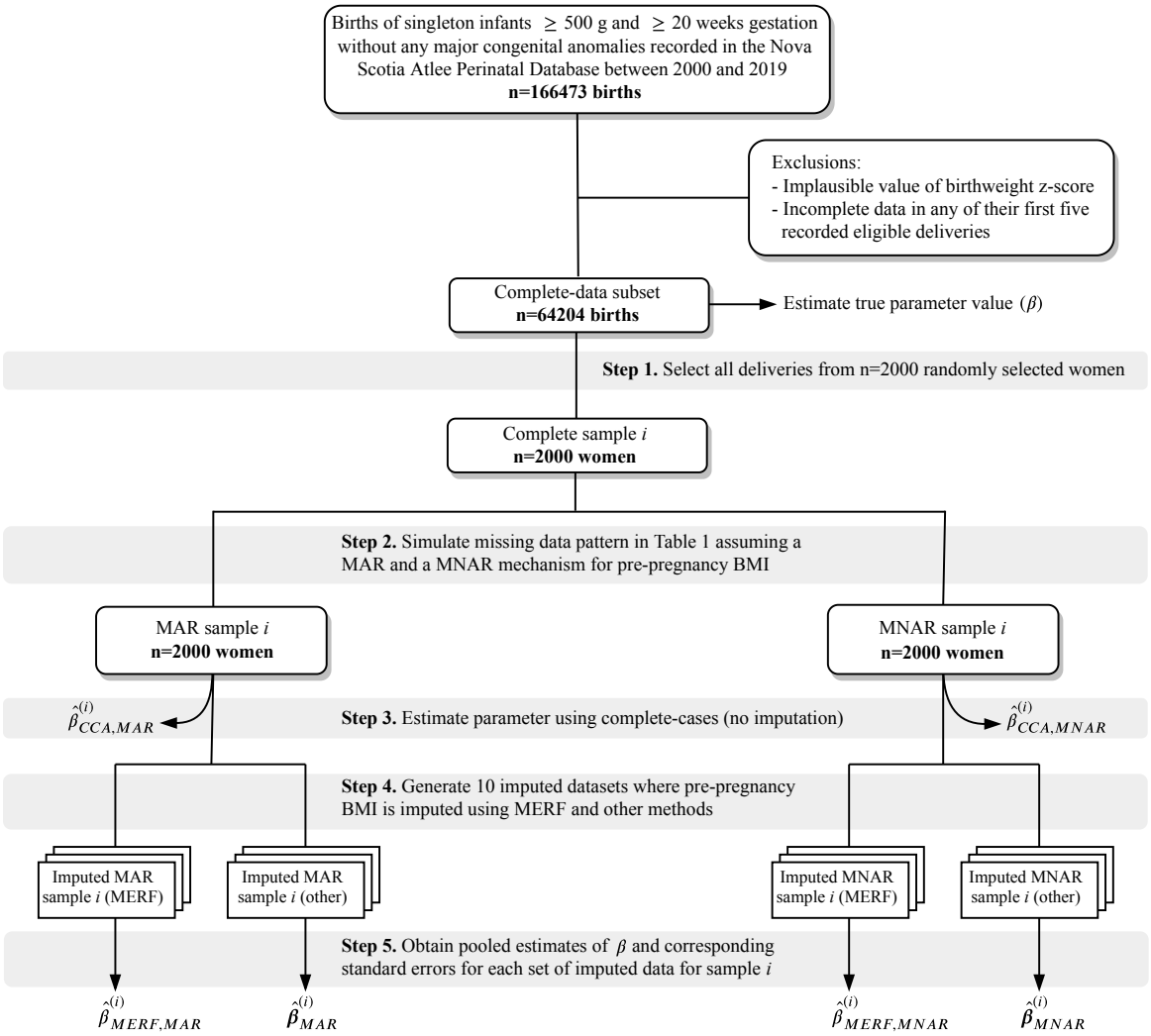
Outcome (true parameter value) and imputation method	**MAR**				**MNAR**			
	Absolute bias	Relative bias (%)	Empirical SE	Coverage	Absolute bias	Relative bias (%)	Empirical SE	Coverage
Birthweight z-score (0.0231)								
None (complete cases)	-0.0006	-2.5	0.0050	0.889	0.0011	4.6	0.0051	0.917
LM	-0.0036	-15.5	0.0039	0.806	-0.0025	-10.9	0.0037	0.833
PMM	-0.0037	-15.8	0.0039	0.694	-0.0028	-12.1	0.0037	0.833
RF	-0.0012	-5.3	0.0036	0.917	-0.0004	-1.6	0.0035	0.972
LMER	-0.0027	-11.7	0.0034	0.806	-0.0028	-11.9	0.0029	0.861
MERF	0.0019	8.3	0.0037	0.950	0.0033	14.1	0.0042	0.840
LGA birth (0.0474)								
None (complete cases)	-0.0015	-3.1	0.0101	0.972	0.0007	1.5	0.0117	0.917
LM	-0.0051	-10.7	0.0091	0.944	-0.0027	-5.7	0.0109	0.917
PMM	-0.0055	-11.7	0.0093	0.944	-0.0039	-8.1	0.0108	0.917
RF	-0.0014	-2.8	0.0078	1.000	0.0000	0.0	0.0096	0.917
LMER	-0.0051	-10.8	0.0079	0.889	-0.0050	-10.6	0.0076	0.917
MERF	0.0047	9.9	0.0087	0.930	0.0075	15.8	0.0116	0.840

Abbreviations: *LGA* large for gestational age; *LM* normal linear model; *LMER* linear mixed-effects model; *MAR* missing at random; *MERF* mixed-effects random forest; *MNAR* missing not at random; *PMM* predictive mean matching; *RF* random forest

Table 4.4: Bias (absolute and relative to the true parameter value), empirical standard error (SE), and coverage probability of the 95% confidence intervals (CI) for the population-averaged estimates of the association of pre-pregnancy BMI and the outcomes of interest obtained using regression with robust standard errors in analyses of complete-cases and imputed data

Outcome (true parameter value) and imputation method	**MAR**				**MNAR**			
	Absolute bias	Relative bias (%)	Empirical SE	Coverage	Absolute bias	Relative bias (%)	Empirical SE	Coverage
Birthweight z-score (0.0231)								
None (complete cases)	-0.0007	-3.1	0.0050	0.889	0.0011	4.5	0.0051	0.944
LM	-0.0014	-6.0	0.0040	0.833	-0.0002	-1.0	0.0039	0.917
PMM	-0.0015	-6.3	0.0040	0.833	-0.0005	-2.2	0.0040	0.972
RF	0.0008	3.2	0.0037	0.917	0.0018	7.4	0.0038	0.972
LMER	-0.0021	-8.7	0.0034	0.917	-0.0019	-8.0	0.0032	0.917
MERF	0.0035	14.7	0.0039	0.830	0.0050	21.1	0.0041	0.690
LGA birth (0.0482)								
None (complete cases)	-0.0016	-3.4	0.0105	0.972	0.0007	1.4	0.0119	0.944
LM	-0.0028	-5.8	0.0092	0.917	-0.0004	-0.7	0.0110	0.917
PMM	-0.0032	-6.7	0.0094	0.944	-0.0014	-2.9	0.0111	0.917
RF	0.0007	1.5	0.0081	0.972	0.0022	4.6	0.0098	0.917
LMER	-0.0045	-9.2	0.0080	0.889	-0.0042	-8.7	0.0081	0.917
MERF	0.0062	12.9	0.0091	0.920	0.0093	19.3	0.0101	0.790

Abbreviations: *LGA* large for gestational age; *LM* normal linear model; *LMER* linear mixed-effects model; *MAR* missing at random; *MERF* mixed-effects random forest; *MNAR* missing not at random; *PMM* predictive mean matching; *RF* random forest



Abbreviations: *BMI* body mass index; *CCA* complete-case analysis; *MAR* missing at random; *MERF* mixed-effects random forest; *MNAR* missing not at random

Figure 4.1: Steps to generate samples with missingness according to Table 4.1 assuming a MAR and a MNAR mechanism for pre-pregnancy BMI and to obtain pooled parameter estimates from multiply imputed datasets after imputing pre-pregnancy BMI using mixed-effects random forest and other widely used imputation techniques (normal linear model, predictive mean matching, random forest, and linear mixed-effects model)

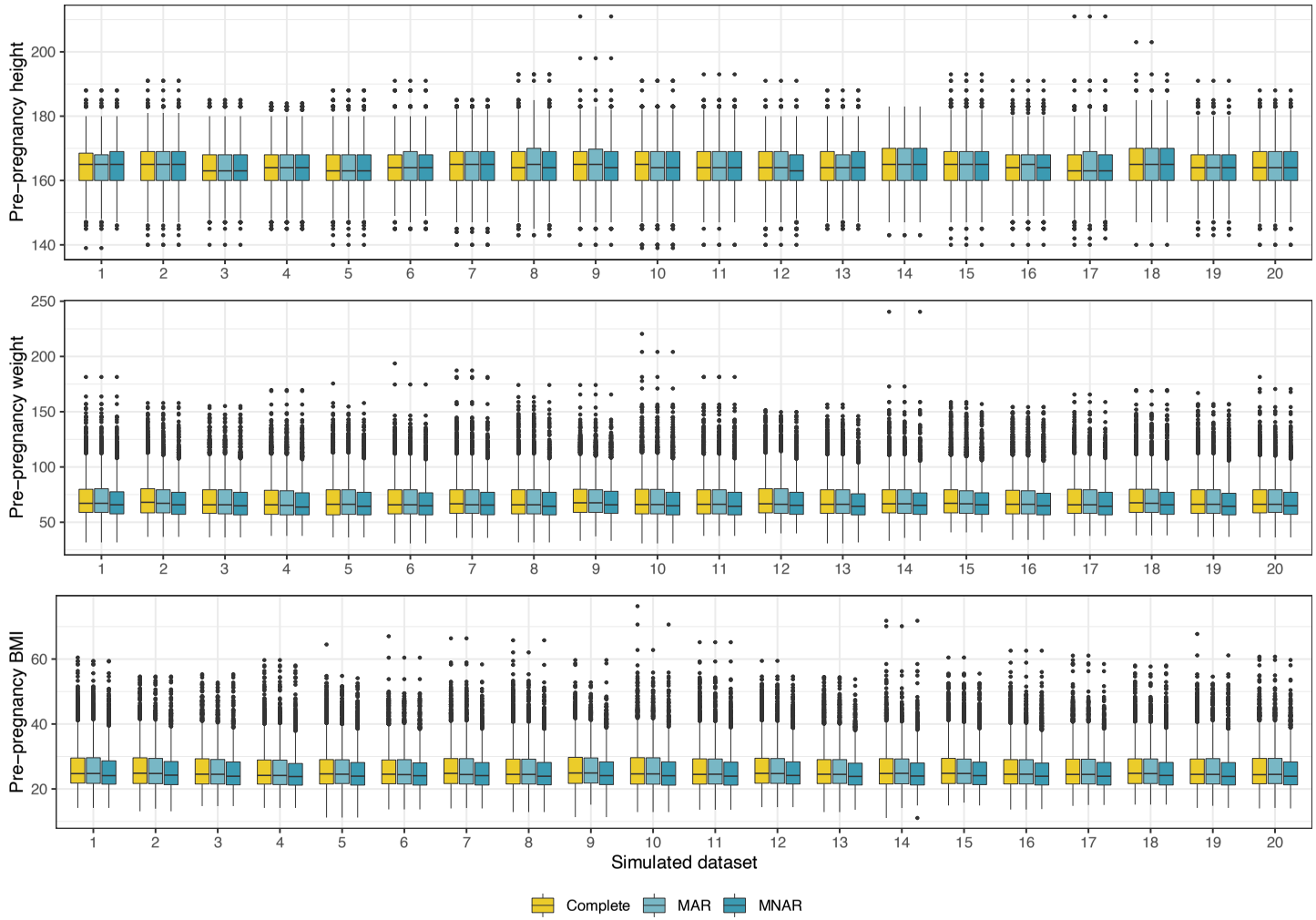


Figure 4.2: Boxplots of pre-pregnancy height, weight, and body mass index (BMI) in 20 of the 100 datasets before inducing missingness (complete) and where pre-pregnancy weight was simulated to be missing at random (MAR) and missing not at random (MNAR). In both MAR and MNAR scenarios, pre-pregnancy height was simulated to be missing completely at random

4.6 Supplementary Methods 1: Mixed-effects random forest algorithm

Algorithm 2: Mixed-effects random forest

Step 0. Set $r = 0$. Let $b_{i(0)} = \overrightarrow{\mathbf{0}}_q$, $\hat{\sigma}^2_{(0)} = 1$, and $\hat{D}_{(0)} = 100^{-1} I_q$.

while $GLL \geq \epsilon$ **do**

Step 1. Set $r = r + 1$.

i) Update $y^*_{i(r)} = y_i - Z_i \hat{b}_{i(r-1)}$, $i = 1, \ldots, n$.

ii) Using $y^*_{ij(r)}$ and x_{ij} for $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, as the full set of training responses and covariates, build *ntree* regression trees using the random forest algorithm, where each tree is built using a bootstrap sample drawn with replacement from $(y^*_{ij(r)}, x_{ij})$.

iii) Obtain estimate $\hat{f}(x_{ij})_{(r)}$ of $f(x_{ij})$ by taking the mean prediction from the subset of trees that are built with the bootstrap samples not containing observation j in cluster i, or the out-of-bag prediction

$$\hat{f}(X_i)_{(r)} = [\hat{f}(x_{i1})_{(r)}, \ldots, \hat{f}(x_{in_i})_{(r)}]^T$$

iv) Update $\hat{b}_{i(r)}$ using

$$\hat{b}_{i(r)} = \hat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} (y_i - \hat{f}(X_i)_{(r)})$$

where $\hat{V}_{i(r-1)} = Z_i \hat{D}_{(r-1)} Z_i^T + \hat{\sigma}^2_{(r-1)} I_{n_i}$ for $i = 1, \ldots, n$.

Step 2. Update $\hat{\sigma}^2_{(r)}$ and $\hat{D}_{(r)}$ using

$$\hat{\sigma}^2_{(r)} = N^{-1} \sum_{i=1}^{n} \{\hat{\epsilon}^T_{i(r)} \hat{\epsilon}_{i(r)} + \hat{\sigma}^2_{(r-1)} [n_i - \hat{\sigma}^2_{(r-1)} \text{trace}(\hat{V}_{i(r-1)})]\}$$

$$\hat{D}_{(r)} = n^{-1} \sum_{i=1}^{n} \{\hat{b}_{i(r)} \hat{b}^T_{i(r)} + [\hat{D}_{(r-1)} - \hat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} Z_i \hat{D}_{(r-1)}]\}$$

where $\hat{\epsilon}_{i(r)} = y_i - \hat{f}(X_i)_{(r)} - Z_i \hat{b}_{i(r)}$.

Step 3. Calculate the generalized log-likelihood (GLL) criterion

$$GLL(f, b_i \mid y) = \sum_{i=1}^{n} \{[y_i - f(X_i) - Z_i b_i]^T R_i^{-1} [y_i - f(X_i) - Z_i b_i] + b_i^T D^{-1} b_i + \log |D| + \log |R_i|\}.$$

end

Chapter 5

Grandmaternal pre-pregnancy body mass index and infant birthweight: a mediation analysis of maternal pre-pregnancy body mass index among Nova Scotians

The target journal for this manuscript is *Obesity*.

MM Brown conceptualized and designed the study, performed the analysis, and wrote the initial manuscript draft. C Woolcott, B Smith, and S Kuhle contributed to the design of the study, and provided supervisory input regarding methodology and continuous feedback on the manuscript. J Payne and V Allen provided content expertise and guidance for implications and design of the study. All authors reviewed and revised the manuscript, and approved the final manuscript as presented in this thesis.

Grandmaternal pre-pregnancy body mass index and infant birthweight: a mediation analysis of maternal pre-pregnancy body mass index among Nova Scotians

Mary M Brown[a,b,c], MSc
Stefan Kuhle[a,b], MD, PhD
Bruce Smith[c], PhD
Victoria M. Allen[a], MD, MSc
Jennifer Payne[d], PhD
Christy G. Woolcott[a,b], PhD

Affiliations:
[a]Dept of Obstetrics & Gynaecology
[b]Dept of Pediatrics
[c]Dept of Mathematics and Statistics
[d]Dept of Diagnostic Radiology
Dalhousie University, Halifax, NS, Canada

Corresponding author: Dr. Christy Woolcott, Perinatal Epidemiology Research Unit, IWK Health Centre, 5980 University Avenue, Halifax, NS B3K 6R8, Canada. Email: christy.woolcott@iwk.nshealth.ca

Funding sources: Mary M. Brown received a Nova Scotia Graduate Scholarship from the Nova Scotia Research and Innovation Trust and a Scotia Scholar Doctoral Award from the Nova Scotia Health Research Foundation. This work was supported by a IWK Health Centre Category B Grant awarded to Stefan Kuhle and Christy Woolcott.

Potential Conflicts of Interest: The authors have no conflicts of interest relevant to this article to disclose.

5.1 Abstract

Objective: The objectives of this study were to examine the total effect of grandmaternal [G0] pre-pregnancy body mass index (BMI) on offspring [G2] birthweight z-score and to quantify the mediation role of maternal [G1] pre-pregnancy BMI.

Methods: Data were extracted from the Nova Scotia 3G Multigenerational Cohort. Path-specific effects were estimated using g-computation with adjustment for confounders identified using a directed acyclic graph. The mean difference in G2 birthweight z-score between observed (no change in G0 BMI) and two counterfactual scenarios was estimated: 1) fixing G0 BMI to 22 kg/m^2; and 2) a 10% gain or loss in G0 with a BMI <18.5 kg/m^2 or >25 kg/m^2, respectively.

Results: 20822 G1-G2 dyads born to 18450 G0 were included. If all G0 had a BMI of 22 kg/m^2, estimated mean G2 birthweight z-score would be 0.016 lower (95% CI -0.034, 0.001) as compared to values from the observed distribution. The estimated mediated effect by G1 pre-pregnancy BMI was -0.012 (95% CI -0.023, -0.001). Estimates of the change in G2 birthweight z-score under the '10% gain/loss' scenario were smaller in magnitude.

Conclusions: This study found no strong evidence for an association between grandmaternal BMI and infant birthweight, but due to inconsistent mediation, maternal BMI may be implicated in the association.

5.2 Introduction

The rising trend in overweight and obesity observed in the Canadian population poses risks for women before, during, and after pregnancy, as well as for their offspring. In 2015, 22% of Canadian women aged 18 to 34 were affected by overweight and 19% by obesity[1]. Women entering pregnancy with overweight or obesity are at increased risk of adverse maternal, fetal, and neonatal outcomes[2], and require specialist care, and additional healthcare services, resulting in higher maternity costs for these women[3]. These issues are especially concerning in Nova Scotia, one of Canada's four Atlantic provinces, where more than half of the women entering pregnancy were either overweight or obese in 2019[4].

The association between maternal pre-pregnancy body mass index (BMI) and the health of the first-generation offspring has been well established. A meta-analysis reported an increased risk of low birthweight in infants born mothers who are underweight, while infants born to mothers with overweight or obesity have increased risk of high birthweight and of becoming overweight and obese themselves in childhood and adolescence[6]. Several studies have reported an association between maternal pre-pregnancy BMI and offspring BMI[164–167] and body composition[164, 165, 168, 169] in young adulthood. These observed associations and the heritability of weight via genetic and epigenetic mechanisms have suggested there may also be an effect of grandmaternal pre-pregnancy BMI on child birthweight.

Evidence for the effects of in utero exposures on second-generation outcomes has been primarily derived from animal studies but multigenerational studies in humans are becoming more common. Investigators looking at the association between grandmaternal body weight measures (BMI, waist circumference, birthweight) and child birthweight have found little to no associations[9–13]. The results of two studies that examined grandmaternal pre-pregnancy BMI specifically suggest no large differences in child birthweight with each unit (kg/m^2) increase (-12 g [p=0.23][12] and 8 g [p=0.32][13]); however both studies were limited by small sample sizes. In addition, one study inappropriately adjusted for mediators of the association[12], thus potentially biasing the total effect estimate.

Two intergenerational studies have also investigated potential mediation of associations between maternal grandmother body weight and child birthweight by maternal

characteristics[10, 13]. Lahti-Pulkkinen et al.[10] examined the association between grandmaternal and child birthweight z-score and found that approximately 35% of the total effect was mediated by maternal birthweight z-score, but the magnitude of the direct and indirect effects were small (indirect effect: 0.06, 95% CI [0.04, 0.08]; direct effect: 0.13, 95% CI [0.07, 0.18]). Shen et al.[13] found that the effect of grandmaternal BMI on child birthweight was largely mediated by maternal body weight (indirect effect via BMI at age 18 and birthweight: 6.6 g per kg/m^2, p=0.04) rather than being a direct effect (1.3 g per kg/m^2, p=0.87). Structural equation modeling (SEM) was used in both studies to assess mediation, and is limited in that it operates under the assumption that all continuous exposure and mediator variables have a linear effect, which may be unrealistic in studies of pre-pregnancy BMI.

The findings of some studies suggest that in utero exposure to under- or overnutrition is associated with an increased risk of obesity throughout the life course including during the reproductive years; ultimately perpetuating the cycle of obesity[63, 170, 171]. Furthermore, evidence for whether the effects of the initial insult persist beyond the first-generation is limited, and the amount by which changes in pre-pregnancy BMI mitigate the risk of adverse outcomes in the second-generation has yet to be adequately estimated. Therefore, the objectives of this study were to estimate the total effect of grandmaternal pre-pregnancy BMI on infant birthweight z-score and to quantify the potential mediated effect by maternal pre-pregnancy BMI using a large sample of women, their mothers, and their offspring in the Canadian providence of Nova Scotia.

5.3 Methods

5.3.1 The 3G Multigenerational Cohort

The Nova Scotia Atlee Perinatal Database (NSAPD) is a population-based database that includes data on all births (delivered infants weighing $\geq$500 g or at a gestational age $\geq$20 weeks) to mothers residing in Halifax County, Nova Scotia, Canada from 1981 to 1987, and to mothers residing anywhere in Nova Scotia thereafter. This database grows by approximately 8000 births annually and contains extensive

information on each delivery including demographics, medical conditions, reproductive history, delivery events, and neonatal outcomes. Information is collected from the first prenatal visit in each pregnancy through to discharge from the hospital after birth admission. Nova Scotia uses a standard prenatal form in addition to forms completed during the hospital stay associated with the delivery to document relevant information[157].

The longstanding nature of the NSAPD enabled the establishment of the 3G Multigenerational Cohort by linking women's information on their own birth with information on their own pregnancies and deliveries. The cohort comprises women whose births had been recorded in the NSAPD and whose pregnancies were subsequently also recorded in the NSAPD. More information on the 3G Multigenerational Cohort can be found elsewhere[30]. For the purposes of the present study, G0 refers to information related to the grandmother at the point of delivery to the mother, G1 refers to information related to mother (i.e., birth and neonatal characteristics, and as an adult, pregnancy and delivery characteristics), and G2 refers to information related to the infant. The present study considered singleton pregnancies and included only the first [G2] live-born infant of the G1 women that survived to seven days.

This study was approved by the IWK Health Centre Research Ethics Board (#1023071) and the Joint Data Access Committee of the Reproductive Care Program of Nova Scotia. This study followed A Guideline for Reporting Mediation Analyses (AGReMA)[172].

5.3.2 Measurements

Outcome

Birthweight was recorded in grams on the birth record. Gestational age was available in days and was estimated using information from a dating ultrasound, the last menstrual period, and, where applicable, embryo transfer; details on the algorithm can be found elsewhere[173]. Gestational age- and sex-specific birthweight z-scores were calculated relative to a Canadian reference population[156].

Exposure and mediator

The exposure of interest was G0 pre-pregnancy BMI. G1 pre-pregnancy BMI was investigated as a potential mediator of the association between G0 pre-pregnancy BMI and G2 birthweight z-score. Height and weight were measured or self-reported at the first prenatal visit. Pre-pregnancy BMI was calculated as weight in kilograms divided by height in metres squared. Where applicable, BMI categories were defined according to World Health Organization (WHO) standards[174] (underweight [<18.5 kg/m^2], normal weight [18.5 to <25 kg/m^2], overweight [25 to <30 kg/m^2], obese [≥30 kg/m^2]).

Confounding variables

Variables that could confound the exposure-outcome, exposure-mediator, and mediator-outcome associations were considered in the analysis (Figure 5.2). Variables from both G0 and G1 included age at delivery (years), area of residence (urban/rural), area-level income quintile, any smoking during pregnancy (yes/no), and pre-existing type 1 or type 2 diabetes (yes/no). Variables from only the G0 pregnancy included parity (0/1/2/≥3 pregnancies that resulted in one or more infants weighing ≥500 g or ≥20 weeks' gestational age), hypertensive disorder of pregnancy (yes/no), gestational diabetes mellitus (yes/no), and G1 birthweight z-score.

Area-level income quintile was used as a measure of socioeconomic status and is derived by linkage of the postal code of the woman's residence (neighborhood income per person equivalent) to national census information[175]. Blood pressure is measured at each prenatal visit and recorded on the prenatal record to screen for pre-existing hypertension (<20 weeks gestation) and hypertensive disorders of pregnancy (≥20 weeks gestation) based on the Society of Obstetrics and Gynaecology Canada Guidelines[176]. All women in Nova Scotia are eligible for gestational diabetes screening according to guidelines set by Diabetes Canada[177].

5.3.3 Statistical analysis

Prior to running analyses, implausible values of birthweight z-score (≥5 in absolute value) and pre-pregnancy weight (BMI<13 kg/m^2, or <35 kg if height was missing)

were set to missing. Descriptive statistics including means and standard deviations and percentages, were used to describe the study sample overall and stratified by G0 pre-pregnancy BMI categories (based on mean BMI value imputed in 25 datasets). For each covariate, pairwise standardized mean differences (SMD)[178] between BMI categories were computed and then averaged.

As per the AGReMA guidelines, the exposure-mediator (i.e., G0 and G1 pre-pregnancy BMI) and mediator-outcome (i.e., G1 pre-pregnancy BMI and G2 birthweight z-score) relationships were examined visually in fitted smooth functions pooled from analyses of 25 imputed datasets. Then unadjusted and adjusted estimates of the association with the independent variable, pre-pregnancy BMI, categorized according to WHO standards[174] were obtained from generalized additive models and results were pooled across 25 imputed datasets using Rubin's rules[70].

A simplified directed acyclic graph (DAG) (Figure 5.1) shows the assumed relationships among the exposure (X; G0 pre-pregnancy BMI), mediator (M; G1 pre-pregnancy BMI), outcome (Y; G2 birthweight z-score), and intermediate confounders ($\boldsymbol{Z}$; G0 hypertensive disorders of pregnancy, G0 gestational diabetes, G1 birthweight z-score, G1 pre-existing diabetes). A detailed DAG including baseline covariates is shown in Figure 5.2. The total effect (TE) is the expected mean change in the outcome (G2 birthweight z-score) if the pre-pregnancy BMI of G0 women could be set to a different value (x) (i.e., hypothetically intervened on) other than what was observed in the population (x^*). The TE of the exposure on the outcome was decomposed into a direct effect and two mediator-specific effects[105]. The direct effect is the effect through neither $\boldsymbol{Z}$ nor M (MS^2-NDE-00); the mediator-specific effect through M is the effect through M but not $\boldsymbol{Z}$ (MS^2-NDE-01); and the mediator-specific effect through $\boldsymbol{Z}$ is all of the effect through $\boldsymbol{Z}$ (MS^2-NDE-11)[105]. Refer to Table 5.1 for the counterfactual definitions and description of the total effect and its three-way decomposition.

Two counterfactual scenarios (i.e., x scenarios) were examined. The first scenario was selected to represent a scenario where every G0 woman had a pre-pregnancy BMI of 22 kg/m^2 ('22 kg/m^2' scenario), which is the midpoint of the normal weight category[174]. The second scenario was selected to represent a less extreme hypothetical shift in BMI where women in the underweight category gained 10% of their

BMI and women in the overweight/obese category lost 10% of their BMI, while the BMI of women in the normal weight category remained unchanged ('10% gain/loss' scenario). These counterfactual scenarios were compared to the 'natural course' scenario[179, 180], which estimates the outcomes that would have been observed had the conditions been that which occurred naturally in the sample.

G-computation via Monte Carlo simulation was used to estimate the total effect and the mediator-specific effects described in Table 5.1. In this estimation approach, regression models fitted to the observed data are used to simulate potential outcome values that would have been observed had the exposure been determined by intervention rather than what was actually observed[105]. The analysis was performed in 1000 bootstrap samples drawn with replacement where, within each, 100 000 observations were simulated using generalized additive models (see Supplementary Methods 1 for more details on the simulation procedure). The final estimates were obtained by averaging across the estimates from the 1000 bootstrap samples. Standard errors estimated from the bootstrap samples were used to construct 95% confidence intervals. A single stochastic imputation using chained equations (10 iterations) was used to account for missingness within the bootstrap samples. This method has been shown to be valid when standard errors are estimated via bootstrapping as opposed to analytically using Rubin's variance estimator[111]. Analyses were performed using R (version 4.1)[160] and RStudio[161].

Identification assumptions

Several assumptions are sufficient to nonparametically identify the direct effect and the mediator-specific effects in Table 5.2 from the observed data. It was assumed that the data were generated from a nonparametric structural equation model with independent errors[104], and that adjustment for the set(s) of observed covariates enable control for confounding of the i) $X-Y$; ii) $M-Y$ and $\boldsymbol{Z}-Y$; iii) $X-\boldsymbol{Z}$ and $X-M$; and iv) $\boldsymbol{Z}-M$ relationships[105]. It was further assumed that there was no variable that was a confounder of the $(M, \boldsymbol{Z})-Y$ relationship that was affected by exposure or a previously occurring mediator (or intermediate confounder)[105]. This last assumption is encoded in Figure 5.2 via a lack of a causal pathway between the node "G0 pre-pregnancy BMI (X)" to the node of G1 baseline characteristics ("G1 age,

..., G1 smoking in pregnancy"), and between the node of intermediate confounders ("G0 GDM, ..., G1 birthweight ($\boldsymbol{Z}$))" to the node of G1 baseline characteristics.

Since the exposure and mediator variables were continuous, positivity was assessed by calculating propensity scores from models with the independent variable, pre-pregnancy BMI, categorized according to WHO standards[174]. Propensity scores for the exposure and mediator values (i.e., the probability of being in each BMI category) were estimated from adjusted multinomial logistic regression models (see Supplementary Methods for list of covariates included in each model). Propensity scores near the 0 and 1 boundaries suggest violations of the positivity assumption. Causal consistency was likely violated and is further discussed in Section 5.5.

5.4 Results

As of April 30, 2021, the 3G cohort included 19583 G0 women (born 1939-1987), 22307 G1 women (born 1981-2006) and 38922 G2 offspring (born 1996-2021). After excluding twin and other non-singleton deliveries of either G0 or G1 (n=1819), G1 deliveries resulting in stillbirths or early-neonatal deaths (n=238), and the offspring of G1 women that were not their first delivery (n=16043), the final analysis sample consisted of 20822 G1 women-offspring dyads born to 18450 G0 women. Characteristics of the analysis sample overall and by G0 pre-pregnancy BMI categories are shown in Table 5.2. Compared to their mothers, G1 women at the point of their first delivery were, on average, younger (SMD 0.42), less likely to smoke during pregnancy (SMD 0.41), and of higher pre-pregnancy weight (SMD 0.33). Furthermore, the G0 women in the overweight or obese BMI categories were more likely to be older, be multiparous, and deliver infants with larger birthweight z-scores compared to G0 women in the underweight or normal weight categories. The means of G0 pre-pregnancy weight and BMI were 63.8 kg and 23.7 kg/m^2, respectively, and the mean G2 birthweight z-score was -0.03.

The exposure-mediator (i.e., G0 and G1 pre-pregnancy BMI) and mediator-outcome (i.e., G1 pre-pregnancy BMI and G2 birthweight z-score) relationships were visually examined (Figure 5.3). G1 pre-pregnancy BMI appeared to increase linearly with increasing G0 pre-pregnancy BMI values, while G2 birthweight z-score appeared to increase linearly with increasing G1 pre-pregnancy BMI for values less

than 30 kg/m^2 and then exponentially thereafter. Estimates of these associations with categorized pre-pregnancy BMI are shown in Table 5.3.

Estimates of the total effect and mediator-specific natural effects in the analysis of G0 pre-pregnancy BMI and G2 birthweight z-score under the '22 kg/m^2' and '10% gain/loss' scenarios relative to the 'natural course' scenario, expressed as the mean change in G2 birthweight z-score, are presented in Table 5.4. Overall, there was no strong evidence of an association between G0 pre-pregnancy BMI and G2 birthweight z-score in adjusted analyses ('22 kg/m^2 scenario: -0.016, 95% CI [-0.034, 0.001]; '10% gain/loss' scenario: -0.006, 95% CI [-0.016, 0.004]). Mediation analyses showed negative point estimates of the mediator-specific effects via maternal pre-pregnancy BMI ('22 kg/m^2' scenario: -0.012, 95% CI [-0.021, -0.003]; '10% gain/loss' scenario: -0.005, 95% CI [-0.013, 0.004]) and positive point estimates of the natural direct effects ('22 kg/m^2' scenario: 0.007, 95% CI [-0.009, 0.023]; '10% gain/loss' scenario: 0.004, 95% CI [-0.019, 0.028]).

5.5 Discussion

Using a large multigenerational cohort of prospectively collected data, we found no strong evidence to suggest an association between G0 pre-pregnancy BMI and G2 birthweight z-score, but a significant estimate of the mediator-specific effect via G1 pre-pregnancy BMI under the '22 kg/m^2' scenario was identified. However, like the other estimates reported in this study, it was of small magnitude. As expected, estimates of the total effect and mediator-specific effects under the '22 kg/m^2' scenario were larger than those in the '10% gain/loss' scenario since the BMI of G0 women is shifted more extremely in the '22 kg/m^2' scenario (all G0 women to 22 kg/m^2) than in the '10% gain/loss' scenario (G0 women in the underweight or overweight/obese categories lost or gained 10%, respectively).

In the current study, if the pre-pregnancy BMI of all G0 women could be set to 22 kg/m^2 as compared to if pre-pregnancy BMI values were random draws from the observed distribution resulted in an estimated average change in G2 birthweight z-score of -0.016 (95% CI: -0.034, 0.001). As an example, this estimated birthweight z-score change is equivalent to a decrease in birthweight of 7.2 g (95% CI: 0.4, 15.2) for male infants born at 37 weeks' gestation. The small and non-significant total

effect estimates observed in this study agree with the findings of other studies examining the intergenerational association of grandmaternal body weight measures and offspring birthweight. Although not directly comparable due to the different estimating procedures, Shen et al.[13] and Harville et al.[12] reported estimates of 8 g (p=0.32, 95% CI or SE not reported) and -12 g (95% CI: -32, 8) in G2 birthweight, respectively, for a 1-kg/m^2 increase in G0 pre-pregnancy BMI. However, these studies had much smaller samples, Shen et al.[13] adjusted only for G0 smoking, G0 SES and G2 sex, and Harville et al.[12] inappropriately adjusted for possible mediators (maternal characteristics) of the association.

The estimated indirect effect via maternal pre-pregnancy BMI under the '22 kg/m^2' scenario was significant in the current study. This is similar to the findings of Shen et al.[13], the only other study to examine the mediation role of maternal BMI in the association of grandmaternal pre-pregnancy BMI and child birthweight. As opposed to the current study, Shen et al.[13] used a SEM approach to mediation, and was limited in its ability to control for some confounders (e.g., G0 age) and whether possible intermediate confounders (e.g., G0 gestational diabetes and hypertension) were considered is unclear. Stepwise variable selection was used that, combined with a small sample size, likely resulted in the exclusion of several important pathways (e.g., maternal birthweight to BMI at age 18) and may have introduced residual confounding. Lastly, as opposed to the methods used in the current study, SEM approaches assume all relationships are linear, which may be unrealistic, and does not provide estimates of the expected mean change in offspring birthweight under different counterfactual scenarios representing hypothetical changes in grandmaternal pre-pregnancy BMI values.

Inconsistent mediation where direct and indirect effects estimates have opposite signs was observed in this study. Since the total effect is the sum of the path-specific effects, inconsistent mediation can lead to situations where the total effect approaches the null and is non-significant, but mediated effects are significant[181]. The results of this three-generation study provide new insights into the mediating role of maternal pre-pregnancy BMI in the association between grandmaternal pre-pregnancy BMI and child birthweight z-score. Hypothetical shifts in the grandmaternal BMI distribution towards values within the normal range may result in a decrease in infant

birthweight z-score indirectly by changing the BMI distribution of the maternal generation. However, the magnitude of the effect is small, and the results suggest that if it were possible to intervene on low and elevated pre-gestational BMI, meaningful changes in birthweight beyond the first-generation are unlikely to occur. Furthermore, although birthweight is a predictor of weight in childhood and adolescence, it is imperfect, and it may be that the impact of grandmaternal BMI only becomes evident later in childhood.

The mechanisms underlying the possible effect of in utero exposure to maternal BMI levels outside of the recommended range and second-generation health outcomes remain unclear. This may be via genetic mechanisms and shared family environment, as well as epigenetic changes in the germ-cells of the developing fetus induced by in utero stressors[182], which can lead to phenotypic changes and increase disease susceptibility in the second-generation offspring. Transmittance of disease risk may occur through both the maternal and paternal germlines, but most studies suggest a stronger relationship along the maternal line[183]. One possible explanation for the maternal transmission of obesity risk is the inheritance of mitochondria along the maternal line, which are membrane-bound cell organelles vital to regulating many biochemical pathways. Animal studies have linked obesity with mitochondrial dysfunction in ooctyes leading to increased risk of metabolic diseases in the subsequent generation[184] and mitochondrial dysfunction in future generations[185]. Further research in humans is needed to clarify these mechanisms and their role in the intergenerational transmission of weight.

Causal interpretation of these results requires strong assumptions of conditional exchangeability (i.e., no unmeasured confounding), positivity, and consistency[186]. Due to the richness of the data source, it is likely that all relevant clinical confounders of the associations under study were accounted for but, like all observational studies, the possibility of residual confounding cannot be ruled out. An examination of propensity scores based on pre-pregnancy BMI categories revealed no major violations to the positivity assumption, but since the exposure and mediator variables were treated as continuous in the analyses, random violations were likely to have occurred[186, 187]. G-computation can consistently estimate causal parameters when the positivity assumption is violated by relying on the parametric models to

interpolate or smooth over regions of nonpositivity; this, however, relies on correct specification of the parametric models[186–189].

Lastly, much debate surrounds the relevancy of the effects of exposures like pre-pregnancy BMI as there may exist multiple ways in which an individual can achieve a BMI of x (e.g., genetics, medical conditions, diet, and lifestyle), and all may have different effects on outcomes[27, 190, 191]. Individuals in the study achieve their BMI via some combination of various mechanisms and so the "intervention" under study is not simply "assign all individuals to a BMI of x" but rather "assign all individuals to a BMI of x by changing the way a BMI of x is achieved to reflect that which is observed in those with a BMI of x in the population."[27] This is problematic as it is not entirely straightforward how this intervention could be applied in practice. In this case, the counterfactual outcome $Y(x)$ is vague and any causal contrasts involving this potential outcome will be ill-defined and thus violate the consistency assumption. Despite these issues, pre-pregnancy BMI is still an important risk factor in pregnancy, and there remains value in exploring its relationship with second-generation birthweight, even if the hypothetical conditions cannot be translated into realistic interventions on the population. In light of this, the findings of this study must be interpreted with caution since the results cannot be linked to a causal effect of a specific intervention targeted at elevating or lowering BMI.

The main strengths of this study are the use of a large sample of prospectively collected data and the ability to adjust for many important confounding factors. Furthermore, the data were collected using standard forms based on obstetric and hospital records, which leads to reliable and accurate measurements. This study also applied modern approaches to mediation analysis, which allowed us to control for intermediate confounding, and estimate causal effects using a flexible modeling approach that accommodates nonlinear relationships. Lastly, this study looked at contrasts of hypothetical conditions on the distribution of pre-pregnancy BMI with a scenario that mimics those that were naturally observed (i.e., no change on the distribution). Rather than contrasting potential outcomes estimated under a scenario where if, for example, all G0 women had a BMI of 30kg/m^2, contrasts with the 'natural course' scenario provide more information on the effects that the hypothetical conditions may have in the population from which the sample was drawn, and may

be a better alternative when examining policy-relevant effects[180].

However, this study is not without limitations. First, the strong assumptions required to interpret the reported mediator-specific natural effect estimates as causal effects were likely to be violated as discussed above. Second, a high proportion of missingness was observed for some variables, including grandmaternal pre-pregnancy height (information needed to derive the primary exposure variable). Maternal height has only been routinely recorded in the NSAPD since 2003 and is missing for the majority of deliveries of the grandmaternal generation. Although a high proportion of missingness was observed, analyses of multiply imputed BMI values were expected to be minimally biased since height is likely to be missing at random, the imputation model contained variables that are both correlated with height (e.g., weight) and related to its hypothesized missingness mechanism[192] (e.g., delivery year), and the coefficient of variation for the square of height is small[81] (the denominator of BMI). Additionally, the 3G cohort does not contain information on the fathers, so only the investigation of the association along the maternal line could be quantified.

In conclusion, the findings of this study suggest no strong association between grandmateral pre-pregnancy BMI and second-generation offspring birthweight. Results of this mediation analysis suggest the possibility of inconsistent mediation where the direct and indirect effect estimates have opposite signs, yielding near-null estimates of the total effect. The results showed that maternal pre-pregnancy BMI may be implicated in the association between grandmaternal pre-pregnancy BMI and offspring birthweight.

Table 5.1: Counterfactual definition and description of the total effect and of the components of its three-way decomposition into mediator-specific (MS) effects where X is the exposure, $\boldsymbol{Z}$ is the joint set of intermediate confounders, M is the mediator, and Y is the outcome

Effect	**Counterfactual definition**	**Description**
TE	$E\left[Y\left(x, \boldsymbol{Z}(x), M(x, \boldsymbol{Z}(x))\right) - Y\left(x^*, L\boldsymbol{Z}a^*\right), M(x^*, \boldsymbol{Z}(x^*))\right)]$	Total effect of X on Y
MS^2-NDE-00	$E\left[Y\left(x, \boldsymbol{Z}(x^*), M(x^*, \boldsymbol{Z}(x^*))\right) - Y\left(x^*, \boldsymbol{Z}(x^*), M(x^*, \boldsymbol{Z}(x^*))\right)\right]$	Direct effect of X on Y though neither M nor $\boldsymbol{Z}$ ($X \rightarrow Y$)
MS^2-NIE-10	$E\left[Y\left(x, \boldsymbol{Z}(x^*), M(x, \boldsymbol{Z}(x^*))\right) - Y\left(x, \boldsymbol{Z}(x^*), M(x^*, \boldsymbol{Z}(x^*))\right)\right]$	Indirect effect via M alone ($X \rightarrow M \rightarrow Y$)
MS^2-NIE-11	$E\left[Y\left(x, \boldsymbol{Z}(x), M(x, \boldsymbol{Z}(x))\right) - Y\left(x, \boldsymbol{Z}(x^*), M(x, \boldsymbol{Z}(x^*))\right)\right]$	Indirect effect via $\boldsymbol{Z}$ ($X \rightarrow \boldsymbol{Z} \rightarrow Y$ and $X \rightarrow \boldsymbol{Z} \rightarrow M \rightarrow Y$)

Abbreviations: *NDE* natural direct effect; *NIE* natural indirect effect; *TE* total effect

Table 5.2: Descriptive statistics of the maternal (G1) female offspring, their mothers (G0), and their offspring (G2) in the full analytical sample and stratified by G0 pre-pregnancy BMI category

	Complete data, n	**Full sample** n=20822	**G0 pre-pregnancy BMI category**[a] Underweight n=1407	Normal weight n=13059	Overweight n=4196	Obese n=2160	**SMD**
G0 pregnancy and delivery characteristics							
Pre-pregnancy weight [kg] (mean (SD))	17196	63.8 (14.1)	45.6 (3.0)	57.8 (5.4)	72.7 (4.6)	92.9 (11.7)	4.07
Pre-pregnancy BMI [kg/m^2] (mean (SD))	930	23.7 (5.2)	17.5 (0.9)	21.7 (1.8)	26.9 (1.4)	34.5 (4.0)	4.47
Age [years] (mean (SD))	20822	25.9 (5.0)	24.3 (4.9)	25.7 (5.1)	26.4 (5.0)	26.6 (4.7)	0.26
Parity (%)	20822						0.13
0		43.8	48.5	45.6	38.9	39.2	
1		35.0	32.8	34.4	36.9	36.5	
2		14.9	13.4	14.1	17.0	16.2	
3+		6.3	5.3	5.9	7.2	8.1	
Rural residence (%)	14720	36.3	33.1	35.5	36.7	41.8	0.10
Area-level income quintile (%)	14642						0.07
Q1 (low)		26.0	25.5	26.1	25.2	27.0	
Q2-Q4 (middle)		63.1	63.9	62.3	64.2	64.0	
Q5 (high)		10.9	10.6	11.5	10.6	9.0	
Any smoking in pregnancy (%)	18431	43.1	54.1	42.9	42.1	38.5	0.16
Pre-existing diabetes (%)	20822	0.3	0.0	0.2	0.4	0.8	0.08
Gestational diabetes (%)	20822	2.1	1.3	1.4	2.6	6.4	0.15
Hypertensive disorders of pregnancy (%)	20822	9.6	6.0	8.0	12.3	16.6	0.19
G1 birthweight z-score (mean (SD))	20294	-0.04 (1.05)	-0.50 (0.96)	-0.11 (0.99)	0.15 (1.09)	0.29 (1.16)	0.42
G1 pregnancy and delivery characteristics							
Pre-pregnancy weight [kg] (mean (SD))	18110	69.3 (18.5)	58.8 (13.9)	67.0 (16.0)	74.7 (20.2)	79.9 (23.6)	0.64
Pre-pregnancy BMI [kg/m^2] (mean (SD))	17529	25.8 (6.5)	22.3 (5.0)	24.8 (5.6)	27.8 (7.1)	29.7 (8.3)	0.64
Age [years] (mean (SD))	20822	23.9 (4.5)	23.1 (4.4)	24.2 (4.6)	23.6 (4.4)	22.7 (3.9)	0.19
Rural residence (%)	20151	34.5	34.2	32.9	35.7	42.5	0.11
Area-level income quintile (%)	20101						0.11
Q1 (low)		22.1	22.4	21.6	22.1	25.7	

Table 5.2: continued

	Complete data, n	**Full sample** n=20822	**G0 pre-pregnancy BMI category**[a] Underweight n=1407	Normal weight n=13059	Overweight n=4196	Obese n=2160	**SMD**
Q2-Q4 (middle)		66.4	66.6	65.9	68.0	66.3	
Q5 (high)		11.5	11.1	12.5	10.2	8.0	
Any smoking in pregnancy (%)	20680	24.3	29.0	22.9	25.0	28.2	0.08
Pre-existing diabetes (%)	20822	0.8	0.6	0.7	0.9	1.3	0.04
Gestational diabetes (%)	20822	3.9	2.9	3.3	5.2	5.4	0.08
Hypertensive disorders of pregnancy (%)	20822	10.2	7.4	9.4	12.4	12.4	0.10
G2 birth characteristics							
Birthweight z-score (mean (SD))	20740	-0.03 (1.00)	-0.16 (0.99)	-0.04 (0.99)	0.02 (1.04)	0.04 (1.01)	0.11

Abbreviations: *SD* standard deviation; *SMD* standardized mean difference; *BMI* body mass index
[a] Calculated from the mean pre-pregnancy BMI value across 25 imputed datasets

Table 5.3: Estimates of the exposure-mediator and mediator-outcome associations

	Unadjusted estimate (95% CI)	**Adjusted estimate (95% CI)**
Exposure-mediator association[a]	Mean difference in G1 pre-pregnancy BMI (kg/m^2)	
G0 pre-pregnancy BMI		
Underweight	-2.51 (-2.98, -2.04)	-2.56 (-3.02, -2.10)
Normal weight	0.00 (ref)	0.00 (ref)
Overweight	1.72 (1.41, 2.03)	1.84 (1.53, 2.15)
Obese	4.30 (3.93, 4.66)	4.52 (4.16, 4.88)
Mediator-outcome association[b]	Mean difference in G2 birthweight z-score	
G1 pre-pregnancy BMI		
Underweight	-0.246 (-0.305, -0.187)	-0.159 (-0.217, -0.102)
Normal weight	0.00 (ref)	0.00 (ref)
Overweight	0.225 (0.187, 0.264)	0.180 (0.143, 0.216)
Obese	0.281 (0.245, 0.318)	0.220 (0.184, 0.256)

Abbreviations: *BMI* body mass index; *CI* confidence interval; *G0* grandmaternal; *G1* maternal; *G2* infant

[a] Adjusted for G0 age, parity, rural residence, area-level income quintile, smoking in pregnancy, pre-existing diabetes, and year of delivery

[b] Adjusted for G0 age, parity, rural residence, area-level income quintile, smoking in pregnancy, pre-existing diabetes, gestational diabetes, hypertensive disorders of pregnancy, and year of delivery, and G1 birthweight z-score, age, rural residence, area-level income quintile, smoking in pregnancy, pre-existing diabetes, and year of delivery

Table 5.4: Total effect and its three-way decomposition into the direct effect and mediator-specific (MS) effects in the analysis of grandmaternal pre-pregnancy BMI and infant birthweight z-score

Effect	**Adjusted Estimate**[a] **(95% CI)**	
	'22 kg/m^2' scenario	'10% gain/loss' scenario
Total effect	-0.016 (-0.034, 0.001)	-0.006 (-0.017, 0.004)
MS^2-NDE-00	0.007 (-0.009, 0.023)	0.004 (-0.019, 0.028)
MS^2-NIE-10 (via G1 pre-pregnancy BMI)	-0.012 (-0.021, -0.003)	-0.005 (-0.013, 0.004)
MS^2-NIE-11 (via pathways involving the joint set of intermediate confounders)	-0.011 (-0.021, -0.002)	-0.006 (-0.027, 0.016)

Abbreviations: *CI* confidence interval; *G1* maternal; *NDE* natural direct effect; *NIE* natural indirect effect

[a] Expressed as the mean difference in birthweight z-score relative to the 'natural course' scenario

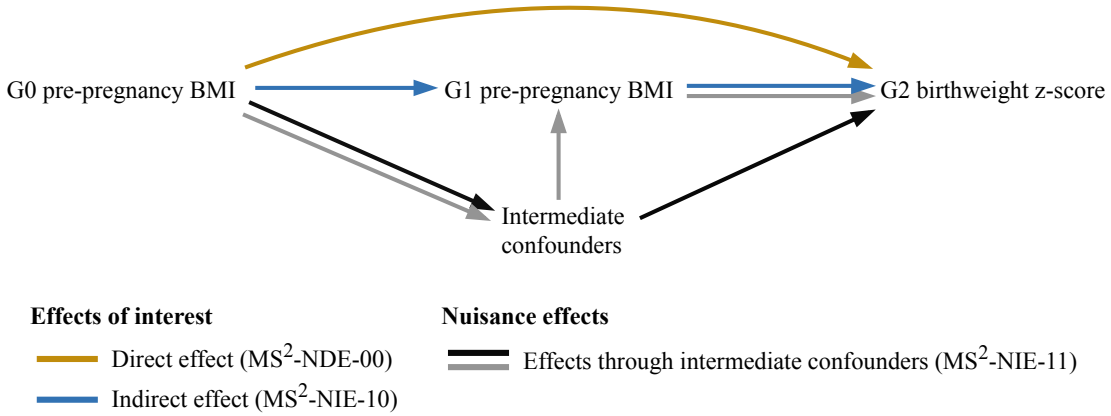
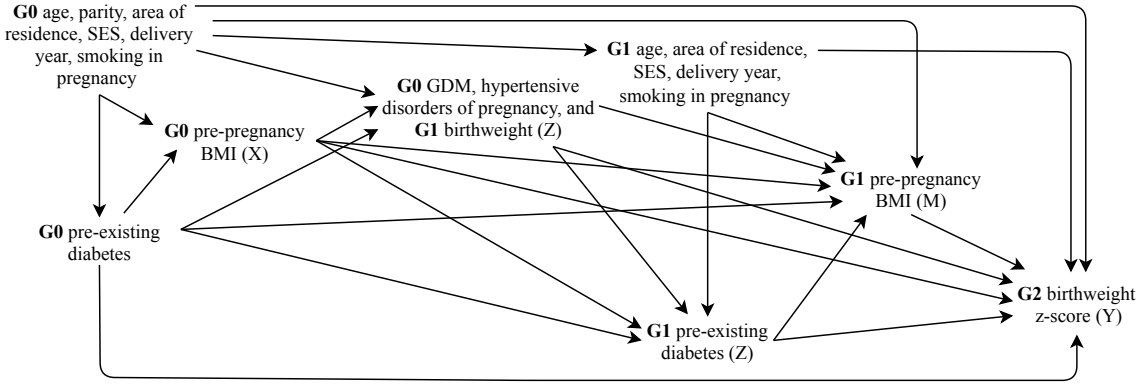
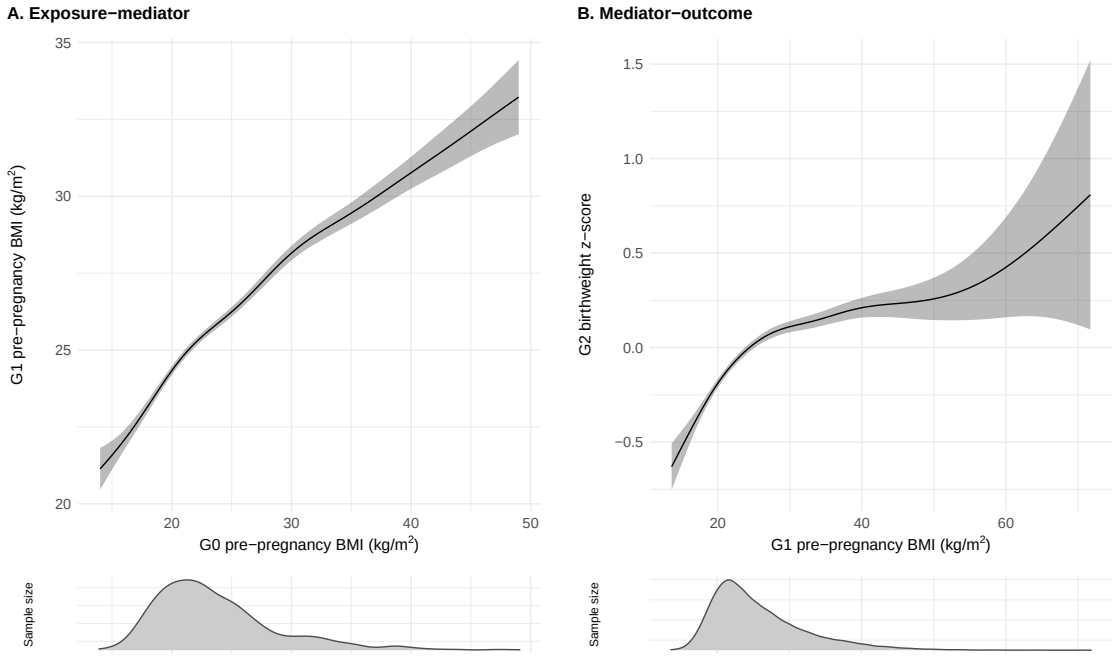


Figure 5.1: Simplified directed acyclic graph showing the three path-specific effects from grandmaternal (G0) pre-pregnancy BMI to infant (G2) birthweight z-score where the direct effect (yellow) and the mediator-specific effect via maternal (G1) pre-pregnancy BMI (blue) are of primary interest



Abbreviations: *GDM* gestational diabetes; *SES* socioeconomic status

Figure 5.2: Directed acyclic graph of the hypothesized relationships among grandmaternal (G0) pre-pregnancy BMI (X), intermediate confounders (Z), maternal (G1) pre-pregnancy BMI (M), and infant (G2) birthweight z-score (Y)



Abbreviations: *BMI* body mass index

Figure 5.3: Fitted smooth function to the A) exposure-mediator (grandmaternal [G0] and maternal [G1] pre-pregnancy BMI) and B) mediator-outcome (G1 pre-pregnancy BMI and infant [G2] birthweight z-score) associations pooled from analyses of 25 imputed datasets. The observed distribution of the independent variables is shown below

5.6 Supplementary Methods 1: Mediation analysis via g-computation

G-computation via Monte Carlo simulation was used to estimate the total effect and the mediator-specific natural effects in Table 5.1. Each half of each of the mediator-specific natural effects is of the form

$$E[Y(x, \boldsymbol{Z}(x^*), M(x^{**}, \boldsymbol{Z}(x^*))],$$

where $x \neq x^* \neq x^{**}$ are different values of the exposure variable X. Under strong identification assumptions, the above can be nonparametrically identified from the observed data by

$$\int_c \int_{\boldsymbol{z}} \int_m E[Y \mid X = x, \boldsymbol{Z} = \boldsymbol{z}, M = m, \boldsymbol{C} = \boldsymbol{c}]$$
$$f_M(m \mid X = x^{**}, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{C} = \boldsymbol{c})\, f_{\boldsymbol{Z}}(\boldsymbol{z} \mid X = x^*, \boldsymbol{C} = \boldsymbol{c}) f_{\boldsymbol{C}}(\boldsymbol{c})\, dm\, d\boldsymbol{z}\, d\boldsymbol{c}. [105]$$

The steps of the g-computation method involve specifying regression models for each density and expectation in the identifying equations, estimating their parameters from the observed data, and then evaluating the integral using Monte Carlo simulation[110].

The analysis was performed in 1000 bootstrap samples drawn with replacement where, within each, 100 000 observations were simulated. Within each bootstrap sample, generalized additive models (with smooth terms for continuous covariates) for the intermediate confounders (G0 gestational diabetes [GDM], G0 hypertensive disorders of pregnancy, G1 birthweight, and G1 pre-existing diabetes), the mediator (G1 pre-pregnancy BMI), and the outcome (G2 birthweight z-score) were fitted. Model covariates are listed below and inclusion was informed by the directed acyclic graph in Figure 5.2:

1. Intermediate confounder 1 model (G0 GDM): G0 pre-pregnancy BMI (X), G0 pre-existing diabetes, and G0 baseline covariates (age, parity, area of residence, socioeconomic status, delivery year, smoking in pregnancy)

2. Intermediate confounder 2 model (G0 hypertensive disorders of pregnancy): G0 pre-pregnancy BMI (X), G0 GDM, G0 pre-existing diabetes, and G0 baseline

covariates

3. Intermediate confounder 3 model (G1 birthweight z-score): G0 pre-pregnancy BMI (X), G0 hypertensive disorders of pregnancy, G0 GDM, G0 pre-existing diabetes, and G0 baseline covariates

4. Intermediate confounder 4 model (G1 pre-existing diabetes): G0 pre-pregnancy BMI (X), G1 baseline covariates (age, area of residence, socioeconomic status, delivery year, smoking in pregnancy), G1 birthweight z-score, G0 hypertensive disorders of pregnancy, G0 GDM, G0 pre-existing diabetes, and G0 baseline covariates

5. Mediator model: G0 pre-pregnancy BMI (X), intermediate confounders ($\boldsymbol{Z}$), G1 baseline covariates, and G0 baseline covariates

6. Outcome model: G0 pre-pregnancy BMI (X), G1 pre-pregnancy BMI (M), intermediate confounders ($\boldsymbol{Z}$), G1 baseline covariates, and G0 baseline covariates

Then G0 and G1 baseline covariate sets were simulated using resampling from the observed baseline covariates. The exposure variables were simulated to be marginally independent of the baseline covariates. The 'natural course' scenario was simulated by drawing from the empirical distribution of G0 pre-pregnancy BMI, the '10% gain/loss' scenario was simulated based on the values of the 'natural course' scenario, and the '22 kg/m^2' scenario was simulated by setting all G0 pre-pregnancy BMI values to 22 kg/m^2.

Then, using a sequential approach, values of the intermediate confounders, mediator, and outcome were simulated under each hypothetical condition of the exposure using the parameters from the models fitted to the observed data. The outcome model additionally contained an interaction term for the exposure and mediator. Contrasts of individual potential outcomes were averaged across observations in the bootstrap sample to estimate the population average estimates of the total effect and mediator-specific natural effects. The final estimates were obtained by averaging across the 1000 bootstrap samples. Standard errors estimated using the bootstrap samples were used to construct 95% confidence intervals.

Chapter 6

Development and validation of a prediction model for second-generation fetal growth abnormalities in the 3G Multigenerational Cohort of Nova Scotian women

The target journal for this manuscript is *International Journal of Obstetrics and Gynaecology*.

MM Brown conceptualized and designed the study, performed the analysis, and wrote the initial manuscript draft. C Woolcott, B Smith, and S Kuhle contributed to the design of the study, and provided supervisory input regarding methodology and continuous feedback on the manuscript. J Payne and V Allen provided content expertise and guidance for implications and design of the study. All authors reviewed and revised the manuscript, and approved the final manuscript as presented in this thesis.

Development and validation of a prediction model for second-generation fetal growth abnormalities in the 3G Multigenerational Cohort of Nova Scotian women

Mary M Brown[a,b,c], MSc
Stefan Kuhle[a,b], MD, PhD
Bruce Smith[c], PhD
Victoria M. Allen[a], MD, MSc
Jennifer Payne[d], PhD
Christy G. Woolcott[a,b], PhD

Affiliations:
[a]Dept of Obstetrics & Gynaecology
[b]Dept of Pediatrics
[c]Dept of Mathematics and Statistics
[d]Dept of Diagnostic Radiology
Dalhousie University, Halifax, NS, Canada

Corresponding author: Dr. Christy Woolcott, Perinatal Epidemiology Research Unit, IWK Health Centre, 5980 University Avenue, Halifax, NS B3K 6R8, Canada. Email: christy.woolcott@iwk.nshealth.ca

Funding sources: Mary M. Brown received a Nova Scotia Graduate Scholarship from the Nova Scotia Research and Innovation Trust and a Scotia Scholar Doctoral Award from the Nova Scotia Health Research Foundation.

Potential Conflicts of Interest: The authors have no conflicts of interest relevant to this article to disclose.

6.1 Abstract

Objective: Because grandmaternal risk factors and maternal birth characteristics (G0 predictors) may improve prediction of fetal growth abnormalities, our objective was to develop and validate a prediction model using these predictors together with maternal pregnancy characteristics (G1 predictors).

Design: Retrospective cohort study.

Setting: Nova Scotia, Canada, between 1981 and 2011.

Population: A total of 9068 pregnancies to first-born, nulliparous women that resulted in singleton live births after 26 weeks' gestation, and their mothers.

Methods: The machine learning ensemble Super Learner was used to develop models for small for gestational age (SGA) and large for gestational age (LGA) using G0 predictors, G1 predictors, and their combination. Models were validated using nested cross-validation, and discrimination and calibration were assessed.

Main outcome measures: Infant birthweight for gestational age and sex: SGA (<10th percentile) and LGA (>90th percentile) relative to a Canadian reference population.

Results: Discriminative performance, measured using the area under the precision-recall curve (AUC-PR), increased with the inclusion of grandmaternal factors and maternal birth characteristics to models fitted using maternal characteristics only and grandmaternal factors and maternal birth characteristics only (0.21 vs. 0.15-0.18 for SGA and 0.22 vs. 0.17-0.18 for LGA). Super Learner models fitted using both sets of predictors were well calibrated.

Conclusions: Grandmaternal factors and maternal birth characteristics modestly improved the prediction of fetal growth abnormalities as compared to models based solely on maternal characteristics; however, prediction remains poor. Maternal birthweight z-score may be a useful predictor of abnormal fetal growth.

6.2 Introduction

Deviations from normal fetal growth are associated with adverse health outcomes in infants[28]. Small for gestational age (SGA) birth is associated with higher rates of perinatal morbidity and mortality and increases the risk of neurodevelopmental deficits, fetal growth restriction, and cardiovascular disease in adulthood[28, 48, 193, 194]. Large for gestational age (LGA) birth is associated with increased risk of birth injury, typically due to the physical size of the fetus, asphyxia, polycythemia, and hypoglycemia, and has been linked to obesity, diabetes and cardiovascular disease in adulthood[28, 48, 49]. Accurate identification of pregnancies at highest risk for fetal growth abnormalities could improve preconception counselling, antenatal assessment, and intrapartum care.

Prediction models for SGA and LGA have been developed using routinely collected data readily available in an antenatal setting, including maternal sociodemographics, pregnancy risk factors, past pregnancy history, and clinical characteristics; however, predictive performance remains relatively poor, especially among nulliparous women. For example, a study validated six prediction models for SGA and LGA using an independent cohort of 1311 nulliparous women and found discriminative performance, measured as the area under the receiver operating characteristic curve (AUC-ROC), to be between 0.50-0.66 for SGA and 0.58-0.67 for LGA[114]. Similarly, in a cohort of 14923 nulliparous Nova Scotian women, prediction using both conventional regression models and machine learning algorithms resulted in AUC-ROC estimates between 0.63-0.70 for both SGA and LGA[139]. Current efforts to improve prediction models for early detection of SGA and LGA include adding ultrasound measurements, biochemical markers, and results of biophysical tests, but only modest improvements have been reported[115–131] and measurement of some predictors may be costly, time-consuming, and inconvenient for pregnant women.

Fetal growth is affected by maternal, fetal and environmental factors, uterine conditions, and placental function. Multigenerational studies on the effects of in utero exposures on second-generation outcomes[8] have also found small to moderate associations between grandparental risk factors and child birthweight, including grandparental birthweight[9, 10], body mass index (BMI)[11], smoking in pregnancy[14–19], socioeconomic status[20–23], and diabetes[24, 25], with most studies focusing

on the maternal line. Despite the well-established associations between maternal and offspring size-at-birth[29], maternal birth characteristics and grandmaternal risk factors have not been used for the predicion of SGA and LGA. Furthermore, since grandmaternal risk factors may act as potential effect modifiers of the maternal-offspring associations, and relationships between predictors and fetal growth may be nonlinear, predictive performance may be improved upon by using machine learning algorithms. Considering that in practice it is impossible to know which machine learning algorithm will perform best in the data under study, multiple algorithms can be combined into a single algorithm called a Super Learner[132].

The primary objective of this study was to develop and validate a Super Learner model for fetal growth abnormalities using grandmaternal risk factors, maternal birth characteristics, and maternal pregnancy characteristics and compare it with prediction models based on grandmaternal risk factors and maternal birth characteristics only and on maternal pregnancy characteristics only in a large sample of nulliparous women in the Canadian province of Nova Scotia. The secondary objective was to compare the predictive performance of the Super Learner model to other parametric and nonparametric algorithms in this context.

6.3 Methods

6.3.1 Study population and design

Data were derived from the 3G Multigenerational Cohort[30], which includes women whose births and their subsequent own pregnancies were recorded in the Nova Scotia Atlee Perinatal Database (NSAPD). The NSAPD is a population-based database that contains extensive information on demographics, medical conditions, reproductive history, delivery events, and neonatal outcomes for each birth to mothers residing in Halifax County, Nova Scotia, Canada, since 1988, and to mothers residing anywhere in the province thereafter, and grows by approximately 8000 deliveries annually. Information is collected from the first prenatal visit in each pregnancy through to discharge from the hospital after birth admission. Nova Scotia uses a standard Prenatal Record in addition to forms completed during the hospital stay associated with the delivery to document information relevant to care and medical research. The

NSAPD is administered and maintained by the Reproductive Care Program of Nova Scotia, and ensures the quality, integrity, and security of the data. The longstanding nature of the NSAPD enabled the establishment of the 3G Multigenerational Cohort by linking women's information on their own birth with information on their own pregnancies and deliveries.

As of April 30th, 2021, the 3G cohort consisted of 19583 grandmothers (G0; born 1939-1987), 22307 mothers (G1; born 1981-2006) and 38922 infants (G2; born 1996-2021). The present study restricted the cohort to singleton pregnancies and the first delivery of a live-born infant in both the G0 and G1 generations. In addition, only G2 infants with complete information on gestational age and birthweight, gestational age $\geq$ 26 weeks, and a plausible value of birthweight z-score (<5 in absolute value) were included in the analysis.

This study was approved by the IWK Health Centre Research Ethics Board (#1023071) and the Joint Data Access Committee of the Reproductive Care Program of Nova Scotia. This study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines[145].

6.3.2 Outcomes

The two outcomes of interest were infant birthweight for gestational age and sex relative to a Canadian reference population[156]: SGA (<10th percentile) and LGA (>90th percentile). Birthweight was recorded in grams on the birth record. Gestational age was available in days and was estimated using information from a dating ultrasound, the last menstrual period, and where applicable, embryo transfer; details of the algorithm can be found elsewhere[173].

6.3.3 Predictors

Two sets of predictors and their combination were considered (Table 6.1). The first set (G1 predictors) included maternal demographic, pre-pregnancy, and pregnancy information that was available at 26 weeks' gestation for the G2 pregnancy. The second set (G0 predictors) included grandmaternal demographic, pregnancy, and delivery characteristics at the time of the mother's birth and the mother's birth

characteristics and neonatal outcomes.

Area-level income quintile was used as a measure of socioeconomic status and was derived from linkage of the woman's residence postal code to national census information[175]. Blood pressure was measured at each prenatal visit to screen for pre-existing hypertension (<20 weeks' gestation) and hypertensive disorders of pregnancy (≥20 weeks' gestation) based on the Society of Obstetrics and Gynaecology Canada Guidelines[176]. All women in Nova Scotia were eligible to undergo screening for gestational diabetes according to guidelines set by Diabetes Canada[177]. Any smoking reported during pregnancy (first prenatal visit, 20 weeks, or birth admission), or any alcohol abuse reported at any point in the pregnancy, were used as proxy measures for smoking and alcohol abuse at 26 weeks. Pre-pregnancy BMI was calculated by dividing pre-pregnancy weight (kg) by the square of height (m), which was recorded at the first prenatal visit. Gestational weight gain in pregnancy was calculated by taking the difference in pre-pregnancy weight and delivery weight (kg), and at 26 weeks was estimated by

$$2 + 13\left(\frac{\text{Delivery weight} - \text{Pre-pregnancy weight} - 2}{\text{Gestational age at birth} - 13}\right)$$

assuming a 2 kg weight gain in the first trimester (13 weeks) and a steady increase in weight thereafter[195].

6.3.4 Statistical analysis

Implausible values of G1 birthweight z-score (≥5 in absolute value) and of G0 and G1 pre-pregnancy weight (BMI<13 kg/m^2, or <35 kg if height was missing) were set to missing. Descriptive statistics including means and standard deviation, and percentages, were used to describe the study sample overall and by SGA and LGA status. Standardized mean differences (SMD)[178] between groups defined by SGA and LGA status were computed for each predictor. Multiple imputation using chained equations was used to account for missingness in predictors[82]. Since pre-pregnancy height information (the denominator of pre-pregnancy BMI) was not collected prior to 2003, it was missing for approximately 90% of births. However, missingness in pre-pregnancy height and other variables was likely missing at random. Ten imputed

datasets (10 iterations) were generated where missing values were imputed using random forest[88]. Analyses were performed on each of the imputed datasets and results were pooled using Rubin's rules[70].

Prediction models for SGA (vs. non-SGA) and LGA (vs. non-LGA) were developed on G1 predictors only (conventional prediction of SGA and LGA), G0 predictors only, and both G0 and G1 predictors. Models were developed using the Super Learner algorithm[132] optimized with respect to the Brier score (equivalent to the mean squared error for predictions). The Super Learner algorithm is a technique that constructs an optimal weighted average of a set of candidate machine learning algorithms and regression models with weights estimated according to a user-specified loss function. The resulting Super Learner model has been shown to perform as well as, or better than, the best algorithm in the ensemble in large samples[135]. The library of candidate learners consisted of linear algorithms [logistic regression (main and interaction term models), generalized additive models (GAM)] and a diverse set of machine learning algorithms [elastic net[146], random forest (RF)[88], tree-based extreme gradient boosting (XGBoost)[147], and kernel-based support vector machine (SVM)[148]].

For each imputed dataset, nested cross-validation (10-fold internal and 5-fold external cross-validation) was used to develop and validate the prediction models. Machine learning algorithms were tuned by creating a weighted ensemble of the learner specified with different hyperparameter values (Table 6.2). Models were evaluated and compared based on discrimination and calibration. Discrimination was assessed using the area under the precision-recall curve (AUC-PR), which is a plot of precision (i.e., positive predictive value) vs. recall (i.e., sensitivity) calculated at all thresholds represented in the data. In the presence of class imbalance ($\sim$10% SGA/LGA vs. $\sim$ 90% non-SGA/LGA), performance based on the AUC-ROC may be misleading because, compared to the PR curve, the false positive rate is less sensitive to changes in the number of correctly classified positive cases due to the large number of negative cases[151]. However, to facilitate comparison to other models in the literature, the AUC-ROC was calculated as a secondary performance metric. The AUC-PR value that indicates no discrimination is the average prevalence of SGA and LGA in training samples, and the AUC-ROC value that indicates no discrimination

is 0.5. Predictive variable importance rankings using the two estimation algorithms with the largest mean weight in the Super Learner ensemble were calculated. Variable importance was measured by the increase in Brier score after permuting the values of each predictor and averaged across imputed datasets.

Calibration, or the agreement between predicted risk from the Super Learner algorithm and the observed risk, was visually assessed using calibration curves fitted using thin plate splines and decile groups. Data for the calibration curves was created by stacking the Super Learner predictions from each imputed dataset when each observation served in the validation fold. Within each imputed dataset, deciles calculated using the stacked Super Learner predictions were used to group observations and within each of the 10 groups, the proportion of observations with the outcome was estimated. Pooled estimates in each group and corresponding standard errors were used to derive calibration point estimates and confidence intervals. Calibration was assessed only for the models fitted using both G0 and G1 predictors.

The robustness of the results to more stringent definitions of the outcomes (<3rd percentile [SGA3] and >97th percentile [LGA97]) was explored in a sensitivity analysis. All analyses were performed in R (version 4.1.2)[160] and RStudio[161] using functions primarily from the *mice*[82], *psfmi*[196], *PRROC*[197, 198], *Super Learner*[199], *wesanderson*[200], and *ck37r*[201] packages and modified R source code developed by Chris Kennedy[202].

6.4 Results

6.4.1 Study population

Between 1981 and 2021, 9111 pregnancies to first-born, nulliparous G1s resulted in singleton live births. After removing G2 infants with missing birthweight or gestational age information (n=27), implausible birthweight z-score values (n=7), and a gestational age less than 26 weeks (n=9), the final analytical sample included 9068 G0-G1-G2 triads. During this time period, 896 (9.9%) and 902 (9.9%) G2 infants were born SGA and LGA, respectively. Characteristics of the sample overall and stratified by SGA and LGA status are shown in Table 6.3. Small differences existed between the grandmothers of infants born SGA and LGA and those of infants born

non-SGA and non-LGA (SMD<0.2). However, the mothers of infants born SGA had lower birthweight z-scores, and, as adults, gained less weight in pregnancy than mothers of infants born non-SGA, while the mothers of infants born LGA had higher birthweight z-scores, and, as adults, gained more weight in pregnancy than mothers of infants born non-LGA.

6.4.2 Discrimination

Cross-validated AUC-PR estimates and 95% confidence intervals for the Super Learner algorithm fitted using G0 predictors only, G1 predictors only, and both sets of predictors are shown in Table 6.4, and additionally with estimates for the individual algorithms in the ensemble in Table 6.5. PR curves for the Super Learner models fitted using the three different sets of predictors are shown in Figure 6.1. Predictions improved when using both G0 and G1 predictors (SGA 0.21; LGA 0.22) compared to using G0 predictors only (SGA 0.15; LGA 0.17) or G1 predictors only (SGA 0.18; LGA 0.18). The same trend was observed in analyses of the cross-validated AUC-ROC (Table 6.4, Table 6.5, Figure 6.2). Compared to the individual machine learning algorithms and regression models, in general, the highest AUC-PR estimates were observed for the Super Learner algorithm across all prediction models (Table 6.5). The predictive performance of GAM was most comparable to that of the Super Learner, while the AUC-PR for logistic regression was typically less, with the largest differences observed in models fitted using both G0 and G1 predictors (Table 6.5).

6.4.3 Calibration

The mean predicted risk of SGA and LGA from the Super Learner ensemble (9.9% for both SGA and LGA, respectively) matched the overall risk in the sample. Calibration plots (Figure 6.3) indicate good agreement between the predicted risk of SGA and LGA from the Super Learner and the smoothed actual risk estimated using thin plate splines. The Super Learner ensemble slightly overestimated the risk of SGA and LGA when the actual risk was small (<5%) (Table 6.8), but predicted risk estimates were within the 95% confidence intervals for the actual risk in all decile groups.

6.4.4 Super Learner weights and variable importance

To assess the contribution of the individual learners to the final Super Learner predictions, coefficients (i.e., weights) of the Super Learner model were calculated and averaged across validation folds and imputed datasets (Table 6.6). Super Learner weights are calculated so that each is non-negative and their sum is equal to one, with larger weights indicating greater contribution to the final Super Learner prediction. Super Learner ensembles often included XGBoost (mean weight 0.40), GAM (mean weight 0.26), and elastic net (mean weight 0.14) for SGA, and GAM (mean weight 0.48), RF (mean weight 0.27), and XGBoost (mean weight 0.18) for LGA. Similar key predictors were identified for each outcome and between estimation methods (Table 6.7), all of which were related to the maternal generation only: birthweight z-score, gestational weight gain at 26 weeks, and pre-pregnancy BMI.

6.4.5 Sensitivity analysis

The prevalence of both SGA3 and LGA97 was 3.0% in training samples, which represented the value of no discrimination for AUC-PR. Super Learner predictions improved when using both G0 and G1 predictors compared to using G0 predictors or G1 predictors only, and discriminative performance was better for the outcome LGA97 than SGA3 (Table 6.9). In general, the key predictors for SGA3 and LGA97 (Table 6.10 and 6.11) were similar to those identified in the primary analyses (i.e., maternal birthweight z-score, weight gain in pregnancy at 26 weeks, and pre-pregnancy BMI). In the sensitivity analysis, however, maternal smoking in pregnancy was identified as a useful predictor of SGA3.

6.5 Discussion

6.5.1 Main findings

The current study used a large sample of prospectively collected data on three generations of Nova Scotians to assess the improvement of prediction models for SGA and LGA in nulliparous women by adding grandmaternal factors and maternal birth characteristics. Prediction models were developed using the Super Learner algorithm and included routinely collected data and information readily available in an

antenatal setting. Predictive performance measured using the AUC-PR and AUC-ROC increased with the inclusion of grandmaternal pregnancy-related factors and maternal birth characteristics to models fitted using only maternal characteristics, but discriminative ability remained poor ($\leq$0.22 AUC-PR and $\leq$0.71 AUC-ROC). Models for SGA and LGA were well calibrated. The strongest predictors identified for both outcomes were all related to the maternal generation, including birthweight z-score, weight gain in pregnancy at 26 weeks, and pre-pregnancy BMI.

6.5.2 Interpretation

Development of prediction models to identify cases of SGA and LGA before the third trimester may assist clinical decision-making for obstetrical care and inform which women may benefit from third trimester ultrasound assessment. Current efforts to improve early prediction of SGA and LGA include adding ultrasound measurements, biochemical biomarkers, and results of biophysical tests. The current study, however, focused on using easily obtainable antenatal information. Other prediction models using the same percentile cut-offs for SGA and LGA based on maternal characteristics alone have reported similar AUC-ROC estimates between 0.59 and 0.75 for SGA and LGA[115, 116, 118, 119, 121, 123, 126, 129–131, 139, 203–205], but few have considered nulliparous women[114, 121, 131, 139, 203], in whom prediction is poorer[114, 139]. The AUC-PR estimates from the present study could not be compared to other studies, as we are the first to report this measure of discrimination in this context.

Prediction of SGA and LGA was moderately improved by the addition of grandmaternal factors and maternal birth characteristics. AUC-PR estimates increased from 0.18 to 0.21 for SGA and 0.17 to 0.22 for LGA when these variables were added to the typical set of maternal predictors. As several studies have shown a significant association between a mother's own birthweight and the intrauterine growth of her offspring[10, 29, 206], the observed increase in performance may be attributed to the addition of maternal birthweight z-score. For instance, a meta-analysis of three studies reported a 2.6 times increase in the odds of having a SGA birth in women who themselves were born SGA compared to women who were born non-SGA[207]. Moreover, the results of a study using the Swedish Birth Register indicated that

women who were born LGA had twice the odds of having an LGA infant in their own pregnancy[208]. In the current study, maternal SGA and LGA status was associated with a two and nearly two and a half times increased risk of having a (first-born) SGA and LGA birth, respectively, compared to mothers born non-SGA and non-LGA.

Only two studies considered maternal birthweight in prediction models for SGA and LGA, but discrimination was poor (AUC-ROC 0.63 for SGA, and 0.59 for LGA)[121, 131]. In the current study, maternal birthweight z-score was consistently identified as an important predictor of both outcomes. In an exploratory analysis, the Super Learner algorithm fitted using maternal predictors and maternal birthweight z-score (i.e., ignoring all other grandmaternal predictors) performed as well as a model that included grandmaternal predictors (AUC-ROC estimates of 0.70 vs. 0.69 for SGA and 0.72 vs. 0.71 for LGA). This suggests that the addition of maternal birthweight z-score is likely the cause of the increase in predictive performance.

Although the use of machine learning algorithms to predict health outcomes is becoming more popular[142, 209], their benefit over traditional regression remains unclear[143]. Machine learning algorithms have the theoretical advantage over logistic regression in that they require no distributional assumptions, no explicit model specification, and can easily accommodate nonlinearities. For example, an AUC-ROC estimate of 0.80 was reported for the prediction of SGA using artificial neural network based on first-trimester maternal characteristics, biochemical and oxidative stress biomarkers, and gestational weight gain[210]. However, the results of a study done in the same population as the current study indicated no improvement in model performance using machine learning algorithms compared to logistic regression for the prediction of SGA and LGA[139]. In the present study, AUC-PR and AUC-ROC estimates from the Super Learner were the same or only minimally higher than those derived from logistic regression models. Since prediction models contained continuous predictors that were likely to have nonlinear associations with the risk of SGA and LGA (e.g., birthweight z-score, BMI, and gestational weight gain), it was expected that the Super Learner algorithm would perform better than logistic regression. Prediction models developed using GAM performed nearly as well as

the Super Learner algorithm, suggesting nonlinear associations were likely accommodated by splines and meaningful interactions among predictors were unlikely to exist. Although the Super Learner algorithm did not substantially improve the prediction of SGA and LGA in the current study, researchers should still consider its use in other datasets that may be more complex or contain a larger number of predictors.

6.5.3 Strengths and limitations

The main strengths of this study are, first, the use of a large sample of prospectively collected data on a diverse set of variables from three generations of Nova Scotians. Secondly, this study used a flexible modeling approach to predict fetal growth abnormalities, which, as opposed to traditional regression-based prediction models, reduces the risk of bias due to model misspecification. This study also has several limitations worth discussing. First, a high proportion of missingness was observed for grandmaternal pre-pregnancy height (information required to calculate BMI). Maternal height has only been routinely collected in the NSAPD since 2003 and was missing for approximately 90% of G1 births. However, analyses of multiply imputed BMI values were expected to be minimally biased since height is likely to be missing at random, and the imputation procedure included variables that are correlated with height. Secondly, this study was limited by the availability of predictors in the NSAPD, and so other early-pregnancy factors such as racial origin and paternal characteristics could not be considered.

6.6 Conclusion

Adding grandmaternal risk factors and maternal birth characteristics modestly improved the prediction of fetal growth abnormalities in nulliparous women as compared to models based solely on maternal characteristics; however, prediction remains poor and more research is needed to identify useful predictors that can be easily obtained early in pregnancy. A novel finding of this study is that the results suggest maternal birthweight z-score to be a useful predictor of abnormal fetal growth.

Table 6.1: Details of candidate predictors of infant (G2) fetal growth abnormalities

Predictors	**Type (units or levels)**	**G0 predictors**	**G1 predictors**
Sociodemographics			
Maternal age	Continuous (years)	X	X
Marital status	Binary (marrried/common-law, other)	X	X
Area-level income quintile	Categorical	X	X
Area of residence	Binary (rural, urban)	X	X
Pregnancy risk factors			
Pre-pregnancy body mass index	Continuous (kg/m^2)	X	X
Pre-existing hypertension	Binary (yes, no)	X	X
Pre-existing diabetes	Binary (yes, no)		X
Pregnancy characteristics			
Weight gain in pregnancy	Continuous (kg)	X	
Any smoking in pregnancy	Binary (yes, no)	X	X
Any alcohol use in pregnancy	Binary (yes, no)	X	X
Gestational diabetes	Binary (yes, no)	X	X
Pregnancy-induced hypertension	Binary (yes, no)	X	X
Mode of delivery	Binary (vaginal, Caesarean section)	X	
Neonatal characteristics			
Maternal [G1] birthweight z-score	Continuous (SD units)	X	
Maternal [G1] gestational age at birth	Continuous (weeks)	X	
Infant [G2] sex	Binary (male, female)		X

Abbreviations: *G0* grandmaternal; *G1* maternal; *SD* standard deviation

Table 6.2: Tuning parameter setting, definition, and grid of values assessed for each base learner included in the Super Learner ensemble

Algorithm	**R package**	**Hyperparameter**		
		Setting	Definition	Values assessed
Generalized additive models	gam	deg.gam	Degrees of freedom	{2, 3, 4}
Elastic net	glmnet	alpha	Elastic net penalty	{0.05, 0.3, 0.7, 0.95}
		lambda	Regularization parameter	Default (100 values)
Random forest[a]	ranger	mtry	Number of covariates (p) sampled for each split	{floor($\sqrt{p}/2$), floor($\sqrt{p}$)), p}
		nodesize	Minimum node size	{5, 100, 455}
Extreme gradient boosting	XGBoost	ntrees	Number of trees	{250, 1000}
		max_depth	Maximum depth of a tree	{2, 4}
		shrinkage	Step size shrinkage	{0.05, 2}
Support vector machine	Kernlab	C	Regularization parameter	{0.1, 1, 10, 100}

[a] Number of trees constructed for each forest (hyperparameter setting *ntree*) set at 2000 for all grid configurations

Table 6.3: Sample characteristics overall and by SGA and LGA status

	Complete data, n	**Overall n=9068**	**SGA (<10th percentile)** No (n=8172)	Yes (n=896)	SMD	**LGA (>90th percentile)** No (n=8166)	Yes (n=902)	SMD
Grandmaternal (G0) characteristics								
Maternal age [years]	9068	23.6 (4.6)	23.6 (4.5)	23.5 (4.7)	0.02	23.5 (4.5)	23.8 (4.6)	0.07
Married or common-law	8748	60.4	60.6	58.9	0.03	60.2	62.5	0.05
Area-level income quintile	6366				0.05			0.06
Low (Q1)		25.8	25.9	25.4		26.0	23.6	
Middle (Q2-Q4)		63.8	63.6	65.3		63.6	65.7	
High (Q5)		10.4	10.6	9.3		10.4	10.7	
Rural residence	6395	35.4	34.8	41.3	0.13	35.5	34.8	0.01
Pre-pregnancy body mass index [kg/m^2]	623	23.2 (4.8)	23.3 (4.8)	22.9 (4.3)	0.08	23.2 (4.8)	23.7 (4.6)	0.12
Pre-existing hypertension	9068	0.7	0.7	0.6	0.02	0.7	0.8	0.01
Weight gain in pregnancy [kg]	7286	15.2 (6.2)	15.3 (6.1)	14.4 (6.8)	0.14	15.1 (6.2)	16.2 (5.8)	0.19
Any smoking in pregnancy	8001	40.8	40.5	43.8	0.07	40.9	40.5	0.01
Any alcohol use in pregnancy	7805	14.5	14.4	15.1	0.02	14.6	13.6	0.03
Gestational diabetes	9068	1.8	1.8	1.9	0.01	1.8	1.9	0.01
Hypertensive disorders of pregnancy	9068	14.3	14.3	15.2	0.03	14.2	15.3	0.03
Caesarean-section	9062	18.3	18.7	15.2	0.09	17.8	23.6	0.14
Maternal (G1) characteristics at birth								
Born large for gestational age	8849	9.5	10.1	4.4	0.22	8.3	20.4	0.35
Born small for gestational age	8849	9.9	9.1	17.9	0.26	10.5	5.2	0.20
Birthweight z-score [SD units]	8849	-0.19 (1.0)	-0.15 (1.0)	-0.59 (1.0)	0.44	-0.24 (1.0)	0.27 (1.1)	0.50
Maternal (G1) pregnancy characteristics								
Maternal age [years]	9068	23.9 (4.5)	24.0 (4.5)	23.5 (4.3)	0.11	23.9 (4.5)	24.1 (4.5)	0.04
Married or common-law	8104	45.5	46.2	40.0	0.13	45.1	49.4	0.08
Area-level income quintile	8756				0.07			0.00
Low (Q1)		22.4	22.4	22.3		22.4	22.4	
Middle (Q2-Q4)		66.6	66.3	68.5		66.6	66.5	

Table 6.3: continued

	Complete data, n	**Overall n=9068**	**SGA (<10th percentile)** No (n=8172)	Yes (n=896)	SMD	**LGA (>90th percentile)** No (n=8166)	Yes (n=902)	SMD
High (Q5)		11.0	11.2	9.2		11.0	11.1	
Rural residence	8777	33.6	32.8	40.9	0.17	33.8	31.0	0.06
Pre-pregnancy body mass index [kg/m^2]	7624	25.9 (6.7)	26.1 (6.7)	24.6 (6.4)	0.22	25.7 (6.6)	27.7 (7.2)	0.29
Pre-existing hypertension	9068	0.9	0.9	0.9	0.00	0.9	1.3	0.05
Pre-existing diabetes	9068	0.9	0.9	0.6	0.04	0.7	2.9	0.17
Weight gain in pregnancy at 26 weeks [kg]	6874	8.9 (4.2)	9.1 (4.2)	7.9 (3.6)	0.30	8.8 (4.1)	10.2 (4.4)	0.33
Smoking in pregnancy at 26 weeks	9000	23.1	21.8	34.8	0.29	24.2	13.1	0.29
Alcohol use in pregnancy at 26 weeks	9068	1.1	1.1	1.3	0.03	1.1	0.8	0.04
Gestational diabetes	9068	4.0	4.0	3.9	0.01	3.9	4.9	0.05
Hypertensive disorders of pregnancy	9068	10.5	10.2	13.2	0.09	10.3	12.1	0.06
Infant (G2) characteristics								
Male sex	9068	51.9	51.2	58.3	0.14	52.0	51.0	0.02

Abbreviations: *LGA* large for gestational age; *SGA* small for gestational age; *SMD* standardized mean difference
*Presented as mean (standard deviation) or frequency (%)

Table 6.4: Cross-validated discriminative performance of the Super Learner algorithm (measured using the AUC-PR and AUC-ROC) for predicting SGA and LGA fitted using grandmaternal (G0) predictors alone, maternal (G1) predictors alone, and both G0 and G1 predictors

	AUC-PR (95% CI)	**AUC-ROC (95% CI)**
SGA (<10th percentile)		
G0 predictors only	0.154 (0.140, 0.170)	0.627 (0.605, 0.648)
G1 predictors only	0.183 (0.164, 0.204)	0.659 (0.637, 0.680)
G0 + G1 predictors	0.213 (0.190, 0.238)	0.688 (0.668, 0.707)
LGA (>90th percentile)		
G0 predictors only	0.167 (0.154, 0.182)	0.641 (0.622, 0.660)
G1 predictors only	0.181 (0.162, 0.201)	0.664 (0.643, 0.683)
G0 + G1 predictors	0.217 (0.200, 0.235)	0.705 (0.686, 0.724)

Abbreviations: *AUC-PR* area under the precision-recall curve; *AUC-ROC* area under the receiver operating characteristic curve; *CI* confidence interval; *LGA* large for gestational age; *SGA* small for gestational age

Table 6.5: Cross-validated AUC-PR and AUC-ROC estimates and 95% confidence intervals

Learner	**SGA(<10th percentile)**			**LGA (>90th percentile)**		
	G0 predictors only	G1 predictors only	G0 + G1 predictors	G0 predictors only	G1 predictors only	G0 + G1 predictors
AUC-PR						
Marginal mean	0.099 (0.099, 0.099)	0.099 (0.099, 0.099)	0.099 (0.099, 0.099)	0.100 (0.099, 0.100)	0.100 (0.099, 0.100)	0.100 (0.099, 0.100)
LR (main)	0.154 (0.141, 0.168)	0.180 (0.162, 0.199)	0.206 (0.187, 0.227)	0.168 (0.155, 0.182)	0.176 (0.160, 0.194)	0.210 (0.194, 0.228)
LR (int.)	0.136 (0.124, 0.149)	0.160 (0.138, 0.185)	0.151 (0.136, 0.168)	0.148 (0.136, 0.161)	0.161 (0.147, 0.177)	0.145 (0.122, 0.172)
SVM	0.110 (0.098, 0.123)	0.110 (0.095, 0.128)	0.147 (0.134, 0.162)	0.116 (0.106, 0.126)	0.134 (0.116, 0.154)	0.150 (0.134, 0.168)
RF	0.146 (0.133, 0.160)	0.170 (0.153, 0.187)	0.191 (0.170, 0.214)	0.162 (0.150, 0.176)	0.176 (0.157, 0.198)	0.207 (0.187, 0.228)
Elastic net	0.154 (0.142, 0.168)	0.180 (0.162, 0.199)	0.207 (0.188, 0.227)	0.169 (0.156, 0.183)	0.177 (0.161, 0.195)	0.214 (0.197, 0.233)
GAM	0.154 (0.141, 0.169)	0.183 (0.164, 0.203)	0.210 (0.189, 0.232)	0.168 (0.155, 0.182)	0.181 (0.163, 0.201)	0.217 (0.199, 0.236)
XGBoost	0.150 (0.133, 0.170)	0.178 (0.157, 0.200)	0.208 (0.186, 0.232)	0.164 (0.148, 0.180)	0.170 (0.15, 0.191)	0.206 (0.189, 0.224)
Discrete SL	0.153 (0.139, 0.168)	0.183 (0.164, 0.203)	0.205 (0.186, 0.225)	0.168 (0.154, 0.183)	0.180 (0.162, 0.200)	0.214 (0.195, 0.235)
SL	0.154 (0.140, 0.170)	0.183 (0.164, 0.204)	0.213 (0.190, 0.238)	0.167 (0.154, 0.182)	0.181 (0.162, 0.201)	0.217 (0.200, 0.235)
AUC-ROC						
Marginal mean	0.500 (0.466, 0.534)	0.500 (0.466, 0.534)	0.500 (0.466, 0.534)	0.500 (0.466, 0.534)	0.500 (0.466, 0.534)	0.500 (0.466, 0.534)
LR (main)	0.624 (0.604, 0.644)	0.655 (0.633, 0.676)	0.683 (0.663, 0.702)	0.641 (0.621, 0.660)	0.661 (0.641, 0.681)	0.699 (0.680, 0.718)
LR (int.)	0.594 (0.571, 0.617)	0.628 (0.590, 0.665)	0.618 (0.591, 0.645)	0.612 (0.591, 0.633)	0.636 (0.616, 0.657)	0.602 (0.557, 0.645)
SVM	0.523 (0.493, 0.553)	0.518 (0.476, 0.560)	0.589 (0.566, 0.611)	0.531 (0.504, 0.558)	0.544 (0.516, 0.571)	0.597 (0.569, 0.623)
RF	0.618 (0.597, 0.639)	0.645 (0.623, 0.666)	0.668 (0.646, 0.689)	0.634 (0.614, 0.653)	0.654 (0.633, 0.674)	0.692 (0.672, 0.711)
Elastic net	0.627 (0.608, 0.647)	0.655 (0.634, 0.676)	0.684 (0.665, 0.703)	0.642 (0.623, 0.662)	0.665 (0.644, 0.684)	0.704 (0.685, 0.723)
GAM	0.625 (0.605, 0.645)	0.658 (0.636, 0.679)	0.686 (0.666, 0.705)	0.641 (0.621, 0.661)	0.664 (0.644, 0.683)	0.703 (0.683, 0.721)
XGBoost	0.620 (0.595, 0.643)	0.652 (0.626, 0.676)	0.681 (0.659, 0.701)	0.631 (0.610, 0.652)	0.651 (0.630, 0.672)	0.696 (0.677, 0.715)
Discrete SL	0.626 (0.605, 0.646)	0.658 (0.636, 0.679)	0.683 (0.662, 0.703)	0.641 (0.622, 0.660)	0.662 (0.642, 0.682)	0.701 (0.681, 0.721)
SL	0.627 (0.605, 0.648)	0.659 (0.637, 0.680)	0.688 (0.668, 0.707)	0.641 (0.622, 0.660)	0.664 (0.643, 0.683)	0.705 (0.686, 0.724)

Abbreviations: *AUC-PR* area under the precision-recall curve; *AUC-ROC* area under the receiver operating characteristics curve; *int.* interaction; *G0* grandmaternal; *G1* maternal; *GAM* generalized additive model; *LGA* large for gestational age; *LR* logistic regression; *SGA* small for gestational age; *SL* Super Learner; *SVM* support vector machine; *RF* random forest; *XGBoost* extreme gradient boosting

Table 6.6: Pooled Super Learner weights across validation folds and corresponding standard errors (SE) from models fitted using both grandmaternal (G0) and maternal (G1) predictors

SGA (<10th percentile)			**LGA (>90th percentile)**		
Learner	Weight	SE	Learner	Weight	SE
XGBoost	0.399	0.160	GAM	0.477	0.148
GAM	0.261	0.202	RF	0.266	0.113
Elastic net	0.143	0.172	XGBoost	0.175	0.153
Logistic regression (main)	0.128	0.161	Elastic net	0.070	0.119
RF	0.039	0.062	Logistic regression (interaction)	0.007	0.012
Logistic regression (interaction)	0.017	0.021	SVM	0.004	0.013
SVM	0.012	0.034	Logistic regression (main)	0.001	0.009
Marginal mean	0.000	0.000	Marginal mean	0.000	0.000

Abbreviations: *GAM* generalized additive model; *LGA* large for gestational age; *RF* random forest; *SGA* small for gestational age; *SVM* support vector machine; *XGBoost* extreme gradient boosting

Table 6.7: Variable importance ranking for the prediction of SGA and LGA using the top two prediction algorithms in the Super Learner ensemble

Rank	Predictor	Mean increase in Brier score	Predictor	Mean increase in Brier score
SGA (<10th percentile)				
	Extreme gradient boosting		**Generalized additive model**	
1	G1 Birthweight z-score	0.0032	G1 Weight gain in pregnancy at 26 weeks	0.0065
2	G1 Weight gain in pregnancy at 26 weeks	0.0019	G1 Birthweight z-score	0.0039
3	G1 Pre-pregnancy BMI	0.0017	G1 Pre-pregnancy BMI	0.0034
4	G1 Smoking in pregnancy at 26 weeks	0.0011	G1 Smoking in pregnancy at 26 weeks	0.0018
5	G1 Area of residence	0.0003	G1 Hypertensive disorders of pregnancy	0.0010
LGA (>90th percentile)				
	Random forest		**Generalized additive model**	
1	G1 Birthweight z-score	0.0042	G1 Birthweight z-score	0.0046
2	G1 Weight gain in pregnancy at 26 weeks	0.0040	G1 Weight gain in pregnancy at 26 weeks	0.0033
3	G1 Pre-pregnancy BMI	0.0033	G1 Pre-pregnancy BMI	0.0019
4	G1 Age	0.0017	G1 Smoking in pregnancy at 26 weeks	0.0018
5	G1 Smoking in pregnancy at 26 weeks	0.0016	G0 Gestational diabetes	0.0004

Abbreviations: *BMI* body mass index; *G0* grandmaternal; *G1* maternal; *LGA* large for gestational age; *SGA* small for gestational age

Table 6.8: Super Learner predicted risk of SGA and LGA fitted using both grandmaternal (G0) and maternal (G1) predictors and observed risk estimated from decile groups and pooled across imputed datasets

Decile	**SGA (<10th percentile)**			**LGA (>90th percentile)**		
	n	Predicted risk (%)	Observed risk (%, 95% CI)	n	Predicted risk (%)	Observed risk (%, 95% CI)
1	907.7	2.9	2.4 (0.4, 4.4)	907.7	3.0	2.4 (0.5, 4.3)
2	907.8	4.3	4.5 (2.5, 6.6)	907.7	4.3	3.4 (1.3, 5.5)
3	907.7	5.4	5.2 (3.1, 7.3)	907.7	5.4	4.8 (2.7, 6.8)
4	907.6	6.5	6.5 (4.1, 9)	907.8	6.4	6.2 (4.1, 8.3)
5	907.8	7.7	7.8 (5.3, 10.2)	907.6	7.6	7.4 (5.1, 9.6)
6	907.6	9.0	9.8 (7.2, 12.4)	907.8	8.9	9.0 (6.4, 11.5)
7	907.8	10.5	10.2 (8, 12.5)	907.7	10.5	10.3 (8.3, 12.3)
8	907.6	12.6	11.7 (8.6, 14.8)	907.6	12.7	13.2 (10.8, 15.6)
9	907.8	15.9	15.5 (12.9, 18.1)	907.8	16.0	17.6 (15.3, 19.9)
10	907.6	24.1	25.3 (23.2, 27.4)	907.6	25.1	25.3 (22.8, 27.7)

Abbreviations: *CI* confidence interval; *LGA* large for gestational age; *SGA* small for gestational age

Table 6.9: Cross-validated AUC-PR and AUC-ROC estimates and 95% confidence intervals from sensitivity analysis

Learner	**SGA(<3rd percentile)**			**LGA (>97th percentile)**		
	G0 predictors only	G1 predictors only	G0 + G1 predictors	G0 predictors only	G1 predictors only	G0 + G1 predictors
AUC-PR						
Marginal mean	0.030 (0.030, 0.030)	0.030 (0.030, 0.030)	0.030 (0.030, 0.030)	0.030 (0.030, 0.030)	0.030 (0.03, 0.03)	0.030 (0.030, 0.030)
LR (main)	0.054 (0.042, 0.070)	0.067 (0.052, 0.087)	0.073 (0.057, 0.094)	0.060 (0.048, 0.074)	0.070 (0.056, 0.087)	0.085 (0.068, 0.105)
LR (int.)	0.044 (0.029, 0.066)	0.050 (0.040, 0.062)	0.044 (0.033, 0.060)	0.048 (0.038, 0.060)	0.056 (0.045, 0.070)	0.048 (0.034, 0.067)
SVM	0.036 (0.025, 0.050)	0.033 (0.026, 0.040)	0.047 (0.036, 0.063)	0.033 (0.026, 0.042)	0.042 (0.032, 0.053)	0.057 (0.044, 0.072)
RF	0.046 (0.038, 0.056)	0.054 (0.044, 0.066)	0.062 (0.047, 0.082)	0.054 (0.044, 0.066)	0.073 (0.059, 0.091)	0.078 (0.066, 0.093)
Elastic net	0.054 (0.042, 0.070)	0.066 (0.050, 0.086)	0.073 (0.057, 0.094)	0.065 (0.051, 0.082)	0.070 (0.057, 0.087)	0.087 (0.070, 0.107)
GAM	0.054 (0.038, 0.074)	0.066 (0.050, 0.086)	0.071 (0.054, 0.093)	0.059 (0.048, 0.073)	0.079 (0.060, 0.104)	0.094 (0.070, 0.126)
XGBoost	0.046 (0.035, 0.060)	0.058 (0.045, 0.075)	0.062 (0.049, 0.077)	0.060 (0.047, 0.076)	0.069 (0.051, 0.092)	0.092 (0.067, 0.126)
Discrete SL	0.054 (0.042, 0.069)	0.065 (0.049, 0.085)	0.072 (0.054, 0.096)	0.064 (0.050, 0.082)	0.072 (0.055, 0.094)	0.081 (0.064, 0.102)
SL	0.052 (0.037, 0.072)	0.065 (0.050, 0.085)	0.073 (0.052, 0.100)	0.062 (0.049, 0.079)	0.076 (0.058, 0.098)	0.093 (0.069, 0.125)
AUC-ROC						
Marginal mean	0.500 (0.440, 0.560)	0.500 (0.440, 0.560)	0.500 (0.440, 0.560)	0.500 (0.440, 0.560)	0.500 (0.440, 0.560)	0.500 (0.440, 0.560)
LR (main)	0.621 (0.585, 0.655)	0.666 (0.629, 0.701)	0.691 (0.655, 0.726)	0.650 (0.615, 0.683)	0.692 (0.659, 0.724)	0.732 (0.698, 0.763)
LR (int.)	0.562 (0.515, 0.609)	0.611 (0.569, 0.651)	0.550 (0.500, 0.599)	0.598 (0.557, 0.638)	0.632 (0.590, 0.672)	0.578 (0.517, 0.637)
SVM	0.512 (0.460, 0.563)	0.496 (0.444, 0.549)	0.578 (0.523, 0.631)	0.513 (0.443, 0.583)	0.540 (0.487, 0.591)	0.627 (0.584, 0.669)
RF	0.614 (0.573, 0.653)	0.639 (0.604, 0.673)	0.665 (0.629, 0.700)	0.625 (0.586, 0.662)	0.705 (0.672, 0.736)	0.738 (0.706, 0.768)
Elastic net	0.622 (0.585, 0.658)	0.663 (0.627, 0.698)	0.692 (0.653, 0.729)	0.664 (0.630, 0.696)	0.701 (0.669, 0.731)	0.744 (0.711, 0.774)
GAM	0.619 (0.581, 0.655)	0.668 (0.632, 0.703)	0.694 (0.657, 0.728)	0.647 (0.613, 0.681)	0.702 (0.669, 0.733)	0.740 (0.706, 0.770)
XGBoost	0.608 (0.567, 0.648)	0.639 (0.603, 0.673)	0.672 (0.629, 0.712)	0.641 (0.603, 0.676)	0.695 (0.660, 0.728)	0.741 (0.708, 0.770)
Discrete SL	0.621 (0.583, 0.657)	0.666 (0.629, 0.701)	0.689 (0.644, 0.730)	0.662 (0.627, 0.696)	0.700 (0.665, 0.734)	0.738 (0.703, 0.770)
SL	0.617 (0.580, 0.652)	0.666 (0.630, 0.700)	0.692 (0.653, 0.729)	0.659 (0.625, 0.692)	0.710 (0.677, 0.740)	0.752 (0.721, 0.781)

Abbreviations: *AUC-PR* area under the precision-recall curve; *AUC-ROC* area under the receiver operating characteristics curve; *int.* interaction; *G0* grandmaternal; *G1* maternal; *GAM* generalized additive model; *LGA* large for gestational age; *LR* logistic regression; *SGA* small for gestational age; *SL* Super Learner; *SVM* support vector machine; *RF* random forest; *XGBoost* extreme gradient boosting

Table 6.10: Pooled Super Learner weights across validation folds and corresponding standard errors (SE) from models fitted using both grandmaternal (G0) and maternal (G1) predictors from sensitivity analysis

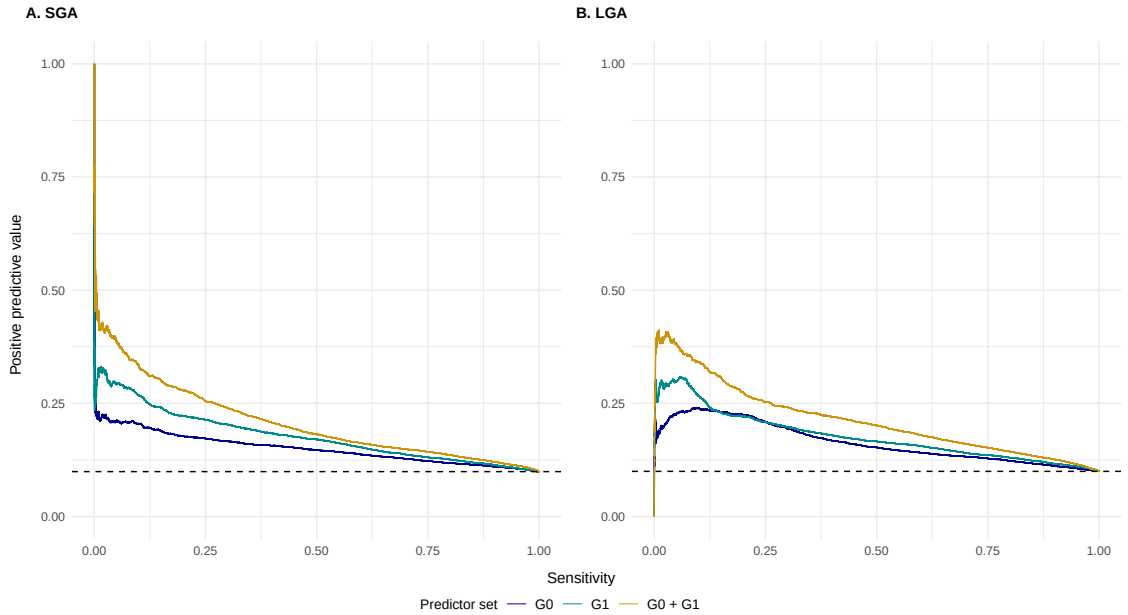
SGA (<3rd percentile)			**LGA (>97th percentile)**		
Learner	Weight	SE	Learner	Weight	SE
Logistic regression (main)	0.286	0.225	RF	0.310	0.133
Elastic net	0.262	0.264	XGBoost	0.297	0.163
RF	0.217	0.133	GAM	0.284	0.147
XGBoost	0.095	0.141	SVM	0.050	0.086
GAM	0.060	0.127	Elastic net	0.039	0.092
SVM	0.040	0.091	Marginal mean	0.011	0.025
Marginal mean	0.026	0.056	Logistic regression (interaction)	0.009	0.016
Logistic regression (interaction)	0.014	0.014	Logistic regression (main)	0.000	0.000

Abbreviations: *GAM* generalized additive model; *LGA* large for gestational age; *RF* random forest; *SGA* small for gestational age; *SVM* support vector machine; *XGBoost* extreme gradient boosting

Table 6.11: Variable importance ranking for the prediction of SGA and LGA using the top two prediction algorithms in the Super Learner ensemble in the sensitivity analysis

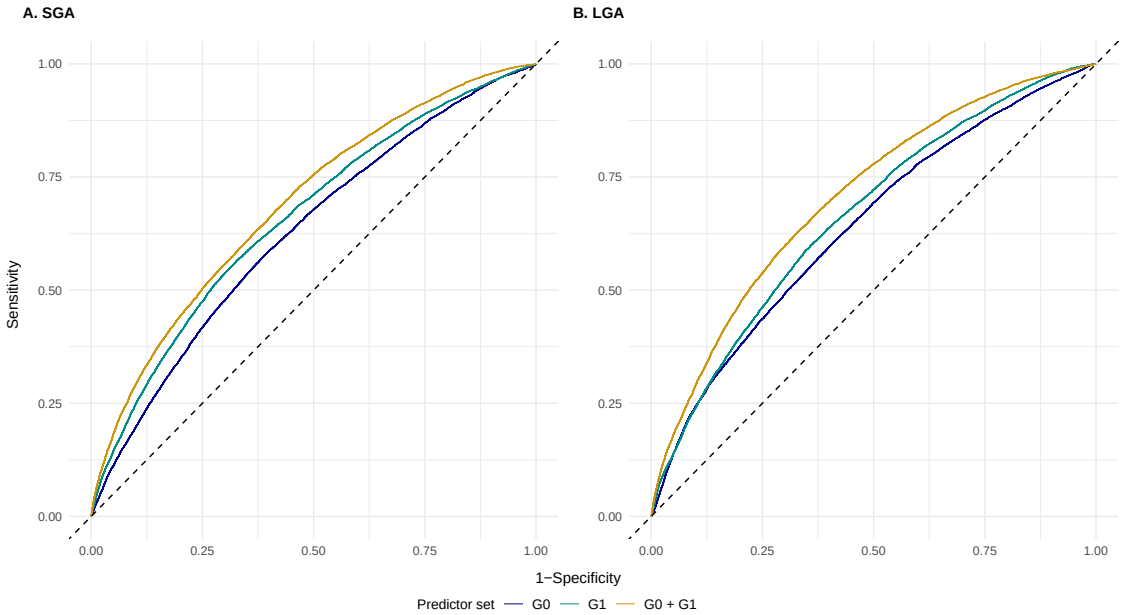
Rank	Predictor	Mean increase in Brier score	Predictor	Mean increase in Brier score
SGA (<3rd percentile)				
	Logistic regression		**Elastic net**	
1	G1 Weight gain in pregnancy at 26 weeks	0.00047	G1 Birthweight z-score	0.0065
2	G1 Birthweight z-score	0.00038	G1 Smoking in pregnancy at 26 weeks	0.0039
3	G1 Pre-pregnancy BMI	0.00037	G1 Weight gain in pregnancy at 26 weeks	0.0034
4	G1 Gestational hypertension	0.00028	G1 Gestational hypertension	0.0018
5	G1 Smoking in pregnancy at 26 weeks	0.00028	G1 Pre-pregnancy BMI	0.0010
LGA (>97th percentile)				
	Random forest		**Extreme gradient boosting**	
1	G1 Pre-pregnancy BMI	0.00186	G1 Weight gain in pregnancy at 26 weeks	0.00076
2	G1 Weight gain in pregnancy at 26 weeks	0.00177	G1 Birthweight z-score	0.00072
3	G1 Birthweight z-score	0.00098	G1 Pre-pregnancy BMI	0.00055
4	G0 Pre-pregnancy BMI	0.00087	G1 Smoking in pregnancy at 26 weeks	0.00012
5	G0 Age	0.00077	G0 Weight gain in pregnancy	0.00003

Abbreviations: *BMI* body mass index; *G0* grandmaternal; *G1* maternal; *LGA* large for gestational age; *SGA* small for gestational age



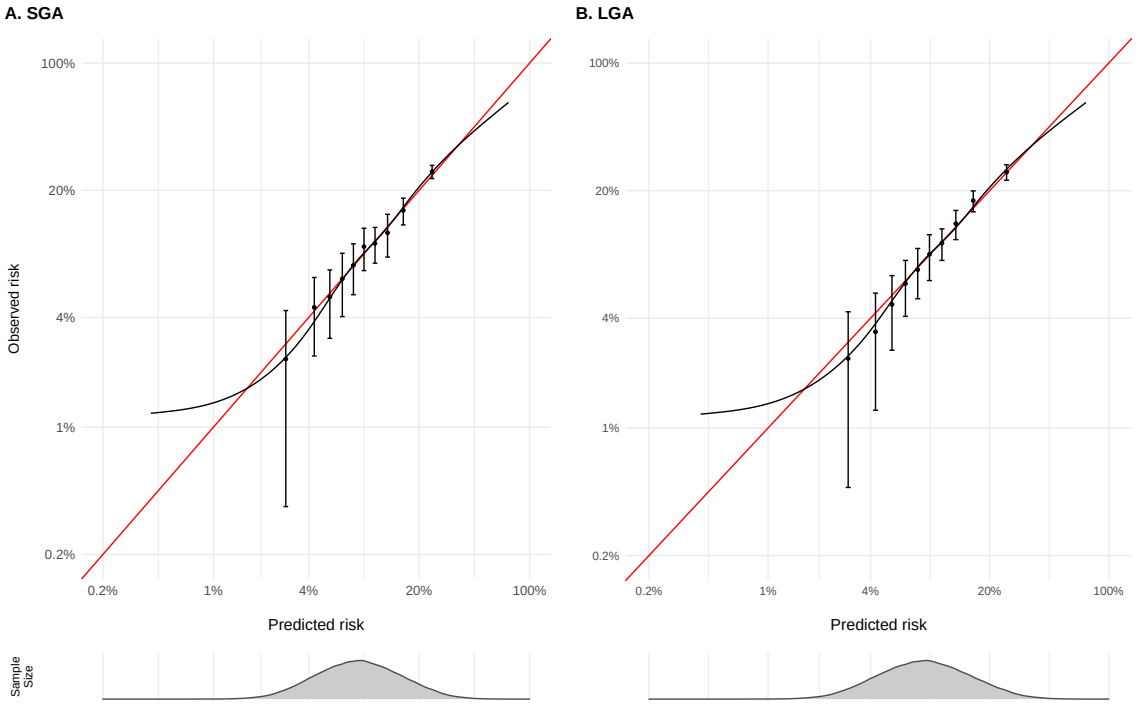
Abbreviations: *G2* infant; *LGA* large for gestational age; *SGA* small for gestational age

Figure 6.1: Cross-validated discriminative performance using precision-recall curves estimated from the Super Learner algorithm fitted using grandmaternal (G0) predictors alone, maternal (G1) predictors alone, and both G0 and G1 predictors for A) SGA and B) LGA. The AUC-PR value that indicates no discrimination is the average prevalence of SGA (9.9%) and LGA (9.9%) in training samples and is indicated by the dotted line



Abbreviations: *AUC-ROC* area under the receive operating characteristic curve; *G2* infant; *LGA* large for gestational age; *SGA* small for gestational age

Figure 6.2: Cross-validated discriminative performance using ROC curves estimated from the Super Learner algorithm fitted using grandmaternal (G0) predictors alone, maternal (G1) predictors alone, and both G0 and G1 predictors for A) SGA and B) LGA. No discrimination is indicated by the dotted line



Abbreviations: *AUC-PR* area under the precision-recall curve; *G2* infant; *LGA* large for gestational age; *SGA* small for gestational age

Figure 6.3: Calibration plots showing the comparison of predicted risk from Super Learner algorithms using both grandmaternal (G0) and maternal (G1) predictors and observed risk plotted on the logarithmic scale for A) SGA and B) LGA. Calibration point estimates and confidence intervals calculated using decile groups. Perfect calibration indicated by the red line

Chapter 7

Discussion

The research presented in this thesis is one of the first to examine the effects of grandmaternal pre-pregnancy BMI and other pregnancy-related factors on child birthweight in a large cohort of prospectively collected data. Obesity is associated with numerous health conditions and morbidity, and the persistent increase in the prevalence of obesity, particularly among young people, is a major health concern. Pediatric obesity tracks into adulthood and tends to worsen with age, leading to adverse health outcomes in adulthood. Furthermore, infants born to parents with obesity are at increased risk of adverse neonatal outcomes, and of becoming obese themselves in childhood and adolescence, thus perpetuating the cycle of obesity. In Canada, the estimated economic burden of obesity between 2000 and 2008 increased by $735 million, from $3.9 to $4.6 billion[211]. This emphasizes the growing urgency of a better understanding of the transmission of weight across generations in the Canadian population in pursuance of more effective intervention strategies.

Although population-based studies of the intergenerational effects of maternal health are clearly needed, multigenerational studies are difficult to conduct[212]. Assessing effects using randomized controlled trials is unethical, and any type of prospective study is unfeasible due to the length of follow-up time. Like the present research, most multigenerational studies rely on birth registers and existing cohorts, which were not created to specifically address multigenerational questions. Studies of this kind not only inherit the limitations of the existing data, but are faced with new challenges, such as those related to clustering and missing data. The clustering feature of birth registers and perinatal databases is akin to a repeated-measures study design but is complicated by individuals with differing numbers of deliveries, variable time between deliveries, and many having only one delivery. This structure, however, presents a unique opportunity to assess new imputation techniques that incorporate the correlation among observations. Few imputation strategies do this,

and those that do may not be well suited to handle this complex structure.

Of particular concern in studies focusing on correlated BMI information, like those in Chapters Five and Six, is that this variable is prone to missingness and may be MNAR. As multiple imputation methods assume data to be MAR, this tends to be the missing mechanism under which new imputation techniques are assessed. However, since MNAR is indistinguishable from MAR in observed data, evaluating the robustness of new techniques to violations of this assumption is valuable. Using data with induced missingness originally sampled from a subset of the NSAPD that consisted of approximately 50% singleton clusters, the results of Chapter Four suggested that imputation using the recently proposed tree-based algorithm MERF was moderately biased in analyses of pre-pregnancy BMI when weight was simulated to be MAR, but was severely biased in MNAR scenarios. This method performed worse than random forest, which was found to be the least biased and most efficient method evaluated in both MAR and MNAR settings. Contrary to other studies in the literature that used a model-based simulation approach, this study induced missingness in data from a real perinatal database to better reflect reality and maintain the natural relationships among the variables. Based on the results of Chapter Four, random forest-based imputation may be the best strategy when faced with missingness in similarly structured datasets.

The results in Chapter Four also add to the evidence supporting machine learning-based imputation algorithms in complex datasets, such as those that have dependent observations or highly correlated variables. With increasing complexity of the data, correctly specifying the imputation model can be difficult, especially when nonlinear effects and interactions are suspected. Unlike parametric-based imputation methods, those that are based on machine learning algorithms, like random forest, do not require explicit specification of the imputation model and, as suggested in Chapter Four, may outperform parametric-based imputation in some settings. Findings from this research guided the imputation of missing data in the studies from Chapters Five and Six, and will help inform multiple imputation procedures in future studies that use the NASPD, the 3G cohort, and other perinatal databases with similar data structures.

The 3G cohort enabled two intergenerational studies of grandmaternal pregnancy-related characteristics and fetal growth: one with the goal of estimating mediator-specific natural effects (Chapter Five), and one with the goal of improving predictions for abnormal fetal growth in the nulliparous population (Chapter Six). The potential benefits of assessing etiologic mechanisms of the transmission of weight can highlight opportunities for intervention that can lessen both the economic and personal burden associated with obesity. In light of this opportunity, research in Chapter Five examined the possible downstream impact on child birthweight under alternative hypothetical conditions on the distribution of grandmaternal pre-pregnancy BMI. The findings of this study suggested no strong evidence for an association between grandmaternal pre-pregnancy BMI and child birthweight after accounting for measured confounders. Moreover, only negligible estimates of the natural direct effects and mediator-specific effects via maternal pre-pregnancy BMI were found.

In studies of grandmaternal and child body weight measures, mediation by maternal factors is seldom examined, and those that have, typically relied on traditional approaches to mediation analysis. Traditional approaches to mediation analysis have been shown to be limited in many settings[26], but remain the most popular approach to mediation in observational studies. For example, in a 2022 meta-analysis of fifty observational studies, only five stated that they applied modern approaches to mediation analysis[112]. The research in Chapter Five, however, used methods based within the counterfactual framework, which enabled more intuitive interpretations of the direct and indirect effects by considering different hypothetical conditions on the distribution of grandmaternal pre-pregnancy BMI. Estimates defined as contrasts of potential outcomes leads to a straightforward understanding of the possible impact on the outcome under different "what if" conditions on the exposure. These definitions are particularly useful when models contain interactions and nonlinear terms; situations where interpreting effect estimates from traditional mediation analyses is difficult. Unlike other studies, this study was able to control for a rich set of exposure-mediator, exposure-outcome, and mediator-outcome confounders, and account for possible intermediate confounders when exploring the possible role of maternal pre-pregnancy BMI in the association between grandmaternal pre-pregnancy BMI and child birthweight.

Causal interpretation of the mediation results presented in Chapter Five requires strong assumptions that were unlikely to have been met. Of particular concern was the violation of the consistency assumption. This topic has been the subject of much debate in the causal inference literature, with many arguing the relevancy of exposure variables like BMI as there may exist multiple ways in which an individual can achieve a BMI of x (e.g., genetics, medical conditions, diet, and lifestyle), and each of these may have a different effect on the outcome[27, 190, 191]. When estimating the causal effect of BMI, causal inference methods use information on individuals with low BMI to predict counterfactual outcomes (e.g., potential outcomes if all individuals had a BMI of 22 kg/m^2) for individuals with high BMI. Contrasts involving these potential outcomes cannnot be interpreted as the effects of an intervention targeted at lowering BMI. Individuals with a low BMI may have achieved this by a combination of mechanisms, and not necessarily by the action of the intervention, and so data on these individuals is not representative of the consequences associated with this intervention. Hernán and Taubman[27] also argue the difficulty of identifying relevant confounders and the increased probability of violations of the positivity assumption as a result of ill-defined exposure variables.

Despite these issues, pre-pregnancy BMI is still an important risk factor in pregnancy. Weight and height information are routinely collected, making BMI readily available in large representative samples. Although hypothetical conditions on the distribution of BMI cannot be translated into meaningful interventions, there remains value in exploring its relationship with first- and second-generation outcomes. Studies like that in Chapter Five may provide insight into the possible mechanisms involved with the transmission of weight, and may help generate hypotheses that could be followed up with further research using more well-defined exposure variables.

Although the results of Chapter Five suggested no strong evidence for an association between grandmaternal pre-pregnancy BMI and child birthweight, findings from Chapter Six demonstrated that intergenerational variables, such as maternal birthweight information, were still useful for predicting abnormal fetal growth in nulliparous women. AUC-PR estimates increased from 0.18 to 0.21 for SGA and 0.17 to 0.22 for LGA when grandmaternal factors and maternal birth characteristics

were added to the typical set of predictors on maternal information. Identifying which infants are at highest risk of abnormal growth may help in clinical decision-making related to obstetrical care, inform which women may benefit from third trimester ultrasound assessment, and may also help with the long-term management of these infants to avoid, or at least mitigate the risk of chronic disease in adulthood. This study was the first to use an ensemble machine learning technique (i.e., Super Learner) for predicting SGA and LGA and will help inform other studies of this kind.

Prediction models based on machine learning algorithms are becoming more common in the field of obstetrics and gynaecology[213], but in general, whether they offer substantial benefit over logistic regression remains unclear[143]. Machine learning algorithms theoretically have advantage over logistic regression in that they make no distributional assumptions, do not require explicit specification of a regression model, and can capture nonlinear associations between the predictors and the outcome. However, Chapter Six found discriminative performance of the Super Learner algorithm for predicting fetal growth abnormalities to be the same or only minimally better than that of logistic regression. The slow uptake of machine learning algorithms in this field may be attributed to their shortcomings. Compared to logistic regression, machine learning algorithms require more skill to implement, have greater computational requirements, and are less easily interpreted. The mechanics underlying these algorithms are harder to understand. This limitation may affect clinicians' trust and acceptability of these models.

An additional aspect of this research was applying machine learning algorithms outside of the typical prediction setting (Chapter Six) for which they were designed. Modern causal inference methods based on the counterfactual framework, like the g-computation procedure used in Chapter Five, involve predicting potential outcomes prior to calculating the desired estimate. Similarly, multiple imputation techniques, like MERF and random forest-based imputation used in Chapter Four, involve prediction to impute missing data values. These pre-final prediction steps in causal inference methods and multiple imputation can be thought of as "nuisance" steps, meaning they are necessary to obtain unbiased results, but do not directly lead to the quantity of interest. Traditionally, regression models have been used to achieve these "nuisance" steps, but, like the research in this thesis suggests, better predictions may

be obtained by exploiting the benefits of machine learning algorithms[214, 215].

It is, however, important to recognize the limitations of incorporating machine learning algorithms in singly-robust methods like the g-computation procedure used in Chapter Five. G-computation, like inverse probability of treatment weighting, is a singly-robust method because, in the context of estimating the average causal effect for example, it relies on only one nuisance function to estimate potential outcomes, namely a model for the outcome mechanism (i.e., $E[Y \mid X = x]$). As a result, singly-robust estimators require fast convergence of the nuisance model, thus limiting the use of algorithms with slower converging rates like machine learning algorithms[216]. Moreover, it is currently unknown whether bootstrapping procedures are valid for g-computation when more data-adaptive methods are used to estimate the nuisance model[216].

These challenges with singly-robust estimators led to the development of several double-robust methods, such as the augmented inverse probability of treatment weighted estimator and targeted maximum likelihood estimation (TMLE)[135, 217]. To estimate the average causal effect, these methods rely on two nuisance functions, one for both the outcome and exposure mechanisms, and if either mechanism is correctly estimated, then the resulting point estimate will be consistent[217–219]. Even when machine learning algorithms are used to model the outcome and treatment mechanisms, these methods have a number of desirable asymptotic properties, including construction of valid CIs[220]. Parametric g-computation was used to perform the mediation analysis in Chapter Five, but predictions may have improved by using a double-robust estimator like TMLE coupled with the Super Learner algorithm. Although double-robust estimators have been integrated into causal mediation analysis[221, 222], estimation in the presence of intermediate confounding and continuous mediators and intermediate confounders pose significant challenges and remains an area of future work[223].

Computational challenges were encountered in this work when implementing the Super Learner algoithm in Chapter Six and investigating the use of MERF to impute pre-pregnancy BMI values in Chapter Four. These challenges highlight the issue that infrastructure and computational resources remain a barrier when attempting to use

machine learning algorithms in large datasets. Some Canadian health care organizations, like the SickKids Research Institute[224], have implemented high powered computing infrastructure using a secure cloud-based platform. However, as high powered computing infrastructure can be costly, and many organizations are hesitant to move to a cloud-based platform due to issues related to cost and security, this lack of infrastructure remains an obstacle, particularly when using personal health records in Nova Scotia. Practicality, feasibility, and a realistic assessment of the possible gain in using computationally intensive methods, like machine learning algorithms, should be considered when faced with limited resources and the choice of a using a simpler model.

The findings of the research presented in this thesis lay the groundwork for future studies using the NSAPD and the 3G cohort. An underlying theme of this work was the intergenerational transmission of weight, but the 3G cohort presents unique opportunities to study other associations between antenatal and prenatal exposures and health outcomes in the first- and second-generation offspring. For example, the 3G cohort can be used to examine a woman's in utero exposure to excessive gestational weight gain, GDM, and hypertension, with her own pregnancy complications and her offspring's birthweight. Furthermore, inclusion of additional variables and linkages to other datasets would allow for the effects of in utero exposures on long-term child health outcomes to be assessed.

Limitations of the individual contributions of this thesis were raised in the discussion section of each manuscript. The primary limitations of this work as a whole will be briefly discussed. First, maternal height, the denominator of BMI, has only been routinely collected in the NSAPD since 2003 and was largely incomplete for the grandmothers. The study in Chapter Four helps to inform how this missing data should be handled, but it is important to recognize that the results of simulation studies may not generalize to all situations. Additionally, we must always remember that the missing values were, in fact, missing in the first place, and imputation is an imperfect tool. Secondly, the NSAPD and 3G cohort do not contain information on the fathers (or grandfathers), so only the maternal line could be examined. Paternal information could have been incorporated in all studies of this thesis, and expanding the scope of these objectives to paternal information is an important area of future

research.

This thesis will inform further investigation into the appropriate treatment of missing data in complex data structures, the examination of the association between maternal exposure to adverse intrauterine conditions and the short- and long-term health of her offspring, and the assessment of the ability of intergenerational information in predicting second-generation health outcomes. Together, these complementary investigations offer notable contributions to the etiological knowledge of the transmission of weight along the maternal line.

Bibliography

[1] Statistics Canada. Measured adult body mass index (BMI) (World Health Organization classification), by age group and sex, Canada and provinces, Canadian Community Health Survey - Nutrition. Government of Canada; 2017. Available from: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310079401.

[2] Vats H, Saxena R, Sachdeva MP, Walia GK, Gupta V. Impact of maternal pre-pregnancy body mass index on maternal, fetal and neonatal adverse outcomes in the worldwide populations: A systematic review and meta-analysis. Obesity Research & Clinical Practice. 2021 Nov;15(6):536–545.

[3] McAuliffe FM, Killeen SL, Jacob CM, Hanson MA, Hadar E, McIntyre HD, et al. Management of pre-pregnancy, pregnancy, and postpartum obesity from the FIGO Pregnancy and Non-Communicable Diseases Committee: A FIGO (International Federation of Gynecology and Obstetrics) guideline. International Journal of Gynecology & Obstetrics. 2020 Sep;151(S1):16–36.

[4] Perinatal Epidemiology Research Unit. Nova Scotia Atlee Perinatal Database Report of Indicators: 2010 - 2019. Dalhousie University: Reproductive Care Program of Nova Scotia; 2019. Available from: http://rcp.nshealth.ca/sites/default/files/publications/APD_Report_2010_2019.pdf.

[5] D'Souza R, Horyn I, Pavalagantharajah S, Zaffar N, Jacob CE. Maternal body mass index and pregnancy outcomes: a systematic review and metaanalysis. American journal of obstetrics & gynecology MFM. 2019 Nov;1(4):100041.

[6] Yu Z, Han S, Zhu J, Sun X, Ji C, Guo X. Pre-Pregnancy Body Mass Index in Relation to Infant Birth Weight and Offspring Overweight/Obesity: A Systematic Review and Meta-Analysis. PLoS ONE. 2013 Apr;8(4):e61627.

[7] Thaker VV. Genetic and epigenetic causes of obesity. Adolescent Medicine: State of the Art Reviews. 2017;28(2):379–405.

[8] Harville EW, Kruse AN, Zhao Q. The Impact of Early-Life Exposures on Women's Reproductive Health in Adulthood. Current Epidemiology Reports. 2021 Dec;8(4):175–189.

[9] De Stavola BL, Leon DA, Koupil I. Intergenerational correlations in size at birth and the contribution of environmental factors: The Uppsala Birth Cohort Multigenerational Study, Sweden, 1915-2002. American Journal of Epidemiology. 2011 Jul;174(1):52–62.

[10] Lahti-Pulkkinen M, Bhattacharya S, Räikkönen K, Osmond C, Norman JE, Reynolds RM. Intergenerational Transmission of Birth Weight Across 3 Generations. American Journal of Epidemiology. 2018 Jun;187(6):1165–1173.

[11] Kelly GE, Murrin C, Viljoen K, O'Brien J, Kelleher C. Body mass index is associated with the maternal lines but height is heritable across family lines in the Lifeways Cross-Generation Cohort Study. BMJ Open. 2014 Dec;4(12):e005732.

[12] Harville EW, Jacobs MB, Qi L, Chen W, Bazzano LA. Multigenerational Cardiometabolic Risk as a Predictor of Birth Outcomes: The Bogalusa Heart Study. The Journal of Pediatrics. 2017 Feb;181:154–162.e1.

[13] Shen Y, Zhang H, Jiang Y, Mzayek F, Arshad H, Karmaus W. Maternal Birth Weight and BMI Mediate the Transgenerational Effect of Grandmaternal BMI on Grandchild's Birth Weight. Obesity. 2020 Mar;28(3):647–654.

[14] Hypponen E. Effects of grandmothers' smoking in pregnancy on birth weight: intergenerational cohort study. BMJ. 2003 Oct;327(7420):898–0.

[15] Misra DP, Astone N, Lynch CD. Maternal Smoking and Birth Weight: Interaction With Parity and Mothers Own In Utero Exposure to Smoking. Epidemiology. 2005 May;16(3):288–293.

[16] Miller LL, Pembrey M, Davey Smith G, Northstone K, Golding J. Is the Growth of the Fetus of a Non-Smoking Mother Influenced by the Smoking of Either Grandmother while Pregnant? PLOS ONE. 2014 Feb;9(2):e86781.

[17] Rillamas-Sun E, Harlow SD, Randolph JF. Grandmothers' Smoking in Pregnancy and Grandchildren's Birth Weight: Comparisons by Grandmother Birth Cohort. Maternal and Child Health Journal. 2014 Sep;18(7):1691–1698.

[18] Ding M, Yuan C, Gaskins AJ, Field AE, Missmer SA, Michels KB, et al. Smoking during pregnancy in relation to grandchild birth weight and BMI trajectories. PLOS ONE. 2017 Jul;12(7):e0179368.

[19] Rumrich IK, Hänninen O, Viluksela M, Vähäkangas K. Effect of Grandmaternal Smoking on Body Size and Proportions at Birth. International Journal of Environmental Research and Public Health. 2021 May;18(9):4985.

[20] Astone NM, Misra D, Lynch C. The effect of maternal socio-economic status throughout the lifespan on infant birthweight. Paediatric and Perinatal Epidemiology. 2007 Jul;21(4):310–318.

[21] Collins JW, David RJ, Rankin KM, Desireddi JR. Transgenerational Effect of Neighborhood Poverty on Low Birth Weight Among African Americans in Cook County, Illinois. American Journal of Epidemiology. 2009 Jan;169(6):712–717.

[22] Gavin AR, Hill KG, Hawkins JD, Maas C. The Role of Maternal Early-Life and Later-Life Risk Factors on Offspring Low Birth Weight: Findings From a Three-Generational Study. Journal of Adolescent Health. 2011 Aug;49(2):166–171.

[23] Huang JY, Gavin AR, Richardson TS, Rowhani-Rahbar A, Siscovick DS, Enquobahrie DA. Are Early-Life Socioeconomic Conditions Directly Related to Birth Outcomes? Grandmaternal Education, Grandchild Birth Weight, and Associated Bias Analyses. American Journal of Epidemiology. 2015 Oct;182(7):568–578.

[24] McCarron P. Type 2 diabetes in grandparents and birth weight in offspring and grandchildren in the ALSPAC study. Journal of Epidemiology & Community Health. 2004 Jun;58(6):517–522.

[25] Naess O, Stoltenberg C, Hoff DA, Nystad W, Magnus P, Tverdal A, et al. Cardiovascular mortality in relation to birth weight of children and grandchildren in 500 000 Norwegian families. European Heart Journal. 2013 Nov;34(44):3427–3436.

[26] Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. International Journal of Epidemiology. 2013;42(5):1511–1519. Publisher: Oxford University Press.

[27] Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. International Journal of Obesity. 2008 Aug;32(S3):S8–S14.

[28] Damhuis SE, Ganzevoort W, Gordijn SJ. Abnormal Fetal Growth. Obstetrics and Gynecology Clinics of North America. 2021 Jun;48(2):267–279.

[29] Ramakrishnan U, Martorell R, Schroeder DG, Flores R. Role of Intergenerational Effects on Linear Growth. The Journal of Nutrition. 1999 Feb;129(2):544S–549S.

[30] Brown MM, Woolcott CG, Dodds L, Ashley-Martin J, Allen VM, Fahey J, et al. The 3G Multigenerational Cohort of Nova Scotian women and their mothers and offspring. Paediatric and Perinatal Epidemiology. 2020 Mar;34(2):214–221.

[31] Enders CK, Mistler SA, Keller BT. Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. Psychological Methods. 2016 Jun;21(2):222–240.

[32] Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. Statistical Methods in Medical Research. 2018 Jun;27(6):1634–1649.

[33] Audigier V, White IR, Jolani S, Debray TPA, Quartagno M, Carpenter J, et al. Multiple Imputation for Multilevel Data with Continuous and Binary Variables. Statistical Science. 2018 May;33(2).

[34] Burgette LF, Reiter JP. Multiple Imputation for Missing Data via Sequential Regression Trees. American Journal of Epidemiology. 2010 Nov;172(9):1070–1076.

[35] Stekhoven DJ, Buhlmann P. MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan;28(1):112–118.

[36] Doove LL, Van Buuren S, Dusseldorp E. Recursive partitioning for missing data imputation in the presence of interaction effects. Computational Statistics & Data Analysis. 2014 Apr;72:92–104.

[37] Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. American Journal of Epidemiology. 2014 Mar;179(6):764–774.

[38] Laqueur HS, Shev AB, Kagawa RMC. SuperMICE: An Ensemble Machine Learning Approach to Multiple Imputation by Chained Equations. American Journal of Epidemiology. 2022 Feb;191(3):516–525.

[39] Slade E, Naylor MG. A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. Statistics in Medicine. 2020 Apr;39(8):1156–1166.

[40] Hajjem A, Bellavance F, Larocque D. Mixed-effects regression trees for clustered data. Statistics & Probability Letters. 2011 Apr;81(4):451–459.

[41] Hajjem A, Bellavance F, Larocque D. Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation. 2014 Jun;84(6):1313–1328.

[42] Sela RJ, Simonoff JS. RE-EM trees: a data mining approach for longitudinal and clustered data. Machine Learning. 2012 Feb;86(2):169–207.

[43] World Health Organization. Body mass index; 2010. Available from: `https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi`.

[44] Institute of Medicine and National Research Council. Weight Gain During Pregnancy: Reexamining the Guidelines. Washington, D.C.: National Academies Press; 2009. Pages: 12584. Available from: `http://www.nap.edu/catalog/12584`.

[45] Lashen H, Fear K, Sturdee DW. Obesity is associated with increased risk of first trimester and recurrent miscarriage: matched case-control study. Human Reproduction (Oxford, England). 2004 Jul;19(7):1644–1646.

[46] Larsen TB, Sørensen HT, Gislum M, Johnsen SP. Maternal smoking, obesity, and risk of venous thromboembolism during pregnancy and the puerperium: a population-based nested case-control study. Thrombosis Research. 2007;120(4):505–509.

[47] Yogev Y, Catalano PM. Pregnancy and Obesity. Obstetrics and Gynecology Clinics of North America. 2009 Jun;36(2):285–300.

[48] Mayer C, Joseph KS. Fetal growth: a review of terms, concepts and issues relevant to obstetrics. Ultrasound in Obstetrics & Gynecology. 2013 Feb;41(2):136–145.

[49] Scifres CM. Short- and Long-Term Outcomes Associated with Large for Gestational Age Birth Weight. Obstetrics and Gynecology Clinics of North America. 2021 Jun;48(2):325–337.

[50] Kanmiki EW, Fatima Y, Mamun AA. Multigenerational transmission of obesity: A systematic review and meta-analysis. Obesity Reviews. 2022 Mar;23(3).

[51] Emanuel I, Filakti H, Alberman E, Evans SJ. Intergenerational studies of human birthweight from the 1958 birth cohort. 1. Evidence for a multigenerational effect. British Journal of Obstetrics and Gynaecology. 1992 Jan;99(1):67–74.

[52] Guillaume M, Lapidus L, Beckers F, Lambert A, Björntorp P. Familial trends of obesity through three generations: the Belgian-Luxembourg child study. International Journal of Obesity and Related Metabolic Disorders: Journal of the International Association for the Study of Obesity. 1995 Sep;19 Suppl 3:S5–9.

[53] Ohta H, Kuroda T, Onoe Y, Nakano C, Yoshikata R, Ishitani K, et al. Familial correlation of bone mineral density, birth data and lifestyle factors among adolescent daughters, mothers and grandmothers. Journal of Bone and Mineral Metabolism. 2010 Nov;28(6):690–695.

[54] van Buuren S. Flexible Imputation of Missing Data. Second edition ed. Chapman and Hall/CRC interdisciplinary statistics series. Boca Raton: CRC Press, Taylor & Francis Group; 2018.

[55] Agius R, Savona-Ventura C, Vassallo J. Transgenerational Metabolic Determinants of Fetal Birth Weight. Experimental and Clinical Endocrinology & Diabetes. 2013 May;121(07):431–435.

[56] Valeri L, Vanderweele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychological Methods. 2013 Jun;18(2):137–150.

[57] VanderWeele TJ. Explanation in Causal Inference: Methods for Mediation and Interaction. New York: Oxford University Press; 2015.

[58] Parlee SD, MacDougald OA. Maternal nutrition and risk of obesity in offspring: the Trojan horse of developmental plasticity. Biochimica Et Biophysica Acta. 2014 Mar;1842(3):495–506.

[59] Schulz LC. The Dutch Hunger Winter and the developmental origins of health and disease. Proceedings of the National Academy of Sciences of the United States of America. 2010 Sep;107(39):16757–16758.

[60] Lustig RH. Obesity Before Birth: Maternal and prenatal influences on the offspring. vol. 30. New York, NY: Springer-Verlag New York Inc.; 2010.

[61] Lopomo A, Burgio E, Migliore L. Epigenetics of Obesity. Progress in molecular biology and translational science. 2016;140:151–184. Publisher: Elsevier.

[62] Opsahl JO, Moen GH, Qvigstad E, Böttcher Y, Birkeland KI, Sommer C. Epigenetic signatures associated with maternal body mass index or gestational weight gain: a systematic review. Journal of Developmental Origins of Health and Disease. 2021 Jun;12(3):373–383.

[63] Bellver J, Mariani G. Impact of parental over- and underweight on the health of offspring. Fertility and Sterility. 2019 Jun;111(6):1054–1064.

[64] Hales CN, Barker DJ. Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. Diabetologia. 1992 Jul;35(7):595–601.

[65] Ojha S, Robinson L, Symonds ME, Budge H. Suboptimal maternal nutrition affects offspring health in adult life. Early Human Development. 2013 Nov;89(11):909–913.

[66] Fernandez-Twinn DS, Constância M, Ozanne SE. Intergenerational epigenetic inheritance in models of developmental programming of adult disease. Seminars in Cell & Developmental Biology. 2015 Jul;43:85–95.

[67] Ding Y, Li J, Liu S, Zhang L, Xiao H, Li J, et al. DNA hypomethylation of inflammation-associated genes in adipose tissue of female mice after multigenerational high fat diet feeding. International Journal of Obesity (2005). 2014 Feb;38(2):198–204.

[68] Chambers TJG, Morgan MD, Heger AH, Sharpe RM, Drake AJ. High-fat diet disrupts metabolism in two generations of rats in a parent-of-origin specific manner. Scientific Reports. 2016 Aug;6:31857.

[69] Green LR, Hester RL, editors. Parental Obesity: Intergenerational Programming and Consequences. New York, NY: Springer New York; 2016.

[70] Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 1987.

[71] Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC Medical Research Methodology. 2012 Dec;12(1):46.

[72] White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011 Feb;30(4):377–399.

[73] Rubin DB. Multiple Imputation after 18+ Years. Journal of the American Statistical Association. 1996 Jun;91(434):473–489.

[74] Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: an overview and case study. Emerging Themes in Epidemiology. 2017 Dec;14(1):8.

[75] Murray JS. Multiple imputation: a review of practical and theoretical findings. Statistical Science. 2018;33(2):142–159. Publisher: Institute of Mathematical Statistics.

[76] Kleinke K. Multiple Imputation Under Violated Distributional Assumptions: A Systematic Evaluation of the Assumed Robustness of Predictive Mean Matching. Journal of Educational and Behavioral Statistics. 2017 Aug;42(4):371–404.

[77] Lee KJ, Carlin JB. Multiple imputation in the presence of non-normal data: Multiple imputation in the presence of non-normal data. Statistics in Medicine. 2017 Feb;36(4):606–617.

[78] Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. BMC medical research methodology. 2014;14(1):1–13. Publisher: Springer.

[79] Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. American Journal of Epidemiology. 2010 Mar;171(5):624–632.

[80] von Hippel PT. How to Impute Interactions, Squares, and other Transformed Variables. Sociological Methodology. 2009 Aug;39(1):265–291.

[81] Morris TP, White IR, Royston P, Seaman SR, Wood AM. Multiple imputation for an incomplete covariate that is a ratio. Statistics in medicine. 2014;33(1):88–104. Publisher: Wiley Online Library.

[82] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3):1–67.

[83] Wagstaff DA, Kranz S, Harel O. A preliminary study of active compared with passive imputation of missing body mass index values among non-Hispanic white youths. The American Journal of Clinical Nutrition. 2009 Apr;89(4):1025–1030.

[84] Schafer JL. Analysis of Incomplete Multivariate Data. Boca Raton: Chapman & Hall/CRC; 1997. OCLC: 123486463.

[85] van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research. 2007 Jun;16(3):219–242.

[86] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification And Regression Trees. 1st ed. Wadsworth, NY: Chapman & Hall; 1984.

[87] Breiman L. Bagging predictors. Machine Learning. 1996 Aug;24(2):123–140.

[88] Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32.

[89] Tilling K, Williamson EJ, Spratt M, Sterne JAC, Carpenter JR. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. Journal of Clinical Epidemiology. 2016 Dec;80:107–115.

[90] Speidel M, Drechsler J, Sakshaug JW. Biases in multilevel analyses caused by cluster-specific fixed-effects imputation. Behavior Research Methods. 2018 Oct;50(5):1824–1840.

[91] Robitzsch A, Grund S. miceadds: Some Additional Multiple Imputation Functions, Especially for 'mice'; 2021.

[92] Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology. 1986 Dec;51(6):1173–1182.

[93] Pearl J. Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann; 2001. p. 411–420.

[94] Robins JM, Greenland S. Identifiability and exchangeability for direct guidnd indirect effects. Epidemiology (Cambridge, Mass). 1992 Mar;3(2):143–155.

[95] Pearl J. The Causal Mediation Formula—A Guide to the Assessment of Pathways and Mechanisms. Prevention Science. 2012 Aug;13(4):426–436.

[96] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology (Cambridge, Mass). 1999 Jan;10(1):37–48.

[97] Miettinen O. Confounding and effect-modification. American Journal of Epidemiology. 1974 Nov;100(5):350–353.

[98] VanderWeele TJ, Shpitser I. On the definition of a confounder. Annals of Statistics. 2013 Feb;41(1):196–220.

[99] Rothman KJ, Lash TL, VanderWeele TJ, Haneuse S. Modern epidemiology. Fourth edition ed. Philadelphia: Wolters Kluwer; 2021.

[100] Gaynor SM, Schwartz J, Lin X. Mediation analysis for common binary outcomes. Statistics in Medicine. 2019 Feb;38(4):512–529.

[101] Samoilenko M, Blais L, Lefebvre G. Comparing logistic and log-binomial models for causal mediation analyses of binary mediators and rare binary outcomes: evidence to support cross-checking of mediation results in practice. Observational Studies. 2018;4(1):193–216.

[102] Doretti M, Raggi M, Stanghellini E. Exact parametric causal mediation analysis for a binary outcome with a binary mediator. Statistical Methods & Applications. 2022 Mar;31(1):87–108. ArXiv:1811.00439 [stat].

[103] Samoilenko M, Lefebvre G. Parametric-Regression–Based Causal Mediation Analysis of Binary Outcomes and Binary Mediators: Moving Beyond the Rareness or Commonness of the Outcome. American Journal of Epidemiology. 2021 Sep;190(9):1846–1858.

[104] Pearl J. Causality: models, reasoning, and inference. 2nd ed. Cambridge, U.K.; New York: Cambridge University Press; 2009.

[105] Daniel RM, De Stavola BL, Cousens SN, Vansteelandt S. Causal mediation analysis with multiple mediators. Biometrics. 2015 Mar;71(1):1–14.

[106] Avin C, Shpitser I, Pearl J. Identifiability of path-specific effects. In: In Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05. Morgan-Kaufmann Publishers; 2005. p. 357–363.

[107] Albert JM, Nelson S. Generalized Causal Mediation Analysis. Biometrics. 2011 Sep;67(3):1028–1038.

[108] Robins JM, Greenland S, Hu FC. Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome. Journal of the American Statistical Association. 1999 Sep;94(447):687–700.

[109] Pearl J. Causal inference in statistics: An overview. Statistics Surveys. 2009 Jan;3:96–146.

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

[111] Daniel RM, De Stavola BL, Cousens SN. Gformula: Estimating Causal Effects in the Presence of Time-Varying Confounding or Mediation using the G-Computation Formula. The Stata Journal: Promoting communications on statistics and Stata. 2011 Dec;11(4):479–517.

[112] Rizzo RRN, Cashin AG, Bagg MK, Gustin SM, Lee H, McAuley JH. A Systematic Review of the Reporting Quality of Observational Studies That Use Mediation Analyses. Prevention Science. 2022 Feb.

[113] Dietz PM, Rizzo JH, England LJ, Callaghan WM, Vesco KK, Bruce FC, et al. Health Care Utilization in the First Year of Life among Small- and Large-for-Gestational Age Term Infants. Maternal and Child Health Journal. 2013 Aug;17(6):1016–1024.

[114] Meertens L, Smits L, Kuijk S, Aardenburg R, Dooren I, Langenveld J, et al. External validation and clinical usefulness of first-trimester prediction models for small- and large-for-gestational-age infants: a prospective cohort study. BJOG: An International Journal of Obstetrics & Gynaecology. 2019 Mar;126(4):472–484.

[115] Boucoiran I, Djemli A, Taillefer C, Rypens F, Delvin E, Audibert F. First-Trimester Prediction of Birth Weight. American Journal of Perinatology. 2013 Jan;30(08):665–672.

[116] Crovetto F, Triunfo S, Crispi F, Rodriguez-Sureda V, Dominguez C, Figueras F, et al. Differential performance of first-trimester screening in predicting small-for-gestational-age neonate or fetal growth restriction. Ultrasound in Obstetrics & Gynecology. 2017 Mar;49(3):349–356.

[117] Frick AP, Syngelaki A, Zheng M, Poon LC, Nicolaides KH. Prediction of large-for-gestational-age neonates: screening by maternal factors and biomarkers in the three trimesters of pregnancy: Screening for LGA. Ultrasound in Obstetrics & Gynecology. 2016 Mar;47(3):332–339.

[118] González-González NL, Plasencia W, González Dávila E, Padrón E, di Renzo GC, Bartha JL. First and second trimester screening for large for gestational age infants. The Journal of Maternal-Fetal & Neonatal Medicine. 2013 Nov;26(16):1635–1640.

[119] González-González NL, González-Dávila E, González Marrero L, Padrón E, Castro-Conde JR, Plasencia W. Value of placental volume and vascular flow indices as predictors of intrauterine growth retardation. European Journal of Obstetrics & Gynecology and Reproductive Biology. 2017 May;212:13–19.

[120] Leal AM, Poon LCY, Frisova V, Veduta A, Nicolaides KH. First-trimester maternal serum tumor necrosis factor receptor-1 and pre-eclampsia. Ultrasound in Obstetrics and Gynecology. 2009 Feb;33(2):135–141.

[121] McCowan LME, Thompson JMD, Taylor RS, Baker PN, North RA, Poston L, et al. Prediction of Small for Gestational Age Infants in Healthy Nulliparous Women Using Clinical and Ultrasound Risk Factors Combined with Early Pregnancy Biomarkers. PLOS ONE. 2017 Jan;12(1):e0169311.

[122] Monari F, Menichini D, Spano' Bascio L, Grandi G, Banchelli F, Neri I, et al. A first trimester prediction model for large for gestational age infants: a preliminary study. BMC Pregnancy and Childbirth. 2021 Sep;21(1):654.

[123] Onwudiwe N, Yu CKH, Poon LCY, Spiliopoulos I, Nicolaides KH. Prediction of pre-eclampsia by a combination of maternal history, uterine artery Doppler and mean arterial pressure. Ultrasound in Obstetrics and Gynecology. 2008 Dec;32(7):877–883.

[124] Papastefanou I, Souka AP, Pilalis A, Eleftheriades M, Michalitsi V, Kassanos D. First trimester prediction of small- and large-for-gestation neonates by an integrated model incorporating ultrasound parameters, biochemical indices and maternal characteristics: First trimester prediction of SGA and LGA newborns. Acta Obstetricia et Gynecologica Scandinavica. 2012 Jan;91(1):104–111.

[125] Plasencia W, Akolekar R, Dagklis T, Veduta A, Nicolaides KH. Placental Volume at 11–13 Weeks' Gestation in the Prediction of Birth Weight Percentile. Fetal Diagnosis and Therapy. 2011;30(1):23–28.

[126] Poon LCY, Karagiannis G, Stratieva V, Syngelaki A, Nicolaides KH. First-Trimester Prediction of Macrosomia. Fetal Diagnosis and Therapy. 2011;29(2):139–147.

[127] Poon LCY, Karagiannis G, Staboulidou I, Shafiei A, Nicolaides KH. Reference range of birth weight with gestation and first-trimester prediction of small-for-gestation neonates. Prenatal Diagnosis. 2011 Jan;31(1):58–65.

[128] Plasencia W, González Dávila E, Tetilla V, Padrón Pérez E, García Hernández JA, González González NL. First-trimester screening for large-for-gestational-age infants. Ultrasound in Obstetrics & Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2012 Apr;39(4):389–395.

[129] Schneuer FJ, Roberts CL, Ashton AW, Guilbert C, Tasevski V, Morris JM, et al. Angiopoietin 1 and 2 serum concentrations in first trimester of pregnancy as biomarkers of adverse pregnancy outcomes. American Journal of Obstetrics and Gynecology. 2014 Apr;210(4):345.e1–345.e9.

[130] Schwartz N, Pessel C, Coletta J, Krieger AM, Timor-Tritsch IE. Early Biometric Lag in the Prediction of Small for Gestational Age Neonates and Preeclampsia. Journal of Ultrasound in Medicine. 2011 Jan;30(1):55–60.

[131] Vieira MC, McCowan LME, Gillett A, Poston L, Fyfe E, Dekker GA, et al. Clinical, ultrasound and molecular biomarkers for early prediction of large for gestational age infants in nulliparous women: An international prospective cohort study. PLOS ONE. 2017 Jun;12(6):e0178484.

[132] van der Laan MJ, Polley EC, Hubbard AE. Super Learner. Statistical Applications in Genetics and Molecular Biology. 2007 Jan;6(1).

[133] Wolpert DH. Stacked generalization. Neural Networks. 1992 Jan;5(2):241–259.

[134] Breiman L. Stacked regressions. Machine Learning. 1996 Jul;24(1):49–64.

[135] van der Laan MJ, Rose S. Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies. 1st ed. Springer Series in Statistics. Cham: Springer International Publishing: Imprint: Springer; 2018.

[136] Zhang M, Differding MK, Benjamin-Neelon SE, Østbye T, Hoyo C, Mueller NT. Association of prenatal antibiotics with measures of infant adiposity and the gut microbiome. Annals of Clinical Microbiology and Antimicrobials. 2019;18(1):18. Publisher: Springer.

[137] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics. 2002 Oct;35(5-6):352–359.

[138] Sakr S, Elshawi R, Ahmed AM, Qureshi WT, Brawner CA, Keteyian SJ, et al. Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project. BMC Medical Informatics and Decision Making. 2017 Dec;17(1):174.

[139] Kuhle S, Maguire B, Zhang H, Hamilton D, Allen AC, Joseph KS, et al. Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study. BMC Pregnancy and Childbirth. 2018 Dec;18(1):333.

[140] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE. 2017;12(4):e0174944.

[141] Feng JZ, Wang Y, Peng J, Sun MW, Zeng J, Jiang H. Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. Journal of Critical Care. 2019 Dec;54:110–116.

[142] Sufriyana H, Husnayain A, Chen YL, Kuo CY, Singh O, Yeh TY, et al. Comparison of Multivariable Logistic Regression and Other Machine Learning Algorithms for Prognostic Prediction Studies in Pregnancy Care: Systematic Review and Meta-Analysis. JMIR Medical Informatics. 2020 Nov;8(11):e16503.

[143] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. 2019 Jun;110:12–22.

[144] van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Medical Research Methodology. 2014 Dec;14:137.

[145] Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Medicine. 2015;13(1):1.

[146] Zou H, Hastie T. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society, Series B. 2005;67:301–320.

[147] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM; 2016. p. 785–794.

[148] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995 Sep;20(3):273–297.

[149] Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996 Jan;58(1):267–288.

[150] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology (Cambridge, Mass). 2010 Jan;21(1):128–138.

[151] Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 2015 Mar;10(3):e0118432.

[152] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning - ICML '06. Pittsburgh, Pennsylvania: ACM Press; 2006. p. 233–240.

[153] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009 Sep;338:b2393.

[154] Meng XL. Multiple-Imputation Inferences with Uncongenial Sources of Input. Statistical Science. 1994 Nov;9(4).

[155] Maxwell C, Gaudet L, Cassir G, Nowik C, McLeod NL, Jacob C et al. Guideline No. 391-Pregnancy and Maternal Obesity Part 1: Pre-conception and Prenatal Care. Journal of Obstetrics and Gynaecology Canada. 2019 Nov;41(11):1623–1640.

[156] Kramer MS, Platt RW, Wen SW, Joseph KS, Allen A, Abrahamowicz M, et al. A New and Improved Population-Based Canadian Reference for Birth Weight for Gestational Age. Pediatrics. 2001 Aug;108(2):e35–e35.

[157] Reproductive Care Program of Nova Scotia. The Nova Scotia Atlee Perinatal Database; 2022. Available from: `http://rcp.nshealth.ca/atlee-database`.

[158] Austin PC, White IR, Lee DS, van Buuren S. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. Canadian Journal of Cardiology. 2021 Sep;37(9):1322–1331.

[159] Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.

[160] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: `https://www.R-project.org/`.

[161] RStudio Team. RStudio: Integrated Development Environment for R. Boston, MA: RStudio, PBC.; 2020. Available from: `http://www.rstudio.com/`.

[162] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in medicine. 2006;25(24):4279–4292. Publisher: Wiley Online Library.

[163] Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. International Journal of Epidemiology. 2019 Aug;48(4):1294–1304.

[164] Drake AJ, Reynolds RM. Impact of maternal obesity on offspring obesity and cardiometabolic disease risk. Reproduction. 2010 Sep;140(3):387–398.

[165] Reynolds RM, Osmond C, Phillips DIW, Godfrey KM. Maternal BMI, Parity, and Pregnancy Weight Gain: Influences on Offspring Adiposity in Young Adulthood. The Journal of Clinical Endocrinology & Metabolism. 2010 Dec;95(12):5365–5369.

[166] Schoppa I, Lyass A, Heard-Costa N, de Ferranti SD, Fox C, Gillman MW, et al. Association of Maternal Prepregnancy Weight with Offspring Adiposity Throughout Adulthood over 37 Years of Follow-up. Obesity. 2019 Jan;27(1):137–144.

[167] Rath S, Marsh JA, Newnham JP, Zhu K, Atkinson HC, Mountain J, et al. Parental pre-pregnancy BMI is a dominant early-life risk factor influencing BMI of offspring in adulthood. Obesity Science & Practice. 2016;2(1):48–57. Publisher: Wiley Online Library.

[168] Chaparro MP, Koupil I, Byberg L. Maternal pre-pregnancy BMI and offspring body composition in young adulthood: the modifying role of offspring sex and birth order. Public Health Nutrition. 2017 Dec;20(17):3084–3089.

[169] Eshriqui I, Valente AMM, Folchetti LD, de Almeida-Pititto B, Ferreira SRG. Pre-pregnancy BMI is associated with offspring body composition in adulthood before adiposity-related disorders: a retrospective cohort. Public Health Nutrition. 2021 Apr;24(6):1296–1303.

[170] Dabelea D, Crume T. Maternal environment and the transgenerational cycle of obesity and diabetes. Diabetes. 2011 Jul;60(7):1849–1855.

[171] Catalano PM. Obesity and Pregnancy—The Propagation of a Viscous Cycle? The Journal of Clinical Endocrinology & Metabolism. 2003 Aug;88(8):3505–3506.

[172] Lee H, Cashin AG, Lamb SE, Hopewell S, Vansteelandt S, VanderWeele TJ, et al. A Guideline for Reporting Mediation Analyses of Randomized Trials and Observational Studies: The AGReMA Statement. JAMA. 2021 Sep;326(11):1045.

[173] Reproductive Care Program of Nova Scotia. Guideline for Assessment of the "Best Estimate" of Gestational Age; 2020. Available from: http://rcp.nshealth.ca/clinical-practice-guidelines/best-estimate-gestational-age-202006.

[174] World Health Organization. Obesity: preventing and managing the global epidemic: report of a WHO consultation. No. 894 in WHO technical report series. Geneva: World Health Organization; 2000.

[175] Wilkins R, Peters PA. PCCF + Version 5K* User's Guide. Automated Geographic Coding Based on the Statistics Canada Postal Code Conversion Files, Including Postal Codes through May 2011. Health Analysis Division, Statistics Canada, Ottawa; 2012. Available from: https://mdl.library.utoronto.ca/sites/default/public/mdldata/open/canada/national/statcan/postalcodes/pccfplus/2006/2011may/MSWORD.PCCF5K.pdf.

[176] Magee LA, Pels A, Helewa M, Rey E, von Dadelszen P, Magee LA, et al. Diagnosis, Evaluation, and Management of the Hypertensive Disorders of Pregnancy: Executive Summary. Journal of Obstetrics and Gynaecology Canada. 2014 May;36(5):416–438.

[177] Feig DS, Berger H, Donovan L, Godbout A, Kader T, Keely E, et al. Diabetes and Pregnancy. Canadian Journal of Diabetes. 2018 Apr;42:S255–S282.

[178] Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research. 2011 May;46(3):399–424.

[179] Keil AP, Edwards JK, Richardson DB, Naimi AI, Cole SR. The Parametric g-Formula for Time-to-event Data: Intuition and a Worked Example. Epidemiology. 2014 Nov;25(6):889–897.

[180] Rudolph JE, Cartus A, Bodnar LM, Schisterman EF, Naimi AI. The role of the natural course in causal analysis. American Journal of Epidemiology. 2021 Oct:kwab248.

[181] O'Rourke HP, MacKinnon DP. Reasons for Testing Mediation in the Absence of an Intervention Effect: A Research Imperative in Prevention and Intervention Research. Journal of Studies on Alcohol and Drugs. 2018 Mar;79(2):171–181.

[182] Lacal I, Ventura R. Epigenetic Inheritance: Concepts, Mechanisms and Perspectives. Frontiers in Molecular Neuroscience. 2018 Sep;11:292.

[183] Murrin CM, Kelly GE, Tremblay RE, Kelleher CC. Body mass index and height over three generations: evidence from the Lifeways cross-generational cohort study. BMC Public Health. 2012 Dec;12(1):81.

[184] Wu LL, Russell DL, Wong SL, Chen M, Tsai TS, St John JC, et al. Mitochondrial dysfunction in oocytes of obese mothers: transmission to offspring and reversal by pharmacological endoplasmic reticulum stress inhibitors. Development (Cambridge, England). 2015 Feb;142(4):681–691.

[185] Saben JL, Boudoures AL, Asghar Z, Thompson A, Drury A, Zhang W, et al. Maternal Metabolic Syndrome Programs Mitochondrial Dysfunction via Germline Changes across Three Generations. Cell Reports. 2016 Jun;16(1):1–8.

[186] Hernán MA, Robins JM. Causal Inference: What if. Boca Raton: Chapman & Hall/CRC;.

[187] Westreich D, Cole SR. Invited commentary: positivity in practice. American Journal of Epidemiology. 2010 March;171(6):674–677; discussion 678–681.

[188] Moore KL, Neugebauer R, van der Laan MJ, Tager IB. Causal inference in epidemiological studies with strong confounding. Statistics in Medicine. 2012 Jun;31(13):1380–1404.

[189] Léger M, Chatton A, Le Borgne F, Pirracchio R, Lasocki S, Foucher Y. Causal inference in case of near-violation of positivity: comparison of methods. Biometrical Journal. 2022 Jan.

[190] Hernán MA, VanderWeele TJ. Compound treatments and transportability of causal inference. Epidemiology (Cambridge, Mass). 2011 May;22(3):368–377.

[191] VanderWeele TJ, Hernan MA. Causal inference under multiple versions of treatment. Journal of Causal Inference. 2013 May;1(1):1–20.

[192] Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinical Epidemiology. 2019 Jun;110:63–73.

[193] Fung C, Zinkhan E. Short- and Long-Term Implications of Small for Gestational Age. Obstetrics and Gynecology Clinics of North America. 2021 Jun;48(2):311–323.

[194] Colella M, Frérot A, Novais ARB, Baud O. Neonatal and Long-Term Consequences of Fetal Growth Restriction. Current Pediatric Reviews. 2018 Dec;14(4):212–218.

[195] Rasmussen KM, Abrams B, Bodnar LM, Butte NF, Catalano PM, Maria Siega-Riz A. Recommendations for Weight Gain During Pregnancy in the Context of the Obesity Epidemic. Obstetrics & Gynecology. 2010 Nov;116(5):1191–1195.

[196] Heymans M. psfmi: Prediction Model Pooling, Selection and Performance Evaluation Across Multiply Imputed Datasets; 2021.

[197] Keilwagen J, Grosse I, Grau J. Area under Precision-Recall Curves for Weighted and Unweighted Data. PLOS ONE. 2014;9(3).

[198] Grau J, Grosse I, Keilwagen J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Bioinformatics. 2015;31(15):2595–2597.

[199] Polley E, LeDell E, Kennedy C, Laan Mvd. SuperLearner: Super Learner Prediction; 2021.

[200] Ram K, Wickham H. wesanderson: A Wes Anderson Palette Generator; 2018.

[201] Kennedy C. ck37r: Chris Kennedy's R Toolkit; 2022.

[202] Kennedy C. Source code from "Development of an ensemble machine learning prognostic model for predicting 60-day risk of major adverse cardiac events in adults with chest pain"; 2021. Available from: `https://github.com/ck37/chest-pain-risk-prediction`.

[203] McCowan LME, Thompson JMD, Taylor RS, North RA, Poston L, Baker PN, et al. Clinical Prediction in Early Pregnancy of Infants Small for Gestational Age by Customised Birthweight Centiles: Findings from a Healthy Nulliparous Cohort. PLOS ONE. 2013 Aug;8(8):e70917.

[204] Macdonald-Wallis C, Silverwood RJ, de Stavola BL, Inskip H, Cooper C, Godfrey KM, et al. Antenatal blood pressure for prediction of pre-eclampsia, preterm birth, and small for gestational age babies: development and validation in two general population cohorts. BMJ. 2015 Nov;351(nov17 5):h5948–h5948.

[205] Erkamp JS, Voerman E, Steegers EAP, Mulders AGMGJ, Reiss IKM, Duijts L, et al. Second and third trimester fetal ultrasound population screening for risks of preterm birth and small-size and large-size for gestational age at birth: a population-based prospective cohort study: Fetal ultrasound screening for common adverse birth outcomes. BMC Medicine. 2020 Dec;18(1):63.

[206] Klebanoff MA, Graubard BI, Kessel SS, Berendes HW. Low birth weight across generations. JAMA. 1984 Nov;252(17):2423–2427.

[207] Shah PS, Shah V. Influence of the maternal birth status on offspring: A systematic review and meta-analysis. Acta Obstetricia et Gynecologica Scandinavica. 2009 Dec;88(12):1307–1318.

[208] Ahlsson F, Gustafsson J, Tuvemo T, Lundgren M. Females born large for gestational age have a doubled risk of giving birth to large for gestational age infants. Acta Paediatrica. 2007 Mar;96(3):358–362.

[209] Bertini A, Salas R, Chabert S, Sobrevia L, Pardo F. Using Machine Learning to Predict Complications in Pregnancy: A Systematic Review. Frontiers in Bioengineering and Biotechnology. 2022 Jan;9:780389.

[210] Perichart-Perera O, Avila-Sosa V, Solis-Paredes JM, Montoya-Estrada A, Reyes-Muñoz E, Rodríguez-Cano AM, et al. Vitamin D Deficiency, Excessive Gestational Weight Gain, and Oxidative Stress Predict Small for Gestational Age Newborns Using an Artificial Neural Network Model. Antioxidants. 2022 Mar;11(3):574.

[211] Public Health Agency of Canada. Obesity in Canada – Health and economic implications [research]; 2011. Last Modified: 2011-06-23. Available from: `https://www.canada.ca/en/public-health/services/health-promotion/healthy-living/obesity-canada/health-economic-implications.html`.

[212] McGee G, Perkins NJ, Mumford SL, Kioumourtzoglou MA, Weisskopf MG, Schildcrout JS, et al. Methodological Issues in Population-Based Studies of Multigenerational Associations. American Journal of Epidemiology. 2020 Dec;189(12):1600–1609.

[213] Shazly SA, Trabuco EC, Ngufor CG, Famuyide AO. Introduction to Machine Learning in Obstetrics and Gynecology. Obstetrics & Gynecology. 2022 Mar;Publish Ahead of Print.

[214] Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. International Journal of Epidemiology. 2021 Jan;49(6):2058–2064.

[215] Balzer LB, Petersen ML. Invited Commentary: Machine Learning in Causal Inference—How Do I Love Thee? Let Me Count the Ways. American Journal of Epidemiology. 2021 Aug;190(8):1483–1487.

[216] Zivich PN, Breskin A. Machine Learning for Causal Inference: On the Use of Cross-fit Estimators. Epidemiology. 2021 May;32(3):393–401.

[217] Bang H, Robins JM. Doubly Robust Estimation in Missing Data and Causal Inference Models. Biometrics. 2005 Dec;61(4):962–973.

[218] Schuler MS, Rose S. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. American Journal of Epidemiology. 2017 Jan;185(1):65–73.

[219] Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. American Journal of Epidemiology. 2011 Apr;173(7):761–767.

[220] Kennedy EH. Semiparametric theory and empirical processes in causal inference. arXiv; 2016. ArXiv:1510.04740 [math, stat].

[221] Tchetgen Tchetgen EJ, Shpitser I. Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. The Annals of Statistics. 2012 Jun;40(3).

[222] Zheng W, van der Laan MJ. Targeted Maximum Likelihood Estimation of Natural Direct Effects. The International Journal of Biostatistics. 2012 Jan;8(1):1–40.

[223] Hejazi NS, Rudolph KE, Van Der Laan MJ, Díaz I. Nonparametric causal mediation analysis for stochastic interventional (in)direct effects. Biostatistics. 2022 Feb:kxac002.

[224] HPC4Health; 2022. Available from: `http://www.hpcforhealth.ca/`.

Appendix A

Supplemental tables

Table A.1: Summary of studies assessing the intergenerational transmission of body-related weight measures

Reference/Study	Exposure measure (G2 weight)	Outcome measure (G0 weight)	Mediator (G1 weight)	Sample size	Adjustment variables	Main findings
Tambs et al., 1991 Individuals over 19 years living in Nord-Trondelag, Norway from 1984-1986.	BMI (mean age not reported). Height and weight measured by investigators.	BMI. No indication of which grandparent(s) was measured. Height and weight measured by investigators.	N/A	1251 G2 and G0 pairs.	G2 age, sex; G0 age, sex.	Correlation of 0.067 between G0 BMI and G2 BMI.
Emanuel et al., 1992 National Child Development Study (NCDS).	Birthweight (grams) Reported by the mother during follow-up interviews.	MGM non-pregnant weight (stones) and coded into class intervals of stones; the midpoint of each interval was converted into kg. Collected by interview and forms from the midwives at the time of G1 birth.	N/A	880 MGM and G2 (most recent birth with complete information) pairs.	MGM height, smoking; MGF social class; mother's birthweight, gestational age, adult height, non-pregnancy weight, age, smoking status; father's social class; G2 sex, birth order.	MGM weight removed in backward elimination. No effect size reported.

Table A.1: continued

Reference/Study	Exposure measure (G2 weight)	Outcome measure (G0 weight)	Mediator (G1 weight)	Sample size	Adjustment variables	Main findings
Guillaume et al., 1995 Children from randomly selected schools in Luxembourg, Belgium.	BMI at age 6-12 years (baseline). Height and weight measured by investigators.	OB problems (yes/no), reported at baseline (mean age not reported). No indication of which grandparent(s) was measured. Problem with OB in MGP and PGP obtained via questionnaires filled out by the student's parents.	N/A	1026 G2s (number of G0s unclear).	G2 age; G0 age.	Significant association between G0 OB status and G2 BMI in girls (p=0.03), but not in boys (p>0.20). Effect size not reported.
Polley et al., 2005 Three-generation families recruited from 10 sites in rural Oklahoma.	BMI percentile at baseline (mean age not reported). Height and weight measured in-home.	BMI, measured at baseline (mean age not reported). No indication of which grandparent(s) was measured. Height and weight measured in-home.	N/A	87 G0 and G2 pairs.	None.	Correlation of 0.22 (p=0.042) between G0 BMI and G2 BMI z-score.

Table A.1: continued

Reference/Study	Exposure measure (G2 weight)	Outcome measure (G0 weight)	Mediator (G1 weight)	Sample size	Adjustment variables	Main findings
Jouret et al., 2007 Kindergarten students from 79 schools in Haute-Garonne, France.	BMI percentile at age 4 years (baseline) based on age- and sex-specific growth curves: ≥90th (OW). Height and weight collected by study physician.	OB status (yes/no), reported at baseline (mean age not reported). Existence of OB in MGP and PGP obtained via questionnaires filled out by the student's parents.	N/A	1242 G2s (number of G0s unclear).	G2 sex; maternal OW; paternal OW; GP diabetes.	Odds of OW in G2s with ≥2 G0s with OB higher than that of those with no G0s with OB (uOR 2.96, 95% CI 1.63-5.38). No association between child OW and having one G0 with OB (uOR 0.89, 95% CI 0.48-1.63). No association in adjusted analyses.
Davis et al., 2008 Panel Study of Income Dynamics (PSID) and children of the families in the PSID, the Child Development Supplement (CDS).	BMI percentile at age 5-19 years (baseline) based on age and sex-specific growth curves: ≥ 95th (OW); 85th-95th (at risk for OW); <5th or UW; 5th-85th (NW). Height and weight measured in-home.	BMI, reported at baseline: ≥ 30 (OB); 25-29.9 (OW); 18.5-24.9 (NW); <18.5 (UW). G0 height and weight information collected by self-report. BMI based on MGM's BMI, or if missing, BMI of MGF, PGM, or PGF.	N/A	1573 G0 and G2 pairs.	None.	Compared to children with both NW parent and GP, the prevalence of child OW was 17.4 ($p<0.001$) for those with NW parent and OB GP, and was 29.2 ($p<0.001$) if both parent and GP were OB.

Table A.1: continued

Reference/Study	Exposure measure (G2 weight)	Outcome measure (G0 weight)	Mediator (G1 weight)	Sample size	Adjustment variables	Main findings
Ohta et al., 2010 Female students from junior and senior high schools in Tokyo, Japan.	Birthweight and age-specific body weight z-score at baseline (mean age 15 years). Birthweight assesssed by interview. Weight measured by investigators.	MGM birthweight and age-specific body weight z-score at baseline (mean age 72 years). Birthweight assesssed by interview. Weight measured by investigators.	N/A	34 G0 (MGM) and G2 (female) pairs.	None.	Correlation of -0.41 (p=0.368) between MGM birthweight and G2 birthweight, and correlation of 0.30 (p=0.095) between MGM and G2 body weight z-score.
De Stavola et al., 2011 Uppsala Birth Cohort Multi-generational Study.	Standardized z-scores for weight and length at birth based on sex- and week-of-gestation-specific means and SDs. Size at birth and length of gestation obtained from the Medical Birth Registry.	MGM standardized z-scores for weight and length at birth based on sex- and week-of-gestation-specific means and SDs. Size at birth and length of gestation obtained from archived hospital obstetric records.	N/A	6169 G0 (MGM) and G2 pairs.	Baseline models adjusted for G0 and G2 year of birth, and motherÕs parity (nulliparous vs. parous). Mediators and modifiers of the size-at-birth correlations including maternal socioeconomic, demographic, and behavioural factors were assessed separately and jointly.	Baseline models reported a correlation of 0.12 (95% CI [0.10, 0.15]) between MGM birthweight and G2 birthweight. No indication of mediation or modification jointly or separately.

Table A.1: continued

Reference/Study	Exposure measure (G2 weight)	Outcome measure (G0 weight)	Mediator (G1 weight)	Sample size	Adjustment variables	Main findings
Saker et al., 2011 Students from 22 elementary schools in Tlemcen, Algeria.	BMI percentile at age 6-8 years (baseline) based on age and sex-specific growth curves: ≥95th (OB); <85th (NW). Height and weight measured by school physicians.	MGM BMI, reported at baseline (mean age not reported): ≥30 (OB). Height and weight collected via questionnaire filled out by student's mother.	N/A	1395 G0 (MGM) and G2 pairs.	None.	Significant association between MGM OB status and G2 OB status (OR 26.1, 95% CI [16.0, 43.0]).
Murrin et al., 2012 Lifeways Cross Generational Cohort Study.	BMI at age 5 years (baseline). Height and weight measured in-home.	BMI, reported at baseline, of at least one randomly selected G0. Only MGM results considered here. Height and weight information collected by self-report.	N/A	147 G0 (MGM) and G2 pairs.	Gender, group (child, mother, MGM, MGF), age, self-rated health, education, medical card holder, fruit & vegetables and fish consumption, physical activity level, and interactions of these variables. Stepwise procedures were used to determine best model.	Correlation of 0.23 (95% CI [0.06, 0.39]) between MGM BMI and G2 BMI adjusted for group and self-rated health.

Table A.1: continued

Reference/Study	Exposure measure (G2 weight)	Outcome measure (G0 weight)	Mediator (G1 weight)	Sample size	Adjustment variables	Main findings
Agius et al., 2013 Women born in 1987 and delivering at the same hospital from 2004 to 2010 in Malta.	Birthweight (kg). Birthweight collected from cllinical notes.	MGM pre-pregnancy BMI (time of G1 birth): >25 (OW/OB); <25 (NW). Height and weight collected from clinical notes.	N/A	182 G0s (MGM) and 233 G2s.	G1 pre-pregnancy BMI.	G2s born to MGMs with pre-pregnancy OW/OB had larger birthweights than those born to NW MGMs (3.4 kg. vs. 3.1 kg, p<0.001). No significant difference in G2 birthweight after adjusting for G1 pre-pregnancy BMI (p=0.61).
Lee et al., 2013 Korea National Heaalth and Nutrition Examination Survey (KNHANES).	BMI percentile at age 2-18 years (baseline) based on age and sex-specific growth curves: ≥95th (OB); 85-95th (OW); <85th (NW). Height and weight collected via the Health Examination Survey.	BMI, reported at baseline (mean age not reported): ≥30 (OB); 25-29.9 (OW); <24.9 (NW). No indication of which grandparent(s) was measured. Height and weight collected via the Health Examination Survey.	N/A	419 G0 and G2 pairs.	G2 gender, age, region of residence, highest education of adult in household, household income.	Prevalence of child OW/OB among those living with GP(s) only was higher if GP(s) was also OW/OB compared to if GP(s) was NW (31% vs. 16%, p<0.01). Among children living with GP(s) only, odds of OW/OB higher if GP(s) also OW/OB, compared with children living with NW GP(s) (OR=2.1, 95% CI [1.06, 4.05]).

Table A.1: continued

Reference/Study	Exposure measure (G2 weight)	Outcome measure (G0 weight)	Mediator (G1 weight)	Sample size	Adjustment variables	Main findings
Shin et al., 2013 Female students from three national universities in Korea.	BMI at age 21 years (baseline): <18.5 (UW); 18.5-22.9 (NW); ≥23.0 (OW). Height and weight collected by self report.	MGM BMI, reported at baseline and when MGM was age 20: ≥23.0 (OW); 18.5-22.9 (NW). Height and weight collected by self-report.	N/A	67 G0 (MGM) and G2 (female) pairs.	G2 age, energy intake, physical activity.	MGM OW at age 20 or when G2 was age 21 years was not significantly associated with G2 OW (aOR 0.59, 95% CI [0.06, 5.61] and aOR 0.66, 95% CI [0.20, 2.17], respectively).
Kelly et al., 2014 Lifeways Cross Generational Cohort Study.	BMI at birth, and ages 5 and 9 years. Weight, and length or height, measured at birth or in-home.	BMI of at least one randomly selected G0 at the birth of the G2. Only MGM results considered here. Height and weight information collected by self-report.	N/A	321, 217, and 128 G0 (MGM) and G2 pairs at birth, 5 years, and 9 years, respectively.	Group (child, mother, father, MGM, MGF, PGM, PGF), age of the family member, and the interaction of these variables.	Significant correlation between MGM BMI at baseline and child BMI at all three time points (0.141, p=0.031; 0.138, p=0.024; and 0.328, p<0.001).
Padechia et al., 2014 Female students in Mumbai, India.	Age- and sex-specific BMI z-score and WC percentiles at age 13-16 years (baseline). Height, weight, and waist circumference measured by trained dietician.	OB status (yes/no), reported at baseline (mean age not reported). No indication of which grandparent(s) was measured. Existence of obesity obtained via questionnaires.	N/A	299 (female) G2s (number of G0s unclear).	None.	No significant association between G0 OB status and G2 WC or BMI (p>0.05). Effect size not reported.

Table A.1: continued

Reference/Study	Exposure measure (G2 weight)	Outcome measure (G0 weight)	Mediator (G1 weight)	Sample size	Adjustment variables	Main findings
Harville et al., 2017 Bogolusa Heart Study (BHS).	Birthweight (g) Birthweight obtained from vital statistics data (or mother's self-report if unavailable).	MGM BMI at the study visit closest in time to the pregnancy used as a continuous variable and categorized: both maternal and MGM OW/OB; both maternal and MGM NW; maternal OW/OB and MGM NW; and maternal NW and MGM OW/OB. Height and weight collected by trained researchers.	N/A	177 G0 (MGM) and 424 G2.	MGM and maternal: age, smoking, race, parity, and time between study visit and pregnancy. Only maternal: BMI, marital status, education, weight gain in pregnancy.	MGM BMI was not significantly associated with G2 birthweight (beta = -12, 95% CI [-32, 8]). Highest G2 birthweight was observed in those with both maternal and MGM OW/OB (beta = 347, 95% CI [80, 614]).

Table A.1: continued

Reference/Study	Exposure measure (G2 weight)	Outcome measure (G0 weight)	Mediator (G1 weight)	Sample size	Adjustment variables	Main findings
Lahti-Pulkkinen et al., 2017 Aberdeen Maternity and Neonatal Databank (AMND).	Sex-, birth order- and gestational age-specific birthweight z- score standardized into SD units with a mean of 0 and a standard deviation of 1. Birthweight and length of gestation measured at birth.	MGM sex-, birth order- and gestational age-specific birthweight z-score standardized into SD units with a mean of 0 and a standard deviation of 1 used as a continuous variable and categorized: $\leq$-2 (SGA); -2-2 (AGA); $\geq$2 (LGA). Birthweight and length of gestation measured at birth.	Maternal sex-, birth order- and gestational age-specific birthweight z-score standardized into SD units with a mean of 0 and a standard deviation of 1.	1,457 G0 (MGM) and G2 pairs.	Partially adjusted models: G2 sex, delivery year; maternal age, SES, parity, hypertensive disorders of pregnancy, preexisting or gestational diabetes, labor type, smoking (each in all three generations). Fully adjusted models: + maternal birthweight, height, BMI, gestational age. No adjustment variables in mediation analysis.	MGM birthweight z-score was associated with higher G2 birthweight z-score in both partially and fully adjusted models (0.17, 95% CI [0.12, 0.23]; and 0.12, 95% CI [0.06, 0.18]). MGM LGA was not significantly associated with G2 birthweight z-score. Direct effect of MGM birthweight z-score on G2 birthweight z-score of 0.13 (95% CI [0.07, 0.18]) and indirect effect through maternal birthweight z-score of 0.06 (95% CI [0.04, 0.08])

Table A.1: continued

Reference/Study	Exposure measure (G2 weight)	Outcome measure (G0 weight)	Mediator (G1 weight)	Sample size	Adjustment variables	Main findings
Reuter et al., 2018 School childen from nine schools in Santa Cruz do Sul, Brazil.	BMI at age 7-17 years (baseline): ≥30 (OB). Height and weight collected by self report from children and validated via data obtained from the school.	MGP and PGP OB status (yes, no) (mean age not reported). Only MGM results considered here. Family history of OB collected via questionnaire completed by the student's parents.	N/A	381 G2s (unclear about number of G0s).	Unclear.	MGM OB status associated with G2 OB status in unadjusted (uPR=1.15, 95% CI [1.02, 1.31]) and adjusted (aPR=1.16, 95% CI [1.02, 1.32]) models.
Schott et al., 2018 Young Lives study.	Birthweight (g) and birthweight z-score (based on WHO 2006 reference distribution). Bithweight collected from birth record or mother's report.	MGM BMI z-score (based on WHO 2007 reference distribution) when mother's were approximately 12 years old (mean age 32.8 years). Height and weight collected by investigators.	N/A	283 G0 (MGM) and G2 pairs.	G2 firstborn; G1 firstborn, early menarche, pregnant at age 15, age at G2 birth; G0 height-for-age z-score, age, urban residence, household wealth index.	No significant association between MGM BMI and G2 birthweight (unadjusted beta=1.5, 95% CI [-66.1, 69.1], adjusted beta=-10.1, 95% CI [-73.3, 53.1]) or G2 birthweight z-score (unadjusted beta=0.04, 95% CI [-0.02, 0.12], adjusted beta=0, 95% CI [-0.14, 0.14]).

Table A.1: continued

Reference/Study	Exposure measure (G2 weight)	Outcome measure (G0 weight)	Mediator (G1 weight)	Sample size	Adjustment variables	Main findings
Somerville et al., 2018 Lifeways Cross Generational Cohort Study.	Age- and gender-specific WC z-scores at ages 5 and 9 years. WC measured in-home.	MGM WC (cm) at G2 birth used for when G2 was age 5 and latest WC measurement used for when G2 was age 9. Measurements done in-home at baseline (G2 birth year) and year 10 (G2 age 9).	WC (cm) measured at year 6 (G2 age 5) and year 10 (G2 age 9) follow-up.	190 and 130 G0 (MGM) and G2 pairs at age 5 and 9 respectively.	Baseline model: G0 age; G1 age, parity, WC. Partially adjusted model: + G0 smoking, SES, self-rated health; G1 smoking, diabetes. Fully adjusted model: + G1 SES, frequency of seeing MGPs, breastfeeding; G2 birthweight z-score. Adjustment variables for mediation analysis unclear.	MGM WC was significantly associated with G2 WC at ages 5 and 9 in partially adjusted models (beta = 0.015, p=0.019; beta = 0.022, p=0.033, respectively). Significant direct effect at both time points (beta = 0.014, p=0.02; beta = 0.024, p=0.01, respectively) and indirect effect via G1 WC (beta = 0.004, 95% CI [0.008, 0.008]; beta = 0.007, 95% CI [0.0024, 0.0144]).

Table A.1: continued

Reference/Study	Exposure measure (G2 weight)	Outcome measure (G0 weight)	Mediator (G1 weight)	Sample size	Adjustment variables	Main findings
Shen et al., 2020 Isle of Wight (IoW) birth cohort.	Birthweight (kg) Birthweight obtained from hospital records.	MGM pre-pregnancy BMI (time of G1 birth). Height and weight measured at first antenatal visit.	Birthweight (kg) and BMI. Birthweight obtained from birth records. Height and weight measured at 18 years during follow-up visit or by self-report.	209 G0 (MGM) and 355 G2.	G0 smoking, SES; G2 sex.	MGM BMI was not significantly associated with G2 birthweight (beta = 8, p=0.32). Non-significant direct effect (beta = 1.3, p=0.87), and significant indirect effect via G1 birthweight (beta=2.3, p-value not reported) and G1 BMI (beta=4.4, p-value not reported). Total indirect effect of beta=6.6 (p=0.04).

Abbreviations: *AGA* appropriate for gestational age; *BMI* body mass index; *G2* infant; *G1* parent; *G0* grandparent; *LGA* large for gestational age; *MGF* maternal grandfather; *MGM* maternal grandmother; *MGP* maternal grandparents; *NW* normal weight; *OB* obese; *OR* odds ratio; *OW* overweight; *PGF* paternal grandfather; *PGM* paternal grandmother; *SGA* small for gestational age; *SD* standard deviation; *UW* underweight

Table A.2: Summary of prediction models for fetal growth abnormalities

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Poon, 2008. Case-control study from London, UK.	Singleton pregnancies attending 1st trimester (11-14w) screening at King's College Hospital (n=296 cases and 609 controls). Cases and controls matched for length of storage of blood samples. Exclusion: none	SGA (<5th percentile for birthweight adjusted for gestational age, sex, maternal ethnic origin, maternal weight, maternal height, parity) vs. non-SGA.	Maternal factors: age, racial origin, any smoking in pregnancy, method of conception, medical history (e.g., chronic hypertension, diabetes), medication, parity, obstetric history (e.g., previous pre-eclampsia), family history of pre-eclampsia (maternal), BMI. Other: serum markers, ultrasound measurements.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: 0.69 (95% CI 0.65-0.72). Maternal factors + PAPP-A (best): 0.74 (95% CI 0.70-0.77).

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Onwudiwe, 2008. Prospective cohort study from London, UK	Singleton pregnancies with 2nd trimester (22-24w) screening at King's College Hospital from Jul 2006 to Oct 2007 (n=3347). Exclusions: women with pre-eclampsia or gestational diabetes, missing outcome data, fetal death or miscarriage <24w, terminated pregnancies.	SGA (<10th percentile for birthweight adjusted for gestational age, sex, maternal ethnic origin, maternal weight, maternal height, parity) vs. non-SGA.	Maternal factors: age, racial origin, any smoking in pregnancy, medical history (e.g., chronic hypertension, diabetes), method of conception, parity, medication, parity, obstetric history (e.g., previous pregnancy with pre-eclampsia), family history of pre-eclampsia (maternal), BMI Other: MAP, ultrasound measurements.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: 0.59 (95% CI 0.57-0.60). Maternal factors + other: 0.65 (95% CI 0.63-0.66).

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Leal, 2009. Case-control study from London, UK	296 cases of SGA matched 1:1 with control that had blood collected and stored on the same day, did not develop any pregnancy complications and resulted in a live birth of phenotypically normal neonates. Cases and controls were women attending King's College Hospital for 1st trimester (11-14w) screening.	SGA (<5th percentile for birthweight adjusted for gestational age, sex, maternal ethnic origin, maternal weight, maternal height, maternal parity) vs. non-SGA.	Maternal factors: age, racial origin, any smoking in pregnancy, method of conception, medical history (e.g., chronic hypertension, diabetes), medication, parity, obstetric history (e.g., previous pregnancy with pre-eclampsia), family history of pre-eclampsia (maternal), BMI Other: serum markers, ultrasound measurements	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: 0.69 (95% CI 0.66-0.72). Maternal factors + other: 0.70 (95% CI 0.67-0.73).
Schwartz, 2011. Prospective cohort study from Philadelphia, USA.	Singleton pregnancies with 1st trimester (11-14w) and 2nd trimester (18-24w) ultrasound measurements (n=245). Exclusions: missing outcome data.	SGA (<10th percentile for birthweight adjusted for gestational age) vs. non-SGA.	Maternal factors: age, parity, ethnicity. Other: ultrasound measurements	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: 0.64 (95% CI not reported). Maternal factors + abdominal circumference lag (best): 0.74 (95% CI not reported).

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Poon, 2011. Prospective cohort study from London, UK.	Singleton pregnancies attending 1st trimester (11-14w) screening at King's College Hospital from Mar 2006 to Sep 2009 (n=33602). Exclusions: missing outcome data, miscarriage <24 weeks, terminated pregnancies, major defects.	LGA (>90th percentile for birthweight adjusted for gestational age) vs. non-LGA.	Maternal factors: age, racial origin, any smoking in pregnancy, parity, previous LGA delivery, method of conception, medical history (e.g., chronic hypertension, diabetes), weight, height. Other: serum markers, ultrasound measurements.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: 0.72 (95% CI 0.71-0.72). Maternal factors + other: 0.73 (95% CI 0.72-0.73) Externally validated in Meertens et al.[114]: 0.52 (95% CI 0.47-0.58) in nulliparous and 0.52 (95% CI 0.46-0.58) in multiparous.
Poon, 2011. Prospective cohort study from London, UK.	Singleton pregnancies attending 1st trimester (11-14w) screening at King's College Hospital from Mar 2006 to Sep 2009 (n=32850). Exclusions: missing outcome data, miscarriage <24 weeks, termination, major defects, women with pre-eclampsia.	SGA (<5th percentile for birthweight adjusted for gestational age) vs. non-SGA.	Maternal factors: age, racial origin, any smoking in pregnancy, parity, birthweight of previous neonates, method of conception, medical history (e.g., chronic hypertension, diabetes), weight, height. Other: serum markers, ultrasound measurements.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: 0.72 (95% CI 0.71-0.73). Maternal factors + other: 0.75 (95% CI 0.74-0.76).

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Nanda, 2011. Nested case-control study within prospective cohort study from London, UK.	Singleton pregnancies attending 1st trimester (11-14w) screening at King's College Hospital from Mar 2006 to Sep 2009 (n=33344) Exclusions: missing outcome data, miscarriage or fetal death <24w, termination, major defects, women with pre-existing diabetes.	LGA (>95th percentile for birthweight adjusted for gestational age) vs. non-LGA.	Maternal factors: age, racial origin, any smoking in pregnancy, parity, previous LGA delivery, method of conception, (e.g., chronic hypertension, diabetes), weight, height. Other: serum markers.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC. Note: Case-control study had serum sample for 50 cases and 300 controls which were used to simulate marker values in full sample.	Maternal factors: 0.72 (95% CI 0.71-0.74) Maternal factors + other: 0.75 (95% CI 0.74-0.76) Externally validated in Meertens et al.[114]: 0.64 (95% CI 0.59-0.69) in nulliparous and 0.73 (95% CI 0.69-0.78) in multiparous.
Plasencia, 2011. Screening study from London, UK	Singleton pregnancies attending 1st trimester (11-14w) screening at King's College Hospital with placental volume measurements (n=3104). Exclusions: none.	SGA (<5th percentile for birthweight adjusted for gestational age) and LGA (>95th percentile for birthweight adjusted for gestational age) vs. AGA (>5th and <95th percentile for birthweight adjusted for gestational age).	Maternal factors: age, racial origin, any smoking in pregnancy, parity, previous LGA delivery, method of conception, weight, height Other: serum markers, placental volume.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: SGA 0.68 (95% CI 0.63-0.73) and LGA 0.70 (95% CI 0.65-0.74). Maternal factors + other: SGA 0.71 (95% CI 0.66-0.75) and LGA 0.72 (95% CI 0.67-0.76).

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Plasencia, 2012 Prospective cohort study from Spain.	Singleton pregnancies attending 1st trimester (11-14w) screening at Hospital Universitario Materno Infantil de Canaria from Mar 2008 and Aug 2009 with UtA Doppler assessment (n=2021). Exclusions: missing outcome data, fetal death or miscarriage <24w, termination.	LGA (>95th percentile for birthweight adjusted for sex and gestational age) vs. non-LGA.	Maternal factors: age, BMI, weight, height, ethnicity, smoking in pregnancy, parity, method of conception. Other: ultrasound measurements, serum markers.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: 0.71 (95% CI 0.68-0.73). Maternal factors + other: 0.72 (95% CI 0.70-0.74). Externally validated in Meertens et al.[114]: 0.66 (95% CI 0.61-0.72) in nulliparous and 0.70 (95% CI 0.65-0.75) in multiparous.

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Papastefanou, 2012. Cross-sectional study from Greece.	Singleton pregnancies attending 1st trimester (11-14w) screening delivering after 24 weeks (n=4702). Exclusions: spontaneous PTB, iatrogenic delivery, women with or had a history of gestational hypertension or gestational diabetes or pre-eclampsia, women with a history of chronic hypertension or diabetes, major defects, miscarriage, intrauterine death.	SGA (<5th percentile for birthweight adjusted for gestational age) and LGA (>95th percentile for birthweight adjusted for gestational age) vs. AGA (>5th and <95th percentile for birthweight adjusted for gestational age).	Maternal factors: weight, height, parity, smoking in pregnancy, method of conception Other: ultrasound measurements, serum markers.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: SGA 0.68 (95% CI 0.64-0.71) and LGA 0.66 (95% CI 0.62-0.76). Maternal factors + other: SGA 0.73 (95% CI 0.69-0.76) and LGA 0.69 (95% CI 0.65-0.72).
Lindell, 2013 Perinatal Revision South (PRS), Sweden.	Singleton pregnancies with ultrasound screenings at 17-19w and 32-34w between 1995 and 2009 (n=48809). Exclusions: PTB, missing data.	LGA (birthweight adjusted for gestational age and sex z-score > 2) vs non-LGA.	Maternal factors: age, parity, BMI, height, smoking in pregnancy, diabetes, gestational diabetes. Other: ultrasound measurements at 32-34w.	Multiple logistic regression. Development sample n=25261 and validation sample n=23548. No stratification by parity. Performance metric: AUC-ROC	Maternal factors: 0.74 (95% CI 0.73-0.76). Maternal factors + other: 0.91 (95% CI 0.90-0.92).

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Gonzalez Gonzalez, 2013. Prospective longitudinal study from Spain.	Singleton pregnancies attending 1st trimester (11-14w) and 2nd trimester (18-23w) screening at the Hospital Universitario Materno Infantil de Canaria between Mar 2008 and Aug 2009 (n=2021) Exclusions: missing outcome data, miscarriage or fetal death <24w, termination.	LGA (>90th percentile for birthweight adjusted for gestational age and sex) vs. non-LGA.	Maternal factors: age, weight, height, parity, smoking in pregnancy. Other: serum markers at 11-14w, ultrasound measurements at 11-14w and 18-23w	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: 0.68 (95% CI 0.66-0.70) Maternal factors + other: 0.77 (95% CI 0.74-0.79) Externally validated in Meertens et al.[114]: 0.67 (95% CI 0.62-0.72) in nulliparous and 0.70 (95% CI 0.64-0.74) in multiparous.

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Boucoiran, 2013. Retrospective cohort study from Quebec, Canada.	Singleton pregnancies attending 1st trimester (11-14w) screening between Jan 2005 and Dec 2009 (n=4901). Exclusions: women under 18 years, major defects, miscarriage or fetal death <24w, missing outcome data.	SGA (<10th percentile for birthweight adjusted for gestational age and sex) vs. non-SGA and LGA (>90th percentile for birthweight adjusted for gestational age and sex) vs. non-LGA.	Maternal factors: age, weight, ethnicity, smoking in pregnancy. Other: serum markers, ultrasound measurements.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: SGA 0.62 (95% CI 0.58-0.66) and LGA 0.66 (95% CI 0.63-0.68). Maternal factors + others: SGA 0.64 (95% CI 0.60-0.68) and LGA 0.67 (95% CI 0.64-0.69). Externally validated in Meertens et al.[114]: 0.64 (95% CI 0.59-0.70) in nulliparous and 0.64 (95% CI 0.58-0.70) in multiparous.

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
McCowan, 2013 Screening for Pregnancy Endpoints (SCOPE) study.	Singleton pregnancies to nulliparous women recruited at 14-16w from Nov 2004 to Feb 2011 (n=5606). Exclusions: women at high risk of pre-eclampsia or SGA or PTB, ≥3 miscarriages or terminations, received interventions that may modify pregnancy outcome, missing outcome data	SGA (<10th percentile for birthweight adjusted for maternal height, maternal weight, maternal ethnicity, gestational age and sex) vs. non-SGA.	Maternal factors at 14-16w: age, ethnicity, gravidity, SES, smoking in pregnancy, BMI, birthweight, blood pressure, family history of obstetric complications and medical disorders, currently attending university, proteinuria, lifestyle factors. Other: ultrasound measurements at 19-21w	Multiple logistic regression. Internal validation using training sample of n=3735 and test sample of n=1871. No stratification by parity. Performance metric: AUC-ROC	Maternal factors: 0.63 (95% CI not reported). Paternal history of CHD and maternal birthweight associated w/ odds of SGA. Maternal factors + other: 0.69 (95% CI not reported).
Schneuer, 2013 New South Wales, Australia using data from Perinatal Data Collection (PDC) and Admitted Patient Data Collection (APDC).	Singleton pregnancies attending 1st trimester (11-14w) screening from July 2006 to June 2007 (n=4621). Exclusions: blood sample taken <10w or >14w, medical abortion, major defects.	SGA10 (<10th percentile for birthweight adjusted for gestational age and sex) vs. non-SGA. Sensitivity analysis of SGA defined as <3rd [SGA3] percentile.	Maternal factors: age, weight, parity, smoking in pregnancy, pre-existing hypertension, previous miscarriage, country of birth, socioeconomic disadvantage. Other: serum markers.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: SGA10 0.68 (95% CI not reported) and SGA3 0.71 (95% CI not reported) Maternal factors + other: SGA10 0.69 (95% CI not reported) and SGA3 0.72 (95% CI not reported).

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Macdonald-Wallis, 2015. Longitudinal Study of Parents and Children (ALSPAC) and Southampton Women's Survey (SWS).	Training data (ALSPAC) consisted of singleton pregnancies resulting in a live birth between Apr 1991 and Dec 1992 (n=12996). External validation data (SWS) contained recruited non-pregnant women between Apr 1998 and Sept 2002 (n=3005 singleton delivery by the end of 2007). Exclusions: missing blood pressure measurements.	SGA (<10th percentile for birthweight adjusted for gestational age) vs. non-SGA.	Maternal factors: age, weight, height, parity, smoking in pregnancy, education, medical history (e.g., chronic hypertension, diabetes), ethnicity, SES, previous gestational diabetes or gestational hypertension, fetal sex. 2. Other: MAP at 11w, 20w, 25w, 28w, 31w, 34w and 36w.	Multiple logistic regression. External validation with development sample of n=12996 and validation sample of n=3005. Performance metric: AUC-ROC	Maternal factors + MAP at 11w: 0.70 (95% CI 0.61-0.79) Maternal factors + MAP at 11w + MAP at 25w: 0.70 (95% CI 0.62-0.79) Externally validated in Meertens et al.[114]: 0.66 (95% CI 0.60-0.71) in nulliparous and 0.67 (95% CI 0.61-0.0.72) in multiparous.

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Frick, 2016. Screening study from London, UK.	Singleton pregnancies attending 1st (11-14w), 2nd (19-25w), and/or 3rd trimester (30-38w) screening at King's College Hospital from Feb 2007 and Dec 2014 (n=72907 [11-14w], n=52573 [19-25w], n=30453 [30-38w]). Exclusions: non-live birth.	LGA95 (>95th percentile for birthweight adjusted for gestational age) vs. non-LGA. Sensitivity analysis of LGA defined as >90th [LGA90] and >97th [LGA97] percentile.	Maternal factors at 11-14w: age, racial origin, method of conception, smoking in pregnancy, medical history of diabetes, family history of diabetes, obstetric history (e.g., parity, previous gestational diabetes, birthweight z-score of previous delivery), BMI, gestational diabetes (assessed 30-34w and 35-37w) Other: serum markers, ultrasound measurements.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: LGA90 0.75 (95% CI 0.74-0.75), LGA95 0.78 (95% CI 0.78-0.78), and LGA97 0.79 (95% CI 0.79-0.79). Maternal factors + other: LGA90 0.76 (95% CI 0.75-0.76), LGA95 0.79 (95% CI 0.79-0.79), and LGA97 0.80 (95% CI 0.80-0.81). Externally validated in Meertens et al.[114]: 0.66 (95% CI 0.61-0.71) in nulliparous and 0.80 (95% CI 0.76-0.0.84) in multiparous.

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Crovetto, 2017. Nested case-control study within prospective cohort from Barcelona, Spain. Note: similar paper Corvetto, 2014 that assessed early- and late-onset SGA (not combined)	Cohort consisted of singleton pregnancies attending 1st trimester (11-14w) screening at Hospital Clinic, Barcelona. Cases and controls selected from May 2007 to Mar 2012 (n=9167). Exclusions: missing outcome data, miscarriage or fetal death <24w, termination, major defects.	SGA (<10th percentile for birthweight adjusted for gestational age and sex) vs. non-SGA neonates.	Maternal factors at 11-14w: age, race, parity, height, weight, smoking in pregnancy, method of conception, medical history (e.g., chronic hypertension, diabetes), obstetric history (e.g., previous fetal growth restricted infant or pre-eclampsia or stillbirth). Other: serum markers, ultrasound measurements, MAP.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC. Note: All cases and 762 controls had serum marker measurements, which were used to estimate the values in remaining controls.	Maternal factors: 0.65 (95% CI 0.63-0.71). Maternal factors + other: 0.68 (95% CI 0.66-0.70).

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Gonzalez Gonzalez, 2017. Prospective cohort study from Spain.	Singleton pregnancies attending 1st trimester (11-14w) screening at the Hospital Universitario Materno Infantil de Canaria between May 2011 and Mar 2014 (n=988). Exclusions: miscarriage or termination <24w.	SGA (<10th percentile for birthweight adjusted for gestational age and sex) vs. non-SGA.	Maternal factors: age, parity, weight, height, BMI, race, smoking in pregnancy, medical history (e.g., chronic hypertension, diabetes), obstetric history (e.g., hypertensive pregnancy complications, gestational diabetes), method of conception. Other: ultrasound measurements serum markers, placental volume, 3D vascular indices.	Multiple logistic regression. Model development using 10-fold CV. No stratification by parity. Performance metric: AUC-ROC	Maternal factors: 0.60 (95% CI not reported). Maternal factors + other: 0.73 (95% CI not reported). Externally validated in Meertens et al.[114]: 0.56 (95% CI 0.52-0.59) in nulliparous and 0.57 (95% CI 0.53-0.61) in multiparous.

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
McCowan, 2017 Screening for Pregnancy Endpoints (SCOPE) study. Note: same cohort as McCowan, 2013 and Vieira, 2017	Singleton pregnancies to nulliparous women recruited at 14-16w from Nov 2004 to Feb 2011 (n=5628). Exclusions: women at high risk of pre-eclampsia or SGA or PTB, ≥3 miscarriages or terminations, received interventions that may modify pregnancy outcome, missing outcome data.	SGA (<10th percentile for birthweight adjusted for maternal height, maternal weight, maternal ethnicity, gestational age, and sex) vs. non-SGA.	Maternal factors at 14-16w: age, ethnicity, gravidity, SES, smoking in pregnancy, BMI, birthweight, PTB, blood pressure, family history of obstetric complications and medical disorders, currently attending university, proteinuria, lifestyle, gestational weight gain between 14-16w and 19-21w. Other: ultrasound measurements at 19-21w, biomarkers at 14-16w.	Multiple logistic regression. Internal validation using training sample of n=3752 and test sample of n=1876. No stratification by parity. Performance metric: AUC-ROC	Maternal factors: 0.59 (95% CI 0.54-0.64). Maternal birthweight associated with a 1.23 times increase in odds of SGA (unadjusted). Maternal factors + other: 0.69 (95% CI 0.65-0.73).

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Vieira, 2017 Screening for Pregnancy Endpoints (SCOPE) study. Note: same cohort as McCowan, 2013 and McCowan, 2017	Singleton pregnancies to nulliparous women recruited at 14-16w from Nov 2004 to Feb 2011 (n=5628). Exclusions: women at high risk of pre-eclampsia or SGA or PTB, ≥3 miscarriages or terminations, received interventions that may modify pregnancy outcome, missing outcome data.	LGA90 (>90th percentile for birthweight adjusted for maternal height, maternal weight, maternal ethnicity, gestational age, and sex, and ≥37w gestation) vs. non-LGA90. Sensitivity analysis of LGA defined as >95th [LGA95] percentile.	Maternal factors at 14-16w: age, ethnicity, gravidity, SES, smoking in pregnancy, BMI, birthweight, PTB, blood pressure, family history of obstetric complications and medical disorders, currently attending university, proteinuria, lifestyle factors, gestational weight gain between 14-16w and 19-21w. Other: ultrasound measurements at 19-21w, biomarkers at 14-16w, blood glucose concentrations at 14-16w and 19-21w.	Multiple logistic regression. Internal validation using training sample of n=3752 and test sample of n=1876. No stratification by parity. Performance metric: AUC-ROC	Maternal factors: LGA90 0.59 (95% CI 0.54-0.64) and LGA95 0.58 (95% CI 0.51-0.64). Maternal factors + other: LGA90 0.69 (95% CI 0.65-0.73) and LGA95 0.70 (95% CI 0.64-0.75).

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Kuhle, 2018 Nova Scotia Atlee Perinatal Database (NSAPD).	Singleton pregnancies from Jan 2009 to Dec 2014 (n=30705). Exclusions: non-live birth, missing outcome or predictor data.	SGA10 (<10th percentile for birthweight adjusted for gestational age and sex) vs. non-SGA10 and LGA90 (>90th percentile for birthweight adjusted for gestational age and sex) vs. non-LGA90. Sensitivity analysis of SGA defined as <3rd [SGA3] and LGA >97th [LGA97] percentile.	Maternal factors: age, marital status, SES, rurality, pre-pregnancy BMI, medical history (e.g., chronic hypertension, diabetes), obstetric history (e.g., gravidity, parity), weight gain in pregnancy at 26w, substance use in pregnancy, smoking in pregnancy, gestational hypertension, psychiatric disorder, fetal sex.	Logistic regression and machine learning algorithms (elastic net [EN], decision trees [DT], random forest [RF], gradient boosting [GB], neural networks [NN]). Internal validation using 80% training and 20% testing data. Model development using 10-fold CV. Stratification by parity. Performance metric: AUC-ROC	Nulliparous: SGA10 from 0.63 to 0.67 and LGA90 from 0.67 to 0.71. Multiparous: SGA10 from 0.72 to 0.77 and LGA90 from 0.70 to 0.75. Key predictors: Smoking, previous low birthweight infant, weight gain at 26w for SGA10 and pre-pregnancy BMI, weight gain at 26w, and previous infant >4080g for LGA90.

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Erkamp, 2020 Generation R Study, Netherlands.	Singleton pregnancies with expected delivery date from Apr 2002 to Jan 2006 (n=7670). Exclusions: non-live birth, no 2nd and 3rd trimester ultrasound data, missing outcome data.	SGA10 (<10th percentile for birthweight adjusted for gestational age) vs. non-SGA10 and LGA90 (>90th percentile for birthweight adjusted for gestational age) vs. non-LGA90. Sensitivity analysis of SGA defined as <3rd [SGA3] and LGA >97th [LGA97] percentile.	Maternal factors: age, BMI, ethnicity, parity, smoking in pregnancy. Other: ultrasound measurements in 2nd and 3rd trimesters.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: SGA10 0.67 (95% CI 0.65-0.69), LGA90 0.68 (95% CI 0.66-0.70), SGA3 0.69 (95% CI 0.66-0.73), and LGA97 0.72 (95% CI 0.69-0.76). Maternal factors + other: SGA10 0.80 (95% CI 0.78-0.82), LGA90 0.77 (0.75-0.78), SGA3 0.85 (95% CI 0.82-0.88), and LGA97 0.82 (95% CI 0.79-0.84).
Monari, 2021. Prospective cohort study from Italy.	Singleton pregnancies attending 1st trimester (11-14w) screening at the University Hospital of Modena from Jun 2018 to Dec 2019 (n=503). Exclusions: crown-rump-length outside of range 45-88mm, miscarriage, termination, major defects.	LGA (>90th percentile for birthweight, body length, and head circumference adjusted for gestational age, sex, and birth order) vs. non-LGA.	Maternal factors: age, education, parity, Italian origin, smoking in pregnancy, BMI, method of conception, medical history (e.g., chronic hypertension, diabetes metabolic syndrome). Other: serum markers, MAP, ultrasound measurements.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors + other: 0.705 (95% CI not reported).

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Perichart-Perera, 2022 Epigenetic and Biochemical Origin of Overweight and Obesity (OBESO) perinatal cohort, Mexico.	Singleton pregnancies between Jan 2017 and Jan 2019 (n=77). Exclusions: comorbidities (e.g., pre-existing diabetes), medication that may affect endocrine metabolism, women who developed gestational diabetes or hypertension or pre-eclampsia during pregnancy, LGA delivery, missing data.	SGA (<10th percentile for birthweight adjusted for gestational age) vs. non-SGA.	Maternal factors at 11-14w: pre-pregnancy weight, height, pre-pregnancy BMI, gestational weight gain during first trimester, fat mass, multivitamin supplementation. Other: serum markers, oxidative stress markers.	Artificial neural network [ANN]. 75% training and 15% testing data. Model development using 10-fold CV. No stratification by parity. Performance metric: AUC-ROC	Maternal factors + other: 0.80 (95% CI not reported). Key predictors: protein oxidation, gestational weight gain, Vitamin D, total antioxidant capacity, lipid oxidation.

Table A.2: continued

Reference/Study	Study population and sample size	Outcome(s)	Predictors	Statistical analysis	Main findings
Wahab, 2022 Generation R Study, Netherlands.	Singleton pregnancies with expected delivery date from Apr 2002 to Jan 2006 (n=8340 mothers and 6062 fathers). Exclusions: non-live birth, missing gestational age or birthweight.	SGA/PTB (<10th percentile for birthweight adjusted for gestational age and sex or gestational age <37w) and LGA (>90th percentile for birthweight adjusted for gestational age and sex) vs. AGA (>10th and <90th percentile for birthweight adjusted for gestational age).	Maternal factors at 13.9w: age, ethnicity, parity, pre-pregnancy weight, height, pre-pregnancy smoking, pre-pregnancy BMI, education, income, marital status, lifestyle factors, medical history (e.g., chronic hypertension, diabetes), obstetric history (e.g., previous stillbirth, miscarriage, pre-eclampsia). Other [maternal]: blood pressure, serum markers at 14.4w and 20.4w. Paternal factors: age, ethnicity, education, smoking, alcohol, height, weight, BMI, blood pressure.	Multiple logistic regression. No internal or external validation performed. No stratification by parity. Performance metric: AUC-ROC.	Maternal factors: SGA/PTB 0.64 (95% CI 0.63-0.66) and LGA 0.65 (95% CI 0.63-0.67). Maternal factors + other: SGA/PTB 0.66 (95% CI0.64-0.67) and LGA 0.67 (95% Ci 0.66-0.69). Maternal factors + other + paternal factors: SGA/PTB 0.66 (95% CI 0.64-0.67) and LGA 0.69 (95% CI 0.67-0.70).

Abbreviations: *AGA* appropriate for gestational age; *AUC-ROC* area under the receiver operating characteristic curve; *BMI* body mass index; *CI* confidence interval; *CV* cross-validation; *LGA* large for gestational age; MAP mean arterial pressure; *PTB* pre-term birth; *SES* socioeconomic status; *SGA* small for gesational age

Table A.3: Definitions and coding of variables in the Nova Scotia Atlee Perinatal Database that were abstracted for the studies in Chapters Four, Five, and Six.

	Definitions	Variable type/coding
Maternal characteristics		
Age [years]	Age at delivery collected at the first prenatal visit	Continuous (truncated at 15 and 45 years)
Parity	Number of pregnancies, excluding the present pregnancy, which resulted in an infant weighting $\geq$500 g or $\geq$20 weeks gestational age, recorded at the first prenatal visit	Categorical (0, I, II, III+)
Marital status	Marital status at the first prenatal visit	Binary (1=married, common-law; 0=otherwise)
Area-level income quintile	Area-based socioeconomic measure of neighborhood income per person equivalent derived by linkage to national census information	Categorical (Q1, Q2, Q3, Q4, Q5)
Area of residence	Determined by woman's postal code recorded on the hospital admission form	Binary (1=urban, 0=rural)
Pre-pregnancy weight [kg]	Weight at the first prenatal visit collected by self-report or measured	Continuous
Pre-pregnancy BMI [kg/m^2]	Derived variable from pre-pregnancy weight (kg) and height (m)	Continuous
Delivery weight [kg]	Weight at delivery reported on the maternal admission assessment	Continuous
Smoking during pregnancy	Any smoking reported at the first prenatal visit, at 20 weeks, or at delivery is considered as smoking during pregnancy	Binary (1=smoking, 0=no smoking)
Pre-existing hypertension	Hypertension at $<$ 20 weeks	Binary (1=yes, 0=no)
Hypertensive disorders of pregnancy	Hypertension at $\geq$ 20 weeks	Binary (1=yes, 0=no)
Pre-existing diabetes	Presence of pre-existing diabetes mellitus (type 1 and type 2) reported on the prenatal record or the hospital discharge form	Binary (1=yes, 0=no)
Gestational diabetes mellitus	Presence of gestational diabetes mellitus derived from results from the glucose challenge test and the oral glucose tolerance test, or from admission forms at the time of delivery	Binary (1=yes, 0=no)

	Definitions	Variable type/coding
Mode of delivery	Type of delivery (Caesarean section or vaginal birth) and the stage of labour at which the Caesarean section was performed is recorded on the birth record	Binary (1=Caesarean section, 0=vaginal birth) or categorical (0=vaginal birth, 1=Caesarean section before onset or 2nd stage of labour, 2=Caesarean section after 2nd stage of labour)
Neonatal characteristics		
Birthweight z-score	Gestational age- and sex-specific birthweight z-scores relative to a Canadian reference population[156]	Continuous
SGA	Infants born with a birthweight z-score <10th percentile according to reference population	Binary (1=SGA, 0=non-SGA)
LGA	Infants born with a birthweight z-score >90th percentile according to reference population	Binary (1=LGA, 0=non-LGA)