Title
Multiple Bias Modeling in a Multi-Center Epidemiologic Study of Endometrial Cancer

Author
Thompson, Caroline Avery

Publication Date
2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Multiple Bias Modeling in a Multi-Center Epidemiologic Study of Endometrial Cancer

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Epidemiology

by

Caroline Avery Thompson

2013

ABSTRACT OF THE DISSERTATION

Multiple Bias Modeling in a Multi-Center Epidemiologic Study

of Endometrial Cancer

by

Caroline Avery Thompson

Doctor of Philosophy in Epidemiology

University of California, Los Angeles, 2013

Professor Onyebuchi Aniweta Arah, Co-Chair

Professor Veronica Wendy Setiawan, Co-Chair

Quantitative treatment of uncontrolled bias in observational research is a neglected matter. In the dawn of the era of "big data", this is of particular concern because systematic error, as a portion of total error, can be greatly magnified when sample sizes increase. Unfortunately, considerable statistical road blocks exist between performing a basic multivariable analysis of an exposure-disease relationship and the thorough consideration of the direction and magnitude of uncontrolled bias. Most published literature points to the use of external formula adjustment for a thorough treatment of bias, but the formulas are often too simple (and thus unrealistic) or too complex (and thus unwieldy). A practical solution might be to perform the bias adjustment in the data, before analysis is performed. This solution would be especially useful in pooled data consortium projects, which are becoming increasingly popular as a way to investigate rare

exposures and disease subtypes in cancer epidemiology, and often employ one shared data source used by multiple investigators simultaneously. Record-level data augmentation for bias analysis is central to a pooling project because it allows for multiple bias parameters to be placed directly in this data source. In this work we utilize causal theory, Monte-Carlo methods, and the missing data framework to contribute the literature of quantitative bias modeling, via flexible algorithms that may be used to translate bias adjustment for unmeasured confounding and non-response directly into the data source, before the analysis stage. We provide proof of concept for these methods via a series of simulation studies, and demonstrate their utility in a large multi-center pooled study of the epidemiology of endometrial cancer, employing both fixed hypothetical and probabilistic empirical priors for our bias parameters. Moving bias adjustment to the pre-analytic stage opens the door for an augmented data set to be analyzed in any conventional way, with no need for a working knowledge of the complex methodology behind existing external formula adjustments. A thorough, accessible, quantitative bias analysis can then serve as a tool to guide qualitative discussions about the impact of systematic error in multi-study data projects.

The dissertation of Caroline Avery Thompson is approved.

Ronald S. Brookmeyer

Leeka I. Kheifets

ZuoFeng Zhang

Veronica Wendy Setiawan, Committee Co-Chair

Oneybuchi Aniweta Arah, Committee Co-Chair

University of California, Los Angeles

2013

DEDICATION

This dissertation is dedicated to the late Dr. Carolyn R. Thompson, PhD, my mother, best friend and hero.

TABLE OF CONTENTS

1. INTRODUCTION 1
1.1. ENDOMETRIAL CANCER 1
1.2. CANCER POOLING PROJECTS 4
1.3. BIAS ANALYSIS 5
1.4. GAPS IN THE BIAS ANALYSIS LITERATURE 7
1.5. THIS DISSERTATION 8
1.5.1. SPECIFIC AIMS OF THE DISSERTATION 9
2. STUDY 1 11
2.1. ABSTRACT 12
2.2. INTRODUCTION 13
2.3. METHODS 14
2.4. RESULTS 23
2.5. DISCUSSION 27
3. STUDY 2 29
3.1. ABSTRACT 30
3.2. INTRODUCTION 31
3.3. NOTATION AND METHODS 32
3.4. RESULTS 40
3.5. DISCUSSION 46
4. STUDY 3 48
4.1. ABSTRACT 49
4.2. INTRODUCTION 50
4.3. MATERIALS AND METHODS 52
4.4. RESULTS 59
4.5. DISCUSSION 73
5. STUDY 4 77
5.1. ABSTRACT 78
5.2. INTRODUCTION 79
5.3. MATERIALS AND METHODS 80

5.4. RESULTS 84
5.5. DISCUSSION 93
6. CONCLUSIONS 95
7. APPENDIX TABLES 101
8. REFERENCES 124

LIST OF TABLES

Table 2.1 Actual characteristics of the trial cohorts (N=10,000) 24

Table 2.2 Bias model performance for scenarios a and b using saturated versus unsaturated models 25

Table 2.3 Bias model performance for selected trials with varying inputs, using reduced model form 26

Table 3.1 Correctly specified priors for adjustment of collider bias in a cohort (N=100,000) defined by the DAGs in figure 3.3 and 3.4 42

Table 3.2 Reduced models with correctly specified priors for adjustment of collider bias in a cohort (N=100,000) defined by the DAGs in figure 3.3 and 3.4 43

Table 3.3 Misspecified priors for adjustment of collider bias in a cohort (N=100,000) defined by the DAG in figure 3.4[1] 44

Table 4.1 Participating studies: location and population description 54

Table 4.2 Subject characteristics 60

Table 4.3 Smoking bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer 64

Table 4.4 Estrogen-only hormone replacement therapy bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer 66

Table 4.5 Diagnosis of diabetes bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer 68

Table 4.6 All three variable (smoking, EHRT, diabetes) bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer 70

Table 4.7 Categorical BMI – with all bias adjustment (ERT, smoking status, diabetes), by study type and white vs. non-white ethnicity 72

Table 5.1 Eligible participating population and hospital based case-control studies: location and population description..... 85

Table 5.2 Subject characteristics..... 86

Table 5.3 Selection bias adjustment fixed by scenario and trial 88

Table 5.4 Scenario A: Selection bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer 90

Table 5.5 Scenario B: Selection bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer 91

Table 5.6 Scenario C: Selection bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer 92

Table 7.1 Study 1: Planned characteristics of the trial cohorts (N=10,000) 102

Table 7.2 Study 1: Actual characteristics of the trial cohorts (N=10,000)..... 104

Table 7.3 Study 1: Full simulation results: imputing a dichotomous confounder (Z_1, Z_2) model (N=10,000, reps=1,000) – Non-saturated models..... 106

Table 7.4 Study 1: Full simulation results for Imputing two dichotomous confounder (Z_1, Z_2) simultaneously model (N=10,000, reps=1,000) – Non-saturated models 107

Table 7.5 Study 1: Full simulation results from imputing a single continuous confounder (Z_3) model (N=10,000, reps=1,000) – Non-saturated models..... 108

Table 7.6 Study 1: Full simulation results for Imputing a single trichotomous confounder (Z_4) model (N=10,000, reps=1,000) – Non-saturated models..... 109

Table 7.7 Study 3: Smoking bias model parameters, based on complete case analysis, by study type[1] 110

Table 7.8 Study 3: Estrogen bias model parameters, based on complete case analysis, by study type[1] 114

Table 7.9 Study 3: Diabetes bias model parameters, based on complete case analysis, by study type[1] 119

LIST OF FIGURES

Figure 2.1 U is an unmeasured variable that confounds the relationship between X and Y. 14

Figure 2.2 Scenario A – A DAG representing marginally independent X and Y with 4 confounding variables Z_1, Z_2, Z_3, and Z_4. 16

Figure 2.3 Scenario B – A DAG representing marginally dependent X and Y with 4 confounding variables Z_1, Z_2, Z_3, and Z_4. 16

Figure 3.1 A DAG representing marginally independent but conditionally (on S=1) dependent X and Y; a simple example of collider bias. 33

Figure 3.2 A DAG representing marginally independent but conditionally (on S=1) dependent X and Y, another example of collider bias as a result of uncontrolled common causes of X-S and Y-S 34

Figure 3.3 Scenario A – A DAG representing marginally independent but conditionally (on S=1) dependent X and Y, with 4 confounding variables Z_1, Z_2, Z_3, and Z_4. 34

Figure 3.4 Scenario B – A DAG representing marginally dependent X and Y with additional conditional (on S=1) dependency (collider biasing path) and 4 confounding variables Z_1, Z_2, Z_3, and Z_4. 34

Figure 4.1 Suspected measured and partially unmeasured confounding variables in the relationship between BMI and type I endometrial cancer (EC) 55

Figure 5.1 Collider bias due to selective non-response 82

Figure 5.2 Signed DAG representing modeled influences of nonresponse in E2C2 case control studies 83

LIST OF ACRONYMS AND ABBREVIATIONS

BMI	Body mass index
CI	Confidence interval
DAG	Directed acyclic graph
E2C2	Epidemiology of Endometrial Cancer Consortium
EC	Endometrial cancer
EHRT	Estrogen-only hormone replacement therapy
HFCA	Health care financing administration
HRT	Hormone replacement therapy
I/O	Input/output
IPCW	Inverse probability of censoring weighting
IPSW	Inverse probability of selection weighting
LCL	Lower 95% confidence limit
Log	Natural logarithm
LSI	Lower 95% simulation interval
MC	Monte-Carlo
OR	Odds ratio
RDD	Random digit dialling
Ref	Reference category
RMSE	Root mean squared error
SD	Standard deviation
SE	Standard error

SI	Simulation interval
UCL	Upper 95% confidence limit
USI	Upper 95% simulation interval

ACKNOLWEDGEMENTS

I wish to give a heartfelt thanks to all who supported me during my 5-year tenure at the UCLA Fielding School of Public Health. I would not be who I am today without the mentorship, attention, expert guidance, and friendship I received from the following people: my dissertation committee chair and valued advisor and mentor, Onyebuchi Arah, my co-advisor, Zuo-Feng Zhang, my co-chair, Wendy Setiawan, my committee members: Leeka Khiefits and Ron Brookmeyer, my master's thesis mentor, Eva Schernhammer, and my supportive and loving friends, Maral DerSarkissian, Aolin Wang, Kaitlin O'Keefe, Heather Pines, and Danny Garcia. I also wish to thank the investigators of the Epidemiology of Endometrial Cancer Consortium (E2C2) for use of their data in studies 3 and 4.

My work on this project was supported by a training grant from the National Cancer Institute: T32 CA09142-031

VITA

EDUCATION

2010	MPH in Epidemiology University of California, Los Angeles, CA, USA
1999	BA in Biology University of North Carolina, Chapel Hill, NC, USA

RESEARCH EXPERIENCE

2011 – present	Junior Researcher EU 7th Framework Program for Research: DUQuE: Deepening our Understanding of Quality Improvement in European Hospitals
2010 – present	Pre-Doctoral Trainee UCLA NCI Training Program in Epidemiology of Cancer UCLA Fielding School of Public Health Los Angeles, CA, USA
2009 – present	Research Assistant Department of Epidemiology UCLA Fielding School of Public Health Los Angeles, CA, USA
Summer 2010	Research Intern Department of Epidemiology University of Vienna Vienna, Austria

PROFESSIONAL EXPERIENCE

2006 – 2008	Data Management Consultant Exelixis, Inc.; Cell Genesys, Inc.; Fibrogen, Inc.; South San Francisco, CA, USA
2004 – 2006	Clinical Operations Manager Pfizer, Inc. La Jolla, CA, USA
2002 – 2004	Sr. Clinical Data Specialist Chugai Pharmaceuticals San Diego, CA, USA

2000 – 2002	Clinical Data Manager PAREXEL International Corporation Research Triangle Park, NC, USA
1999 – 2000	Clinical Data Coordinator Quintiles Transnational Corporation Research Triangle Park, NC, USA

PUBLICATIONS

Thompson CA, Zhang ZF, Arah, OA (2013). Competing risk bias to explain the inverse relationship between smoking and malignant melanoma. *European Journal of Epidemiology*. [Epub ahead of print. DOI: 10.1007/s10654-013-9812-0].

Thompson CA, Waldhör T, Schernhammer ES, Hackl M, Vutuc C, Haidinger G. (2012). Smoking and lung cancer: current trends in Austria. *Wiener klinische Wochenschrift,* 124(15-16), 493-9.

Schernhammer ES, **Thompson CA**. (2011). Light at night and health: the perils of rotating shift work. *Occupational and Environmental Medicine*, 68(5), 327-31.

PRESENTATIONS

Thompson CA, Zhang ZF, Arah OA. "Competing risk bias to explain the inverse relationship between smoking and melanoma." The 3rd North American Congress for Epidemiology. Montreal, Canada. June 2011. (Oral Presentation)

Arah OA, **Thompson CA**. "Global impact of different health states on self-assessed general health among individuals in 68 countries." The 3rd North American Congress for Epidemiology. Montreal, Canada. June 2011. (Poster Presentation)

Thompson CA, Zhang ZF, Setiawan VW, Arah OA. "Handling uncontrolled confounding in the comparative effectiveness research of cancer." Annual Research Meeting for Academy Health. Walt Disney World, Florida. June 2012. (Oral Presentation)

Thompson CA, DerSarkissian M, Stronks K, Arah OA. "Global health spending and noncommunicable disease co-occurrence." Annual Research Meeting for Academy Health. Walt Disney World, Florida. June 2012. (Poster Presentation)

Thompson CA, Setiawan VW, Zhang ZF, Arah OA. "Record-level bias analysis for uncontrolled confounding in cancer pooling projects with multiple investigators". 45th Annual Society for Epidemiologic Research Meeting. Minneapolis, MN. June 2012. (Poster Presentation)

Thompson CA, Arah OA. "Using DAGs to guide the translation of priors for record-level analysis of bias due to unmeasured confounding." 45th Annual Society for Epidemiologic Research Meeting. Minneapolis, MN. June 2012. (Poster Presentation)

1. INTRODUCTION

1.1. ENDOMETRIAL CANCER

Endometrial cancer is the most common gynecological cancer in the western world, and it affects more than 40,000 women in a year in the US. Even so, it is still a rare cancer, accounting for only 4% of all cancers worldwide. Most of the time, prognosis is very good, with a 5-year survival rate of 70-80% in western countries [1]. Incidence rates as well as survival rates are highest in Caucasian women in developed countries. Endometrial cancer incidence is highest in peri- and post-menopausal women, between the ages of 45 and 70. Black women have lower incidence than Caucasian women, but their 5-year case survival rates are lower. One explanation is that black women may be diagnosed with later stage tumors on average, but it is possible that their prognosis profile is substantially different from white women as well [2]. Less is known about how the risk profiles may differ among minorities such as Asians, Pacific Islanders, and Native Americans, due to typically low representation of these groups in observational studies.

Generally endometrial cancer is understood to be caused by unopposed estrogen exposure, either through early age of menarche, nulliparity, late age at first birth, estrogen only hormone therapy, or high dose estrogen in oral contraceptive pills [3]. Obesity, as measured by BMI as well as anamorphic measurements, is also a major risk factor for endometrial cancer, but the molecular mechanism is different for pre- and post-menopausal women. In pre-menopausal women, obesity leads to increased insulin, progesterone deficiency, and thus a reduced ability to oppose free estrogens [4]. In post-menopausal women obesity leads to endometrial cancer through increases in free floating estrogens. Endometrial cancer is one of the only neoplasms for which

smoking is protective. The biological mechanism for this is thought to be related to increases in estrogen-opposing progesterone [5]. Other protective factors include normal weight, weight loss, physical activity, grand multiparity and exogenous hormones that include a cycle of progesterone, such as combination hormone replacement therapy and modern low dose estrogen and progesterone combination oral contraceptive pills. In addition to being risk factors for disease, exogenous hormones are strong modifiers of the BMI effect, often with paradoxical results. Overweight women who have previously taken exogenous hormones tend to experience less risk in higher BMI categories than overweight women who have not taken exogenous hormones [6]. Other risk factors (not related to the estrogen pathway) include family history of endometrial cancer and increasing age.

The most common histological types of endometrial tumors are endometrioid adenocarcinomas which constitute 80% of all endometrial cancer, and are collectively referred to as Type I. Type I tumors are generally considered to be estrogen-dependent, and they are associated with endometrial hyperplasia, hyperlipidemia, and obesity. Type II tumors are mainly serous and clear cell carcinomas, as well as sarcomas, and some mixed tumor types. Type II tumors are more aggressive clinically, arising from the atrophic endometrium in elderly women [7-10]. Due to the small number of cases, historically, all endometrial cancer histology types have been considered together in epidemiologic research. Recent work has found evidence that the risk profiles of cases with type II tumors may differ from those women with type I, especially with regard to sex steroids and body size [11]. More research is needed to fully characterize the risk profile of type II endometrial cancer as it compares to type I.

Since 2005, there has been a slight increase in incidence of endometrial cancer in the US. While more data is needed before this can be confirmed as an upward trend, there is some indication that the western lifestyle, in particular lack of physical activity and excess body weight may be causing a rise in this type of cancer. Obesity is by far the strongest risk factor for Endometrial cancer, with reported relative risk of 2-5 times that of normal body weight in both pre- and post-menopausal women [12]. In many studies, the risk of cancer increases linearly with BMI, but in younger, pre-menopausal women the risk has occasionally been reported only in obese weight categories [13, 14]. The prevailing hypothesis to link excess body weight to endometrial cancer is through endogenous hormones of estrogen, progesterone, and insulin, exposure to which are also independent risk factors for endometrial cancer. These hormones each play a role in the regulation of cell proliferation, differentiation, and apoptosis. Increased proliferation can result in increases in cell mutations. Decreased differentiation and apoptosis improves the local environment for a growing tumor cell. Excess body weight alters the balance of these endogenous sex steroids. As would be expected, other exogenous factors that affect the balance of sex steroids and are also related to body weight are considered to be confounders of the BMI and endometrial cancer relationship. For example, exposure to estrogen-only hormone replacement therapy is related to an increase in body weight, and an increase in endometrial cancer risk. Oral contraception with a progesterone component is slightly protective for endometrial cancer, but is often associated with weight changes. Smoking causes reduced body weight, and is protective for endometrial cancer, both via the endogenous estrogen pathway as well as reduced inflammation.

Due to the rare nature of this disease, there are a lot of knowledge gaps concerning its complex etiology. The genetic basis for the established risk factors has not yet been clearly defined, nor

has much work been done to explore gene-gene interactions or gene-environment interactions. Additionally, the risk factors for rarer histological types (type II) cancers are largely unknown, and for some ethnic minorities the risk profile may also be different. The major limitation to exploring these gaps has always been lack of sample size. There are a lot of small studies evaluating risk of endometrial cancer, but few have the statistical power to perform these investigations due to low cases numbers and/or homogenous populations.

1.2. CANCER POOLING PROJECTS

Cancer is a rare disease. Because of its rarity, observational researchers often resort to pooling data to achieve the power needed for studies of exposure-disease associations. A major example of this is the Genome Wide Association Study (GWAS) trend, which has become very popular in the past 5 years. A pooled data analysis is the combining of raw data from multiple studies, recoding to fit a single pooled "standardized" database, and then analyzing all studies as one. This is different from a traditional meta-analysis of point estimates, which just involves extracting point estimates from published literature and then summarizing over the extracted point estimates. The advantages of pooling are vast. Consortium projects that combine multiple studies can vastly increase power and efficiency of their desired investigations. This allows for less prevalent exposures to be studied as risk factors, as well as the investigation of interactions between multiple risk factors in predicting disease risk. Additionally, pooling projects allows for the study of risk factors in populations that are often underrepresented in epidemiology, such as minority ethnicities. When compared with the meta-analysis of point estimates, standardized pooling allows for better control of systematic biases, as well as careful investigation and explanation of the heterogeneity within and across studies. Pooling can also provide an

opportunity to estimate effect measures that may not have been reported originally for all the individual studies [15]. Pooling data from various studies and sources is not without its limitations, however. Studies of different design types are often combined and treated as the same design, methods of variable measurement may vary from study to study, and important risk factors or confounding variables may not be collected in all of the pooled studies. The creation of a standardized database may result in the loss of information when partially unmeasured variables are dropped across studies, or when categories are collapsed to fit a standardized framework of variable definition. All of these limitations are likely to result in bias.

1.3. BIAS ANALYSIS

Almost all sources of bias can be subsumed under the bias due to uncontrolled confounding, selective non-response bias, and measurement error [16]. Confounding is the mixing of effects, when the apparent effect is distorted by the effect of extraneous factors. Specifically, when one or more variables affects both the exposure and the outcome, but is not itself affected by either the exposure or the outcome, the presence of such a variable can bias an estimate upward or downward, or produce an non-null estimate when the exposure and outcome are unrelated [16]. Selection bias is the distortion of effect that results from subject selection procedures or participation rates. Specifically, if diseased and non-diseased persons participate, or choose to stop participating at different rates, selection bias can occur. This bias can be insurmountable if participation is also decided as a direct result of the exposure of interest, such bias is referred to as differential selection bias [16]. Measurement error, or classification error, occurs when one or more variables (exposure, outcome, or covariates) are measured with error. This type of error can result in an information bias that can distort effect estimates, often unpredictably depending

on the magnitude of error and the variables affected. While it is an important consideration when evaluating the validity of a study, bias from measurement error will not be a focus of this dissertation.

Historically, most published epidemiologic studies include a qualitative treatment of how potential bias sources may be influencing the effect estimates in the discussion section of the paper. While this is often sufficient for small or low powered studies, larger studies and/or pooled studies warrant a quantitative treatment of potential biasing factors. Quantitative bias analysis is the quantitative (as opposed to qualitative description) treatment of uncertainty in nonrandomized research. It is often used to hypothesize the magnitude and direction of bias, and produce bias adjusted point estimates and confidence intervals that represent both the random error as well as the hypothesized systematic error in the model. There are a number of ways, ranging from simple to very complex, to carry out a bias analysis. The traditional "simple" sensitivity analysis aims to determine if the treatment-outcome association could be explained by one or more confounding variables. This sensitivity analysis is performed on the resulting association estimates for a fixed level of bias and then repeated over and over to determine the bias required to approach a null point estimate, for instance [16-19]. Adding a probabilistic component to the simple sensitivity analysis allows the prior specifications (or hypotheses about the bias effect) to take on distributions in the way real observational data would behave, and allow adjustment of systematic error as well as random error. Probabilistic sensitivity analysis can be performed using Monte Carlo methods, or in a semi- or full Bayesian analysis, if all parameters are specified by prior distributions [20-22]. Most published methods of bias analysis are based on external adjustment – which often requires either simple (but unrealistic) calculations, or very complex (although more realistic) formulas and must be repeated for each

target parameter, stratification level, treatment categorization, and so on. Additionally, bias formulas for external adjustment are always model specific – each modeling strategy requires a slightly different form of the adjustment formula [21, 23-25].

1.4. GAPS IN THE BIAS ANALYSIS LITERATURE

There are very few published instances of quantitative bias modeling in a cancer epidemiology pooling project, and none that employ a unified bias model across a consortium of investigators. Indeed, some of the methods (especially complex external adjustment) are very cumbersome and implementing them across multiple investigators in a consortium might prove difficult, especially because each study in the pool might contribute a slightly different level of bias depending on the population or study conduct. One possible solution for this would be a data augmentation approach, where the bias analysis parameters are used to simulated record-level data on missing confounding variables or selection probabilities in the database before any statistical modeling is performed. This could be performed using fixed or probabilistic bias modeling techniques, using internal or external priors for the bias parameters and Monte Carlo methods, but instead of adjusting model-based point estimates *post hoc*, each individual in the study gets *a priori* a set of augmented variable(s) that can be used in any subsequent statistical analysis. This type of "record-level" variable data imputation for bias analysis is very rare in the current epidemiologic literature. It has been briefly described [16, 26] and demonstrated as a method to adjust for misclassification bias [27, 28]. It has not been applied in a bias model for uncontrolled confounding or non-response bias; it also has not been formalized using missing data methods or the graphical language of causal theory.

1.5. THIS DISSERTATION

Large pooling projects are one of the most important areas of epidemiology for a thorough bias analysis. Unfortunately, bias analysis is uncommon in pooling projects; this is worrisome because larger sample sizes increase the chance of systematic error while decreasing those of sampling errors [16]. Additionally, accessible methodologic literature on how to handle bias in pooling projects is limited. Small p-values and tight confidence limits that come from a pooled analysis can often lead to overconfidence in results that may actually be a reflection of serious systematic bias rather than true effect estimates. Even quantitative treatment of one bias at a time might not suffice when multiple biases are acting congruently, as the combined bias often exceeds that of the individual biases. Thus, it is very important to express uncertainties in pooled estimates given the limitations associated with forming a standardized database for use in large pooling projects, and to explore these biases as acting simultaneously. Further, the bias analysis methods should be accessible to the Epidemiology community and integrated with the record level data used routinely, i.e., in the standardized database, before any statistical modeling is performed.

The risk factor profiles for rare types of endometrial cancer and for high-risk populations are not well characterized. The data source for this dissertation, the Epidemiology of Endometrial Cancer Consortium (E2C2) has been established to contribute to this characterization, yet it is likely that it may suffer from some of the limitations inherent in rare cancer pooling projects. It will be important not only to evaluate the risk factors discovered, but also to evaluate them for sensitivity to biasing factors that are a result of the pooling structure as well as the limitations of the individual contributing studies. In a large consortium such as the E2C2, with multiple

investigators using the same standardized database, it would also be beneficial to incorporate bias analysis parameters into the standardized database so that such sensitivity analyses can be decentralized and easily accessible to all contributing investigators. Therefore, the objectives of this research project will be to develop, and formalize bias modeling methods for probabilistic data imputation for analysis of uncontrolled confounding and non-response bias in rare cancer pooling projects, and to demonstrate these methods using data from the Epidemiology of Endometrial Cancer Consortium (E2C2). The project will be broken into four studies, each with its own specific aim:

1.5.1. SPECIFIC AIMS OF THE DISSERTATION

Study 1: To develop and formalize bias modeling methods by probabilistic data imputation for analysis of **uncontrolled confounding** using a missing data framework, joint probabilities, directed acyclic graphs, and simulation studies.

Study 2: To develop and formalize bias modeling methods by probabilistic data imputation for analysis of **selection (non-response) bias** using a missing data framework, joint probabilities, directed acyclic graphs, and simulation studies.

Study 3: To demonstrate the aforementioned method for addressing **bias due to uncontrolled confounding** using the Epidemiology of Endometrial Cancer Consortium (E2C2) data, by investigating the relationship between BMI and type I endometrial Cancer and potential confounding by partially measured smoking status, comorbid diagnosis of diabetes, and ever use of exogenous estrogen-only hormones.

Study 4: To demonstrate the aforementioned method to correct for **bias due to non-response** using the Epidemiology of Endometrial Cancer Consortium (E2C2) data, by investigating the relationship between BMI and type I endometrial Cancer and varying response rates in cases and non-cases for contributing case-control studies.

2. STUDY 1

SENSITIVITY ANALYSIS FOR UNCONTROLLED CONFOUNDING, WITHOUT BIAS FORMULAS

2.1. ABSTRACT

Background: Quantitative analysis of the bias due to uncontrolled confounding in observational research is becoming increasingly important, especially given the rise of "big data" in which such bias or systematic error can become greatly magnified with increasing sample sizes. Unfortunately, considerable statistical obstacles exist between performing a multivariable pre-adjusted analysis of an exposure-outcome association and a thorough quantitative bias analysis for uncontrolled confounding. Most published literature points to the use of bias formulas for external adjustment of such bias, but the formulas are often too simple or complex. We introduce and demonstrate an alternative approach that simulates or imputes the unmeasured confounding variable(s) with the working dataset for use in any subsequent analysis.

Methods: We used directed acyclic graphs, probability language, and Monte Carlo simulations to recast and impute unmeasured, uncontrolled confounding variable(s) which can be seen as missing data used to augment the observed data. We illustrate the technique for unmeasured dichotomous, trichotomous, and continuous confounding variables singly and together which were imputed using observed data and prior information of the plausible associations between the unmeasured variables and the observed data. The new augmented dataset could subsequently be used in the planned analysis stage. We illustrate this method via a series of simulation studies.

Results: In a series of illustrative simulation studies, we could recreate and control for hypothetically unmeasured confounding variables using the "true" priors for their conditional associations with the other study variables. When used with the observed data, we could estimate fully adjusted effect estimates. Once simulation is set up as imputation equations, the input priors could easily be varied or updated to reflect other and new background information.

Conclusion: Moving bias adjustment to the pre-analytic stage by first imputing the unmeasured confounder(s) using observed data and plausible priors opens the door for the augmented data set to be analyzed in any conventional way, without resorting to complex bias formulas for external adjustment.

2.2. INTRODUCTION

Quantitative analysis of the bias due to uncontrolled confounding in observational research is becoming increasingly important, especially given the rise of "big data" in which such bias or systematic error can become greatly magnified with increasing sample sizes. Unfortunately, considerable statistical obstacles exist between performing a multivariable pre-adjusted analysis of an exposure-outcome association and a thorough quantitative bias analysis for uncontrolled confounding. Most published literature points to the use of bias formulas for external formula adjustment of uncontrolled confounding [17-19, 21, 25, 29-31], but the formulas are often either simple (and thus unrealistic) or too complex (and thus unwieldy). There is thus a need for a solution that neatly integrates any planned data analysis with bias analysis, preferably within the working dataset. Moving bias adjustment to the pre-analytic stage opens the door for an augmented data set to be analyzed in any conventional way, with no need for a working knowledge of the complex methodology behind existing bias formulas for external adjustments.

We used directed acyclic graphs, probability language, and Monte Carlo simulations to recast and impute unmeasured, uncontrolled confounding variable(s) which can be seen as missing data used to augment the observed data. We illustrate the technique for unmeasured dichotomous, trichotomous, and continuous confounding variables singly and together which were imputed using observed data and prior information of the plausible associations between the unmeasured

variables and the observed data. The new augmented dataset could subsequently be used in the planned analysis stage. We illustrate this method via a series of simulation studies. The method described in this paper avoids external adjustment formulas; bias parameters are used to generate record-level data in the database before any planned statistical analysis is performed.

2.3. METHODS

We demonstrate record-level imputation for uncontrolled confounding variables. First we will define and formalize the problem of uncontrolled confounding using directed acyclic graphs (DAGs). The use of DAGs to express these causal relationships imparts a basic set of rules, which have been extensively described for use in causal analysis elsewhere [32-36]. Referring to figure 2.1, where X is an exposure, Y is the outcome of interest, and U is an unmeasured confounding variable, the relationship of interest is the effect of X on Y conditional on U.

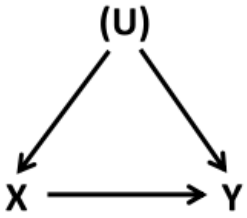

Figure 2.1 U is an unmeasured variable that confounds the relationship between X and Y.

The joint probability that can be read off the DAG in figure 2.1 is given by

(1) $P(y, x, u) = P(y|x, u)P(x|u)P(u)$

However, since U is unmeasured, we can use probability rules to rewrite the joint probability recasting U as a conditional probability involving the observed x and y:

(2) $P(y, x, u) = P(u|y, x)P(y|x)P(x)$

In equation 2 we have reordered the conditional probability equations to isolate what we need - P(u|x,y) - from what we know - P(y|x) and P(x). Having not measured U, we can redefine P(u|x, y) as a modeled expectation or probability using our measured variables X and Y, depending on the variable type. For a binary variable, the imputation would take on the following parameterization:

$$(3)\ P(u|x,y) = \text{expit}(\beta_u + \beta_{ux}x + \beta_{uy}y + \beta_{uxy}xy)$$

Where β_u is the background prevalence of U (setting X=Y=0), β_{ux} is the log odds ratio relating the U and X among Y=0, β_{uy} is the log odds ratio relating U and Y when X=0, and β_{uxy} is the logarithm of the ratio of the odds ratios (such as the ratio of the U-Y association when X=1 to the ratio of the U-Y association when X=0). Equation (3) requires a fully saturated model because of the invocation of Bayes' theorem. As we will see in the illustration, this latter point becomes important when there are multiple confounder variables in the assumed causal structure.

2.3.1. ILLUSTRATION

2.3.1.1. MONTE CARLO SIMULATIONS

We performed a series of simulation studies to demonstrate and evaluate the bias model performance under varying confounding scenarios. For each study, we simulated a large cohort (N=10,000) with one dichotomous exposure variable (X), two dichotomous confounding variables (Z_1 and Z_2), one continuous confounding variable (Z_3), one trichotomous confounding variable (Z_4), and a dichotomous outcome (Y). The data generating mechanism was based on the relationships between these variables as depicted in the causal structures in figures 2.2 and 2.3. In scenario A (figure 2.2), Y is independent of X conditional on Z_1-Z_4, and in scenario B (figure 2.3), X causes Y.

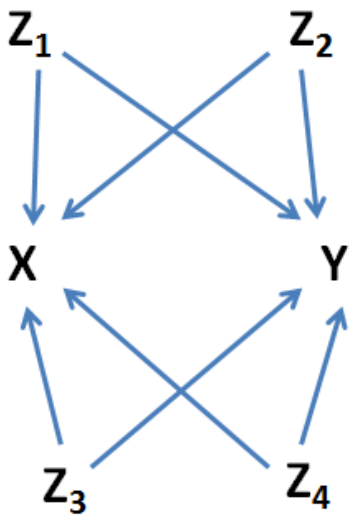

Figure 2.2 Scenario A – A DAG representing marginally independent X and Y with 4 confounding variables Z_1, Z_2, Z_3, and Z_4.

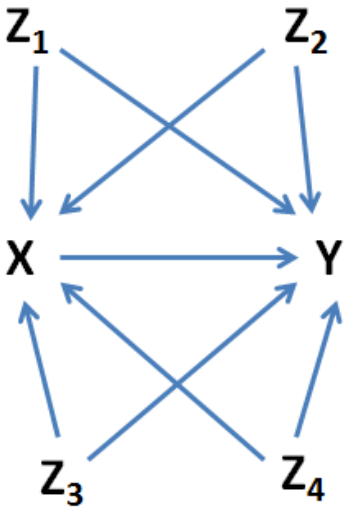

Figure 2.3 Scenario B – A DAG representing marginally dependent X and Y with 4 confounding variables Z_1, Z_2, Z_3, and Z_4.

Z_1 and Z_2 were generated by random draws from independent Bernoulli distributions with probabilities of success of $P(Z_1=1)$ and $P(Z_2=1)$ respectively. We varied $P(Z_1=1)$ and $P(Z_2=1)$ depending on the simulation trial. Z_3 was generated from the normal distribution such that $Z_3 \sim N(0, 1)$. Z_4 was generated from two conditional Bernoulli distributions such that the resulting two indicator variables combined made an exclusive trichotomous categorization with mean population distributions depending on the simulation trial. The probability of exposure was generated as a function of variables Z_1-Z_4, and the exposure variable was generated from random draws from a corresponding Bernoulli distribution. The outcome variable was generated from random draws from a Bernoulli distribution as a function of the background risk of outcome (varying according to simulation trial), the exposure status, and Z_1-Z_4.

The simulation studies were conducted in two stages: derivation of empirical priors from the simulated cohort, and then imputation of the "missing" confounder and bias modeling. For the first stage, we used the complete simulated cohort to derive prior inputs for the bias model. In the second stage, we deleted one or more known confounder variables from our dataset and attempted to re-create those data using the variable imputation method, and our empirical priors from stage 1. We then compared outcome model results between those using the true confounder and those using the imputed confounder. This was repeated for each imputation strategy (i.e., imputing a dichotomous, continuous, or trichotomous confounding variable.) There were a total of 14 trials per scenario (28 studies in all). The trials were designed to evaluate the bias model performance under varying background prevalence of exposure and outcome and prevalence and strength of confounding.

2.3.1.2. IMPUTING A DICHOTOMOUS CONFOUNDING VARIABLE

In order to derive empirical priors from our simulated cohort, Z_1 and Z_2 were each modeled twice, as a function of the other DAG variables (X, Y, Z_1, Z_3, and Z_4) according to the following fully saturated logistic regression equations.

For investigating our bias model by imputing a single dichotomous confounder:

$$
\begin{aligned}
(1)\ \mathrm{logit}(P(Z_1 = 1|x, y, z_2, z_3, z_4) \\
&= \beta_{Z_1} + \beta_{Z_1X}x + \beta_{Z_1Y}y + \beta_{Z_1Z_2}z_2 + \beta_{Z_1Z_3}z_3 + \beta_{Z_1Z_4}z_4 + \beta_{Z_1XY}xy + \beta_{Z_1XZ_2}xz_2 \\
&+ \beta_{Z_1XZ_3}xz_3 + \beta_{Z_1XZ_4}xz_4 + \beta_{Z_1YZ_2}yz_2 + \beta_{Z_1YZ_3}yz_3 + \beta_{Z_1YZ_4}yz_4 + \beta_{Z_1Z_2Z_3}z_2z_3 \\
&+ \beta_{Z_1Z_2Z_4}z_2z_4 + \beta_{Z_1Z_3Z_4}z_3z_4 + \beta_{Z_1XYZ_2}xyz_2 + \beta_{Z_1XYZ_3}xyz_3 + \beta_{Z_1XYZ_4}xyz_4 \\
&+ \beta_{Z_1XZ_2Z_3}xz_2z_3 + \beta_{Z_1XZ_2Z_4}xz_2z_4 + \beta_{Z_1XZ_3Z_4}xz_3z_4 + \beta_{Z_1YZ_2Z_3}yz_2z_3 \\
&+ \beta_{Z_1YZ_2Z_4}yz_2z_4 + \beta_{Z_1YZ_3Z_4}yz_3z_4 + \beta_{Z_1Z_2Z_3Z_4}z_2z_3z_4 + \beta_{Z_1XYZ_2Z_3}xyz_2z_3 \\
&+ \beta_{Z_1XYZ_2Z_4}xyz_2z_4 + \beta_{Z_1XYZ_3Z_4}xyz_3z_4 + \beta_{Z_1XZ_2Z_3Z_4}xz_2z_3z_4 + \beta_{Z_1YZ_2Z_3Z_4}yz_2z_3z_4 \\
&+ \beta_{Z_1XYZ_2Z_3Z_4}xyz_2z_3z_4
\end{aligned}
$$

$$
\begin{aligned}
(2)\ \mathrm{logit}(P(Z_2 = 1|x, y, z_1, z_3, z_4) \\
&= \beta_{Z_2} + \beta_{Z_2X}x + \beta_{Z_2Y}y + \beta_{Z_2Z_1}z_1 + \beta_{Z_2Z_3}z_3 + \beta_{Z_2Z_4}z_4 + \beta_{Z_2XY}xy + \beta_{Z_2XZ_1}xz_1 \\
&+ \beta_{Z_2XZ_3}xz_3 + \beta_{Z_2XZ_4}xz_4 + \beta_{Z_2YZ_1}yz_1 + \beta_{Z_2YZ_3}yz_3 + \beta_{Z_2YZ_4}yz_4 + \beta_{Z_2Z_1Z_3}z_1z_3 \\
&+ \beta_{Z_2Z_1Z_4}z_1z_4 + \beta_{Z_2Z_3Z_4}z_3z_4 + \beta_{Z_2XYZ_1}xyz_1 + \beta_{Z_2XYZ_3}xyz_3 + \beta_{Z_2XYZ_4}xyz_4 \\
&+ \beta_{Z_2XZ_1Z_3}xz_1z_3 + \beta_{Z_2XZ_1Z_4}xz_1z_4 + \beta_{Z_2XZ_3Z_4}xz_3z_4 + \beta_{Z_2YZ_1Z_3}yz_1z_3 \\
&+ \beta_{Z_2YZ_1Z_4}yz_1z_4 + \beta_{Z_2YZ_3Z_4}yz_3z_4 + \beta_{Z_2Z_1Z_3Z_4}z_1z_3z_4 + \beta_{Z_2XYZ_1Z_3}xyz_1z_3 \\
&+ \beta_{Z_2XYZ_1Z_4}xyz_1z_4 + \beta_{Z_2XYZ_3Z_4}xyz_3z_4 + \beta_{Z_2XZ_1Z_3Z_4}xz_1z_3z_4 + \beta_{Z_2YZ_1Z_3Z_4}yz_1z_3z_4 \\
&+ \beta_{Z_2XYZ_1Z_3Z_4}xyz_1z_3z_4
\end{aligned}
$$

And for imputing Z_1 and Z_2 simultaneously, we performed two steps sequentially:

$$(3)\ \text{logit}(P(Z_1 = 1|x, y, z_3, z_4)$$

$$= \beta_{Z_1} + \beta_{Z_1 X} x + \beta_{Z_1 Y} y + \beta_{Z_1 z_3} z_3 + \beta_{Z_1 z_4} z_4 + \beta_{Z_1 XY} xy + \beta_{Z_1 XZ_3} xz_3$$

$$+ \beta_{Z_1 XZ_4} xz_4 + \beta_{Z_1 YZ_3} yz_3 + \beta_{Z_1 YZ_4} yz_4 + \beta_{Z_1 z_3 z_4} z_3 z_4 + \beta_{Z_1 XYZ_3} xyz_3$$

$$+ \beta_{Z_1 XYZ_4} xyz_4 + \beta_{Z_1 XZ_3 z_4} xz_3 z_4 + \beta_{Z_1 YZ_3 z_4} yz_3 z_4 + \beta_{Z_1 XYZ_3 z_4} xyz_3 z_4$$

$$(4)\ \text{logit}(P(Z_2 = 1|x, y, z_1, z_3, z_4)$$

$$= \beta_{Z_2} + \beta_{Z_2 X} x + \beta_{Z_2 Y} y + \beta_{Z_2 z_1} z_1 + \beta_{Z_2 z_3} z_3 + \beta_{Z_2 z_4} z_4 + \beta_{Z_2 XY} xy + \beta_{Z_2 XZ_1} xz_1$$

$$+ \beta_{Z_2 XZ_3} xz_3 + \beta_{Z_2 XZ_4} xz_4 + \beta_{Z_2 YZ_1} yz_1 + \beta_{Z_2 YZ_3} yz_3 + \beta_{Z_2 YZ_4} yz_4 + \beta_{Z_2 z_1 z_3} z_1 z_3$$

$$+ \beta_{Z_2 z_1 z_4} z_1 z_4 + \beta_{Z_2 z_3 z_4} z_3 z_4 + \beta_{Z_2 XYZ_1} xyz_1 + \beta_{Z_2 XYZ_3} xyz_3 + \beta_{Z_2 XYZ_4} xyz_4$$

$$+ \beta_{Z_2 XZ_1 z_3} xz_1 z_3 + \beta_{Z_2 XZ_1 z_4} xz_1 z_4 + \beta_{Z_2 XZ_3 z_4} xz_3 z_4 + \beta_{Z_2 YZ_1 z_3} yz_1 z_3$$

$$+ \beta_{Z_2 YZ_1 z_4} yz_1 z_4 + \beta_{Z_2 YZ_3 z_4} yz_3 z_4 + \beta_{Z_2 z_1 z_3 z_4} z_1 z_3 z_4 + \beta_{Z_2 XYZ_1 z_3} xyz_1 z_3$$

$$+ \beta_{Z_2 XYZ_1 z_4} xyz_1 z_4 + \beta_{Z_2 XYZ_3 z_4} xyz_3 z_4 + \beta_{Z_2 XZ_1 z_3 z_4} xz_1 z_3 z_4 + \beta_{Z_2 YZ_1 z_3 z_4} yz_1 z_3 z_4$$

$$+ \beta_{Z_2 XYZ_1 z_3 z_4} xyz_1 z_3 z_4$$

Using Monte-Carlo methods, we generated new versions of each confounding variable, $\widehat{Z_1}$ and $\widehat{Z_2}$, by repetitively (repetitions=1,000) drawing from a Bernoulli distribution defined by functions of the retained β coefficients from equations (1-4) combined with the actual values each variable.

The resulting replicate dataset consisted of 1,000 copies of the known variables, and imputes of $\widehat{Z_1}$ and $\widehat{Z_2}$. To investigate the imputed variables performance as substitutes for the original confounder variables Z_1 and Z_2, we modeled the outcome Y as a logistic function of $\widehat{Z_1}$ and $\widehat{Z_2}$ as well as the other known variables (X, Z_3, and Z_4) by replicate.

2.3.1.3. IMPUTING A CONTINUOUS CONFOUNDING VARIABLE

In order to derive empirical priors from the simulated cohort, Z_3 was modeled as a linear function of the other DAG variables (X, Y, Z_1, Z_2, and Z_4) according to the following fully saturated expectation:

$$(5)\ z_3 = \beta_{Z_3} + \beta_{Z_3X}x + \beta_{Z_3Y}y + \beta_{Z_3XY}xy + \beta_{Z_3Z_1}z_1 + \beta_{Z_3Z_2}z_2 + \beta_{Z_3Z_4}z_4 + \beta_{Z_3XY}xy + \beta_{Z_3XZ_1}xz_1 + \beta_{Z_3XZ_2}xz_2 + \beta_{Z_3XZ_4}xz_4 + \beta_{X_3YZ_1}yz_1 + \beta_{Z_3YZ_2}yz_2 + \beta_{Z_3YZ_4}yz_4 + \beta_{Z_3Z_1Z_2}z_1z_2 + \beta_{Z_3Z_1Z_4}z_1z_4 + \beta_{Z_3Z_2Z_4}z_2z_4 + \beta_{Z_3XYZ_1}xyz_1 + \beta_{Z_3XYZ_2}xyz_2 + \beta_{Z_3XYZ_4}xyz_4 + \beta_{Z_3XZ_1Z_2}xz_1z_2 + \beta_{Z_3XZ_1Z_4}xz_1z_4 + \beta_{Z_3XZ_2Z_4}xz_2z_4 + \beta_{Z_3YZ_1Z_2}yz_1z_2 + \beta_{Z_3YZ_1Z_4}yz_1z_4 + \beta_{Z_3YZ_2Z_4}yz_2z_4 + \beta_{Z_3Z_1Z_2Z_4}z_1z_2z_4 + \beta_{Z_3XYZ_1Z_2}xyz_1z_2 + \beta_{Z_3XYZ_1Z_4}xyz_1z_4 + \beta_{Z_3XYZ_2Z_4}xyz_2z_4 + \beta_{Z_3XZ_1Z_2Z_4}xz_1z_2z_4 + \beta_{Z_3YZ_1Z_2Z_4}yz_1z_2z_4 + \beta_{Z_3XYZ_1Z_2Z_4}xyz_1z_2z_4 + \varepsilon_{Z_3}$$

Using Monte-Carlo methods, we generated new versions of the continuous confounding variable $\widehat{Z_3}$ by repetitively (repetitions=1,000) drawing from a normal distribution defined by the sum of the retained beta coefficients from equation (5) and individual data values.

The resulting replicate dataset consisted of 1,000 copies of the known variables, and the imputes of $\widehat{Z_3}$. To check the imputed variables performance as substitutes for the original confounder variable Z_3, we modeled the outcome Y as a logistic function of $\widehat{Z_3}$ as well as the other known variables (X, Z_1, Z_2, and Z_4) by replicate.

2.3.1.4. IMPUTING A TRICHOTOMOUS CONFOUNDING VARIABLE

In order to derive empirical priors from the simulated cohort, two mutually exclusive indicator variables, Z_{41} and Z_{42} were modeled as a function of the other DAG variables (X, Y, Z_1, Z_2, and Z_3) according to the following fully saturated logistic regression equations, first for all individuals:

(6) $\text{logit}(\text{P}(\text{Z}_{4_1} = 1 | \text{x}, \text{y}, \text{z}_1, \text{z}_2, \text{z}_3)$

$$= \beta_{\text{Z}_{4_1}} + \beta_{\text{Z}_{4_1}\text{X}}\text{x} + \beta_{\text{Z}_{4_1}\text{Y}}\text{y} + \beta_{\text{Z}_{4_1}\text{z}_1}\text{z}_1 + \beta_{\text{Z}_{4_1}\text{z}_2}\text{z}_2 + \beta_{\text{Z}_{4_1}\text{z}_3}\text{z}_3 + \beta_{\text{Z}_{4_1}\text{XY}}\text{xy}$$

$$+ \beta_{\text{Z}_{4_1}\text{XZ}_1}\text{xz}_1 + \beta_{\text{Z}_{4_1}\text{XZ}_2}\text{xz}_2 + \beta_{\text{Z}_{4_1}\text{XZ}_3}\text{xz}_3 + \beta_{\text{Z}_{4_1}\text{YZ}_1}\text{yz}_1 + \beta_{\text{Z}_{4_1}\text{YZ}_2}\text{yz}_2$$

$$+ \beta_{\text{Z}_{4_1}\text{YZ}_3}\text{yz}_3 + \beta_{\text{Z}_{4_1}\text{z}_1\text{z}_2}\text{z}_1\text{z}_2 + \beta_{\text{Z}_{4_1}\text{z}_1\text{z}_3}\text{z}_1\text{z}_3 + \beta_{\text{Z}_{4_1}\text{z}_2\text{z}_3}\text{z}_2\text{z}_3 + \beta_{\text{Z}_{4_1}\text{XYZ}_1}\text{xyz}_1$$

$$+ \beta_{\text{Z}_{4_1}\text{XYZ}_2}\text{xyz}_2 + \beta_{\text{Z}_{4_1}\text{XYZ}_3}\text{xyz}_3 + \beta_{\text{Z}_{4_1}\text{XZ}_1\text{Z}_2}\text{xz}_1\text{z}_2 + \beta_{\text{Z}_{4_1}\text{XZ}_1\text{Z}_3}\text{xz}_1\text{z}_3$$

$$+ \beta_{\text{Z}_{4_1}\text{XZ}_2\text{Z}_3}\text{xz}_2\text{z}_3 + \beta_{\text{Z}_{4_1}\text{YZ}_1\text{Z}_2}\text{yz}_1\text{z}_2 + \beta_{\text{Z}_{4_1}\text{YZ}_1\text{Z}_3}\text{yz}_1\text{z}_3 + \beta_{\text{Z}_{4_1}\text{YZ}_2\text{Z}_3}\text{yz}_2\text{z}_3$$

$$+ \beta_{\text{Z}_{4_1}\text{z}_1\text{z}_2\text{z}_3}\text{z}_1\text{z}_2\text{z}_3 + \beta_{\text{Z}_{4_1}\text{XYZ}_1\text{Z}_2}\text{xyz}_1\text{z}_2 + \beta_{\text{Z}_{4_1}\text{XYZ}_1\text{Z}_3}\text{xyz}_1\text{z}_3 + \beta_{\text{Z}_{4_1}\text{XYZ}_2\text{Z}_3}\text{xyz}_2\text{z}_3$$

$$+ \beta_{\text{Z}_{4_1}\text{XZ}_1\text{z}_2\text{z}_3}\text{xz}_1\text{z}_2\text{z}_3 + \beta_{\text{Z}_{4_1}\text{YZ}_1\text{z}_2\text{z}_3}\text{yz}_1\text{z}_2\text{z}_3 + \beta_{\text{Z}_{4_1}\text{XYZ}_1\text{z}_2\text{z}_3}\text{xyz}_1\text{z}_2\text{z}_3$$

And then restricted to those individuals who were not classified as Z_{4l}=1:

(7) $\text{logit}(\text{P}(\text{Z}_{4_2} = 1 | \text{x}, \text{y}, \text{z}_1, \text{z}_2, \text{z}_3)$

$$= \beta_{\text{Z}_{4_2}} + \beta_{\text{Z}_{4_2}\text{X}}\text{x} + \beta_{\text{Z}_{4_2}\text{Y}}\text{y} + \beta_{\text{Z}_{4_2}\text{z}_1}\text{z}_1 + \beta_{\text{Z}_{4_2}\text{z}_2}\text{z}_2 + \beta_{\text{Z}_{4_2}\text{z}_3}\text{z}_3 + \beta_{\text{Z}_{4_2}\text{XY}}\text{xy}$$

$$+ \beta_{\text{Z}_{4_2}\text{XZ}_1}\text{xz}_1 + \beta_{\text{Z}_{4_2}\text{XZ}_2}\text{xz}_2 + \beta_{\text{Z}_{4_2}\text{XZ}_3}\text{xz}_3 + \beta_{\text{Z}_{4_2}\text{YZ}_1}\text{yz}_1 + \beta_{\text{Z}_{4_2}\text{YZ}_2}\text{yz}_2$$

$$+ \beta_{\text{Z}_{4_2}\text{YZ}_3}\text{yz}_3 + \beta_{\text{Z}_{4_2}\text{z}_1\text{z}_2}\text{z}_1\text{z}_2 + \beta_{\text{Z}_{4_2}\text{z}_1\text{z}_3}\text{z}_1\text{z}_3 + \beta_{\text{Z}_{4_2}\text{z}_2\text{z}_3}\text{z}_2\text{z}_3 + \beta_{\text{Z}_{4_2}\text{XYZ}_1}\text{xyz}_1$$

$$+ \beta_{\text{Z}_{4_2}\text{XYZ}_2}\text{xyz}_2 + \beta_{\text{Z}_{4_2}\text{XYZ}_3}\text{xyz}_3 + \beta_{\text{Z}_{4_2}\text{XZ}_1\text{Z}_2}\text{xz}_1\text{z}_2 + \beta_{\text{Z}_{4_2}\text{XZ}_1\text{Z}_3}\text{xz}_1\text{z}_3$$

$$+ \beta_{\text{Z}_{4_2}\text{XZ}_2\text{Z}_3}\text{xz}_2\text{z}_3 + \beta_{\text{Z}_{4_2}\text{YZ}_1\text{Z}_2}\text{yz}_1\text{z}_2 + \beta_{\text{Z}_{4_2}\text{YZ}_1\text{Z}_3}\text{yz}_1\text{z}_3 + \beta_{\text{Z}_{4_2}\text{YZ}_2\text{Z}_3}\text{yz}_2\text{z}_3$$

$$+ \beta_{\text{Z}_{4_2}\text{z}_1\text{z}_2\text{z}_3}\text{z}_1\text{z}_2\text{z}_3 + \beta_{\text{Z}_{4_2}\text{XYZ}_1\text{Z}_2}\text{xyz}_1\text{z}_2 + \beta_{\text{Z}_{4_2}\text{XYZ}_1\text{Z}_3}\text{xyz}_1\text{z}_3 + \beta_{\text{Z}_{4_2}\text{XYZ}_2\text{Z}_3}\text{xyz}_2\text{z}_3$$

$$+ \beta_{\text{Z}_{4_2}\text{XZ}_1\text{z}_2\text{z}_3}\text{xz}_1\text{z}_2\text{z}_3 + \beta_{\text{Z}_{4_2}\text{YZ}_1\text{z}_2\text{z}_3}\text{yz}_1\text{z}_2\text{z}_3 + \beta_{\text{Z}_{4_2}\text{XYZ}_1\text{z}_2\text{z}_3}\text{xyz}_1\text{z}_2\text{z}_3$$

Using Monte-Carlo methods, we generated new versions of the indicator variables $\widehat{\text{Z}_{41}}$ and $\widehat{\text{Z}_{42}}$ using retained coefficients from equation (7) combined with individual level data, over 1,000 repetitions.

The resulting replicate dataset consisted of 1,000 copies of the known variables, and imputes of $\widehat{Z_{41}}$ and $\widehat{Z_{42}}$. To investigate the imputed variables performance as substitutes for the original confounder variable Z_4, we modeled the outcome Y as a logistic function of $\widehat{Z_{41}}$ and $\widehat{Z_{42}}$ as well as the other known variables (X, Z_1, Z_2, and Z_3), by replicate.

2.3.2. EVALUATING MODEL PERFORMANCE

Bias adjusted estimates were extracted by finding the median of the estimated coefficients of the target X-Y relation. These were compared to results from the true outcome model, and the unadjusted (biased) outcome model. All bias-adjusted estimates were compared to the true fully adjusted estimates to calculate bias and root mean squared error (RMSE) and 95% confidence interval coverage of the bias modeling simulations. All simulations and statistical analyses were performed using SAS (version 9.3; SAS Institute, Cary NC).

2.3.3. A NOTE ON OUR USE OF FULLY SATURATED MODELS

Non-parametrically, the fully saturated bias model form is the most accurate in recreating each unmeasured variable. However, due complexities of fitting the fully saturated models such as computational limitations, and the practical considerations of supplying external bias parameters to for 2-way, 3-way, 4-way, and 5-way product terms, we also demonstrate that a reduced model form, such as the less flexible model containing only XY product term, will be sufficient to control for bias in non-extreme scenarios. As such, the results of this paper will first demonstrate comparability between the two model forms, and subsequently, all further results will use the reduced model form.

2.4. RESULTS

Here we provide results from trials 1-5 for two scenarios, which correspond to figures 2.2 and 2.3. Table 2.1 provides actual characteristics of each cohort. Table 2.2 provides the bias model performance statistics (bias and root mean squared error) for the selected trials comparing fully saturated to unsaturated models. Selected trials using unsaturated models are provided in table 2.3. Full simulation results, including planned characteristics of each cohort and true, biased, and bias adjusted point estimates for all 24 trials (using non-saturated models) are available the appendix tables 7.1-7.6. The results were as would be expected mathematically – accurate and unbiased. Reduced model forms performed as well as the fully saturated forms, although the latter were less biased, results were comparable in simulation trials. Across all simulations, no bias levels were recorded above 0.10, and coverage was 100% for all studies performed.

Table 2.1 Actual characteristics of the trial cohorts (N=10,000)

Trial	$P_{(Z1)}$	$P_{(Z2)}$	Mean (Z3)	$SD_{(Z3)}$	$P_{(Z4)}$		$P(X_0=1)$[1]	$P(Y_0=1)$[2]	$OR_{YX\|Z1, Z2, Z3, Z4}$	$OR_{XZ1\|Y, Z2, Z3, Z4}$	$OR_{XZ2\|Y, Z1, Z3, Z4}$	$OR_{XZ3\|Y, Z1, Z2, Z4}$	$OR_{XZ4\|Y, Z1, Z2, Z3}$		$OR_{YZ1\|X, Z2, Z3, Z4}$	$OR_{YZ2\|X, Z1, Z3, Z4}$	$OR_{YZ3\|X, Z1, Z2, Z4}$	$OR_{YZ4\|X, Z1, Z2, Z3}$	
					Z4=1	Z4=2							Z4=1	Z4=2				Z4=1	Z4=2
1a	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.31	0.95	3.20	1.83	2.11	1.92	2.93	2.75	5.15	2.02	1.95	3.07
2a	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.48	0.97	3.18	1.81	2.10	1.91	2.91	2.96	5.30	1.97	2.14	2.93
3a	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.70	0.96	3.18	1.81	2.10	1.91	2.91	3.06	5.72	1.98	2.14	3.15
4a	0.29	0.30	0.01	1.00	0.40	0.30	0.49	0.31	0.97	3.06	2.00	2.07	2.15	3.54	2.73	5.13	2.01	1.94	3.06
5a	0.29	0.30	0.01	1.00	0.40	0.30	0.72	0.27	1.05	3.37	2.06	2.04	2.10	3.22	2.70	5.09	1.99	1.92	3.02
1b	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.31	1.91	2.88	1.58	1.96	1.77	2.60	2.86	4.98	1.97	1.98	2.96
2b	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.48	1.93	2.98	1.67	2.01	1.82	2.71	3.03	5.53	2.01	2.17	3.20
3b	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.70	1.94	2.63	1.60	1.89	1.94	3.07	3.35	5.91	1.99	2.13	3.08
4b	0.29	0.30	0.01	1.00	0.40	0.30	0.49	0.31	2.01	2.90	1.65	1.86	1.87	2.77	2.90	5.10	1.91	1.97	2.74
5b	0.49	0.30	0.01	1.00	0.40	0.30	0.72	0.27	2.12	2.76	4.21	1.91	1.87	2.70	2.86	5.22	1.92	2.14	2.97

[1] $P(X_0=1)$ is the background exposure prevalence $P(X=1|Z_1=Z_2=Z_3=Z_4=0)$

[2] $P(Y_0=1)$ is the background risk of disease $P(Y=1|X=Z_1=Z_2=Z_3=Z_4=0)$

Table 2.2 Bias model performance for scenarios a and b using saturated versus unsaturated models

Model form	Trial	Imputing for Z_1 Only		Imputing for Z_2 Only		Imputing for Z_1 and Z_2 simultaneously		Imputing for Z_3		Imputing for Z_4	
		Bias[1]	RMSE[2]	Bias[1]	RMSE[2]	Bias[1]	RMSE[2]	Bias[1]	RMSE[2]	Bias[1]	RMSE[2]
Saturated	1a	-0.0006	0.0500	-0.0000	0.0499	0.0267	0.0564	0.0017	0.0500	0.0006	0.0500
Unsaturated	1a	0.0020	0.0500	0.0001	0.0500	-0.0019	0.0500	0.0027	0.0500	0.0039	0.0501
Saturated	1b	-0.0024	0.0512	-0.0024	0.0512	0.0371	0.0630	0.0064	0.0515	0.0015	0.0512
Unsaturated	1b	0.0031	0.0512	-0.0052	0.0514	-0.0086	0.0519	0.0016	0.0512	0.0044	0.0513

[1] Bias=True OR – Bias adjusted OR

[2] $\text{RMSE} = \sqrt{(\text{Bias})^2 + (\text{Median SE obtained from the bias adjusted OR distributes})^2}$

Table 2.3 Bias model performance for selected trials with varying inputs, using reduced model form

Trial	Imputing for Z_1 Only		Imputing for Z_2 Only		Imputing for Z_1 and Z_2 simultaneously		Imputing for Z_3		Imputing for Z_4	
	Bias[1]	RMSE[2]	Bias[1]	RMSE[2]	Bias[1]	RMSE[2]	Bias[1]	RMSE[2]	Bias[1]	RMSE[2]
1a	0.0020	0.0500	0.0001	0.0500	-0.0019	0.0500	0.0027	0.0500	0.0039	0.0501
2a	0.0027	0.0552	-0.0028	0.0553	-0.0008	0.0552	0.0049	0.0553	0.0049	0.0554
3a	0.0054	0.0680	-0.0032	0.0680	0.0024	0.0679	0.0091	0.0683	0.0055	0.0681
4a	0.0037	0.0539	0.0051	0.0540	0.0035	0.0539	0.0052	0.0541	0.0123	0.0551
5a	0.0039	0.0650	0.0031	0.0649	0.0012	0.0649	0.0053	0.0650	0.0089	0.0654
1b	0.0031	0.0512	-0.0052	0.0514	-0.0086	0.0519	0.0016	0.0512	0.0044	0.0513
2b	0.0087	0.0605	-0.0112	0.0610	-0.0060	0.0602	0.0083	0.0604	0.0066	0.0602
3b	0.0047	0.0779	-0.0167	0.0797	-0.0057	0.0780	0.0082	0.0782	0.0082	0.0782
4b	0.0097	0.0549	0.0011	0.0541	0.0021	0.0541	0.0122	0.0554	0.0174	0.0567
5b	0.0104	0.0656	0.0016	0.0648	0.0044	0.0651	0.0154	0.0666	0.0156	0.0667

[1] Bias=True OR – Bias adjusted OR

[2] $\text{RMSE} = \sqrt{(\text{Bias})^2 + (\text{Median SE obtained from the bias adjusted OR distributes})^2}$

2.5. DISCUSSION

We have introduced a simple method of bias adjustment that is based on imputation via simulation of missing data on the unmeasured confounder(s). We used the assumed causal structure and a variety of choices for priors about the bias parameters to impute missing data on each unmeasured confounding variable for every individual in the dataset. This method is based on repairing the joint distribution using what is known about the individual to simulate (impute) what is unknown. The priors can be derived from a complete case sub-population, a prior data analysis that included the unmeasured confounding variable, or literature sources that could even be integrated to reflect differing opinions of multiple collaborating investigators. Under normal (not extreme) examples of unmeasured confounding and compatible prior specification, we demonstrated this method as valid in providing adjusted point estimates, without external adjustment formulae or other complex model-specific bias adjustment approaches. The validity was extended to imputation models that were not fully saturated for all variables. This represents a typical roadblock that may be encountered during routine use of this type of algorithm: when fully saturated models may be unwieldy or sample sizes prohibit using them, under some basic assumptions of no extreme non-null product terms, the more restrictive model forms will perform as well as the fully flexible. This algorithm provides the advantage of predicting the unmeasured confounding variable in the source dataset, before any outcome models are run, and the augmented dataset can be used in any statistical package or with any modeling strategy. This moves bias analysis tasks to the end user, and can be particularly helpful in research settings where multiple analysts are working with the same data.

Bias modeling is an exercise in quantitative skepticism that is wholly dependent on latent characteristics of a causal system – namely the assumed causal structure, prevalence of the unmeasured variable, and the magnitude and direction of the confounding bias. In this study, we used hypothetical data generated directly from our assumed causal structure and prior inputs empirically derived from the hypothetical cohort. Outside of a simulation study, these items would be specified and varied through careful consideration of expert opinion, literature review, and (ideally) statistical modeling using similar studies for comparison. Indeed, the method described here was developed for initial use in a large pooled epidemiologic study – where prior distributions can be derived by study, study type, geographical location, demographic distribution of the study population, etc., which will allow for a realistic treatment of the potential bias over a series of probabilistic sensitivity analyses. In the absence of such data this method will perform as well as external adjustment under realistic hypothetical priors – with the added benefit of being approachable to the end-user, as the complexities of this method are embedded in the imputation step, before any statistical analysis is undertaken. As with any bias modeling strategy, unrealistic priors will produce unrealistic results.

3. STUDY 2

SELECTION BIAS MODELING USING OBSERVED DATA AUGMENTED WITH IMPUTED RECORD-LEVEL PROBABILITIES

3.1. ABSTRACT

Introduction: Selection bias is a form of systematic error that can be severe in compromised study designs as in case-control studies with inappropriate selection of cases or control series (e.g., Berksonian bias or non-response bias) or in follow-up studies that suffer from extensive loss of contact with participants (e.g., loss to follow-up, follow-up bias). External adjustment for selection bias in the form of a sensitivity analysis is commonly undertaken when such bias is suspected, but methods to perform such an analysis are often complex and unwieldy. In this work, we introduce a flexible method of record-level data augmentation that can be used in both case-control and follow-up studies in order to perform sensitivity analysis for selection bias without the use of external formula.

Methods: Through a series of simulation studies, we demonstrate how investigators can use externally obtained bias parameters in easy-to-implement equations combined with data on respondents or uncensored to simulate or impute the corresponding selection probability for each respondent under the assumed selection and data generating mechanism, as would be depicted in a directed acyclic graph (DAG). Selection bias can then be adjusted using inverse probability of selection weighted fitting of any planned outcome regression.

Results: In simulation studies, we successfully demonstrated the ability to recapture the true odds ratio in an observational study analyzing only those in the selected strata (responders) by assigning weights based the probability of selection, which were simulated on the basis of an assumed causal structure.

Conclusion: We elucidated a flexible method of selection bias modeling that uses existing data and internal or external bias parameters to simulate selection. This record-level technique is

applicable to any type of observational study. It is especially desirable for use in pooled studies that combine studies that may be affected by varying levels of selection bias with other studies that may not be affected by selection bias.

3.2. INTRODUCTION

Selection bias is a form of systematic error that can be severe in compromised study designs as in case-control studies with inappropriate selection of cases or control series (e.g., Berksonian bias or non-response bias) or in follow-up studies that suffer from extensive loss of contact with participants (e.g., loss to follow-up, follow-up bias). Adjusting for selection bias in a study requires knowledge of or plausible assumptions about the factors that affect the selection mechanism. If the parameters of the selection mechanism are known or can be assumed reasonably, a selection factor can be used to adjust the biased measure of association, typically the sample odds ratio [29, 37-39]. This method is formulaic, requiring external adjustment to each outcome model in a sensitivity analysis. In studies affected by follow-up bias, inverse probability of censoring weighted (IPCW) fitting of the target model can be used to create a pseudo-population that mimics the underlying cohort (including those who were lost to follow up) [40]. This entails modeling censoring as a function of last fully observed exposure and measured risk factors that affect both censoring and the endpoint under study, which requires having the said factors measured for both the censored and uncensored. This method generates record-level selection probability and its inverse can be used as a weighting factor incorporated into the analytical dataset before any outcome models are run. A distinct advantage to record-level estimation of the selection probabilities is the possibility for a variety of bias parameters to be applied to a single or combined dataset, such as one which is composed of multiple studies. Additionally, record-level data augmentation for bias analysis can allow end-users to conduct

different analyses without and with adjustment for selection bias for different association or effect measures of interest using shared datasets, own statistical software and regression methods of choice, without resorting to cumbersome and repetitive formula-based external adjustments. Nonetheless, it is often harder for investigators who do not have additional data on variables that predict selection among the censored or non-respondents to conduct meaningful bias analysis, without resorting simplistic or unwieldy bias formulas.

In this paper, we demonstrate how investigators can use externally obtained bias parameters in easy-to-implement equations combined with data on respondents or uncensored to simulate or impute the corresponding selection probability for each respondent under the assumed selection and data generating mechanism, as would be depicted in a directed acyclic graph (DAG). Selection bias can then be adjusted using IPCW fitting of any planned outcome regression. This record-level technique is applicable to any observational study. It is especially desirable for use in pooled studies that combine studies that may be affected by varying levels of selection bias with other studies that may not be affected by selection bias. We formalize this technique using DAGs and probability and illustrate its use with a series of simulation studies.

3.3. NOTATION AND METHODS

Let X be a binary exposure, Y a binary disease outcome, **Z** be a set of confounding variables that are common causes of both X and Y, and S be a binary selection factor affected by both X and Y, such that exposure in the population can be represented by the probability of $P(X=1|\mathbf{Z}=\mathbf{z})$, prevalence of disease among the unexposed can be represented by the probability $P(Y=1|X=x, \mathbf{Z}=\mathbf{z})$, and those selected into the study population can be represented by the probability

P(S=1|X=x, Y=y, **Z=z**). Assuming no unmeasured confounding, the causal odds ratio of Y on X can be represented by the conditional odds ratio, $OR_{YX|Z}$.

In the language of DAGs, selection bias is the result of collider bias, which occurs when the exposure (or cause of the exposure) and outcome (or cause of the outcome) both directly or indirectly affect selection into the study. The use of DAGs to express these causal relationships imparts a basic set of rules that have been extensively described elsewhere [32-36]. The minimal structure for collider bias is depicted in figure 3.1.

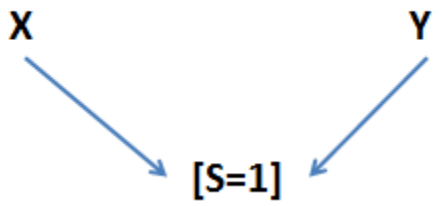

Figure 3.1 A DAG representing marginally independent but conditionally (on S=1) dependent X and Y; a simple example of collider bias.

This figure shows that the marginally independent exposure X and outcome Y can become conditionally dependent given selection S=1. Figure 3.2 shows another example.

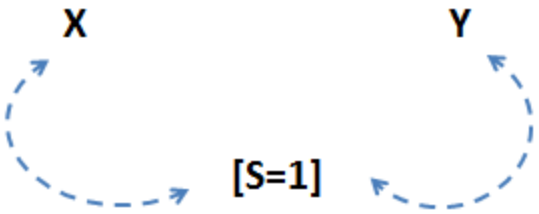

Figure 3.2 A DAG representing marginally independent but conditionally (on S=1) dependent X and Y, another example of collider bias as a result of uncontrolled common causes of X-S and Y-S

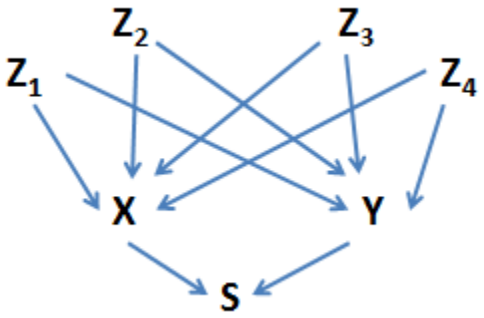

Figure 3.3 Scenario A – A DAG representing marginally independent but conditionally (on S=1) dependent X and Y, with 4 confounding variables Z_1, Z_2, Z_3, and Z_4.

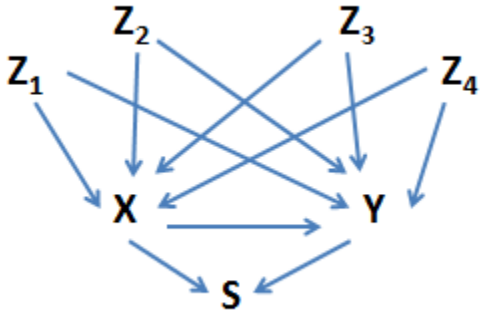

Figure 3.4 Scenario B – A DAG representing marginally dependent X and Y with additional conditional (on S=1) dependency (collider biasing path) and 4 confounding variables Z_1, Z_2, Z_3, and Z_4.

Figures 3.3 and 3.4 show scenarios where the selection caused by exposure X, confounding variable set $\mathbf{Z}$ (= Z_1, Z_2, Z_3, and Z_4) and outcome Y. In either scenario, the joint probability of S=1, y, x, and $\mathbf{z}$ is given by:

(1) $P(S = 1, y, x, \mathbf{z}) = P(S = 1|y, x, \mathbf{z})P(y|x, \mathbf{z})P(x|\mathbf{z})P(\mathbf{z})$

The term P(S=1|y,x,$\mathbf{z}$) is the probability of selection given the observed data on Y, X and $\mathbf{Z}$. To obtain the selection-bias-free joint probability P(y, x, $\mathbf{z}$) or P(S=1)P(y, x, $\mathbf{z}$), we weight the observed P(S=1, y, x , $\mathbf{z}$) by the inverse of P(S=1|y,x, $\mathbf{z}$) or P(S=1|y,x, $\mathbf{z}$)/P(S=1). This entails weighting all records in the S=1 sample by either 1/P(S=1|y,x, $\mathbf{z}$) or P(S=1)/P(S=1|y,x, $\mathbf{z}$) in a procedure known as inverse-probability-weighting. We will call this procedure inverse-probability-of-selection-weighting (IPSW) to generalize the notion of IPCW.

The conditional probability of selection P(S=1|y,x, $\mathbf{z}$) is unknown, but it can be modeled using a logistic equation with bias parameter set $\boldsymbol{\beta}$ as follows:

(2) $\text{logit}(P(S = 1|y, x, \mathbf{z};\ \boldsymbol{\beta}) = \beta_S + \beta_{SY}y + \beta_{SX}x + \beta_{SZ}\mathbf{z} + \beta_{SYX}yx + \beta_{SYZ}y\mathbf{z} + \beta_{SXZ}x\mathbf{z} + \beta_{SYXZ}yx\mathbf{z}$

where β_S is the log odds of selection S=1 when Y=0, X=0 and $\mathbf{Z}$=0 (indicating a degree of selection that is independent of Y, X, and $\mathbf{Z}$); β_{SY} is the log odds ratio (OR) relating selection S and Y when X=$\mathbf{Z}$=0; β_{SX} is the log odds ratio relating S and X when Y=$\mathbf{Z}$=0; β_{SZ} is the log odds ratio relating S and $\mathbf{Z}$ when Y=X=0; β_{SYX} is the logarithm of the ratio of (i) the odds ratio relating S and Y among X=1 and $\mathbf{Z}$=0 to (ii) the odds ratio relating S and Y among X=0 and $\mathbf{Z}$=0 (that is, $\log(OR_{SY|X=1,\mathbf{Z}=0}/OR_{SY|X=0,\mathbf{Z}=0}) = \log(OR_{SX|Y=1,\mathbf{Z}=0}/OR_{SX|Y=0,\mathbf{Z}=0})$, by the symmetry of the odds ratio); $\beta_{SY\mathbf{Z}}$ is the logarithm of the ratio of (i) the odds ratio relating S and Y when $\mathbf{Z}$=1 and

X=0 to (ii) the odds ratio relating S and Y when **Z**=0 and X=0 (that is, $\log(OR_{SY|\mathbf{Z}=1,X=0}/ OR_{SY|\mathbf{Z}=0,X=0}) = \log(OR_{S\mathbf{Z}|Y=1,X=0}/OR_{S\mathbf{Z}|Y=0,X=0})$); $\beta_{\mathbf{SXZ}}$ is the logarithm of the ratio of (i) the odds ratio relating S and X when **Z**=1 and Y=0 to (ii) the odds ratio relating S and X when **Z**=0 and Y=0 (that is, $\log(OR_{SX|\mathbf{Z}=1,Y=0}/OR_{SX|\mathbf{Z}=0,Y=0}) = \log(OR_{S\mathbf{Z}|X=1,Y=0}/OR_{S\mathbf{Z}|X=0,Y=0})$); and $\beta_{\mathbf{SYXZ}}$ is the logarithm of the ratio of two ratios, namely the ratio of (i) the ratio of the odds ratio relating S and Y when X=1 and **Z**=1 and the odds ratio relating S and Y when X=0 and **Z**= 1 to (ii) the ratio of the odds ratio relating S and Y when X=1 and **Z**=0 and the odds ratio relating S and Y when X=0 and **Z**= 0 (that is, $\log[(OR_{SY|X=1,\mathbf{Z}=1}/OR_{SY|X=0,\mathbf{Z}=1})/(OR_{SY|X=1,\mathbf{Z}=0}/OR_{SY|X=0,\mathbf{Z}=0})]$). This $\beta_{\mathbf{SYXZ}}$ is alternatively given by $\log[(OR_{SX|Y=1,\mathbf{Z}=1}/OR_{SX|Y=0,\mathbf{Z}=1})/(OR_{SX|Y=1,\mathbf{Z}=0}/OR_{SX|Y=0,\mathbf{Z}=0})] = \log[(OR_{S\mathbf{Z}|Y=1,X=1}/OR_{S\mathbf{Z}|Y=0,X=1})/(OR_{S\mathbf{Z}|Y=1,X=0}/OR_{S\mathbf{Z}|Y=0,X=0})]$.

The expit transform expit(logit(P(S=1|y,x, **z**))) yields the selection probability P(S=1|y,x, **z**) for each actually selected (S=1) record in the dataset conditional on their Y, X and Z values and given the externally obtained $\boldsymbol{\beta}$ above. An important advantage of using the logistic model to estimate the selection probability is that it will be bounded by 0 and 1, as a probability should. In some scenarios, the product term parameters might be presumed to be null, but misspecification of it as null in a selection mechanism that involves product terms might result in insufficient bias adjustment. These bias parameters should be defined using knowledge of the selection process, or the underlying source population. In most cases, these parameters will not be known, and any selection bias adjustment attempt will need to use a range of plausible priors for the bias parameters to conduct robust sensitivity analysis. We reiterate that the key difference between this technique (IPSW) and the established IPCW used in longitudinal data with censoring is that the betas or bias parameters are supplied to the dataset in our technique while they are estimated from the observed data in IPCW.

3.3.1. ILLUSTRATION 1: PROOF OF CONCENT SIMULATION USING "CORRECT" BIAS PARAMETERS

The aim of illustration 1 was to provide a proof of principle using a valid, empirically derived set of bias parameters from a hypothetical cohort in which both strata S=1 and S=0 were simulated. Using the equation in expression (2), and IPSW techniques, we demonstrate the ability to recovery of the true OR_{YX} in an analysis involving only the S=1 stratum. To do this, we simulated a large cohort study (N=100,000) with one dichotomous exposure variable (X), two dichotomous confounder variables (Z_1 and Z_2), one continuous confounder variable (Z_3), one trichotomous confounder variable (Z_4), and a dichotomous disease outcome (Y). The data generating mechanism was based on the relationships between these variables as depicted in the causal structures, figure 3.3 and 3.4. In scenario A (figure 3.3) Y is marginally independent of X, and in scenario B (figure 3.4), X causes Y.

Z_1 and Z_2 were generated by random draws from independent Bernoulli distributions with success probability of $P(Z_1=1) = 0.3$ and $P(Z_2=1)=0.3$. Z_3 was generated from the normal distribution such that $Z_3 \sim N(0, 1)$. Z_4 was generated from two conditional Bernoulli distributions such that the resulting two indicator variables combined made an exclusive categorization with mean population distributions $P(Z_4=1)=0.4$, $P(Z_4=2)=0.3$ and $P(z=0)=0.3$. The probability of exposure was generated as a function of variables Z_1-Z_4, and the exposure variable was generated from random draws from a corresponding Bernoulli distribution.

The disease variable was generated from random draws from a Bernoulli distribution as a function of the background risk of disease ($P(Y=1|X=0,Z_1=0,Z_2=0,Z_3=0,Z_4=0)=0.3$), the exposure status, and Z_1-Z_4.

Finally S was generated as by drawing from a Bernoulli distribution as a function of X and Y according to the expression (1) with varying levels of P(S=1|Y=0,X=0, Z_1=0,Z_2=0,Z_3=0,Z_4=0) according to the simulation trial.

Next, we ran logistic regression of Y on X, Z_1, Z_2, Z_3, and Z_4 for the entire cohort to determine the "true" OR relating Y and X conditional on Z_1, Z_2, Z_3, and Z_4 ($OR_{YX|\mathbf{z}}$). We then fit a binary logistic model for S=1 as a function of the other DAG variables in the full cohort, including all 2-way, 3-way, 4-way and 5-way product terms according to expression (2). We then restricted the cohort to only subjects where S=1 and ran logistic regression of Y on X, Z_1, Z_2, Z_3, Z_4 to determine the biased OR relating Y and X conditional on Z among the S=1 records, $OR_{YX|\mathbf{z},S=1}$. Finally, we generated each selected records' P(S=1|y, x, **z**) using the bias parameters (**β**) estimated from the full data as described above.

We then ran logistic regression of Y on X, Z_1, Z_2, Z_3, and Z_4 using data on the S=1 records, with 1/ P(S=1|y, x, **z**) as the regression weight to estimate the "adjusted" $OR_{YX|z,S\text{-}adj}$. We repeated this illustration for different hypothetical data with different selection bias scenarios by varying the effect of X and Y on selection. Trials 1a-8a correspond to figure 3.3, trials 1b-8b correspond to figure 3.4 with no modification by X on the S-Y relationship, and trials 1c-4c correspond to figure 3.4 with an added parameter for the modification by X on the S-Y relationship in the data generation process. We evaluated model performance by calculating bias and RMSE comparing "true" $OR_{YX|\mathbf{z}}$ and "adjusted" $OR_{YX|\mathbf{z},S\text{-}adj}$.

3.3.2. ILLUSTRATION 2: PERFORMANCE OF A REDUCED ALGORITHM

The aim of illustration 2 was to evaluate the performance of the algorithm applied in illustration 1 under less flexible equations that do not account for any 2-way, 3-way-, 4-way, 5-way, or 6-

way interaction coefficients other than β_{SYX} in the bias parameter set ($\boldsymbol{\beta}$). To this, we repeated the DAG-directed simulation of our probability selection weights for the hypothetical population described in illustration 1, excluding all interaction terms in our modeling of P(S=1) from the full cohort, using the following modified version of equation (2):

$$(3)\ \text{logit}(P(S = 1|y, x, \mathbf{z};\ \boldsymbol{\beta_r}) = \beta_S + \beta_{SY}y + \beta_{SX}x + \mathbf{z}\boldsymbol{\beta_{SZ}} + \beta_{SYX}yx$$

This resulted in a reduced bias parameter set $\boldsymbol{\beta_r}$ which was used in the IPSW process to weight the outcome model in the S=1 stratum. As in illustration 2, we ran logistic regression of Y on X, Z_1, Z_2, Z_3, and Z_4 using data on the S=1 records, with 1/ P(S=1|y, x, $\mathbf{z}$) as the regression weight to estimate the "adjusted" $OR_{YX|z,S\text{-}adj}$. We repeated this illustration for different hypothetical data with different selection bias scenarios by varying the effect of X and Y on selection. Trials 1a-8a correspond to figure 3.3, trials 1b-8b correspond to figure 3.4 with no modification by X on the S-Y relationship, and trials 1c-4c correspond to figure 3.4 with an added parameter for the modification by X on the S-Y relationship in the data generation process. We evaluated the reduced model algorithm performance by calculating bias and RMSE comparing "true" $OR_{YX|\mathbf{z}}$ and "adjusted" $OR_{YX|\mathbf{z},S\text{-}adj}$.

3.3.3. ILLUSTRATION 3: MISSPECIFIED PRIORS

The objective of illustration 3 was to demonstrate the performance of the algorithm using external bias parameters that are an imperfect measure of the true bias. We repeated the DAG-directed simulation of our probability of selection weights for a hypothetical population (N=100,000) corresponding to the DAG in figure 3.4, assuming a true causal relationship between X and Y ($OR_{YX|\mathbf{z}} = 2$). This time we applied bias parameters with slight misspecification (-20% to +20%) of the true empirical bias parameters. For this illustration, true

prevalences in the hypothetical population were held constant as follows: P(S=1) = 0.2, P(X=1|**z**=0) = 0.3 and P(Y=1|x=0, **z**=0) =0.5. The trials were performed twice, once with a strong level of selection bias: $e^{\beta_{SX}} = 5$, $e^{\beta_{SY}} = 5$ and $e^{\beta_{SYX}} = 5$ (scenario d), and once with a weak to moderate level of selection bias: $e^{\beta_{SX}} = 2$ and $e^{\beta_{SY}} = 2$ and $e^{\beta_{SYX}} = 0.8$ (scenario e). In both scenarios, these 3 parameters were "misspecified" by multiplying or dividing by 0.1 and 0.2 to represent the bias adjustment under incorrect externally applied bias parameters. This resulted in a total of 34 trials, which comprised scenarios d and e. As in scenarios a-c, we evaluated model performance by calculating bias and RMSE comparing "true" $OR_{YX|\mathbf{z}}$ and "adjusted" $OR_{YX|\mathbf{z},S\text{-adj}}$.

3.4. RESULTS

In table 3.1 we simulated populations from the DAGs pictured in figures 3.3 and 3.4 and attempted to use inverse probability weighting to correct for the selection bias effect that was the result of conditioning on the collider at the S node. All prior inputs were correctly specified from the underlying hypothetical population. Generally, we observed a downward bias in any model that included a positive relationship between exposure and selection and disease and selection. If a negative parameter were included for one of these direct effects, the bias was upward. Bias adjustment was adequate (all bias levels were <.10) in all models. Variation in the population characteristics P(S=1), P(X=1|**Z**=0), and P(Y=1|X=0, **Z**=0) did not result in any discernible pattern of bias adjustment accuracy. Increasing the $e^{\beta_{SX}}$ and $e^{\beta_{SY}}$ resulted in slightly reduced accuracy of the bias adjustment. Addition of the interaction parameter, $e^{\beta_{SYX}}$, also slightly degraded bias adjustment performance.

In table 3.2 we carried forward the simulation from DAG 3b, this time including a varying degree of misspecification of the prior inputs (-20% to +20%). We did this twice, once for a

strong selection bias (scenario d) and once for a moderate to weak selection bias (scenario e). In both scenarios, misspecification of the β_S or the $e^{\beta_{SYX}}$ parameters did not greatly inhibit bias adjustment. Misspecification of the $e^{\beta_{SX}}$ and $e^{\beta_{SY}}$ resulted in inadequate bias adjustment in the presence of strong selection bias (scenario d).

Table 3.1 Correctly specified priors for adjustment of collider bias in a cohort (N=100,000) defined by the DAGs in figure 3.3 and 3.4

Trial	P(S=1)	P(X=1|z=0)	P(y|x=0, z=0)	$e^{\beta_{SX}}$	$e^{\beta_{SY}}$	$e^{\beta_{SYX}}$	True $OR_{YX\|z}$	Biased $OR_{YX\|z,S=1}$	Bias adjusted $OR_{YX\|z,S\text{-}adj}$	Bias[1]	RMSE[2]
1a	0.10	0.3	0.5	5	5	1	1.01 (0.97, 1.04)	0.59 (0.55, 0.63)	1.01 (0.98, 1.05)	0.0037	0.0178
2a	0.20	0.3	0.5	5	5	1	1.01 (0.97, 1.04)	0.57 (0.54, 0.60)	1.01 (0.98, 1.04)	0.0014	0.0174
3a	0.50	0.3	0.5	5	5	1	1.01 (0.97, 1.04)	0.70 (0.67, 0.73)	1.01 (0.97, 1.04)	-0.0002	0.0174
4a	0.70	0.3	0.5	5	5	1	1.01 (0.97, 1.04)	0.81 (0.78, 0.84)	1.01 (0.97, 1.04)	-0.0008	0.0174
5a	0.10	0.3	0.5	0.7	0.7	1	1.01 (0.97, 1.04)	0.99 (0.88, 1.12)	1.01 (0.98, 1.05)	0.0046	0.0180
6a	0.10	0.5	0.5	10	0.5	1	1.00 (0.96, 1.03)	1.25 (1.13, 1.39)	0.99 (0.96, 1.03)	-0.0042	0.0186
7a	0.10	0.5	0.5	10	5	1	1.00 (0.96, 1.03)	0.44 (0.41, 0.47)	0.99 (0.96, 1.03)	-0.0018	0.0182
8a	0.10	0.5	0.5	10	10	1	1.00 (0.96, 1.03)	0.33 (0.30, 0.35)	0.99 (0.96, 1.03)	-0.0032	0.0183
1b	0.10	0.3	0.5	5	5	1	1.97 (1.90, 2.05)	1.14 (1.06, 1.23)	1.98 (1.90, 2.05)	0.0026	0.0190
2b	0.20	0.3	0.5	5	5	1	1.97 (1.90, 2.05)	1.10 (1.04, 1.17)	1.97 (1.90, 2.05)	0.0017	0.0189
3b	0.50	0.3	0.5	5	5	1	1.97 (1.90, 2.05)	1.37 (1.31, 1.43)	1.97 (1.90, 2.05)	-0.0001	0.0188
4b	0.70	0.3	0.5	5	5	1	1.97 (1.90, 2.05)	1.58 (1.52, 1.65)	1.97 (1.90, 2.05)	-0.0015	0.0189
5b	0.10	0.3	0.5	0.7	0.7	1	1.97 (1.90, 2.05)	1.96 (1.71, 2.24)	1.99 (1.92, 2.06)	0.0144	0.0237
6b	0.10	0.5	0.5	10	0.5	1	1.97 (1.90, 2.04)	2.47 (2.22, 2.75)	1.97 (1.90, 2.04)	0.0009	0.0189
7b	0.10	0.5	0.5	10	5	1	1.97 (1.90, 2.04)	0.86 (0.79, 0.93)	1.96 (1.89, 2.04)	-0.0037	0.0193
8b	0.10	0.5	0.5	10	10	1	1.97 (1.90, 2.04)	0.64 (0.59, 0.69)	1.96 (1.89, 2.04)	-0.0050	0.0196
1c	0.20	0.3	0.5	5	5	0.4	1.97 (1.90, 2.05)	0.91 (0.86, 0.96)	1.97 (1.90, 2.05)	-0.0003	0.0188
2c	0.20	0.3	0.5	5	5	0.8	1.97 (1.90, 2.05)	1.07 (1.00, 1.13)	1.97 (1.90, 2.05)	0.0018	0.0189
3c	0.20	0.3	0.5	5	5	2.0	1.97 (1.90, 2.05)	1.19 (1.12, 1.26)	1.97 (1.90, 2.05)	0.0017	0.0189
4c	0.20	0.3	0.5	5	5	5.0	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.97 (1.90, 2.05)	0.0018	0.0189

[1] Bias=True OR – Bias adjusted OR

[2] $\text{RMSE} = \sqrt{(\text{Bias})^2 + (\text{Median SE obtained from the bias adjusted OR distributes})^2}$

Table 3.2 Reduced models with correctly specified priors for adjustment of collider bias in a cohort (N=100,000) defined by the DAGs in figure 3.3 and 3.4

Trial	P(S=1)	P(x=1|z=0)	P(Y=1|X=0, Z=0)	$e^{\beta_{SX}}$	$e^{\beta_{SY}}$	$e^{\beta_{SYX}}$	True $OR_{YX\|z}$	Biased $OR_{YX\|z,S=1}$	Bias adjusted $OR_{YX\|z,S\text{-}adj}$	Bias[1]	RMSE[2]
1a	0.10	0.3	0.5	5	5	1	1.01 (0.97, 1.04)	0.59 (0.55, 0.63)	1.00 (0.97, 1.04)	-0.0047	0.0181
2a	0.20	0.3	0.5	5	5	1	1.01 (0.97, 1.04)	0.57 (0.54, 0.60)	1.00 (0.97, 1.04)	-0.0055	0.0183
3a	0.50	0.3	0.5	5	5	1	1.01 (0.97, 1.04)	0.70 (0.67, 0.73)	1.01 (0.98, 1.04)	0.0014	0.0174
4a	0.70	0.3	0.5	5	5	1	1.01 (0.97, 1.04)	0.81 (0.78, 0.84)	1.01 (0.98, 1.05)	0.0042	0.0179
5a	0.10	0.3	0.5	0.7	0.7	1	1.01 (0.97, 1.04)	0.99 (0.88, 1.12)	1.04 (1.00, 1.07)	0.0265	0.0317
6a	0.10	0.5	0.5	10	0.5	1	1.00 (0.96, 1.03)	1.25 (1.13, 1.39)	0.98 (0.94, 1.01)	-0.0162	0.0243
7a	0.10	0.5	0.5	10	5	1	1.00 (0.96, 1.03)	0.44 (0.41, 0.47)	0.97 (0.94, 1.01)	-0.0230	0.0293
8a	0.10	0.5	0.5	10	10	1	1.00 (0.96, 1.03)	0.33 (0.30, 0.35)	0.97 (0.94, 1.01)	-0.0244	0.0304
1b	0.10	0.3	0.5	5	5	1	1.97 (1.90, 2.05)	1.14 (1.06, 1.23)	1.94 (1.87, 2.01)	-0.0315	0.0367
2b	0.20	0.3	0.5	5	5	1	1.97 (1.90, 2.05)	1.10 (1.04, 1.17)	1.95 (1.88, 2.03)	-0.0191	0.0268
3b	0.50	0.3	0.5	5	5	1	1.97 (1.90, 2.05)	1.37 (1.31, 1.43)	1.98 (1.91, 2.05)	0.0042	0.0193
4b	0.70	0.3	0.5	5	5	1	1.97 (1.90, 2.05)	1.58 (1.52, 1.65)	1.98 (1.91, 2.05)	0.0040	0.0192
5b	0.10	0.3	0.5	0.7	0.7	1	1.97 (1.90, 2.05)	1.96 (1.71, 2.24)	2.05 (1.97, 2.12)	0.0731	0.0755
6b	0.10	0.5	0.5	10	0.5	1	1.97 (1.90, 2.04)	2.47 (2.22, 2.75)	1.93 (1.86, 2.01)	-0.0335	0.0385
7b	0.10	0.5	0.5	10	5	1	1.97 (1.90, 2.04)	0.86 (0.79, 0.93)	1.90 (1.83, 1.97)	-0.0672	0.0698
8b	0.10	0.5	0.5	10	10	1	1.97 (1.90, 2.04)	0.64 (0.59, 0.69)	1.91 (1.84, 1.98)	-0.0576	0.0607
1c	0.20	0.3	0.5	5	5	0.4	1.97 (1.90, 2.05)	0.91 (0.86, 0.96)	1.91 (1.84, 1.98)	-0.0605	0.0634
2c	0.20	0.3	0.5	5	5	0.8	1.97 (1.90, 2.05)	1.07 (1.00, 1.13)	1.92 (1.85, 1.99)	-0.0530	0.0563
3c	0.20	0.3	0.5	5	5	2.0	1.97 (1.90, 2.05)	1.19 (1.12, 1.26)	1.92 (1.85, 2.00)	-0.0490	0.0525
4c	0.20	0.3	0.5	5	5	5.0	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.93 (1.86, 2.01)	-0.0405	0.0447

[1] Bias=True OR – Bias adjusted OR

[2] $\text{RMSE} = \sqrt{(\text{Bias})^2 + (\text{Median SE obtained from the bias adjusted OR distributes})^2}$

Table 3.3 Misspecified priors for adjustment of collider bias in a cohort (N=100,000) defined by the DAG in figure 3.4[1]

Trial	True Bias Parameters				Mis-specified Parameter	Degree of mis-specification	True $OR_{YX|z}$	Biased $OR_{YX|z,\ S=1}$	Bias adjusted $OR_{YX|z,S\text{-}adj}$	Bias[2]	RMSE[3]
	β_S	$e^{\beta_{SX}}$	$e^{\beta_{SY}}$	$e^{\beta_{SYX}}$							
1d	0.20	5	5	5	None	None	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.97 (1.90, 2.05)	0.0018	0.0189
2d	0.20	5	5	5	β_S	-20%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.98 (1.91, 2.06)	0.0106	0.0227
3d	0.20	5	5	5	β_S	-10%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.98 (1.91, 2.06)	0.0114	0.0225
4d	0.20	5	5	5	β_S	+10%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.95 (1.89, 2.03)	-0.0187	0.0260
5d	0.20	5	5	5	β_S	+20%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.92 (1.86, 1.99)	-0.0502	0.0532
6d	0.20	5	5	5	$e^{\beta_{SX}}$	-20%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.72 (1.66, 1.79)	-0.2498	0.2505
7d	0.20	5	5	5	$e^{\beta_{SX}}$	-10%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.85 (1.78, 1.92)	-0.1237	0.1251
8d	0.20	5	5	5	$e^{\beta_{SX}}$	+10%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	2.10 (2.02, 2.18)	0.1253	0.1268
9d	0.20	5	5	5	$e^{\beta_{SX}}$	+20%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	2.22 (2.14, 2.30)	0.2455	0.2462
10d	0.20	5	5	5	$e^{\beta_{SY}}$	-20%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.72 (1.66, 1.78)	-0.2546	0.2553
11d	0.20	5	5	5	$e^{\beta_{SY}}$	-10%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.85 (1.78, 1.92)	-0.1260	0.1274
12d	0.20	5	5	5	$e^{\beta_{SY}}$	+10%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	2.10 (2.02, 2.18)	0.1274	0.1288
13d	0.20	5	5	5	$e^{\beta_{SY}}$	+20%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	2.22 (2.14, 2.31)	0.2493	0.2500
14d	0.20	5	5	5	$e^{\beta_{SYX}}$	-20%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	2.00 (1.93, 2.08)	0.0305	0.0358
15d	0.20	5	5	5	$e^{\beta_{SYX}}$	-10%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.99 (1.92, 2.06)	0.0147	0.0239
16d	0.20	5	5	5	$e^{\beta_{SYX}}$	+10%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.96 (1.89, 2.04)	-0.0087	0.0207
17d	0.20	5	5	5	$e^{\beta_{SYX}}$	+20%	1.97 (1.90, 2.05)	1.24 (1.17, 1.32)	1.96 (1.89, 2.04)	-0.0087	0.0207
1e	0.20	2	2	0.8	None	None	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	1.98 (1.91, 2.06)	0.0112	0.0219
2e	0.20	2	2	0.8	β_S	-20%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	1.97 (1.90, 2.05)	0.0009	0.0206
3e	0.20	2	2	0.8	β_S	-10%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	1.98 (1.90, 2.06)	0.0067	0.0208
4e	0.20	2	2	0.8	β_S	+10%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	1.99 (1.92, 2.06)	0.0144	0.0230
5e	0.20	2	2	0.8	β_S	+20%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	1.99 (1.92, 2.06)	0.0165	0.0237
6e	0.20	2	2	0.8	$e^{\beta_{SX}}$	-20%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	1.95 (1.88, 2.02)	-0.0218	0.0284
7e	0.20	2	2	0.8	$e^{\beta_{SX}}$	-10%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	1.97 (1.90, 2.04)	-0.0055	0.0193
8e	0.20	2	2	0.8	$e^{\beta_{SX}}$	+10%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	2.00 (1.93, 2.08)	0.0283	0.0341

Trial	True Bias Parameters				Mis-specified Parameter	Degree of mis-specification	True $OR_{YX|z}$	Biased $OR_{YX|z,\ S=1}$	Bias adjusted $OR_{YX|z,S\text{-}adj}$	Bias[2]	RMSE[3]
	β_S	$e^{\beta_{SX}}$	$e^{\beta_{SY}}$	$e^{\beta_{SYX}}$							
9e	0.20	2	2	0.8	$e^{\beta_{SX}}$	+20%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	2.02 (1.94, 2.10)	0.0457	0.0496
10e	0.20	2	2	0.8	$e^{\beta_{SY}}$	-20%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	1.95 (1.88, 2.03)	-0.0190	0.0266
11e	0.20	2	2	0.8	$e^{\beta_{SY}}$	-10%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	1.97 (1.90, 2.04)	-0.0041	0.0191
12e	0.20	2	2	0.8	$e^{\beta_{SY}}$	+10%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	2.00 (1.93, 2.08)	0.0268	0.0328
13e	0.20	2	2	0.8	$e^{\beta_{SY}}$	+20%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	2.02 (1.94, 2.09)	0.0427	0.0467
14e	0.20	2	2	0.8	$e^{\beta_{SYX}}$	-20%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	1.92 (1.85, 2.00)	-0.0495	0.0529
15e	0.20	2	2	0.8	$e^{\beta_{SYX}}$	-10%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	1.95 (1.88, 2.03)	-0.0196	0.0272
16e	0.20	2	2	0.8	$e^{\beta_{SYX}}$	+10%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	2.02 (1.94, 2.09)	0.0428	0.0467
17e	0.20	2	2	0.8	$e^{\beta_{SYX}}$	+20%	1.97 (1.90, 2.05)	1.63 (1.53, 1.75)	2.02 (1.94, 2.09)	0.0428	0.0467

[1]Simulated prevalences (probabilities) in the hypothetical population were held constant as follows: P(S=1) = 0.2, P(X=1|**Z**=0) = 0.3 and P(Y=1|X=0, **Z**=0) = 0.5. The true $OR_{YX|Z}$ was simulated as 2.0.
[2]Bias=True OR – Bias adjusted OR
[3] $\text{RMSE} = \sqrt{(\text{Bias})^2 + (\text{Median SE obtained from the bias adjusted OR distributes})^2}$

3.5. DISCUSSION

We have demonstrated a method of selection bias control using record level data augmentation instead of external formula adjustment, which is based on principals of DAGs and the recoverability of the odds ratio from an identified causal structure [41]. We used inverse probability of selection weighting (IPSW) to create a pseudo population that resembled both the selected and non-selected strata, and were able to produce unbiased estimates of the causal odds ratio using only the selected stratum. This method is distinct from IPCW because it need not be based on data from censored individuals in the underlying cohort, and thus may be applicable to case-control studies. Another benefit to this method is that the individual level data augmentation is flexible to allow for varying bias parameters, which is especially advantageous in combined data sources where some, but not all, studies are suspected to be affected by selection bias.

Simulation scenarios demonstrate adequate performance of this bias adjustment method under empirically derived priors, but the simple framework of the method lends itself easily to use of external bias parameters. Performance is best using fully saturated models, but the reduced model forms perform comparably, and with much simpler computational execution. Application of this method under prior misspecification demonstrated that (as would be expected intuitively) reweighting the population according to bias parameters that are slightly invalid produces invalid results.

Although this method performs adequately in our simulation scenarios, it is highly dependent on accurate or plausible characterization of the magnitude and direction of the bias, most of which we derived empirically from the underlying source population. Under extreme levels of

selection bias, upon even slight misspecification of these parameters the bias adjustment degrades, and although we did not present examples of it, gross misspecification, or misspecification of multiple priors simultaneously could result in entirely invalid adjusted estimates. Even in a detailed sensitivity analysis, inference would be speculative at best if little is known about the selection forces in the underlying population.

We detected a discernible pattern of bias direction in our simulations. When both the exposure and disease were positively associated with selection, the bias direction was downward. When one was positive and the other was negative, the bias direction was upward. If the overall magnitude of bias was small, this rule of directionality was not as evident. A thorough evaluation of the expected magnitude and direction of selection bias has not yet been published in the epidemiologic literature. Suspected examples of severe Berksonian bias have been shown to cause extreme downward bias, to 10-fold decrease in effect estimate [42]. Exploration of the potential impact of selection bias in the EMF leukemia literature demonstrated that this type of bias could result in a 2-fold increase in effect estimates [43]. Further research and simulation studies may be warranted in this area, especially to uncover the pre-requisites for downward vs. upward bias in collider-type selection bias.

4. STUDY 3

CONFOUNDING BIAS DUE TO UNMEASURED LIFESTYLE AND HORMONAL CHARACTERISTICS IN THE INVESTIGATION OF BMI AND TYPE I ENDOMETRIAL CANCER: THE EPIDEMIOLOGY OF ENDOMETRIAL CANCER CONSORTIUM (E2C2)

4.1. ABSTRACT

Introduction: Modern epidemiologic investigations of the etiology of rare cancers have given rise to a flurry of cancer pooling projects, in which multiple studies are combined to enhance precision and investigate rare exposures and sub-types of disease. Pooling data from various studies can introduce multiple forms of systematic error as a result of combining studies of different design types, which utilize different methods of variable measurement, or including one or more studies with missing important confounding variables. As such, when conducting a large consortium studies, some attention should be paid to conducting a meaningful quantitative sensitivity analysis for bias as a way to guide qualitative interpretation of results. In this paper, we consider the issue of unmeasured confounding in the Epidemiology of Endometrial Cancer Consortium (E2C2), and demonstrate the use of a simple algorithm that allows for a quantitative bias model to be implemented at the record-level in a study database, rather than by employing complex, unwieldy bias formulas for external adjustment.

Methods: We used directed acyclic graphs (DAGs) and Monte-Carlo methods to perform empirical prior model-based simulation and subsequent imputation of three partially measured important confounding variables (smoking status, ever use of estrogen-only hormone replacement therapy, and diabetes) for the relationship between BMI and type I endometrial cancer (EC) in the E2C2 pooled database. We performed routine, complete case statistical analysis of the relationship between BMI and type I EC and compared this to bias adjusted estimates generated using the imputed data.

Results: After adjustment for each confounding variable individually, and all three simultaneously, we noted small to moderate attenuations in the odds ratio estimates for the BMI

type I EC relationship in pooled and by-study analyses. These attenuations were most notable in studies with larger proportions of missing confounding variables.

Conclusion: We provide a very detailed demonstration of the bias model performance in a large, multi-study epidemiologic investigation of endometrial cancer. We also found evidence that, assuming a valid causal model, the relationship between BMI and type I EC in this study is robust to the influence of partially measured confounding by smoking, estrogen-only hormone replacement therapy, and diagnosis of diabetes.

4.2. INTRODUCTION

Given the rarity of some cancers, epidemiologists and other health researchers often resort to pooling data to achieve the power needed for studies of rare outcomes associations. A pooled data analysis is the combining of raw data from multiple studies, recoding to fit a single pooled "standardized" database, and then analyzing all studies as one. This is different from a traditional meta-analysis of point estimates, which involves extracting point estimates from published literature and then summarizing over the extracted point estimates. The advantages of pooling are vast. Consortium projects that combine multiple studies can vastly increase power and efficiency of their desired investigations. This allows for less prevalent outcomes to be studied as risk factors, as well as the investigation of effect heterogeneity. Additionally, pooling projects allow for the study of risk factors in populations that are often underrepresented in epidemiology, such as minority ethnicities. When compared with the meta-analysis of point estimates, standardized pooling allows for better control of systematic biases, as well as careful investigation and explanation of the heterogeneity within and across studies. Pooling can also provide an opportunity to estimate effect measures that may not have been reported originally for

all the individual studies [15]. Pooling data from various studies and sources is not without its limitations, however. Studies of different design types are often combined and treated as the same design, methods of variable measurement may vary from study to study, and important risk factors or confounding variables may not be collected in all of the pooled studies. The creation of a standardized database may result in the loss of information when partially unmeasured variables are dropped across studies, or when categories are collapsed to fit a standardized framework of variable definition. All of these limitations are likely to result in bias.

Historically, most published epidemiologic studies include a qualitative treatment of how potential bias sources may be influencing the association estimates in the discussion section of the papers. While this is often sufficient for small or low powered studies, larger or pooled studies can benefit from a quantitative treatment of potential bias due to uncontrolled confounding, measurement error or selective (non)response. Unfortunately, accessible methodologic literature on how to handle bias in pooling projects is limited. Indeed, some of the bias formulas for external adjustment are very cumbersome and implementing them across multiple investigators in a consortium might prove difficult, especially because each study in the pool might contribute a slightly different level of bias depending on the population or study conduct.

In this paper, we consider the issue of uncontrolled confounding in cancer pooling studies. Recognizing that uncontrolled confounding can be cast a missing data problem [5], one possible solution to using flexible but unwieldy external adjustment bias formulas is to create an imputation or simulation model that generates the unmeasured confounding variables for each record in the standardized database that can then be used like other variables in subsequent statistical analysis undertaken by the consortium. Previous work has described a simple Monte

Carlo algorithm based on assumed underlying causal structure and 'repair' of the joint probability distribution of the unmeasured confounding variables given the observed data and assumed relationships between the unmeasured and measured variables. The technique is flexible to the varying study designs and levels of bias that may be encountered in a data pooling project and can be applied to the shared database prior to running any statistical models [Thompson CA, Arah OA: Sensitivity analysis for uncontrolled confounding, without bias formulas (manuscript)]. The purpose of this paper is to demonstrate this novel method of adjustment for uncontrolled confounding, using data from the Epidemiology of Endometrial Cancer Consortium (E2C2), and to evaluate the relationship between BMI and type I endometrial cancer, before and after adjustment for unmeasured confounding by smoking status, ever use of estrogen-only hormone replacement therapy, and the comorbid diagnosis of diabetes.

4.3. MATERIALS AND METHODS

4.3.1. DATA SOURCE: THE EPIDEMIOLOGY OF ENDOMETRIAL CANCER CONSORTIUM (E2C2)

The Epidemiology of Endometrial Cancer Consortium (E2C2) is a formally designated NCI consortium to study rare cancer. The objective of the consortium is to pool as many studies of endometrial cancer as possible, with goals to study genetic variation, gene-gene interactions, gene-environment interactions, diet, and risk profiles of rarer histologic subtypes and underrepresented minorities [44]. This project includes 24 studies, 10 prospective cohort studies and 14 case-control studies. Participating cohort studies were included in the database via nested case-control sampling: each case and up to four randomly selected controls (who had an intact uterus and no evidence of endometrial cancer at the date of the index case diagnosis) were matched on age, and in some cases other variables such as race. In total, there are 17,000 endometrial cancer cases and 39,384 control subjects. Studies vary in their design, population

size, matching variables, sampling strategies, inclusion criteria, population ethnicities, participant ages, matching criteria, and recruitment periods. The dataset utilized for this study is the core risk factors database, which includes all E2C2 study participants and values for all major risk factors for endometrial cancer (BMI, weight change, reproductive variables, menopausal status, use of hormone replacement therapy and oral contraception, smoking, chronic comorbidities, physical activity), as well as demographic characteristics. Some important risk factors are not collected in every study however, which makes this pooling project an ideal example for exploration of new bias analysis techniques. For example one large case control study does not have information on smoking, and several studies do not have data on three important confounding variables: comorbidity of diabetes, or ever use of estrogen-only hormone replacement therapy. A list of participating studies including study type, location, sample sizes, ethnicity distributions, mean age of cases and controls, and extent of missingness for these three confounding variables is provided in table 4.1.

Table 4.1 Participating studies: location and population description

Study Site	Location	% White	Total No Cases	Mean age cases (SD)	Total no Controls	Mean age controls	% missing Smoking	% missing EHRT	% missing Diabetes
Cohort Studies									
MEC/USC	Hawaii, California	26.7	515	65.5 (8.6)	2623	65.9 (8.7)	2.1	5.5	0.0
NHS	11 US states	94.9	581	61.1 (8.3)	1641	60.7 (8.1)	2.6	0.0	0.0
CSDLH	Canada	96.8	643	60.0 (8.1)	3072	59.9 (6.4)	0.0	24.3	100.0
CPS-II	21 US states	98.7	572	69.0 (6.4)	2664	63.3 (6.2)	1.9	3.2	14.0
NLCS	Netherlands	100.0	402	70.0 (6.0)	864	69.6 (6.0)	0.0	4.4	0.0
CTS	California	91.2	682	65.6 (10.7)	3010	65.7 (10.6)	0.4	6.8	0.0
IWHS	Iowa	98.4	466	71.5 (6.4)	2212	71.7 (6.5)	1.5	29.6	0.8
NIH-AARP	8 US areas	93.5	1506	67.5 (5.8)	7400	67.5 (5.8)	2.9	19.0	0.0
SMC	Sweden	100.0	329	70.0 (9.3)	1412	70.3 (9.3)	2.8	46.5	94.7
BCDDP	29 US clinics	93.2	423	64.4 (7.7)	2418	65.6 (8.3)	14.0	49.7	19.5
Case control Studies									
Edge	New Jersey	90.2	418	60.6 (9.8)	467	64.2 (11.4)	0.0	0.0	0.0
WISE	Philadelphia	79.2	552	62.6 (8.1)	1583	61.3 (8.1)	0.1	28.3	0.0
Hawaii case-control	Hawaii	23.5	432	57.5 (12.0)	511	56.6 (12.2)	2.9	37.3	2.9
SECS	China	0.0	1071	54.5 (8.5)	1212	54.6 (8.5)	0.0	0.0	0.9
PECS	Poland	100.0	435	60.9 (8.1)	1925	56.3 (10.2)	0.0	0.1	0.9
US Case-control	5 US clinics	92.2	332	59.3 (10.1)	320	57.9 (10.5)	3.7	13.5	0.8
Alberta	Canada	94.8	474	58.3 (9.5)	1032	58.1 (10.1)	0.0	45.0	0.1
BAWHS	California	90.3	429	61.7 (9.9)	470	61.6 (10.7)	100.0	5.9	100.0
USC LA	Los Angeles	100.0	787	63.0 (5.3)	791	63.1 (5.4)	0.0	0.0	0.0
ANECS	Australia	88.3	740	60.7 (9.4)	1125	61.1 (9.9)	0.1	14.8	0.2
PEDS	New York	97.0	541	62.4 (11.3)	468	63.2 (11.1)	0.2	12.1	89.1
WNYDS	New York	100.0	232	63.5 (9.4)	639	55.9 (10.6)	0.0	0.0	0.1
Turin	Italy	100.0	249	61.3 (7.4)	307	60.4 (7.7)	4.7	24.5	4.9
CECS	Connecticut	93.4	588	60.2 (9.6)	665	61.5 (10.8)	0.0	2.8	0.0

The primary objective of this study to demonstrate the use of record-level bias adjustment in the E2C2 risk factors database. We have chosen BMI as the exposure of interest, and we will focus on type I endometrial cancer cases only. We will attempt to adjust for three partially, measured confounders: smoking status (never smoker, previous smoker, current smoker), ever use estrogen-only hormone replacement (EHRT) therapy and diagnosis of the comorbid condition of diabetes. Based on our understanding of the disease mechanism, we expect the data generating mechanism to resemble the DAG in figure 4.1.

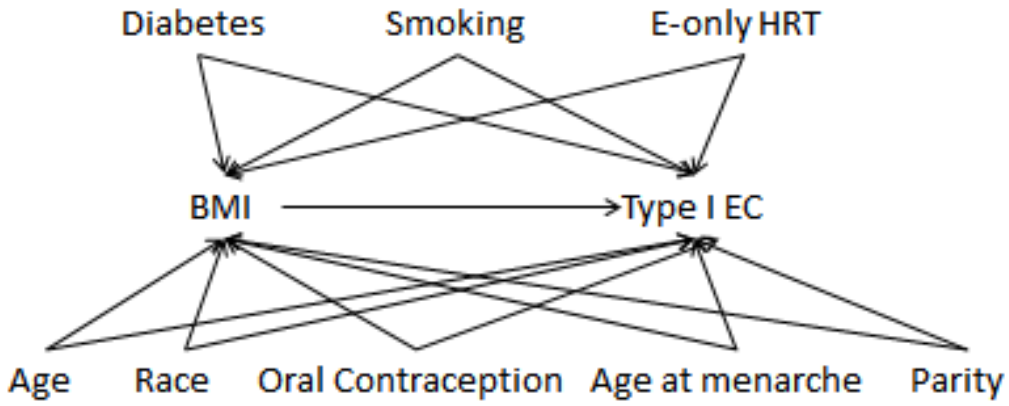

Figure 4.1 Suspected measured and partially unmeasured confounding variables in the relationship between BMI and type I endometrial cancer (EC)

4.3.2. SIMULATING UNMEASURED CONFOUNDING VARIABLES

The bias adjustment process used in this study is based on probabilistic missing data imputation algorithms based on variable simulation and resampling, for sensitivity analysis of uncontrolled confounding. The method is based on the idea that uncontrolled confounding can be seen as missing data [16, 45]. Uncontrolled confounding is an example of all study participants missing data on one or more variables needed for confounding control. Selective nonresponse is an example of missing all data for invited study participants who did not respond or dropped out. Extending this, bias analysis can be recast as a missing data problem, and approached through

common methods for data imputation or simulation, provided the analysis is performed under the right assumptions [30]. Conventional statistical analysis involves data reduction to equations in order to estimate their parameters, but it is also possible to use those equations and parameters to produce the data that would have yielded these equations had they not been missing. This perspective allows us to generate the missing data that would have been observed had unmeasured confounders been absent under the assumed causal structure. This is done by repetitively sampling from the "repaired" joint distribution of the observed data and the unmeasured confounding variables. This joint distribution is "repaired" using the pre-specified equations based on the bias parameters linking the missing data to the non-missing data.

The underlying causal structure used to guide imputation of the missing confounding variables will be visualized using directed acyclic graphs, or DAGs for short. DAGs are directed, not cyclic, path diagrams that depict the variables and relationships between those variables in a causal epidemiologic investigation. The use of DAGs to express these causal relationships imparts a basic set of rules, which have been extensively described for use in causal analysis elsewhere [32-36]. Briefly, in a DAG, nodes are occupied by variables. An arrow originating from a node or variable X and pointing to another node Y indicates a direct causal relationship between X and Y. Y is also a child or consequence of X. If variables X and Y are caused by another variable Z, Z is a common cause of X and Y and thus confounding variable for the relationship between X and Y. This common cause path starting from X through Z and on to Y is a biasing path between X and Y as it does not represent a (direct or indirect) causal effect of X on Y. Without controlling for the confounding variable Z, the path between X and Y is open and biasing.

The data source for this study, the E2C2, lends itself particularly well to the record-level variable simulation for eventual bias adjustment because it is a very large database, and all target

confounding variables are partially unmeasured. As such, the biasing parameters can be derived empirically from the complete case subjects in the database using a carefully specified model designed to capture as much of the variability in the confounding variables as possible. The parameters from this model can then be utilized in a Monte-Carlo simulation step that fills in values for the unmeasured confounders repetitively based on the empirical bias parameters and the values of the measured variables (i.e., using what has been measured to work back to what was unmeasured). Averaging across the repetitions will allow us to repair the joint distribution of the full covariate list, as if the partially measured confounding variables had been measured in full.

4.3.3. STATISTICAL ANALYSES

We restricted the data to type I endometrial cancer cases and controls. After describing the analytical dataset by study, and participant, including details on the extent of missingness in the imputation variables estrogen-only hormone replacement therapy, smoking status and diagnosis of diabetes, we generated priors empirically from the data on complete subjects using the following DAG variables (age at reference, BMI, race, age at menarche, parity, ever use of estrogen-only hormone replacement therapy, ever use of oral contraception, smoking status, and case/control status) and all 2-way interactions involving categorical or binary variables to model our three imputation variables: smoking status, ever use of estrogen-only hormone replacement therapy, and diagnosis of diabetes. All three models were fit by study type (case-control or cohort) because we assumed the bias parameters to be distinct by this variable, and they included a random effect for study site. For ever use of estrogen-only hormone replacement therapy and diabetes status, we used a mixed binomial logistic regression model, for smoking status we used a mixed multinomial cumulative logistic regression model. Parameters from these models were used to supply probabilistic priors for our variable imputation step, which were then applied via

Monte Carlo simulation (replicates=1,000) to fill in missing values for smoking status (first), then estrogen-only hormone replacement therapy, then diabetes diagnosis, according to the methods described in paper 1 [Thompson CA, Arah OA: Sensitivity analysis for uncontrolled confounding, without bias formulas (manuscript)]. Estrogen-only hormone replacement therapy was imputed using smoking status imputes when applicable (i.e., if a subject was missing both the smoking and EHRT status), and diabetes was imputed using smoking and EHRT imputes when applicable (i.e., if a subject was missing all three variables).

After imputation, the complete case restricted dataset was analyzed for the effect of BMI on type I endometrial cancer using conditional logistic regression (matching factors of study site and reference age in 5-year increments) controlling for race, age at menarche, parity, ever use of estrogen-only hormone replacement therapy, ever use of oral contraception, and smoking status. This was then compared to the same analysis using the imputed variables for smoking status estrogen-only hormone replacement therapy and diabetes, using lognormal prior distributions of the log odds ratios generated from the models for smoking, EHRT and diabetes from complete case data; these were considered the "bias-adjusted" estimates. Since the imputation was done using MC methods, the model was fit by repetition, and odds ratio estimates are accompanied by simulation intervals (2.5, 50, 97.5 percentiles) of the estimate distributes were generated. Complete case (which included only those subjects who were not missing the imputation initially) odds ratios and 95% confidence intervals versus "bias-adjusted" (including complete case subjects as well as all subjects for whom imputation was performed) odds ratios and simulation intervals are shown per 5 kg/m^2 increase in BMI, overall and by study, as well as by categories of BMI (underweight versus normal, overweight versus normal, obese class I versus normal and obese class II versus normal). The latter categorical BMI analysis was also stratified

by study type (case versus control) as well as race (whites versus nonwhites). All statistical analyses were performed in SAS version 9.3 (SAS Institute; Cary, NC).

4.4. RESULTS

24 studies were included in the analysis, 10 cohort studies and 14 case control studies. These studies varied significantly in the % white population and levels of missingness of the target confounding variables (Table 4.1).

In total, 13,711 cases and 38,519 were eligible for analysis, among them, 140 (1.0%) cases and 174 (0.5%) controls were excluded for missing race, 239 (1.7%) cases and 713 (1.9%) controls were excluded for missing BMI, 159 (1.2%) cases and 421 (1.1%) controls were excluded for missing age at menarche, 328 cases (2.4%) and 805 controls (2.1%) were excluded for missing parity, and 144 cases (1.1%) and 422 controls (1.1%) were excluded for missing ever use of oral contraception. Among cases, we observed higher percentages of women in obese class I (17.4% compared to 11.3% among controls), obese class II&III (18.3% compared to 5.3% among controls), nulliparity (17.8% compared to 11.9% among controls), higher percentage of ever use of estrogen only hormone replacement therapy (12.0% compared to 7.7% among controls), slightly lower history of ever use of oral contraception (35.1% compared to 39.4% among controls), and a higher percentage of diabetes diagnosis (13.8% compared to 7.8% among controls) (Table 4.2).

Table 4.2 Subject characteristics

Characteristic	Cases	Controls
	N= 13,711	**N=38,519**
Race, n (%)		
White	11,317 (82.5)	33,120 (85.9)
Black	309 (2.3)	1,518 (3.9)
Asian	1,543 (11.3)	2,494 (6.5)
Hawaiian/Pacific Islander	100 (0.7)	319 (0.8)
Other	302 (2.2)	894 (2.3)
Missing/unknown	140 (1.0)	174 (0.5)
BMI, mean (SD)	29.0 (7.5)	25.8 (5.2)
BMI categories, n (%)		
Underweight (<18.5)	176 (1.3)	787 (2.0)
Normal (18.5-<25)	4,561 (33.3)	18,889 (49.0)
Overweight (25 to <30)	3,843 (28.0)	11,741 (30.5)
Obese Class I (30 to <35)	2,386 (17.4)	4,333 (11.3)
Obese Class II & III (35+)	2,506 (18.3)	2,056 (5.3)
Missing/unknown	239 (1.7)	713 (1.9)
Age at menarche, n (%)		
<11	918 (6.7)	1,775 (4.6)
11-12	2,781 (20.3)	7,687 (20.0)
13-14	8,127 (59.3)	23,451 (60.9)
15+	1,726 (12.6)	5,185 (13.5)
Missing/unknown	159 (1.2)	421 (1.1)
Parity, n (%)		
0	2,440 (17.8)	4,592 (11.9)
1	2,113 (15.4)	4,880 (12.7)
2	3,828 (27.9)	10,911 (28.3)
3-4	4,015 (29.3)	13,120 (34.1)
5+	987 (7.2)	4,211 (10.9)
Missing/unknown	328 (2.4)	805 (2.1)
HRT ever users, n (%)		
Yes	5,033 (36.7)	13,813 (35.9)
No	8,269 (60.3)	23,580 (61.2)
missing/unknown	409 (3.0)	1,126 (2.9)
ET ever users, n (%)		
Yes	1,648 (12.0)	2,970 (7.7)
No	9,977 (72.8)	29,094 (75.5)
Missing/unknown	2,086 (15.2)	6,455 (16.8)
OC ever users, n (%)		
Yes	4,818 (35.1)	15,173 (39.4)
No	8,749 (63.8)	22,924 (59.5)
Missing/unknown	144 (1.1)	422 (1.1)

Characteristic	Cases	Controls
Diabetes diagnosis, n (%)		
Yes	1,895 (13.8)	3,017 (7.8)
No	9,811 (71.6)	29,211 (75.8)
Missing/unknown	2,005 (14.6)	6,291 (16.3)
Smoking, n (%)		
Never	8,104 (59.1)	20,022 (52.0)
Past	3,761 (27.4)	11,485 (29.8)
Current	1,242 (9.1)	5,694 (14.8)
Missing/unknown	604 (4.4)	1,318 (3.4)

In complete case analysis for all studies combined, we observed an OR = 1.62 (95% CI 1.59–1.65) for type I endometrial cancer per 5 kg/m2 increase in BMI. Adjustment for missing smoking status (Table 4.3) did not affect the pooled point estimate, OR = 1.61 (95% SI 1.58-1.64). In study specific estimates of this bias adjustment when smoking missingness was high, we observed slight attenuation of results: for the cohort study BCDDP (16.9% missing) the OR decreased from 1.63 (95% CI 1.38-1.92) to 1.59 (95% SI 1.36-1.86). In one study in which smoking status was never measured, we observed an OR = 1.25 (9%% SI 1.12-1.39) following imputation.

Adjustment for missing estrogen-only hormone replacement therapy (Table 4.4) in the pooled analysis also slightly attenuated the per 5 kg/m^2 OR = 1.58 (95% SI 1.56-1.62). This attenuation was consistently observed in studies with higher levels of estrogen missingness, in the cohort studies: CSDLH (24% missing) the per 5 kg/m2 OR was adjusted from 1.74 (95% CI 1.56-1.71) to 1.52 (95% SI 1.40-1.68), in IWHS (29.8% missing) the per 5 kg/m2 OR was adjusted from 1.97 (95% CI 1.76-2.21) to 1.68 (95% SI 1.53-1.84), in SMC (26.6% missing) the per 5 kg/m2 OR was adjusted from 2.03 (95% CI 1.66-1.92) to 1.67 (95% SI 1.45 to 1.93), and in BCDDP (51.7% missing) the per 5 kg/m2 OR was adjusted from 1.63 (95% CI 1.38-1.92) to 1.24 (95%

SI 1.11-1.38); in case-control studies: WISE (28.4% missing) the per 5 kg/m2 OR was adjusted from 1.75 (95% CI 1.54-1.99) to 1.68 (95% SI 1.51-1.87), in Alberta (44.6% missing) the per 5 kg/m2 OR was adjusted from 1.75 (95% CI 1.54-1.96) to 1.62 (95% SI 1.48-1.77), in ANECS (14.5% missing) the per 5 kg/m2 OR was adjusted from 1.82 (95% CI 1.65-2.01) to 1.72 (95% SI 1.58-1.88), and in Turin (20.5% missing) the per 5 kg/m2 OR was adjusted from 1.70 (95% CI 1.36-2.11) to 1.65 (95% SI 1.35-2.01). In a comparison analysis using ever use of any type of hormone replacement therapy as a confounder substitute for ever use of estrogen-only hormone replacement therapy, complete case analysis results were comparable to the bias adjusted results.

In the diabetes adjustment analysis (table 4.5), sample sizes and complete case results were distinct from those presented in tables 4.3 and 4.4 because of the added required covariate of diabetes diagnosis. In the pooled analysis, additional adjustment for diabetes in the complete case population resulted in a per 5 kg/m^2 OR of 1.58 (95% CI 1.55-1.61) which was shifted slightly upwards, OR = 1.60 (95% SI 1.57-1.63) after bias adjustment for 10% missingness in the diabetes status variable. However, in studies with extreme levels of diabetes missingness, the pattern of bias adjustment was clearly that of attenuation: in the cohort study SMC (93.8% missing) the OR was adjusted from 3.84 (95% CI 1.31-11.21) to 2.03 (95% SI 1.64-2.50), and in the case control study PEDS (88.7% missing) the OR was adjusted from 1.85 (95% CI 1.13-3.01) to 1.66 (95% SI 1.46-1.89).

After adjusting for all 3 confounding variables (table 4.6), some marked attenuation in the per 5 kg/m^2 odds ratios were observed in studies with high levels of missingness in one or more variables. In the cohorts studies: for IWHS (31.1% missing) the OR shifted from 1.96 (95% CI 1.74-2.21) to 1.67 (95% SI 1.52-1.84), for BCDDP (63.4% missing) the OR shifted from 1.70 (95% CI 1.41-2.05) to 1.27 (95% SI 1.14-1.41); in the case-control studies: for WISE (28.5%

missing) the OR shifted from 1.72 (95% CI 1.51-1.95) to 1.63 (95% SI 1.46-1.82), for Alberta (44.7% missing) the OR shifted from 1.74 (95% CI 1.54-1.97) to 1.61 (95% SI 1.47-1.77), and for PEDS (90.1% missing), the OR shifted from 1.85 (95% CI 1.13-3.01) to 1.67 (95% SI 1.47-1.89). In one case control study from Hawaii, which had 35.8% subjects missing one or more imputation variables, the bias adjustment resulted in an increased point estimate of 1.70 (95% SI 1.48-1.96) compared to the complete case OR 1.46 (95% CI 1.22-1.75). Some marked changes after bias adjustment were also observed in the BMI categorical analysis (table 4.7), especially in the high BMI categories. In the obese class II/III, the pooled OR shifted from 5.98 (95% CI 5.50-6.50) to 5.79 (95% SI 5.38-6.23), and similar patterns of attenuation were seen in the study type stratification analysis and among white women. Stratification among non-white women in the bias adjusted estimates resulted in an increase of effect in both obese categories: in obese category I the OR shifted from 2.09 (95% CI 2.54-3.76) to 3.25 (95% SI 2.71-3.93) and in obese category II/II the OR shifted from 7.36 (95% CI 5.77-9.37) to 7.48 (95% SI 5.99-9.35).

Full descriptions of the models used for empirical prior generation from complete case data as well as the parameters generated from them are provided in the appendix tables 7.7-7.9.

Table 4.3 Smoking bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer

Study	Complete case analysis sample size (cases / controls)	% Missing Smoking[4]	OR[1] per 5 kg/m^2 increase Complete case analysis	Bias adjusted sample size (cases / controls)	OR[2] per 5 kg/m^2 increase Bias Adjusted[3] for Smoking
All studies combined	10,558 / 29,353	3.4	1.62 (1.59, 1.65)	11,025 / 30,306	1.61 (1.58, 1.64)
Cohort Studies					
MEC/USC	468 / 2,352	0.7	1.57 (1.45, 1.71)	470 / 2370	1.57 (1.45, 1.70)
NHS	538 / 1,541	2.2	1.48 (1.35, 1.61)	552 / 1,573	1.49 (1.37, 1.63)
CSDLH	314 / 1,944	0.0	1.74 (1.56, 1.94)	314 / 1,944	1.74 (1.56, 1.94)
CPS-II	508 / 2,377	1.7	1.56 (1.42, 1.71)	511 / 2,423	1.55 (1.41, 1.70)
NLCS	351 / 766	0.0	1.50 (1.28, 1.76)	351 / 766	1.50 (1.28, 1.76)
CTS	580 / 2,516	0.2	1.34 (1.23, 1.45)	582 / 2,521	1.34 (1.24, 1.45)
IWHS	276 / 1,523	1.4	1.97 (1.76, 2.21)	279 / 1,545	1.96 (1.75, 2.12)
NIH-AARP	1,124 / 5,588	2.5	1.53 (1.46, 1.60)	1,146 / 5,739	1.52 (1.45, 1.60)
SMC	159 / 692	2.2	2.03 (1.66, 2.49)	163 / 707	2.06 (1.68, 2.52)
BCDDP	125 / 1036	16.9	1.63 (1.38, 1.92)	135 / 1262	1.59 (1.36, 1.86)
Case control Studies					
Edge	414 / 464	0.0	1.60 (1.42, 1.79)	414 / 464	1.59 (1.42, 1.79)
WISE	402 / 1,115	0.1	1.75 (1.54, 1.99)	402 / 1,117	1.75 (1.54, 1.99)
Hawaii case-control	282 / 305	0.0	1.55 (1.30, 1.85)	282 / 305	1.55 (1.30, 1.85)
SECS	1,060 / 1,207	0.0	1.78 (1.59, 1.98)	1,060 / 1,207	1.77 (1.59, 1.98)
PECS	427 / 1,832	0.0	1.45 (1.30, 1.62)	427 / 1,832	1.45 (1.30, 1.61)
US Case-control	275 / 265	2.7	1.56 (1.37, 1.78)	282 / 273	1.56 (1.37, 1.78)
Alberta	259 / 567	0.0	1.74 (1.54, 1.96)	259 / 567	1.74 (1.54, 1.96)
BAWHS	0 / 0	100.0	-	426 / 399	1.25 (1.12, 1.39)
USC LA	787 / 791	0.0	1.46 (1.33, 1.60)	791 / 787	1.46 (1.33, 1.60)
ANECS	884 / 556	0.0	1.82 (1.65, 2.01)	884 / 556	1.82 (1.65, 2.01)
PEDS	365 / 437	0.3	1.66 (1.46, 1.88)	366 / 438	1.66 (1.47, 1.88)

Study	Complete case analysis sample size (cases / controls)	% Missing Smoking[4]	OR[1] per 5 kg/m^2 increase Complete case analysis	Bias adjusted sample size (cases / controls)	OR[2] per 5 kg/m^2 increase Bias Adjusted[3] for Smoking
WNYDS	232 / 639	0.0	1.74 (1.52, 2.00)	232 / 639	1.74 (1.52, 2.00)
Turin	160 / 209	0.3	1.70 (1.36, 2.11)	160 / 210	1.70 (1.36, 2.12)
CECS	568 / 631	0.0	1.68 (1.52, 1.84)	568 / 631	1.67 (1.52, 1.84)

[1]Conditional logistic regression, adjusted for age, race, parity, smoking status, age at menarche, ever use of oral contraception, ever use of estrogen-only hormone replacement therapy.
[2]Conditional logistic regression, adjusted for age, race, parity, imputed smoking status, age at menarche, ever use of oral contraception, and ever use of estrogen-only hormone replacement therapy.
[3]Smoking imputation using probabilistic priors based on study, study type, case/control status, BMI, age, race, parity smoking status, age at menarche, ever use of oral contraception.
[4]% missing calculated out of complete case sample size for all other covariates.

Table 4.4 Estrogen-only hormone replacement therapy bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer

Study	Complete case analysis sample size (cases / controls)	% Missing EHRT[5]	OR[1] per 5 kg/m^2 increase Complete case analysis	Bias adjusted sample size (cases / controls)	OR[2] per 5 kg/m^2 increase Bias Adjusted for EHRT	OR[2] per 5 kg/m^2 increase Adjusted for HRT (confounder substitute)
All studies combined	10,558 / 29,353	16.0	1.62 (1.59, 1.65)	12,395 / 35,123	1.58 (1.56, 1.62)	1.58 (1.55, 1.61)
Cohort Studies						
MEC/USC	468 / 2,352	2.7	1.57 (1.45, 1.71)	480 / 2,419	1.56 (1.44, 1.69)	1.60 (1.48, 1.74)
NHS	538 / 1,541	0.0	1.48 (1.35, 1.61)	538 / 1,541	1.48 (1.35, 1.61)	1.50 (1.36, 1.66)
CSDLH	314 / 1,944	24.0	1.74 (1.56, 1.94)	473 / 2,497	1.53 (1.40, 1.68)	1.56 (1.42, 1.72)
CPS-II	508 / 2,377	3.2	1.56 (1.42, 1.71)	527 / 2,452	1.53 (1.39, 1.67)	1.58 (1.44, 1.74)
NLCS	351 / 766	4.1	1.50 (1.28, 1.76)	365 / 800	1.51 (1.29, 1.77)	1.50 (1.28, 1.76)
CTS	580 / 2,516	5.7	1.34 (1.23, 1.45)	612 / 2,671	1.34 (1.24, 1.45)	1.34 (1.34, 1.45)
IWHS	276 / 1,523	29.8	1.97 (1.76, 2.21)	451 / 2,111	1.68 (1.53, 1.84)	1.74 (1.59, 1.91)
NIH-AARP	1,124 / 5,588	19.0	1.53 (1.46, 1.60)	1,404 / 6,884	1.46 (1.40, 1.52)	1.49 (1.42, 1.55)
SMC	159 / 692	46.6	2.03 (1.66, 2.49)	300 / 1,294	1.67 (1.45, 1.93)	1.68 (1.45, 1.94)
BCDDP	125 / 1,036	51.7	1.63 (1.38, 1.92)	384 / 2,020	1.24 (1.11, 1.38)	1.31 (1.17, 1.46)
Case control Studies						
Edge	414 / 464	0.0	1.60 (1.42, 1.79)	414 / 464	1.59 (1.42, 1.79)	1.60 (1.43, 1.79)
WISE	402 / 1,115	28.4	1.75 (1.54, 1.99)	546 / 1,574	1.68 (1.51, 1.87)	1.66 (1.49, 1.85)
Hawaii case-control	282 / 305	35.7	1.55 (1.30, 1.85)	403 / 510	1.79 (1.56, 2.05)	1.53 (1.30, 1.81)
SECS	1,060 / 1,207	0.0	1.78 (1.59, 1.98)	1,060 / 1,207	1.77 (1.59, 1.98)	1.77 (1.59, 1.98)
PECS	427 / 1,832	0.1	1.45 (1.30, 1.62)	427 / 1,833	1.45 (1.30, 1.61)	1.46 (1.31, 1.63)
US Case-control	275 / 265	12.8	1.56 (1.37, 1.78)	314 / 305	1.60 (1.41, 1.82)	1.59 (1.40, 1.81)
Alberta	259 / 567	44.6	1.74 (1.54, 1.96)	468 / 1,024	1.62 (1.48, 1.77)	1.62 (1.48, 1.77)
BAWHS	399 / 426	5.9		423 / 451		
USC LA	787 / 791	0.0	1.46 (1.33, 1.60)	787 / 791	1.46 (1.33, 1.60)	1.51 (1.37, 1.65)

Study	Complete case analysis sample size (cases / controls)	% Missing EHRT[5]	OR[1] per 5 kg/m^2 increase Complete case analysis	Bias adjusted sample size (cases / controls)	OR[2] per 5 kg/m^2 increase Bias Adjusted for EHRT	OR[2] per 5 kg/m^2 increase Adjusted for HRT (confounder substitute)
ANECS	884 / 556	14.5	1.82 (1.65, 2.01)	1016 / 669	1.72 (1.58, 1.88)	1.74 (1.59, 1.90)
PEDS	365 / 437	12.5	1.66 (1.46, 1.88)	420 / 496	1.62 (1.44, 1.82)	1.63 (1.45, 1.83)
WNYDS	232 / 639	0.0	1.74 (1.52, 2.00)	232 / 639	1.74 (1.52, 2.00)	1.74 (1.52, 2.00)
Turin	160 / 209	20.5	1.70 (1.36, 2.11)	197 / 267	1.65 (1.35, 2.01)	1.66 (1.36, 2.03)
CECS	568 / 631	2.7	1.68 (1.52, 1.84)	577 / 655	1.70 (1.55, 1.87)	1.67 (1.52, 1.84)

[1]Conditional logistic regression, adjusted for age, race, parity, smoking status, age at menarche, ever use of oral contraception, ever use of estrogen-only hormone replacement therapy.
[2]Conditional logistic regression, adjusted for age, race, parity, smoking status, age at menarche, ever use of oral contraception, and imputed ever use of estrogen-only hormone replacement therapy.
[3]Estrogen imputation using fixed priors based on study, study type, case/control status, BMI, age, race, parity, smoking status, age at menarche, ever use of oral contraception; imputed values for smoking status used when necessary.
[4]Estrogen imputation using probabilistic priors based on study, study type, case/control status, BMI, age, race, parity, smoking status, age at menarche, ever use of oral contraception.
[5]% missing calculated out of complete case sample size for all other covariates.

Table 4.5 Diagnosis of diabetes bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer

Study	Complete case analysis sample size (cases / controls)	% Missing Diabetes[4]	OR[1] per 5 kg/m^2 increase Complete case analysis	Bias adjusted sample size (cases / controls)	OR[2] per 5 kg/m^2 increase Bias Adjusted for Diabetes[3]
All studies combined	9,686 / 25,885	10.9	1.58 (1.55, 1.61)	10,558 / 29,353	1.60 (1.57, 1.63)
Cohort Studies					
MEC/USC	468 / 2,352	0.0	1.57 (1.45, 1.71)	468 / 2,352	1.57 (1.45, 1.71)
NHS	538 / 1,541	0.0	1.45 (1.33, 1.59)	538 / 1,541	1.45 (1.33, 1.59)
CSDLH	0 / 0	100.0	-	314 / 1,944	1.73 (1.55, 1.93)
CPS-II	446 / 2061	13.1	1.54 (1.34, 1.75)	508 / 2,377	1.55 (1.41, 1.71)
NLCS	351 / 766	0.0	1.51 (1.28, 1.77)	351 / 766	1.51 (1.28, 1.77)
CTS	580 / 2,516	0.0	1.33 (1.23, 1.44)	580 / 2,516	1.33 (1.23, 1.44)
IWHS	271 / 1,517	0.6	1.96 (1.74, 2.21)	276 / 1,523	1.95 (1.73, 2.19)
NIH-AARP	1,124 / 5,588	0.0	1.52 (1.44, 1.59)	1,124 / 5,588	1.52 (1.44, 1.59)
SMC	17 / 36	93.8	3.84 (1.31, 11.21)	159 / 692	2.03 (1.64, 2.50)
BCDDP	111 / 908	12.2	1.70 (1.41, 2.05)	125 / 1,036	1.61 (1.36, 1.92)
Case control Studies					
Edge	414 / 464	0.0	1.57 (1.40, 1.76)	414 / 464	1.57 (1.40, 1.76)
WISE	402 / 1,115	0.0	1.72 (1.51, 1.95)	402 / 1,115	1.72 (1.51, 1.95)
Hawaii case-control	281 / 305	0.2	1.46 (1.22, 1.75)	282 / 305	1.46 (1.22, 1.75)
SECS	1,049 / 1,200	0.8	1.72 (1.54, 1.93)	1,060 / 1,207	1.73 (1.54, 1.93)
PECS	420 / 1,820	0.8	1.40 (1.52, 1.57)	427 / 1,832	1.39 (1.24, 1.55)
US Case-control	275 / 265	0.0	1.53 (1.34, 1.75)	275 / 265	1.53 (1.34, 1.75)
Alberta	259 / 566	0.1	1.74 (1.54, 1.97)	259 / 567	1.74 (1.54, 1.97)
BAWHS	0 / 0	100.0	-	399 / 426	
USC LA	787 / 791	0.0	1.45 (1.32, 1.59)	787 / 791	1.45 (1.32, 1.59)
ANECS	883 / 555	0.1	1.80 (1.63, 1.99)	884 / 556	1.78 (1.61, 1.97)
PEDS	40 / 51	88.7	1.85 (1.13, 3.01)	365 / 437	1.66 (1.46, 1.89)
WNYDS	231 / 639	0.11	1.68 (1.46, 1.94)	232 / 639	1.68 (1.46, 1.94)

Study	Complete case analysis sample size (cases / controls)	% Missing Diabetes[4]	OR[1] per 5 kg/m^2 increase Complete case analysis	Bias adjusted sample size (cases / controls)	OR[2] per 5 kg/m^2 increase Bias Adjusted for Diabetes[3]
Turin	160 / 209	0.0	1.76 (1.40, 2.20)	160 / 209	1.76 (1.40, 2.20)
CECS	568 / 631	0.0	1.67 (1.51, 1.84)	568 / 631	1.67 (1.51, 1.84)

[1]Conditional logistic regression, adjusted for age, race, parity, smoking status, age at menarche, ever use of oral contraception, ever use of estrogen-only hormone replacement therapy.
[2]Conditional logistic regression, adjusted for age, race, parity, smoking status, age at menarche, ever use of oral contraception, ever use of estrogen-only hormone replacement therapy, and imputed diabetes status.
[3]Diabetes imputation based on study, study type, case/control status, BMI, age, race, parity smoking status, age at menarche, ever use of oral contraception, ever use of estrogen-only hormone replacement therapy.
[4]% missing calculated out of complete case sample size for all other covariates.

Table 4.6 All three variable (smoking, EHRT, diabetes) bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer

Study	Complete case analysis sample size (cases / controls)	% Missing one or more variables[4]	OR[1] per 5 kg/m^2 increase Complete case analysis	Bias adjusted sample size (cases / controls)	OR[2] per 5 kg/m^2 increase Bias Adjusted for EHRT, smoking, and diabetes
All studies combined	9,686 / 25,885	27.7	1.58 (1.55, 1.61)	12,918 / 36,309	1.56 (1.53, 1.59)
Cohort Studies					
MEC/USC	468 / 2,352	3.6	1.57 (1.45, 1.71)	484 / 2,442	1.56 (1.44, 1.69)
NHS	538 / 1,541	2.2	1.45 (1.33, 1.59)	552 / 1,573	1.47 (1.34, 1.61)
CSDLH	0 / 0	100.0	-	473 / 1,833	1.54 (1.40, 1.70)
CPS-II	446 / 2061	17.3	1.54 (1.34, 1.75)	530 / 2,501	1.54 (1.41, 1.69)
NLCS	351 / 766	4.1	1.51 (1.28, 1.77)	365 / 800	1.51 (1.29, 1.77)
CTS	580 / 2,516	6.0	1.33 (1.23, 1.44)	615 / 2,677	1.34 (1.24, 1.45)
IWHS	271 / 1,517	31.1	1.96 (1.74, 2.21)	455 / 2,141	1.67 (1.52, 1.84)
NIH-AARP	1,124 / 5,588	21.2	1.52 (1.44, 1.59)	1,431 / 7083	1.45 (1.39, 1.51)
SMC	17 / 36	96.8	3.84 (1.31, 11.21)	308 / 1,322	1.69 (1.46, 1.96)
BCDDP	111 / 908	63.4	1.70 (1.41, 2.05)	410 / 2,375	1.27 (1.14, 1.41)
Case control Studies					
Edge	414 / 464	0.0	1.57 (1.40, 1.76)	414 / 464	1.57 (1.40, 1.76)
WISE	402 / 1,115	28.5	1.72 (1.51, 1.95)	546 / 1,576	1.63 (1.46, 1.82)
Hawaii case-control	281 / 305	35.8	1.46 (1.22, 1.75)	403 / 510	1.70 (1.48, 1.96)
SECS	1,049 / 1,200	0.8	1.72 (1.54, 1.93)	1,060 / 1,207	1.73 (1.54, 1.93)
PECS	420 / 1,820	0.9	1.40 (1.52, 1.57)	427 / 1,833	1.39 (1.24, 1.55)
US Case-control	275 / 265	15.4	1.53 (1.34, 1.75)	324 / 314	1.55 (1.37, 1.76)
Alberta	259 / 566	44.7	1.74 (1.54, 1.97)	468 / 1,024	1.61 (1.47, 1.77)
BAWHS	0 / 0	100.0	-	423 / 451	1.25 (1.12, 1.40)
USC LA	787 / 791	0.0	1.45 (1.32, 1.59)	787 / 791	1.45 (1.32, 1.59)
ANECS	883 / 555	14.7	1.80 (1.63, 1.99)	669 / 1,016	1.68 (1.53, 1.83)

Study	Complete case analysis sample size (cases / controls)	% Missing one or more variables[4]	OR[1] per 5 kg/m^2 increase Complete case analysis	Bias adjusted sample size (cases / controls)	OR[2] per 5 kg/m^2 increase Bias Adjusted for EHRT, smoking, and diabetes
PEDS	40 / 51	90.1	1.85 (1.13, 3.01)	421 / 497	1.67 (1.47, 1.89)
WNYDS	231 / 639	0.1	1.68 (1.46, 1.94)	232 / 639	1.68 (1.46, 1.94)
Turin	160 / 209	20.7	1.76 (1.40, 2.20)	197 / 268	1.74 (1.42, 2.14)
CECS	568 / 631	2.7	1.67 (1.51, 1.84)	577 / 655	1.70 (1.54, 1.87)

[1]Conditional logistic regression, adjusted for age, race, parity, smoking status, diabetes, age at menarche, ever use of oral contraception, ever use of estrogen-only hormone replacement therapy.
[2]Conditional logistic regression, adjusted for age, race, parity, imputed smoking status, diabetes, age at menarche, ever use of oral contraception, and ever use of estrogen-only hormone replacement therapy.
[4]% missing calculated out of complete case sample size for all other covariates.

Table 4.7 Categorical BMI – with all bias adjustment (ERT, smoking status, diabetes), by study type and white vs. non-white ethnicity

BMI category	Complete case analysis sample size (cases / controls)	Complete case OR[1] (95% CI)	Bias adjusted sample size (cases / controls)	Bias Adjusted for EHRT, smoking, and diabetes OR[2] (95% SI)
All studies				
Underweight (<18.5)	122 / 546	0.84 (0.69, 1.03)	165 / 752	0.83 (0.70, 0.99)
Normal (18.5 to <25)	3,115 / 12,610	1.00	4,363 / 18,109	1.00
Overweight (25 to <30)	2,794 / 8,118	1.49 (1.40, 1.58)	3,705 / 11,319	1.44 (1.37, 1.51)
Obese Class I (30 to <35)	1,747 / 3,076	2.60 (2.42, 2.80)	2,281 / 4,156	2.53 (2.38, 2.70)
Obese Class II & III (35+)	1,908 / 1,535	5.98 (5.50, 6.50)	2,404 / 1,973	5.79 (5.38, 6.23)
Cohort Studies				
Underweight (<18.5)	43 / 281	1.01 (0.73, 1.41)	64 / 428	0.92 (0.70, 1.20)
Normal (18.5 to <25)	1,103 / 7,284	1.00	1,817 / 11,431	1.00
Overweight (25 to <30)	995 / 4,837	1.39 (1.26, 1.53)	1,482 / 7,208	1.32 (1.22, 1.42)
Obese Class I (30 to <35)	663 / 1,833	2.45 (2.18, 2.74)	909 / 2,585	2.28 (2.08, 2.50)
Obese Class II & III (35+)	656 / 989	4.59 (4.06, 5.20)	821 / 1,258	4.35 (3.91, 4.84)
Case-control studies				
Underweight (<18.5)	79 / 265	0.75 (0.57, 0.98)	101 / 324	0.79 (0.62, 1.00)
Normal (18.5 to <25)	2,012 / 5,326	1.00	2,254 / 6,678	1.00
Overweight (25 to <30)	1,799 / 3,281	1.49 (1.37, 1.61)	2,223 / 4,111	1.46 (1.36, 1.57)
Obese Class I (30 to <35)	1,084 / 1,243	2.51 (1.37, 1.61)	1,372 / 1,571	2.49 (2.28, 2.73)
Obese Class II & III (35+)	1,252 / 546	6.87 (6.08, 7.76)	1,583 / 715	6.48 (5.82, 7.22)
White women				
Underweight (<18.5)	90 / 406	1.01 (0.80, 1.28)	126 / 586	0.95 (0.78, 1.17)
Normal (18.5 to <25)	2,373 / 10,537	1.00	3,551 / 15,700	1.00
Overweight (25 to <30)	2,158 / 6,779	1.43 (1.34, 1.53)	3,020 / 9,798	1.38 (1.31, 1.46)
Obese Class I (30 to <35)	1,460 / 2,590	2.52 (2.33, 2.74)	1,953 / 3,602	2.44 (2.28, 2.61)
Obese Class II & III (35+)	1,660 / 1,277	5.79 (5.29, 6.34)	2,101 / 1,669	5.59 (5.17, 6.04)
Non-white women				
Underweight (<18.5)	32 / 140	0.54 (0.36, 0.81)	39 / 166	0.58 (0.40, 0.85)
Normal (18.5 to <25)	742 / 2,073	1.00	812 / 2,409	1.00
Overweight (25 to <30)	636 / 1,339	1.80 (1.57, 2.07)	685 / 1,521	1.79 (1.57, 2.04)
Obese Class I (30 to <35)	287 / 486	2.09 (2.54, 3.76)	328 / 554	3.25 (2.71, 3.91)
Obese Class II & III (35+)	248 / 258	7.36 (5.77, 9.37)	303 / 304	7.48 (5.99, 9.35)

[1]Conditional logistic regression, adjusted for age, race, parity, smoking status, diabetes, age at menarche, ever use of oral contraception, ever use of estrogen-only hormone replacement therapy.
[2]Conditional logistic regression, adjusted for age, race, parity, imputed smoking status, imputed diabetes, age at menarche, ever use of oral contraception, and imputed ever use of estrogen-only hormone replacement therapy

4.5. DISCUSSION

We utilized a large pooled database to demonstrate a novel bias modeling technique that uses prior parameters from complete subjects and available data from all subjects to impute missing confounding variables based on the underlying assumed causal structure. After deriving empirical priors from the complete pooled database by taking into account important modifiers of the causal relationships such as study type and the chosen model covariates, we applied these priors in an imputation step that resulted in some measurable sample size gains and, mostly, attenuated results compared to the complete case odds ratios.

Generally endometrial cancer is understood to be caused by unopposed estrogen exposure, either through early age of menarche, nulliparity, late age at first birth, estrogen only hormone therapy, or high dose estrogen in oral contraceptive pills [3]. Obesity, as measured by BMI as well as anamorphic measurements, is one of the most important modifiable risk factors for endometrial cancer, but the molecular mechanism is different for pre- and post-menopausal women. In pre-menopausal women, obesity leads to increased insulin, progesterone deficiency, and thus a reduced ability to oppose free estrogens [4]. In post-menopausal women obesity leads to endometrial cancer through increases in free floating estrogens. Endometrial cancer is one of the only neoplasms for which smoking is protective. The biological mechanism for this is thought to be related to increases in estrogen-opposing progesterone [5]. Other protective factors include normal weight, weight loss, physical activity, grand multiparity and exogenous hormones that include a cycle of progesterone, such as combination hormone replacement therapy and modern low dose estrogen and progesterone combination oral contraceptive pills. In addition to being risk factors for disease, exogenous hormones are strong modifiers of the BMI effect, often with paradoxical results. Overweight women who have previously taken exogenous hormones tend to experience less risk in higher BMI categories than overweight women who have not taken

exogenous hormones [6]. Other risk factors (not related to the estrogen pathway) include family history of endometrial cancer and increasing age.

Since high BMI is a very strong predictor of type I EC and the confounding variables we examined are not as strong risk/protective factors for this disease, the bias adjustment we observed was mostly subtle, and towards the null. According to the DAG structure and direction and magnitude of the relationships described within, the bias adjustment was in line with our expectations. Smoking is a protective factor for type I EC, and it also reduces (or maintains low) BMI. Unmeasured smoking would thus produce a positive bias away from the null because it is a protective common cause of both BMI and type I EC. Estrogen-only hormone replacement therapy increases risk of type I EC, and may be associated with decreased BMI, although the direction of the relationship between these latter two variables is debatable. Increased BMI results in increased endogenous estrogen production, which may alleviate the need for hormone replacement therapy due to menopausal symptoms [REF]. If the DAG edge connecting these two variables points from BMI to EHRT, this variable would be considered an intermediate on the causal pathway. Under a certain set of assumptions, controlling for an intermediate could result in an attenuation of the point estimate, as was seen in our bias adjusted estimates. However these assumptions are often violated when there is unmeasured confounding of the intermediate-disease relationship [46, 47]. Diabetes diagnosis and BMI are closely associated within the constellation of diseases known as metabolic syndrome. Many of the individual components of metabolic syndrome, including diabetes have been shown to be associated with higher incidence of type I EC [48, 49]. Diabetes is also independently associated and with higher BMI [50], but, like EHRT, the directionality of the relationship between BMI and diabetes is debatable, and it may differ between the type of diabetes diagnosis (type I vs. type II), the details of which were unavailable in the pooled database used for this study. This points to an important consideration

in this study, and in pooling projects in general, which is that misclassification of the confounding variables may produce bias that is difficult to predict in magnitude or direction [51]. Indeed, ever use of a medication is often a poor classifier for actual exposure, and self-reported exposure (which is usually the data collection method in epidemiologic studies) adds an additional caveat to interpreting these results. If the complete case population measure of EHRT or diagnosis of diabetes was a poorly classified one, the bias adjustment parameters may suffer from inadequacy to control for the unmeasured confounding as a consequence. Use of the random effect models for empirical prior generation would help to alleviate this concern, especially if one study was severely misclassified, the clustering of such misclassification would be taken into account.

Some caution should be taken when interpreting the study-specific bias-adjusted estimates presented in this paper. Because the bias model is designed to take study-level and participant-level characteristics into account, we have provided the study-specific estimates as instructive for the bias model demonstration. However, in cases where missingness is extreme (e.g., table 4.6, study SMC) the level of missingness would render the imputations performed inappropriate for reporting (and interpreting) anything other than the bias-adjusted pooled estimates.

Another important consideration for the results of this study is that missing data bias in general may be stronger than any confounding bias we observed. Because BMI is such a strong predictor of type I EC, the sample size gains in some of the individual studies may be the sole reason for observing marked shifts in the odds ratio estimates. However, this is clearly an argument in favor of data imputation for large pooling projects as missingness can be pervasive and complete case results may not reflect the true effect estimates, sometimes in counterintuitive ways. The benefits of the method we have demonstrated include record-level imputation that reflects a complex structure bias adjustment algorithm that can be easily implemented in a

standardized database for analysis by multiple consortium investigators simultaneously. We utilized internal parameters for our bias adjustment models, but the algorithm can also be fit using external parameters, based on expert knowledge or validation sub-studies.

5. STUDY 4

A SENSITIVITY ANALYSIS FOR SELECTION BIAS IN A LARGE POOLED STUDY OF BMI AND TYPE I ENDOMETRIAL CANCER: THE EPIDEMIOLOGY OF ENDOMETRIAL CANCER CONSORTIUM (E2C2)

5.1. ABSTRACT

Introduction: Selection or non-response bias occurs when participation in an observational study is directly (or indirectly) affected by the exposure and outcome. In population-based and hospital-based case-control studies, a relationship between the disease outcome and participation is unavoidable, so even a weak relationship between exposure and selection, either via uncontrolled common causes or a direct relationship can introduce bias into effect estimates. External formula guided sensitivity analysis for the influence of selection bias can be useful in etiologic studies, especially if a selection bias is suspected based on study conduct or recruitment of participants. When more than one study is combined, however, such as in a pooled database, using external adjustment techniques can be a challenge, especially if selection bias is suspected in some, but not all contributing studies.

Methods: In this study we demonstrate the use of record-level techniques for the simulation of selection probabilities to facilitate adjustment for non-response bias using a large multi-study pooled data source, the Epidemiology of Endometrial Cancer Consortium (E2C2). Using as an example the well characterized relationship between BMI and type I endometrial cancer, we use known data values under an assumed causal structure to recreate selection probabilities reflective of the underlying source population in a full sensitivity analysis.

Results: We found that the BMI and endometrial cancer relationship, one of the strongest predictors of type I EC, is robust to small to moderate levels of the type of selection bias we simulated, but in the presence of larger effects, or substantial heterogeneity, unadjusted results may underreport the true magnitude of the effect.

Conclusion: Record-level data augmentation for selection bias modeling, in comparison to external formula adjustment, is a flexible approach that provides the option for bias parameters to be applied in a study-specific fashion, taking into account suspected selection bias as a result of control recruitment, refusal rates, and the ethnic distribution of the source population. It also incorporates bias parameters into the source database, prior to analysis, improving accessibility to quantitative bias modeling for observational research.

5.2. INTRODUCTION

Selection or non-response bias occurs when participation in an observational study is directly (or indirectly) affected by the exposure and outcome. In population-based and hospital-based case-control studies, a relationship between the disease outcome and participation is unavoidable, so even a weak relationship between exposure and selection, either via uncontrolled common causes or a direct relationship can introduce bias into effect estimates. Sensitivity analysis for the influence of selection bias can be useful in etiologic studies, especially if a selection bias is suspected based on study conduct or recruitment of participants. In its most rudimentary form, selection bias sensitivity analysis may be accomplished using external adjustment formulas to adjust the outcome models in a given analysis [26, 29, 37, 38]. When more than one study is combined, however, such as in a pooled database, using external adjustment techniques can be a challenge, especially if selection bias is suspected in some, but not all contributing studies. Internal adjustment for selection bias, or the record-level weighting known as inverse probability of censoring weighting (IPCW) is commonly used in follow-up studies. This technique uses individualized weights to create a pseudo-population that mimics the underlying cohort, including those who were censored, in order to produce estimates reflective of the entire cohort rather than just the selected strata [40]. IPCW is not an appropriate technique for retrospective

case control studies because it relies on information collected on both censored and retained cohort members, over time. If some basic knowledge of the selection mechanism is available, however, the basic premise behind this technique can be applied using a reasonable set of prior assumptions in a data augmentation step that can be used in all or some studies in a pooled database. The details of this method were described in full in a previous work [Thompson, Arah. Selection bias modeling using observed data augmented with imputed record-level probabilities (manuscript)]. The purpose of this study is to demonstrate this method using a multi-study epidemiologic project designed to investigate the etiology of endometrial cancer: The Epidemiology of Endometrial Cancer Consortium (E2C2).

5.3. MATERIALS AND METHODS

5.3.1. DATA SOURCE: THE EPIDEMIOLOGY OF ENDOMETRIAL CANCER CONSORTIUM (E2C2)

The Epidemiology of Endometrial Cancer Consortium (E2C2) is a formally designated NCI consortium to study rare cancer. The objective of the consortium is to pool as many studies of endometrial cancer as possible, with goals to study genetic variation, gene-gene interactions, gene-environment interactions, diet, and risk profiles of rarer histologic subtypes and underrepresented minorities[44]. The E2C2 project includes 24 studies, 9 prospective cohort studies and 15 case-control studies. For the purpose of this study, we restricted to hospital and population-based case-control studies only, and excluded one study that did not measure an important confounding variable, smoking status. This left 13 studies with a total of 7,163 endometrial cancer cases and 10,733 control subjects. These studies vary in population size, sampling strategies, inclusion criteria, population ethnicities, participant ages, case and control response rates, and recruitment periods. A list of analyzed studies including study type, location, case and control recruitment strategy, case and control response rates, and ethnicity distributions,

is provided in table 5.1. The dataset utilized for this study is the core risk factors database, which includes all E2C2 study participants and values for all major risk factors for endometrial cancer (BMI, weight change, reproductive variables, menopausal status, use of hormone replacement therapy and oral contraception, smoking, chronic comorbidities, physical activity), as well as demographic characteristics. The primary objective of this study is to demonstrate the use of record-level techniques for the simulation of selection probabilities to facilitate adjustment for non-response bias in the E2C2 risk factors database. We have chosen BMI as the exposure of interest, and we will focus on type I endometrial cancer cases only. The adjustment technique will be applied with a variety of hypothetical bias parameters in a sensitivity analysis.

5.3.2. DIRECTED ACYCLIC GRAPHS

We will use directed acyclic graphs (DAGs) visualize the bias model using. DAGs are directed, not cyclic, path diagrams that depict the variables and relationships between those variables in causal epidemiologic investigations. The use of DAGs to express these causal relationships imparts a basic set of rules, which have been extensively described elsewhere [32-36]. The two major sources of biasing paths in DAGs are uncontrolled confounding, that results from failing to control for a confounding (such as a common cause) variable, and conditioning on a collider, thus opening up a previously blocked path. Selection or non-response bias is a well-known example of collider bias (figure 5.1). Confounding variables may also directly affect selection; adequate control of such variables will succeed in blocking the open path that is created by conditioning on the collider, but background knowledge of their effect on selection will improve simulation of selection probabilities in the record-level data augmentation.

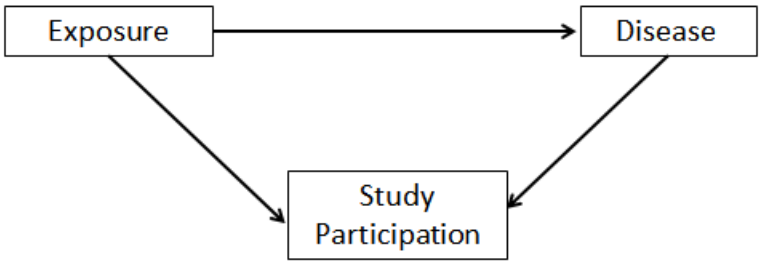

Figure 5.1 Collider bias due to selective non-response

5.3.3. BIAS MODEL

Let X be the exposure BMI (per 5 kg/m^2 unit increase), Y be the outcome, type I endometrial cancer, and S=1 be study participation (where S=0 is non-participation). In our bias model, in addition to BMI and type I EC, we also take into account two additional variables that are suspected to directly affect selection: Z_1 represents white race (reference being non-white) and Z_2 represents ever use of hormone replacement therapy (HRT). The selection mechanism is displayed in figure 5.2.

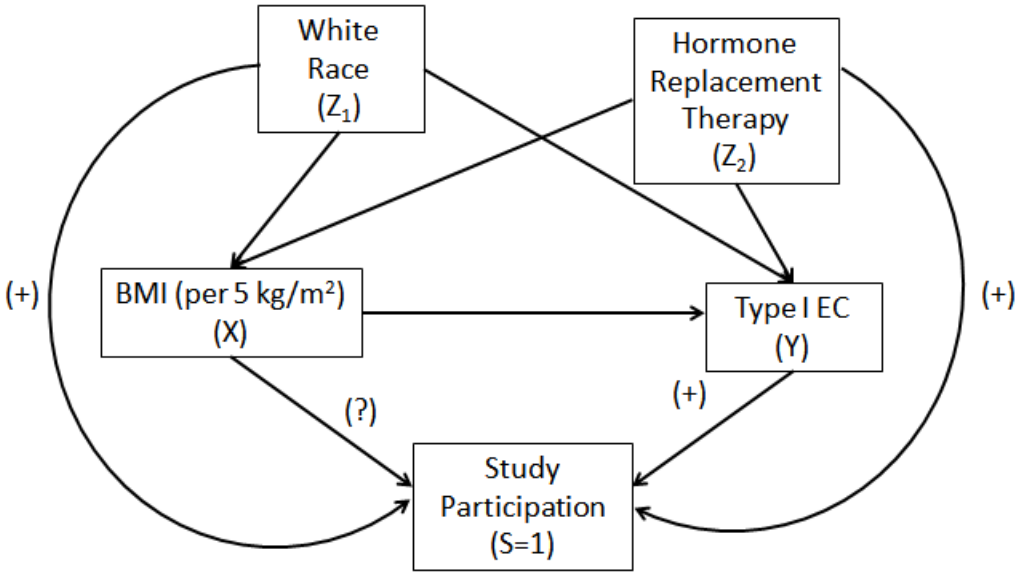

Figure 5.2 Signed DAG representing modeled influences of nonresponse in E2C2 case control studies

The probability of S as a function of the DAG variables can be obtained from the expit function of the logistic equation:

$$(1)\ \text{logit}(P(S=1|y,x,z_1,z_2)) = \beta_S + \beta_{SY}y + \beta_{SX}x + \beta_{Sz_1}z_1 + \beta_{Sz_2}z_2 + \beta_{SYX}yx +$$
$$\beta_{SYz_1}yz_1 + \beta_{SYz_2}yz_2 + \beta_{SXz_1}xz_1 + \beta_{SXz_2}xz_2 + \beta_{Sz_1z_2}z_1z_2 + \beta_{SYXz_1}yxz_1 +$$
$$\beta_{SYXz_2}yxz_2 + \beta_{SYz_1z_2}yz_1z_2 + \beta_{SXz_2z_2}xz_1z_2 + \beta_{SYXz_1z_2}yxz_1z_2$$

We defined a range of fixed prior distributions for the βs in expression (1). For the non-product terms, we have chosen small to moderate in magnitude prior that assign a range of possibilities and directionality to the forces predicting selection into the study (S=1) to demonstrate how the results might change under each set of priors. For the YX product terms, we assigned a very small positive and a very slight negative interaction coefficient in two scenarios. For all other product terms, we set an uninformative null prior, indicating that we strongly doubt the existence

of heterogeneity in the influences of the X, Y, Z1 and Z2 on selection. We set the P(S=1|X=0,Y=0, **Z**=0) to 0.05, indicating that probability of (S=1|X=0,Y=0,**Z**=0) was very low. Our priors were combined with the existing data in this model to generate the probability of selection for each participant in the analyzed case-control studies. The stabilized weight, P(S=1|X=0, Y=0, **Z**=0)/(P(S=1|X=x, Y=y, **Z**=**z**), which is the background probability of selection multiplied by the inverse of the conditional probability of selection for each participant, was used as a weight in the conditional logistic regression model for the relationship between BMI (per 5 kg/m^2) and type I endometrial cancer, adjusted for Z_1, Z_2, and additional appropriate confounding variables, all of which will be represented by the set **Z** in further notation. We present customary unadjusted results, odds ratios, with 95% confidence intervals and the selection bias adjusted results odds ratios, with adjusted 95% confidence intervals (without further random error adjustment). All statistical analyses were performed in SAS version 9.3 (SAS Institute; Cary, NC).

5.4. RESULTS

Of the 13 case-control studies included, only one (WNYDS) had hospital controls. Others had neighborhood or population registry controls and four studies had control series selected by random-digit-dialing. The case and control response rates varied from 42% to 83%, and 39% to 73%, respectively. Sample sizes, including ratio of matched controls to cases varied widely. Two studies, Hawaii case control, and SECS had very low (or no) white race participants. Recruitment periods varied also.

Table 5.1 Eligible participating population and hospital based case-control studies: location and population description

Study	Location	Sample size Cases / Controls	% White	Case Mean age (SD)	Control Mean age (SD)	Case source	Case response rate	Control Source	Control response rate	Recruitment period
EDGE [52]	New Jersey	418 / 467	90.2	60.6 (9.8)	64.2 (11.3)	Cancer registry	42%	RDD, HCFA, neighborhood	39%	2001-2005
WISE [53]	Philadelphia	552 / 1583	79.2	62.6 (8.1)	61.3 (8.1)	Population based	52%	RDD	58%	1999-2002
Hawaii case control [54]	Hawaii	432 / 511	23.5	57.5 (12.0)	56.6 (12.2)	Cancer registry	66%	Dept. of Health random sample	73%	1988-1993
SECS[55]	China	1,071 / 1,212	0.0	54.5 (8.5)	54.6 (8.5)	Cancer registry	83%	Random sample from the Shanghai Resident Registry	74%	1997-2004
PECS[56]	Poland	435 / 1925	100.0	60.9 (8.1)	56.3 (10.2)	Case-control	79%	Population register	68%	2000-2003
US case control[57]	5 US clinics	332 / 320	92.2	59.3 (10.1)	57.9 (10.5)	5 US clinics		RDD and HCFA files		1987-1990
Alberta	Canada	474 / 1,032	94.8	58.3 (9.5)	58.1 (10.1)	Cancer registry		Alberta province (population-based)		2002-2006
USC LA[58]	Los Angeles	787 / 791	100.0	63.0 (5.3)	63.1 (5.4)	Cancer registry	77%	Neighborhood	60%	1987-1993
ANECS[59]	Australia	1,125 / 740	88.3	60.7 (9.4)	61.1 (9.9)	Major treatment centers and cancer registries	67%	Australian Electoral roll (~95% complete)	53%	2005-2007
PEDS [60]	New York	468 / 541	97.0	62.4 (11.3)	63.2 (11.1)	Roswell Park cancer institute patients	50%	Roswell Park cancer institute patients	50%	1982-1998
WNYDS [61]	New York	232 / 639	100.0	63.5 (9.4)	55.9 (10.6)	Identified at WNYDS hospitals	51%	Randomly selected from driver's license and HCFA	51%	1986-1991
Turin	Italy	249 / 307	100.0	61.3 (7.4)	60.4 (7.7)					
CECS[62]	Connecticut	588 / 665	93.4	60.2 (9.6)	61.5 (10.8)	28 hospitals in Connecticut	60%	RDD	65%	2004-2008

Subject characteristics were generally comparable between cases and controls with the exception of BMI (higher on average in cases), nulliparity (lower on average in cases), and ever smoking (lower on average in cases). The pooled (unadjusted) odds ratio reflecting a BMI increase of 5 kg/m^2 on the odds of type I endometrial cancer was 1.58 (95% CI 1.54-1.63). This estimate varied in by study analyses, from a low of 1.51 (95% CI 1.37-1.65) in the USC LA study, to a high of 1.77 (95% CI 1.59-1.98) in the SECS study.

Table 5.2 Subject characteristics

Characteristic	Cases	Controls
	N=7,163	**N=10,733**
Race, n (%)		
White	5,413 (75.6)	8,486 (79.1)
Black	120 (1.7)	463 (4.3)
Asian	1,348 (18.8)	1,542 (14.4)
Hawaiian/Pacific Islander	48 (0.7)	89 (0.8)
Other	131 (1.8)	85 (0.8)
Missing/unknown	103 (1.4)	68 (0.6)
BMI, mean (SD)	29.5 (7.8)	25.7 (5.1)
BMI categories, n (%)		
Underweight (<18.5)	92 (1.3)	283 (2.6)
Normal (18.5-<25)	2,251 (31.4)	5,283 (49.2)
Overweight (25 to <30)	1,994 (27.8)	3,252 (30.3)
Obese Class I (30 to <35)	1,247 (17.4)	1,249 (11.6)
Obese Class II & III (35+)	1,475 (20.6)	582 (5.4)
Missing/unknown	104 (1.5)	84 (0.8)
Age at menarche, n (%)		
<11	505 (7.1)	463 (4.3)
11-12	1,259 (17.6)	1,494 (13.9)
13-14	4,163 (58.1)	6,565 (61.2)
15+	1,156 (16.1)	2,124 (19.8)
Missing/unknown	80 (1.1)	87 (0.8)
Parity, n (%)		
0	1,308 (18.3)	1,152 (10.7)
1	1,394 (19.5)	2,019 (18.8)
2	2,052 (28.7)	3,491 (32.5)
3-4	1,904 (26.6)	3,140 (29.3)
5+	408 (5.7)	877 (8.2)
Missing/unknown	97 (1.4)	54 (0.5)

Characteristic	Cases	Controls
HRT ever users, n (%)		
Yes	2,099 (29.3)	3,127 (29.1)
No	4,914 (68.6)	7,385 (68.8)
Missing/unknown	150 (2.1)	221 (2.1)
OC ever users, n (%)		
Yes	2,631 (36.73)	4,408 (41.07)
No	4,477 (62.5)	6,276 (58.5)
Missing/unknown	55 (0.8)	49 (0.5)
Smoking, n (%)		
Never	4,617 (25.3)	5,804 (54.1)
Past	1,811 (25.3)	2,835 (26.4)
Current	675 (9.4)	2,071 (19.3)
Missing/unknown	60 (0.8)	23 (0.2)

Table 5.3 Selection bias adjustment fixed by scenario and trial

Trial	$e^{\beta_{SY}}$	$e^{\beta_{SX}}$	$e^{\beta_{SZ_1}}$	$e^{\beta_{SZ_2}}$	$e^{\beta_{SYX}}$
Scenario A: No Y*X interaction					
1a	1.5	1.1	1.5	1.5	1
2a	2.0	1.2	2.0	2.0	1
3a	5.0	1.3	5.0	5.0	1
4a	2.0	0.9	2.0	2.0	1
5a	2.0	0.8	2.0	2.0	1
Scenario B: Positive Y*X product term					
1b	1.5	1.1	1.5	1.5	1.1
2b	2.0	1.2	2.0	2.0	1.1
3b	5.0	1.3	5.0	5.0	1.1
4b	2.0	0.9	2.0	2.0	1.1
5b	2.0	0.8	2.0	2.0	1.1
Scenario B: Negative Y*X product term					
1c	1.5	1.1	1.5	1.5	0.9
2c	2.0	1.2	2.0	2.0	0.9
3c	5.0	1.3	5.0	5.0	0.9
4c	2.0	0.9	2.0	2.0	0.9
5c	2.0	0.8	2.0	2.0	0.9

In scenario A, we introduced low to moderate strength fixed priors for the log odds relating endometrial cancer to the probability of selection (β_{SY}), the log odds relating BMI (per 5 kg/m^2 increase) to the probability of selection (β_{SX}), the log odds relating white race to the probability of selection (β_{SZ_1}), and the log odds relating ever use of hormone replacement therapy to the probability of selection (β_{SZ_2}). The pooled odds ratio estimate increased from 1.58 (95% CI 1.54-1.63) to 1.74 (95% CI 1.58-1.92) in the most extreme trial for this scenario (trial 3a), which assigned OR of 5.0 to β_{SY}, β_{SZ_1}, β_{SZ_2}, and an OR of 1.3 to β_{SX}. All other trials saw modest increases in the OR (less that 0.1). The by-study analysis reflected the results of the pooled adjustment, with the most extreme change detected in trial 3a. In trials 4a and 5a, which included a negative prior for β_{SX}, results were attenuated towards the null in several of the smaller case-control studies.

In scenario B, we used the same low to moderate strength fixed priors as were applied in scenario A, this time introducing a slight positive value (OR=1.1) for the β_{SYX}, the modification by X on the effect of Y on S (or the modification by X for the effect of X on S, by the symmetry property of the odds ratio). When this term is introduced, adjusted results remain similar to those in scenario A, with some exaggeration of adjusted OR's in smaller studies; the pooled estimates are however, were almost identical to those from scenario A. As in scenario A, trial 3 resulted in the most extreme shifts in odds ratios after adjustment.

In scenario C, we introduced a slight negative value (OR=0.9) for the β_{SYX}. As in scenarios A and B, the pooled estimate shifted upwards, but this time it was most marked in trial 5, which imposed a negative relationship between X and S (OR=0.8). In by-study analysis, all point estimates were shifted upwards, some by more than 30%.

Table 5.4 Scenario A: Selection bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer

Study	Sample size (cases / controls)	OR-adj[1] (95% CI) per 5 kg/m^2 increase					
		Unadjusted analysis	Trial 1a	Trial 2a	Trial 3a	Trial 4a	Trial 5a
Pooled	6,836 / 10,287	1.58 (1.54, 1.63)	1.62 (1.54, 1.69)	1.65 (1.54, 1.77)	1.74 (1.58, 1.92)	1.64 (1.59, 1.70)	1.65 (1.61, 1.68)
Edge	415 / 464	1.62 (1.45, 1.82)	1.65 (1.36, 2.01)	1.69 (1.27, 2.25)	1.80 (1.17, 2.76)	1.69 (1.49, 1.93)	1.71 (1.56, 1.87)
WISE	546 / 1,574	1.66 (1.49, 1.85)	1.67 (1.39, 2.01)	1.70 (1.31, 2.22)	1.77 (1.19, 2.63)	1.65 (1.44, 1.88)	1.64 (1.49, 1.80)
Hawaii case-control	313 / 337	1.57 (1.35, 1.83)	1.64 (1.31, 2.06)	1.70 (1.24, 2.33)	1.84 (1.16, 2.91)	1.58 (1.38, 1.81)	1.52 (1.39, 1.66)
SECS	1,060 / 1,207	1.77 (1.59, 1.98)	1.78 (1.53, 2.06)	1.79 (1.48, 2.17)	1.91 (1.45, 2.52)	1.79 (1.61, 1.98)	1.80 (1.67, 1.95)
PECS	427 / 1,832	1.46 (1.31, 1.63)	1.48 (1.21, 1.80)	1.51 (1.13, 2.02)	1.57 (1.02, 2.44)	1.48 (1.27, 1.72)	1.48 (1.33, 1.65)
US Case-control	313 / 305	1.56 (1.38, 1.77)	1.54 (1.25, 1.90)	1.56 (1.15, 2.11)	1.66 (1.05, 2.60)	1.59 (1.38, 1.82)	1.64 (1.50, 1.80)
Alberta	472 / 1,029	1.62 (1.48, 1.77)	1.62 (1.38, 1.91)	1.66 (1.30, 2.11)	1.72 (1.20, 2.46)	1.63 (1.45, 1.82)	1.63 (1.51, 1.76)
USC LA	787 / 791	1.51 (1.37, 1.65)	1.54 (1.31, 1.82)	1.58 (1.24, 2.02)	1.65 (1.14, 2.39)	1.65 (1.46, 1.87)	1.71 (1.56, 1.87)
ANECS	704 / 1,071	1.74 (1.59, 1.90)	1.84 (1.59, 2.13)	1.90 (1.53, 2.34)	1.97 (1.44, 2.70)	1.82 (1.64, 2.01)	1.79 (1.67, 1.92)
PEDS	436 / 506	1.63 (1.45, 1.83)	1.65 (1.36, 2.01)	1.68 (1.27, 2.24)	1.78 (1.16, 2.75)	1.60 (1.41, 1.82)	1.56 (1.44, 1.70)
WNYDS	232 / 639	1.74 (1.52, 2.00)	1.74 (1.37, 2.21)	1.77 (1.24, 2.53)	1.88 (1.10, 3.21)	1.79 (1.50, 2.13)	1.81 (1.60, 2.05)
Turin	267 / 197	1.66 (1.36, 2.03)	1.65 (1.18, 2.30)	1.68 (1.03, 2.75)	1.76 (0.83, 3.73)	1.56 (1.24, 1.96)	1.49 (1.28, 1.75)
CECS	567 / 632	1.67 (1.52, 1.83)	1.70 (1.44, 2.01)	1.74 (1.37, 2.22)	1.81 (1.26, 2.59)	1.66 (1.49, 1.86)	1.62 (1.51, 1.74)

[1]Conditional logistic regression, adjusted for age (in 5 year intervals), white vs. nonwhite race, parity, smoking status, age at menarche, ever use of oral contraception, and ever use of any type of hormone replacement therapy.

Table 5.5 Scenario B: Selection bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer

Study	Sample size (cases / controls)	OR-adj[1] (95% CI) per 5 kg/m^2 increase					
		Unadjusted analysis	Trial 1a	Trial 2a	Trial 3a	Trial 4a	Trial 5a
Pooled	6,836 / 10,287	1.58 (1.54, 1.63)	1.51 (1.43, 1.59)	1.60 (1.49, 1.72)	1.76 (1.61, 1.93)	1.50 (1.44, 1.56)	1.49 (1.44, 1.54)
Edge	415 / 464	1.62 (1.45, 1.82)	1.55 (1.25, 1.92)	1.65 (1.21, 2.23)	1.81 (1.24, 2.66)	1.55 (1.33, 1.79)	1.60 (1.38, 1.85)
WISE	546 / 1,574	1.66 (1.49, 1.85)	1.57 (1.28, 1.94)	1.66 (1.24, 2.22)	1.84 (1.28, 2.66)	1.52 (1.30, 1.78)	1.43 (1.23, 1.66)
Hawaii case-control	313 / 337	1.57 (1.35, 1.83)	1.52 (1.18, 1.95)	1.61 (1.15, 2.26)	1.77 (1.13, 2.76)	1.44 (1.24, 1.68)	1.39 (1.25, 1.54)
SECS	1,060 / 1,207	1.77 (1.59, 1.98)	1.65 (1.40, 1.95)	1.70 (1.37, 2.10)	1.90 (1.41, 2.56)	1.63 (1.45, 1.84)	1.64 (1.50, 1.80)
PECS	427 / 1,832	1.46 (1.31, 1.63)	1.38 (1.10, 1.74)	1.47 (1.06, 2.03)	1.62 (1.08, 2.41)	1.35 (1.13, 1.62)	1.33 (1.09, 1.62)
US Case-control	313 / 305	1.56 (1.38, 1.77)	1.45 (1.15, 1.82)	1.53 (1.10, 2.11)	1.70 (1.13, 2.55)	1.45 (1.24, 1.69)	1.49 (1.28, 1.73)
Alberta	472 / 1,029	1.62 (1.48, 1.77)	1.53 (1.27, 1.84)	1.62 (1.25, 2.12)	1.78 (1.29, 2.46)	1.49 (1.30, 1.70)	1.45 (1.27, 1.65)
USC LA	787 / 791	1.51 (1.37, 1.65)	1.44 (1.20, 1.74)	1.54 (1.18, 2.00)	1.67 (1.21, 2.31)	1.51 (1.30, 1.74)	1.52 (1.30, 1.78)
ANECS	704 / 1,071	1.74 (1.59, 1.90)	1.72 (1.47, 2.01)	1.83 (1.47, 2.29)	2.00 (1.52, 2.64)	1.66 (1.48, 1.85)	1.65 (1.49, 1.83)
PEDS	436 / 506	1.63 (1.45, 1.83)	1.53 (1.24, 1.90)	1.63 (1.20, 2.20)	1.81 (1.23, 2.65)	1.46 (1.27, 1.68)	1.42 (1.24, 1.63)
WNYDS	232 / 639	1.74 (1.52, 2.00)	1.63 (1.23, 2.15)	1.73 (1.17, 2.55)	1.91 (1.18, 3.10)	1.63 (1.32, 2.00)	1.62 (1.30, 2.03)
Turin	267 / 197	1.66 (1.36, 2.03)	1.53 (1.05, 2.21)	1.63 (0.96, 2.76)	1.80 (0.93, 3.48)	1.42 (1.10, 1.84)	1.36 (1.04, 1.78)
CECS	567 / 632	1.67 (1.52, 1.83)	1.60 (1.33, 1.91)	1.70 (1.31, 2.21)	1.86 (1.35, 2.57)	1.52 (1.35, 1.72)	1.52 (1.35, 1.71)

[1]Conditional logistic regression, adjusted for age (in 5 year intervals), white vs. nonwhite race, parity, smoking status, age at menarche, ever use of oral contraception, and ever use of any type of hormone replacement therapy.

Table 5.6 Scenario C: Selection bias adjusted estimates for the effect of BMI (per 5 kg/m^2 increase) on type I endometrial cancer

Study	Sample size (cases / controls)	OR-adj[1] (95% CI) per 5 kg/m^2 increase					
		Unadjusted analysis	Trial 1a	Trial 2a	Trial 3a	Trial 4a	Trial 5a
Pooled	6,836 / 10,287	1.58 (1.54, 1.63)	1.77 (1.70, 1.85)	1.76 (1.66, 1.87)	1.77 (1.61, 1.95)	1.83 (1.78, 1.88)	1.84 (1.80, 1.88)
Edge	415 / 464	1.62 (1.45, 1.82)	1.81 (1.52, 2.15)	1.80 (1.39, 2.33)	1.82 (1.20, 2.75)	1.88 (1.67, 2.12)	1.90 (1.75, 2.06)
WISE	546 / 1,574	1.66 (1.49, 1.85)	1.82 (1.55, 2.13)	1.80 (1.42, 2.27)	1.78 (1.22, 2.59)	1.81 (1.62, 2.03)	1.82 (1.68, 1.97)
Hawaii case-control	313 / 337	1.57 (1.35, 1.83)	1.82 (1.48, 2.25)	1.86 (1.39, 2.48)	1.92 (1.24, 2.97)	1.77 (1.56, 2.01)	1.70 (1.56, 1.84)
SECS	1,060 / 1,207	1.77 (1.59, 1.98)	1.95 (1.71, 2.23)	1.94 (1.64, 2.31)	1.99 (1.55, 2.55)	1.98 (1.81, 2.17)	2.00 (1.87, 2.14)
PECS	427 / 1,832	1.46 (1.31, 1.63)	1.62 (1.37, 1.91)	1.60 (1.24, 2.06)	1.58 (1.04, 2.41)	1.65 (1.46, 1.87)	1.66 (1.52, 1.81)
US Case-control	313 / 305	1.56 (1.38, 1.77)	1.69 (1.40, 2.04)	1.65 (1.25, 2.18)	1.66 (1.07, 2.57)	1.76 (1.56, 1.99)	1.82 (1.68, 1.97)
Alberta	472 / 1,029	1.62 (1.48, 1.77)	1.77 (1.54, 2.04)	1.75 (1.41, 2.17)	1.73 (1.22, 2.44)	1.81 (1.64, 1.99)	1.82 (1.70, 1.94)
USC LA	787 / 791	1.51 (1.37, 1.65)	1.70 (1.46, 1.97)	1.68 (1.34, 2.10)	1.66 (1.16, 2.38)	1.83 (1.64, 2.05)	1.90 (1.75, 2.06)
ANECS	704 / 1,071	1.74 (1.59, 1.90)	2.02 (1.76, 2.31)	2.02 (1.66, 2.47)	2.00 (1.47, 2.72)	2.02 (1.85, 2.22)	2.00 (1.88, 2.13)
PEDS	436 / 506	1.63 (1.45, 1.83)	1.83 (1.53, 2.18)	1.80 (1.39, 2.35)	1.80 (1.18, 2.75)	1.80 (1.60, 2.03)	1.76 (1.63, 1.90)
WNYDS	232 / 639	1.74 (1.52, 2.00)	1.91 (1.55, 2.35)	1.88 (1.38, 2.57)	1.88 (1.12, 3.17)	1.99 (1.71, 2.30)	2.02 (1.82, 2.24)
Turin	267 / 197	1.66 (1.36, 2.03)	1.82 (1.35, 2.47)	1.81 (1.15, 2.83)	1.78 (0.86, 3.69)	1.76 (1.43, 2.16)	1.69 (1.47, 1.95)
CECS	567 / 632	1.67 (1.52, 1.83)	1.86 (1.60, 2.17)	1.85 (1.48, 2.32)	1.82 (1.28, 2.58)	1.86 (1.68, 2.05)	1.82 (1.70, 1.94)

[1]Conditional logistic regression, adjusted for age (in 5 year intervals), white vs. nonwhite race, parity, smoking status, age at menarche, ever use of oral contraception, and ever use of any type of hormone replacement therapy.

5.5. DISCUSSION

We have demonstrated a flexible sensitivity analysis technique using record-level simulation of selection probabilities to perform IPSW for adjustment of selection bias. We assumed that the exposure, BMI, case/control status, white race (vs. non-white race) and ever use of hormone replacement therapy would directly affect the probability of participation in the case-control study, we additionally controlled for participant age, parity, age at menarche, and ever use of oral contraception in both unadjusted and selection bias-adjusted results. We found that the BMI and endometrial cancer relationship, one of the strongest predictors of type I EC, is robust to small to moderate levels of the type of collider bias we simulated, but in the presence of larger effects, or substantial heterogeneity, unadjusted results may underreport the true magnitude of the effect. By-study analysis demonstrated that the impact of the bias model varied greatly, from study to study; this is likely a reflection of varying exposure distributions, case to control ratios, and prevalence of non-white race and ever use of hormone replacement therapy by study. We considered that most of our model variables would positively influence selection; for example, we hypothesized that heavier women, white women, and women who had previously been exposed to hormone replacement therapy would be more likely to participate in a study. But, but in two trials per scenario (trials 4 and 5) we reverse the direction of this prior to assume a negative relationship between BMI and participation. These two trials are compatible with the idea that heavier women might be less likely to participate in a case-control trial compared to women of normal weight. While it is difficult to know whether overweight women would be less responsive to case-control studies because extensive data on non-responders is rare, there is evidence that overweight women are less likely to participate in screening programs for endometrial and cervical cancers [63, 64]. A range of plausible priors, when paired with an

adequate sensitivity analysis are useful to examine a bias mechanism in the absence of definitive information about the direction or magnitude of one or more of the biasing effects. In this case, an argument could be made for obese women being more or less likely to participate, and we have provided quantitative information that can be used for both sides of this discussion.

We also found that the introduction of a negative product term β_{SYX} , which implies the additional hypothesis that overweight women with endometrial cancer would be even less likely to participate, resulted in the most exaggerated (away from the null) selection bias adjusted estimates, especially when combined with these small negative parameters for the relationship between BMI and participation (trials 4c and 5c). If these priors are compatible with the true causal effects, the presence of selection bias in these studies would be shifting results towards the null. Our use of stabilized weights helped to minimize the difference between the numerator and the denominator of the weight, eliminating extreme weights and achieving better, more robust performance of the weighting procedure. [65]

This study has a number of limitations. Our use of fixed hypothetical priors is instructive with regard to potential performance of the bias model, but not necessarily reflective of the level of uncertainty one may desire when approaching a selection bias adjustment. More realistic priors, with adequate distributions could be obtained from data collected on subjects who refused, or extracted from a follow-up study that suffered from loss related to background characteristics of the subjects enrolled (such as race or ever use of hormone replacement therapy). Additionally, application of the same bias parameters for all studies is potentially unrealistic, especially considering the variation in study design and recruitment successes. A benefit of the record-level adjustment, in comparison to external formula adjustment is the option for bias parameters to be

applied in a study-specific fashion, taking into account suspected selection bias as a result of control recruitment, refusal rates, and the ethnic distribution of the source population.

6. CONCLUSIONS

The purpose of this dissertation was to develop and carefully demonstrate novel probabilistic imputation methods for analysis of bias due to unmeasured confounding and non-response that make fuller use of the observed record level data. Our bias modeling techniques employ probability equations that are used to simulate or impute unmeasured confounding variables or population weights for adjustment of non-response bias. These models are fit using bias parameters, and actual values of the individual-level data in the source database, yielding augmented data that can be used to conduct bias-naïve analysis and bias adjusted sensitivity analyses.

In studies 1 and 2 we elucidated the general methodology behind these algorithms and conducted extensive simulation studies to support the proof of concept in hypothetical data sources. In all simulation studies, the objective was to demonstrate our ability to recapture the joint distribution in the unbiased source population, either via imputing unmeasured confounding variables or re-weighting the selected population to produce effect estimates that would be achieved by analyzing the entire source population. When employing true empirical priors, our methods were completely accurate and unbiased, as would be expected mathematically. In study 2, we also demonstrated performance under slight misspecification of priors. In both studies 1 and 2, we discovered and emphasized the importance of fully saturated probability equations when simulating the confounding or selection probability variables. Liberal invocation of Bayes theorem in these algorithms explains this need, but its importance can also be visualized via

DAGs. In a DAG, conditioning on any collider variable will introduce dependencies between the parents of said variable. Since our simulation equations control for all DAG variables, including the outcome, these conditional dependencies are unavoidable and thus must be circumvented by using fully saturated models. Use of fully saturated models can grow to be quite unwieldy however, as we saw in applied settings. Exercises in studies 1 and 2 comparing results from the fully saturated to the unsaturated model forms were designed to demonstrate the differences in performance that one might expect when choosing a flexible versus a more restrictive model form. Exclusion of interaction parameters in the model is tantamount to assuming a sharp null (i.e., a very strong prior assumption that the interaction does not exist). We noted that the confounding variable simulation technique is much more sensitive to exclusion of these parameters than the selection probability simulations. This finding can also be visualized via the DAG structure for collider bias compared to that of the structure for confounding bias. Confounding variables are parent variables to the exposure and outcome (both of which become colliders in the presence of 2 or more confounders), the selection probability is itself a collider, and conditioning its parents may introduce dependencies between other confounding variables in the models, but the selection variable itself is not necessarily included in these additional dependencies. DAG structures, including our ability to visualize the behavior of the bias as well as formalizing it through probability equations based on these diagrams, were a key component of this work.

In studies 3 and 4 we used the Epidemiology of Endometrial Cancer Consortium (E2C2) risk factors database to demonstrate the use of our algorithms to support quantitative bias modeling in a large, multi-center epidemiologic study. For the confounding variable imputation (paper 3), we demonstrated a very comprehensive use of empirical priors derived from the complete subject

subset and applied to the full E2C2 database in order to fill values for 3 partially measured variables, in sequence. In paper 4, we demonstrated the use of fixed hypothetical priors in exploring the robustness of the BMI-EC relationship to significant levels of non-response bias. The E2C2 consortium is an ideal study to explore bias analysis techniques specifically designed for studying rare cancer etiology in a data pooling project. Our algorithms are highly relevant to a pooling project because they allow for multiple bias parameters to be placed directly in the dataset. As we were able to successfully demonstrate in studies 3 and 4, these bias parameters can vary by derivation (empirical or hypothetical), by distribution (fixed or probabilistic), by characteristics at the study level (study type, matching criteria, non-response rates, study location), and by characteristics at the participant level (ethnicity, age, other exposures, and so on.) Furthermore, the flexibility of the algorithms lends particularly well to consortium investigators re-specifying the priors for the bias parameters based on subject matter knowledge, validation studies, etc. before performing any bias analysis.

By utilizing the risk of BMI on type I endometrial cancer in studies 3 and 4, we were able to evaluate our bias analysis methods using an exposure-disease relationship that has been well characterized in the literature. The reason for using a well-characterized relationship, rather than a novel one, was to aid in the qualitative evaluation of the bias model performance (in light of published findings). BMI is one of the strongest risk factors for endometrial cancer, and its causal effect is not under debate. This sets our work apart from the typical quantitative bias analysis, because it is not often the main objective to adjust for unmeasured confounding in well-characterized relationships, but to evaluate the robustness of associations detected in new etiological research. However, especially in study 3, background knowledge of the magnitude and direction of the confounding bias in this relationship improved our ability to discuss and

frame findings from the bias adjusted models. And in study 4, the strength and robustness of the BMI-EC relationship was supported by a sensitivity analysis for non-response bias.

One of the most daunting limitations of this project was the computational power that was needed to perform the MC simulations in a database with over 50,000 participants. For probabilistic bias analysis, every observation in the database had to first be copied at least 1,000 times. This resulted in enormous size working files: for paper 3 alone there were over 500 gigabytes of data files to be processed. We employed raid hard drive configurations and some shortcut strategies in our programming to reduce computational times, but we hit many roadblocks including those related to implementing full probabilistic bias analysis in study 4 (we substituted fixed priors) and achieving the multiple bias algorithm that was proposed last year (as yet undone). Quantitative researchers who plan to use this type of modeling should ensure adequate computational and data storage resources, including systems optimized for fast I/O sequential reads that can be gained from high capacity solid state drives or raided hard drive configurations.

Other limitations include model assumptions – all simulated data were based on causal structures and our prior specifications. While it is always impossible to account for all sources of bias in observational studies, in order for the application section of this work to be adequately demonstrative, we assumed that no further biases, other than the ones modeled, were present. Prior specifications in our application sections (studies 3 and 4) varied in complexity and generalizability. In some cases, the priors may have been outside of the scope of plausibility or subject to doubt by experts in the field. This points to an important aspect of (semi-) Bayesian statistics which is that all the models are sensitive to the inputs, but these inputs are entirely open

for debate, and a good sensitivity analysis will include a range of inputs to satisfy a full consideration of the biasing sources. Additionally, and this is extremely important, all bias-naïve analyses (not adjusted for bias) that we presented assume that there is absolutely no systematic error in our data models; assuming no systematic error is a very extreme prior, much more extreme than many of the priors we applied.

Through this work, we contributed a novel method, as well exploration of this method in a number of hypothetical scenarios and in an applied setting with a very large multi-study database. The application component of this work included demonstrating adjustment for unmeasured confounding and non-response bias in multiple studies at a time, employing both empirical and hypothetical priors, as well as fixed and probabilistic priors, and taking into account varying degrees of bias at both the study-level and participant-level. Additionally we intend to contribute working algorithms to the E2C2 committee, so that these methods can be used in the important task of uncovering the etiology of rare subtypes of endometrial cancer.

Bias modeling is a tool to guide the qualitative discussion of epidemiologic findings in a quantitative fashion. As a departure from traditional discussion of bias using generalizing statements, these methods will continue to gain relevance and importance in health research, especially as researchers design and build larger and more rich combined datasets (from multiple observational studies or routine hospital sources). External formula adjustment has been the major vehicle for bias modeling in the epidemiologic literature but its accessibility is limited due to weighty and complex formulas. The record-level approach that we described and demonstrated in this work is a flexible, transparent, and easy-to-understand approach that can put quantitative bias modeling in the hands of any analyst, without substantive knowledge of the

complex statistics behind external formula adjustment, and additionally allows for bias parameters to be integrated into a shared data source.

7. APPENDIX TABLES

Table 7.1 Study 1: Planned characteristics of the trial cohorts (N=10,000)

Trial	Target model[1]	$P_{(Z1)}$	$P_{(Z2)}$	Mean (Z3)	SD (Z3)	$P_{(Z4)}$		$P(X_0=1)$[2]	$P(Y_0=1)$[3]	$OR_{YX\|Z1, Z2, Z3, Z4}$	$OR_{XZ1\|Y, Z2, Z3, Z4}$	$OR_{XZ2\|Y, Z1, Z3, Z4}$	$OR_{XZ3\|Y, Z1, Z2, Z4}$	$OR_{XZ4\|Y, Z1, Z2, Z3}$		$OR_{YZ1\|X, Z2, Z3, Z4}$	$OR_{YZ2\|X, Z1, Z3, Z4}$	$OR_{YZ3\|X, Z1, Z2, Z4}$	$OR_{YZ4\|X, Z1, Z2, Z3}$	
						Z4=1	Z4=2							Z4=1	Z4=2				Z4=1	Z4=2
1a	All	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.3	1	3	2	2	2	3	3	5	2	2	3
2a	All	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.5	1	3	2	2	2	3	3	5	2	2	3
3a	All	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.7	1	3	2	2	2	3	3	5	2	2	3
4a	All	0.3	0.3	0.0	1.0	0.4	0.3	0.5	0.3	1	3	2	2	2	3	3	5	2	2	3
5a	All	0.3	0.3	0.0	1.0	0.4	0.3	0.7	0.3	1	3	2	2	2	3	3	5	2	2	3
6a	z1, z2	0.5	0.3	0.0	1.0	0.4	0.3	0.3	0.3	1	3	5	2	2	3	3	5	2	2	3
7a	z1, z2	0.5	0.3	0.0	1.0	0.4	0.3	0.3	0.3	1	5	2	2	2	3	5	5	2	2	3
8a	z1, z2	0.7	0.3	0.0	1.0	0.4	0.3	0.3	0.3	1	7	2	2	2	3	7	5	2	2	3
9a	z3	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.3	1	3	2	1.5	2	3	3	5	1.5	2	3
10a	z3	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.3	1	3	2	5	2	3	3	5	5	2	3
11a	z3	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.3	1	3	2	1.2	2	3	3	5	5	2	3
12a	z4	0.3	0.3	0.0	1.0	0.2	0.2	0.3	0.3	1	3	2	2	2	3	3	5	2	2	3
13a	z4	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.3	1	3	2	2	5	3	3	5	2	2	5
14a	z4	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.3	1	3	2	2	7	3	3	5	2	2	2
1b	All	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.3	2	3	2	2	2	3	3	5	2	2	3
2b	All	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.5	2	3	2	2	2	3	3	5	2	2	3
3b	All	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.7	2	3	2	2	2	3	3	5	2	2	3
4b	All	0.3	0.3	0.0	1.0	0.4	0.3	0.5	0.3	2	3	2	2	2	3	3	5	2	2	3
5b	All	0.3	0.3	0.0	1.0	0.4	0.3	0.7	0.3	2	3	2	2	2	3	3	5	2	2	3
6b	z1, z2	0.5	0.3	0.0	1.0	0.4	0.3	0.3	0.3	2	3	5	2	2	3	3	5	2	2	3
7b	z1, z2	0.5	0.3	0.0	1.0	0.4	0.3	0.3	0.3	2	5	2	2	2	3	5	5	2	2	3
8b	z1, z2	0.7	0.3	0.0	1.0	0.4	0.3	0.3	0.3	2	7	2	2	2	3	7	5	2	2	3
9b	z3	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.3	2	3	0.7	1.5	2	3	3	5	1.5	2	3
10b	z3	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.3	2	3	0.7	5	2	3	3	5	5	2	3
11b	z3	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.3	2	3	0.7	1.2	2	3	3	5	5	2	3

Trial	Target model[1]	$P_{(Z1)}$	$P_{(Z2)}$	Mean (Z3)	SD (Z3)	$P_{(Z4)}$		$P(X_0=1)^2$	$P(Y_0=1)^3$	$OR_{YX\|Z1, Z2, Z3, Z4}$	$OR_{XZ1\|Y, Z2, Z3, Z4}$	$OR_{XZ2\|Y, Z1, Z3, Z4}$	$OR_{XZ3\|Y, Z1, Z2, Z4}$	$OR_{XZ4\|Y, Z1, Z2, Z3}$		$OR_{YZ1\|X, Z2, Z3, Z4}$	$OR_{YZ2\|X, Z1, Z3, Z4}$	$OR_{YZ3\|X, Z1, Z2, Z4}$	$OR_{YZ4\|X, Z1, Z2, Z3}$	
12b	z4	0.3	0.3	0.0	1.0	0.2	0.2	0.3	0.3	2	3	0.7	2	2	3	3	5	2	2	3
13b	z4	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.3	2	3	0.7	2	5	3	3	5	2	2	5
14b	z4	0.3	0.3	0.0	1.0	0.4	0.3	0.3	0.3	2	3	0.7	2	7	3	3	5	2	2	2

[1]Trials 1-5 were designed to check all imputation models. Trials 6-8 were designed for the dichotomous confounder model. Trials 9-11 were designed for the continuous confounder model. Trials 12-14 were designed for the trichotomous confounder model.

[2]$P(X_0=1)$ is the background exposure prevalence $P(X=1|Z_1=Z_2=Z_3=Z_4=0)$

[3]$P(Y_0=1)$ is the background risk of disease $P(Y=1|X=Z_1=Z_2=Z_3=Z_4=0)$

Table 7.2 Study 1: Actual characteristics of the trial cohorts (N=10,000)

Trial	Target model[1]	$P_{(Z1)}$	$P_{(Z2)}$	Mean (Z3)	$SD_{(Z3)}$	$P_{(Z4)}$		$P(X_0=1)^2$	$P(Y_0=1)^3$	$OR_{YX\|Z1, Z2, Z3,Z4}$	$OR_{XZ1\|Y, Z2, Z3,Z4}$	$OR_{XZ2\|Y, Z1, Z3,Z4}$	$OR_{XZ3\|Y, Z1, Z2,Z4}$	$OR_{XZ4\|Y, Z1, Z2,Z3}$		$OR_{YZ1\|X, Z2, Z3,Z4}$	$OR_{YZ2\|X, Z1, Z3,Z4}$	$OR_{Y Z3\|X, Z1, Z2,Z4}$	$OR_{YZ4\|X, Z1, Z2,Z3}$	
						Z4=1	Z4=2							Z4=1	Z4=2				Z4=1	Z4=2
1a	All	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.31	0.95	3.20	1.83	2.11	1.92	2.93	2.75	5.15	2.02	1.95	3.07
2a	All	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.48	0.97	3.18	1.81	2.10	1.91	2.91	2.96	5.30	1.97	2.14	2.93
3a	All	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.70	0.96	3.18	1.81	2.10	1.91	2.91	3.06	5.72	1.98	2.14	3.15
4a	All	0.29	0.30	0.01	1.00	0.40	0.30	0.49	0.31	0.97	3.06	2.00	2.07	2.15	3.54	2.73	5.13	2.01	1.94	3.06
5a	All	0.29	0.30	0.01	1.00	0.40	0.30	0.72	0.27	1.05	3.37	2.06	2.04	2.10	3.22	2.70	5.09	1.99	1.92	3.02
6a	z1, z2	0.49	0.30	0.01	1.00	0.40	0.30	0.31	0.28	0.97	3.17	5.16	2.09	2.06	3.14	2.84	5.11	2.01	2.05	3.13
7a	z1, z2	0.49	0.30	0.01	1.00	0.40	0.30	0.31	0.28	0.96	5.28	1.95	2.05	1.86	3.06	4.62	5.32	2.01	2.07	3.23
8a	z1, z2	0.71	0.30	0.01	1.00	0.40	0.30	0.26	0.30	0.97	7.09	2.00	2.05	2.06	2.92	7.59	5.31	2.02	1.97	3.19
9a	z3	0.29	0.30	0.01	1.00	0.40	0.30	0.29	0.31	0.91	3.12	1.94	1.55	1.96	2.97	2.87	5.68	1.50	2.04	3.32
10a	z3	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.30	0.95	3.08	1.98	5.20	2.05	3.03	2.95	4.78	5.27	2.02	3.27
11a	z3	0.29	0.30	0.01	1.00	0.40	0.30	0.29	0.32	0.99	3.03	1.99	1.21	1.99	3.06	2.92	4.75	5.18	2.00	3.24
12a	z4	0.29	0.30	0.00	1.00	0.20	0.20	0.29	0.30	0.97	2.93	1.99	2.10	2.05	3.31	2.99	5.17	2.05	1.92	2.82
13a	z4	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.31	0.95	3.02	2.06	2.07	5.04	2.93	2.81	5.29	2.02	1.97	4.85
14a	z4	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.31	0.99	3.02	2.06	2.04	6.97	2.88	2.70	5.14	2.04	1.95	2.12
1b	All	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.31	1.91	2.88	1.58	1.96	1.77	2.60	2.86	4.98	1.97	1.98	2.96
2b	All	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.48	1.93	2.98	1.67	2.01	1.82	2.71	3.03	5.53	2.01	2.17	3.20
3b	All	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.70	1.94	2.63	1.60	1.89	1.94	3.07	3.35	5.91	1.99	2.13	3.08
4b	All	0.29	0.30	0.01	1.00	0.40	0.30	0.49	0.31	2.01	2.90	1.65	1.86	1.87	2.77	2.90	5.10	1.91	1.97	2.74
5b	All	0.49	0.30	0.01	1.00	0.40	0.30	0.72	0.27	2.12	2.76	4.21	1.91	1.87	2.70	2.86	5.22	1.92	2.14	2.97
6b	z1, z2	0.49	0.30	0.01	1.00	0.40	0.30	0.31	0.28	1.95	4.43	1.64	1.89	1.70	2.69	2.86	5.34	1.96	2.11	3.15
7b	z1, z2	0.71	0.30	0.01	1.00	0.40	0.30	0.31	0.28	1.88	5.81	1.73	1.93	1.96	2.64	4.53	5.21	1.99	2.09	3.04
8b	z1, z2	0.29	0.30	0.01	1.00	0.40	0.30	0.26	0.30	1.83	2.71	1.56	1.47	1.76	2.54	7.76	5.64	2.03	1.83	3.16
9b	z3	0.29	0.30	0.01	1.00	0.40	0.30	0.29	0.31	1.88	2.70	1.63	4.25	1.88	2.63	2.82	5.53	1.48	2.13	3.10
10b	z3	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.30	1.98	2.74	1.70	1.03	1.86	2.75	2.83	5.05	5.22	1.96	3.23
11b	z3	0.29	0.30	0.01	1.00	0.40	0.30	0.29	0.32	1.94	2.55	1.61	1.92	1.89	2.87	2.84	5.10	5.25	1.98	2.99
12b	z4	0.29	0.30	0.00	1.00	0.20	0.20	0.29	0.30	1.91	2.63	1.67	1.90	4.56	2.38	2.97	5.49	1.99	1.91	3.00

Trial	Target model[1]	$P_{(Z1)}$	$P_{(Z2)}$	Mean $_{(Z3)}$	$SD_{(Z3)}$	$P_{(Z4)}$		$P(X_0=1)^2$	$P(Y_0=1)^3$	$OR_{YX\|Z1, Z2, Z3,Z4}$	$OR_{XZ1\|Y, Z2, Z3,Z4}$	$OR_{XZ2\|Y, Z1, Z3,Z4}$	$OR_{XZ3\|Y, Z1, Z2,Z4}$	$OR_{XZ4\|Y, Z1, Z2,Z3}$		$OR_{YZ1\|X, Z2,}$	$OR_{YZ2\|X, Z1, Z3,Z4}$	$OR_{Y Z3\|X,}$	$OR_{YZ4\|X, Z1, Z2,Z3}$	
13b	z4	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.31	2.00	2.62	1.66	1.86	6.29	2.63	2.75	5.27	1.93	2.04	4.76
14b	z4	0.29	0.30	0.01	1.00	0.40	0.30	0.30	0.31	1.99	2.88	1.58	1.96	1.77	2.60	2.78	5.07	1.96	2.08	1.96

[1]Trials 1-5 were designed to check all imputation models. Trials 6-8 were designed for the dichotomous confounder model. Trials 9-11 were designed for the continuous confounder model. Trials 12-14 were designed for the trichotomous confounder model.
[2]$P(X_0=1)$ is the background exposure prevalence $P(X=1|Z_1=Z_2=Z_3=Z_4=0)$
[3]$P(Y_0=1)$ is the background risk of disease $P(Y=1|X=Z_1=Z_2=Z_3=Z_4=0)$

Table 7.3 Study 1: Full simulation results: imputing a dichotomous confounder (Z_1, Z_2) model (N=10,000, reps=1,000) – Non-saturated models

Trial	OR true[1] (LCL, UCL)	Imputing for Z_1 Only					Imputing for Z_2 Only				
		OR unadj[2] (LCL, UCL)	OR adj[3] (LSI, USI)	Bias[4]	RMSE[5]	Coverage[6]	OR unadj[2] (LCL, UCL)	OR adj[3] (LSI, USI)	Bias[4]	RMSE[5]	Coverage[6]
1a	0.95 (0.86, 1.05)	1.18 (1.07, 1.29)	0.95 (0.86, 1.05)	0.0020	0.0500	100%	1.13 (1.03, 1.24)	0.95 (0.86, 1.05)	0.0001	0.0500	100%
2a	0.97 (0.87, 1.08)	1.21 (1.09, 1.35)	0.98 (0.88, 1.09)	0.0027	0.0552	100%	1.15 (1.03, 1.27)	0.97 (0.87, 1.08)	-0.0028	0.0553	100%
3a	0.95 (0.84, 1.09)	1.19 (1.05, 1.35)	0.96 (0.84, 1.10)	0.0054	0.0680	100%	1.12 (0.98, 1.27)	0.95 (0.83, 1.09)	-0.0032	0.0680	100%
4a	0.97 (0.87, 1.08)	1.18 (1.07, 1.31)	0.98 (0.88, 1.08)	0.0037	0.0539	100%	1.19 (1.08, 1.31)	0.98 (0.88, 1.09)	0.0051	0.0540	100%
5a	1.05 (0.92, 1.19)	1.27 (1.12, 1.44)	1.05 (0.92, 1.19)	0.0039	0.0650	100%	1.28 (1.13, 1.44)	1.05 (0.92, 1.19)	0.0031	0.0649	100%
6a	0.98 (0.89, 1.08)	1.28 (1.16, 1.41)	0.98 (0.89, 1.09)	0.0025	0.0519	100%	1.48 (1.35, 1.63)	0.98 (0.89, 1.09)	0.0013	0.0518	100%
7a	0.96 (0.86, 1.06)	1.63 (1.48, 1.79)	0.96 (0.86, 1.07)	0.0033	0.0549	100%	1.15 (1.04, 1.27)	0.96 (0.86, 1.07)	0.0022	0.0548	100%
8a	0.96 (0.85, 1.09)	2.13 (1.91, 2.37)	0.98 (0.87, 1.11)	0.0154	0.0656	100%	1.17 (1.04, 1.31)	0.97 (0.86, 1.10)	0.0099	0.0644	100%
1b	1.91 (1.73, 2.11)	2.32 (2.10, 2.55)	1.91 (1.73, 2.11)	0.0031	0.0512	100%	2.13 (1.93, 2.34)	1.90 (1.72, 2.11)	-0.0052	0.0514	100%
2b	1.92 (1.71, 2.16)	2.35 (2.09, 2.63)	1.93 (1.72, 2.17)	0.0087	0.0605	100%	2.16 (1.93, 2.42)	1.91 (1.70, 2.15)	-0.0112	0.0610	100%
3b	1.94 (1.67, 2.26)	2.40 (2.07, 2.78)	1.95 (1.67, 2.27)	0.0047	0.0779	100%	2.20 (1.90, 2.56)	1.93 (1.65, 2.24)	-0.0167	0.0797	100%
4b	2.01 (1.81, 2.23)	2.39 (2.16, 2.65)	2.02 (1.82, 2.24)	0.0097	0.0549	100%	2.28 (2.06, 2.52)	2.01 (1.81, 2.24)	0.0011	0.0541	100%
5b	2.11 (1.86, 2.40)	2.52 (2.23, 2.85)	2.13 (1.87, 2.41)	0.0104	0.0656	100%	2.41 (2.14, 2.72)	2.12 (1.86, 2.40)	0.0016	0.0648	100%
6b	1.93 (1.74, 2.15)	2.48 (2.24, 2.75)	1.93 (1.74, 2.15)	0.0012	0.0540	100%	2.80 (2.53, 3.09)	1.92 (1.73, 2.14)	-0.0069	0.0544	100%
7b	1.88 (1.68, 2.10)	2.99 (2.70, 3.31)	1.88 (1.69, 2.10)	0.0065	0.0567	100%	2.13 (1.91, 2.36)	1.87 (1.67, 2.09)	-0.0067	0.0567	100%
8b	1.83 (1.61, 2.08)	3.77 (3.36, 4.23)	1.86 (1.63, 2.11)	0.0280	0.0717	100%	2.11 (1.87, 2.39)	1.85 (1.62, 2.10)	0.0197	0.0688	100%

[1]True odds ratio adjusting for simulated Z
[2]Unadjusted odds ratio, without control for Z
[3]Odds ratio adjusting for $\hat{Z}$ (bias model): 50th percentile of the median of the estimates, lower and upper simulation interval (LSI, USI)
[4]Bias=True OR – Bias adjusted OR
[5] $\text{RMSE} = \sqrt{(\text{Bias})^2 + (\text{Median SE obtained from the bias adjusted OR distributes})^2}$
[6]95% confidence interval coverage as defined by the % of times OR true is between the lower and upper limit of the 95% confidence interval of the repetition OR.

Table 7.4 Study 1: Full simulation results for Imputing two dichotomous confounder (Z_1, Z_2) simultaneously model (N=10,000, reps=1,000) – Non-saturated models

Trial	OR true[1] (LCL, UCL)	OR unadj[2] (LCL, UCL)	OR adj[3] (LSI, USI)	Bias[4]	RMSE[5]	Coverage[6]
1a	0.95 (0.86, 1.05)	1.36 (1.25, 1.49)	0.95 (0.86, 1.05)	-0.0019	0.0500	100%
2a	0.97 (0.87, 1.08)	1.39 (1.26, 1.54)	0.97 (0.87, 1.08)	-0.0008	0.0552	100%
3a	0.95 (0.84, 1.09)	1.36 (1.20, 1.54)	0.96 (0.84, 1.09)	0.0024	0.0679	100%
4a	0.97 (0.87, 1.08)	1.40 (1.27, 1.54)	0.98 (0.88, 1.08)	0.0035	0.0539	100%
5a	1.05 (0.92, 1.19)	1.51 (1.34, 1.70)	1.05 (0.92, 1.19)	0.0012	0.0649	100%
6a	0.98 (0.89, 1.08)	1.80 (1.64, 1.97)	0.98 (0.88, 1.09)	0.0091	0.0535	100%
7a	0.96 (0.86, 1.06)	1.82 (1.66, 2.00)	0.96 (0.86, 1.07)	0.0009	0.0549	100%
8a	0.96 (0.85, 1.09)	2.33 (2.10, 2.59)	0.99 (0.87, 1.12)	0.0228	0.0676	100%
1b	1.91 (1.73, 2.11)	2.52 (2.30, 2.77)	1.90 (1.72, 2.10)	-0.0086	0.0519	100%
2b	1.92 (1.71, 2.16)	2.59 (2.32, 2.89)	1.92 (1.71, 2.16)	-0.0060	0.0602	100%
3b	1.94 (1.67, 2.26)	2.68 (2.32, 3.11)	1.94 (1.66, 2.26)	-0.0057	0.0780	100%
4b	2.01 (1.81, 2.23)	2.66 (2.41, 2.93)	2.01 (1.81, 2.24)	0.0021	0.0541	100%
5b	2.11 (1.86, 2.40)	2.81 (2.50, 3.17)	2.12 (1.87, 2.41)	0.0044	0.0651	100%
6b	1.93 (1.74, 2.15)	3.34 (3.03, 3.68)	1.97 (1.77, 2.19)	0.0191	0.0574	100%
7b	1.88 (1.68, 2.10)	3.22 (2.92, 3.55)	1.87 (1.68, 2.09)	-0.0026	0.0564	100%
8b	1.83 (1.61, 2.08)	4.04 (3.61, 4.51)	1.87 (1.64, 2.12)	0.0363	0.0754	100%

[1]True odds ratio adjusting for simulated Z_1 and Z_2.
[2]Unadjusted odds ratio, without control for Z
[3]Odds ratio adjusting for $\hat{Z}$ (bias model): 50th percentile of the median of the estimates, lower and upper simulation interval (LSI, USI)
[4]Bias=True OR – Bias adjusted OR
[5] $\text{RMSE} = \sqrt{(\text{Bias})^2 + (\text{Median SE obtained from the bias adjusted OR distributes})^2}$
[6]95% confidence interval coverage as defined by the % of times OR true is between the lower and upper limit of the 95% confidence interval of the repetition OR.

Table 7.5 Study 1: Full simulation results from imputing a single continuous confounder (Z_3) model (N=10,000, reps=1,000) – Non-saturated models

Trial	OR true[1] (LCL, UCL)	OR unadj[2] (LCL, UCL)	OR adj[3] (LSI, USI)	Bias[4]	RMSE[5]	Coverage[6]
1a	0.95 (0.86, 1.05)	1.47 (1.34, 1.60)	0.95 (0.86, 1.05)	0.0027	0.0500	100%
2a	0.97 (0.87, 1.08)	1.49 (1.35, 1.65)	0.98 (0.88, 1.09)	0.0049	0.0553	100%
3a	0.95 (0.84, 1.09)	1.48 (1.31, 1.67)	0.96 (0.84, 1.10)	0.0091	0.0683	100%
4a	0.97 (0.87, 1.08)	1.49 (1.35, 1.64)	0.98 (0.88, 1.09)	0.0052	0.0541	100%
5a	1.05 (0.92, 1.19)	1.59 (1.42, 1.80)	1.05 (0.93, 1.19)	0.0053	0.0650	100%
9a	0.91 (0.83, 1.00)	1.08 (0.98, 1.18)	0.91 (0.83, 1.00)	0.0009	0.0484	100%
10a	0.95 (0.84, 1.06)	3.79 (3.47, 4.14)	0.99 (0.88, 1.10)	0.0394	0.0701	99%
11a	0.99 (0.89, 1.10)	1.22 (1.12, 1.33)	0.99 (0.89, 1.10)	-0.0042	0.0542	100%
1b	1.91 (1.73, 2.11)	2.77 (2.52, 3.04)	1.91 (1.73, 2.11)	0.0016	0.0512	100%
2b	1.92 (1.71, 2.16)	2.85 (2.55, 3.18)	1.93 (1.72, 2.17)	0.0083	0.0604	100%
3b	1.94 (1.67, 2.26)	2.93 (2.53, 3.38)	1.95 (1.68, 2.27)	0.0082	0.0782	100%
4b	2.01 (1.81, 2.23)	2.85 (2.58, 3.14)	2.02 (1.82, 2.25)	0.0122	0.0554	100%
5b	2.11 (1.86, 2.40)	3.00 (2.66, 3.38)	2.13 (1.88, 2.42)	0.0154	0.0666	100%
9b	1.88 (1.70, 2.07)	2.16 (1.96, 2.41)	1.88 (1.70, 2.07)	0.0007	0.0497	100%
10b	1.98 (1.76, 2.22)	6.68 (6.07, 7.35)	2.06 (1.84, 2.31)	0.0824	0.1014	99%
11b	1.94 (1.74, 2.16)	1.95 (1.78, 2.13)	1.89 (1.70, 2.11)	-0.0514	0.0753	100%

[1]True odds ratio adjusting for simulated Z_3
[2]Unadjusted odds ratio, without control for Z_3
[3]Odds ratio adjusting for $\widehat{Z_3}$ (bias model): simulation interval derived from 50th percentile of the median of the estimates, lower and upper simulation interval (LSI, USI)
[4]Bias=True OR – Bias adjusted OR
[5] $\text{RMSE} = \sqrt{(\text{Bias})^2 + (\text{Median SE obtained from the bias adjusted OR distributes})^2}$
[6]95% confidence interval coverage as defined by the % of times OR true is between the lower and upper limit of the 95% confidence interval of the repetition OR.

Table 7.6 Study 1: Full simulation results for Imputing a single trichotomous confounder (Z_4) model (N=10,000, reps=1,000) – Non-saturated models

Trial	OR true[1] (LCL, UCL)	OR unadj[2] (LCL, UCL)	OR adj[3] (LSI, USI)	Bias[4]	RMSE[5]	Coverage[6]
1a	0.95 (0.86, 1.05)	1.13 (1.03, 1.24)	0.95 (0.86, 1.05)	0.0039	0.0501	100%
2a	0.97 (0.87, 1.08)	1.16 (1.04, 1.29)	0.98 (0.88, 1.09)	0.0049	0.0554	100%
3a	0.95 (0.84, 1.09)	1.15 (1.01, 1.31)	0.96 (0.84, 1.10)	0.0055	0.0681	100%
4a	0.97 (0.87, 1.08)	1.20 (1.08, 1.32)	0.98 (0.89, 1.09)	0.0123	0.0551	100%
5a	1.05 (0.92, 1.19)	1.26 (1.12, 1.43)	1.05 (0.93, 1.20)	0.0089	0.0654	100%
12a	0.97 (0.88, 1.06)	1.18 (1.08, 1.30)	0.97 (0.88, 1.07)	0.0051	0.0500	100%
13a	0.95 (0.85, 1.05)	1.18 (1.07, 1.30)	0.96 (0.86, 1.06)	0.0102	0.0540	100%
14a	0.99 (0.89, 1.09)	1.22 (1.10, 1.34)	0.99 (0.89, 1.10)	0.0027	0.0528	100%
1b	1.91 (1.73, 2.11)	2.21 (2.00, 2.44)	1.91 (1.73, 2.12)	0.0044	0.0513	100%
2b	1.92 (1.71, 2.16)	2.26 (2.02, 2.54)	1.93 (1.72, 2.17)	0.0066	0.0602	100%
3b	1.94 (1.67, 2.26)	2.30 (1.98, 2.67)	1.95 (1.68, 2.27)	0.0082	0.0782	100%
4b	2.01 (1.81, 2.23)	2.38 (2.15, 2.64)	2.03 (1.82, 2.25)	0.0174	0.0567	100%
5b	2.11 (1.86, 2.40)	2.50 (2.21, 2.83)	2.13 (1.88, 2.42)	0.0156	0.0667	100%
12b	1.91 (1.73, 2.11)	2.29 (2.08, 2.52)	1.92 (1.74, 2.12)	0.0119	0.0521	100%
13b	2.00 (1.80, 2.22)	2.40 (2.17, 2.65)	2.01 (1.80, 2.23)	0.0055	0.0546	100%
14b	1.99 (1.79, 2.20)	2.46 (2.23, 2.72)	1.99 (1.79, 2.21)	0.0058	0.0536	100%

[1]True odds ratio adjusting for simulated Z_4
[2]Unadjusted odds ratio, without control for Z_4
[3]Odds ratio adjusting for imputed $\widehat{Z_4}$ (bias model), simulation interval derived from 50th percentile of the median of the estimates, lower and upper simulation interval (LSI, USI)
[4]Bias=True OR – Bias adjusted OR
[5] $\text{RMSE} = \sqrt{(\text{Bias})^2 + (\text{Median SE obtained from the bias adjusted OR distributes})^2}$
[6]95% confidence interval coverage as defined by the % of times OR true is between the lower and upper limit of the 95% confidence interval of the repetition OR.

Table 7.7 Study 3: Smoking bias model parameters, based on complete case analysis, by study type[1]

Model term	Cohort studies			Case-control studies		
	Estimate	**SE**	**Pr > |t|**	**Estimate**	**SE**	**Pr > |t|**
Intercept						
Current smoking	-4.901	0.710	0.0001	-2.476	0.710	0.004
Past smoking	-3.134	0.710	0.0022	-0.886	0.710	0.2342
BMI (per kg/m^2)	0.015	0.012	0.21	0.002	0.013	0.8718
Race						
White	Ref	Ref	Ref	Ref	Ref	Ref
Black	-0.762	0.696	0.2735	1.309	0.818	0.1096
Asian	-1.009	0.743	0.174	-3.224	0.777	<.0001
Hawaiian or Pacific Islander	-1.964	1.238	0.1125	0.085	1.342	0.9495
Other	-3.143	0.814	0.0001	-2.536	1.193	0.0335
Age at menarche						
<11 years	Ref	Ref	Ref	Ref	Ref	Ref
11-12 years	1.387	0.569	0.0149	0.649	0.574	0.2578
13-14 years	1.080	0.536	0.0439	0.780	0.519	0.1327
15+ years	1.770	0.637	0.0055	1.750	0.606	0.0039
Parity						
Nulliparous	Ref	Ref	Ref	Ref	Ref	Ref
1 child	0.205	0.478	0.6686	0.914	0.489	0.0617
2 children	0.377	0.400	0.3455	1.065	0.416	0.0104
3-4 children	0.666	0.392	0.0893	0.866	0.422	0.04
5+ children	2.025	0.536	0.0002	1.057	0.649	0.1033
Ever use of oral contraception	1.553	0.280	<.0001	0.899	0.289	0.0018
Case status	-0.181	0.334	0.5877	-0.717	0.307	0.0196
Age in years	0.046	0.010	<.0001	0.008	0.010	0.4363
BMI*Race interactions						
BMI*Black	0.002	0.010	0.8574	-0.023	0.013	0.0827
BMI*Asian	0.039	0.016	0.0176	0.041	0.016	0.01
BMI*Hawaiian or pacific islander	0.020	0.017	0.2316	-0.003	0.022	0.8937
BMI*Other	0.048	0.013	0.0003	0.026	0.020	0.1782
BMI*Age at menarche interactions						
BMI*11-12 years	-0.021	0.008	0.0137	-0.008	0.009	0.3783
BMI*13-14 years	-0.018	0.008	0.0282	-0.015	0.008	0.0753
BMI*15+ years	-0.020	0.010	0.0489	-0.012	0.011	0.2638
BMI*Parity interactions						
BMI*1 child	0.007	0.008	0.3694	-0.001	0.009	0.9144
BMI*2 children	-0.004	0.007	0.5186	0.001	0.007	0.9065
BMI*3-4 children	-0.009	0.007	0.158	0.003	0.007	0.632
BMI*5+ children	-0.023	0.008	0.0058	-0.005	0.010	0.5912
BMI*Ever use of oral contraception	-0.004	0.005	0.33	-0.005	0.005	0.2887
BMI*Case status	0.004	0.005	0.4179	0.011	0.005	0.0219

Model term	Cohort studies			Case-control studies		
	Estimate	SE	Pr > |t|	Estimate	SE	Pr > |t|
Black Race*Age at menarche interactions						
Black*11-12 years	0.020	0.252	0.9379	-0.161	0.342	0.6384
Black*13-14 years	-0.053	0.250	0.8318	-0.058	0.306	0.8502
Black*15+years	-0.517	0.296	0.0807	0.195	0.356	0.5845
Black race*Parity interactions						
Black*1 child	-0.248	0.244	0.31	-0.270	0.317	0.393
Black*2 children	-0.431	0.220	0.0501	-0.475	0.297	0.1101
Black*3-4 children	-0.485	0.203	0.0169	-0.379	0.296	0.2003
Black*5+ children	-0.372	0.222	0.0947	-0.711	0.327	0.0297
Black race*Ever use of oral contraception	-0.051	0.139	0.7136	0.328	0.180	0.0673
Black race*Case status	0.016	0.188	0.9313	-0.238	0.227	0.2959
Black race*Age in years	0.022	0.009	0.0158	-0.005	0.010	0.6076
Asian race*Age at menarche interactions						
Asian*11-12 years	-0.296	0.280	0.2892	0.392	0.483	0.4172
Asian*13-14 years	-0.387	0.277	0.1631	0.214	0.438	0.6243
Asian*15+years	-0.380	0.338	0.2606	0.512	0.457	0.2628
Asian race*Parity interactions						
Asian*1 child	0.022	0.263	0.9335	-0.543	0.265	0.0404
Asian*2 children	-0.213	0.218	0.3282	-0.255	0.258	0.3228
Asian*3-4 children	-0.277	0.219	0.2063	0.063	0.251	0.8028
Asian*5+ children	-0.145	0.340	0.6704	0.536	0.316	0.0897
Asian race*Ever use of oral contraception	-0.276	0.162	0.0891	0.109	0.173	0.53
Asian race*Case status	0.376	0.184	0.0411	0.133	0.165	0.4194
Asian race*Age in years	0.001	0.009	0.9218	0.012	0.009	0.1635
Hawaiian race*Age at menarche interactions						
Hawaiian*11-12 years	0.757	0.476	0.1113	-0.741	0.710	0.2965
Hawaiian*13-14 years	0.602	0.496	0.2247	-0.515	0.631	0.4147
Hawaiian*15+years	0.871	0.562	0.1208	0.436	0.748	0.5594
Hawaiian race*Parity interactions						
Hawaiian*1 child	0.857	0.595	0.1503	0.312	0.714	0.6618
Hawaiian*2 children	0.008	0.553	0.989	0.338	0.718	0.6373
Hawaiian*3-4 children	-0.353	0.508	0.4873	0.570	0.630	0.3658
Hawaiian*5+ children	0.283	0.514	0.5816	0.759	0.646	0.2399
Hawaiian race*Ever use of oral contraception	0.091	0.269	0.7342	1.046	0.445	0.0189
Hawaiian race*Case status	0.080	0.332	0.8096	-0.037	0.415	0.9282
Hawaiian race*Age in years	0.014	0.016	0.3587	-0.027	0.018	0.144
Other race*Age at menarche interactions						
Other*11-12 years	-0.042	0.295	0.888	0.485	0.594	0.4139
Other*13-14 years	0.400	0.289	0.1659	0.706	0.556	0.204
Other*15+years	0.056	0.339	0.8694	1.251	0.660	0.058
Other race*Parity interactions						
Other*1 child	0.328	0.330	0.3216	-0.165	0.509	0.746

Model term	Cohort studies			Case-control studies		
	Estimate	SE	Pr > |t|	Estimate	SE	Pr > |t|
Other*2 children	0.188	0.276	0.4961	-0.266	0.433	0.5395
Other*3-4 children	0.096	0.254	0.7064	-0.010	0.448	0.9816
Other*5+ children	-0.057	0.264	0.8304	0.985	0.543	0.0698
Other race*Ever use of oral contraception	-0.137	0.156	0.3801	-0.050	0.311	0.8718
Other race*Case status	-0.027	0.193	0.8892	0.588	0.335	0.0795
Other race*Age in years	0.020	0.009	0.0316	0.010	0.015	0.5032
Age at menarche*Parity interactions						
11-12 years*1 child	-0.269	0.215	0.2112	-0.213	0.252	0.3996
11-12 years*2 children	-0.130	0.179	0.4684	-0.413	0.213	0.052
11-12 years*3-4 children	-0.156	0.171	0.3617	-0.128	0.214	0.55
11-12 years*5+ children	0.050	0.229	0.8286	0.029	0.309	0.9251
11-12 years*Ever use of oral contraception	-0.030	0.121	0.8044	-0.014	0.145	0.922
11-12 years*Case status	-0.021	0.143	0.8806	-0.073	0.153	0.6336
11-12 years*Age in years	-0.012	0.008	0.1323	-0.004	0.008	0.6307
Age at menarche*Parity interactions						
13-14 years*1 child	-0.036	0.205	0.8622	0.076	0.228	0.7388
13-14 years*2 children	-0.085	0.171	0.6195	-0.146	0.194	0.4537
13-14 years*3-4 children	-0.223	0.164	0.1751	0.090	0.194	0.6432
13-14 years*5+ children	-0.046	0.220	0.8324	0.160	0.283	0.5706
13-14 years*Ever use of oral contraception	-0.099	0.116	0.3915	-0.112	0.131	0.3962
13-14 years*Case status	0.045	0.136	0.7381	-0.054	0.139	0.6995
13-14 years*Age in years	-0.007	0.007	0.3364	-0.006	0.007	0.4305
Age at menarche*Parity interactions						
15+ years*1 child	-0.331	0.245	0.1756	-0.365	0.266	0.1693
15+ years*2 children	-0.489	0.205	0.017	-0.570	0.230	0.0133
15+ years*3-4 children	-0.436	0.196	0.0262	-0.290	0.231	0.2095
15+ years*5+ children	-0.206	0.254	0.4177	-0.265	0.326	0.4169
15+ years*Ever use of oral contraception	-0.098	0.137	0.4764	-0.266	0.154	0.0852
15+ years*Case status	0.084	0.167	0.6144	-0.075	0.164	0.6464
15+ years*Age in years	-0.012	0.009	0.1471	-0.014	0.008	0.0904
Parity interactions						
1 child*Ever use of oral contraception	0.025	0.111	0.8181	-0.189	0.123	0.1259
1 child*Case status	-0.337	0.122	0.0057	-0.170	0.120	0.1548
1 child*Age in years	-0.004	0.006	0.5188	-0.006	0.006	0.2874
2 children*Ever use of oral contraception	-0.038	0.090	0.6705	-0.248	0.106	0.0187
2 children*Case status	-0.130	0.097	0.1803	-0.005	0.103	0.9617
2 children*Age in years	-0.002	0.005	0.7311	-0.006	0.005	0.1956
3-4 children*Ever use of oral contraception	0.070	0.087	0.4206	-0.159	0.107	0.1372
3-4 children*Case status	-0.076	0.094	0.42	-0.185	0.105	0.0765
3-4 children*Age in years	-0.006	0.005	0.1949	-0.010	0.005	0.0598
5+ children*Ever use of oral contraception	0.060	0.108	0.5818	-0.285	0.142	0.0448
5+ children*Case status	-0.111	0.131	0.3936	-0.282	0.155	0.0679

Model term	Cohort studies			Case-control studies		
	Estimate	SE	Pr > |t|	Estimate	SE	Pr > |t|
5+ children*Age in years	-0.024	0.007	0.0003	-0.007	0.008	0.3549
Ever use of oral contraception*Case status	-0.040	0.069	0.5598	-0.260	0.071	0.0003
Ever use of oral contraception*Age in years	-0.025	0.003	<.0001	-0.010	0.003	0.0031
Case status*Age in years	0.000	0.004	0.962	0.011	0.004	0.0045

[1]Cumulative logistic regression model for smoking status (current or previous vs. never smoker), stratified by study type, with random intercept by study site and fixed effects as appear in the table.

Table 7.8 Study 3: Estrogen bias model parameters, based on complete case analysis, by study type[1]

Model term	Cohort studies			Case-control studies		
	Estimate	**SE**	**Pr > |t|**	**Estimate**	**SE**	**Pr > |t|**
Intercept	-10.078	1.852	0.0006	-7.564	1.566	0.0003
BMI (per kg/m^2)	-0.034	0.031	0.2716	0.054	0.027	0.0438
Race						
White	Ref	Ref	Ref	Ref	Ref	Ref
Black	-2.746	1.678	0.1018	-0.523	1.911	0.7842
Asian	-2.289	1.502	0.1276	-0.399	1.463	0.7851
Hawaiian or Pacific Islander	-5.442	6.056	0.3689	-0.782	26.499	0.9765
Other	0.641	1.432	0.6544	-1.610	2.859	0.5733
Age at menarche						
<11 years	Ref	Ref	Ref	Ref	Ref	Ref
11-12 years	0.499	1.284	0.6975	0.498	1.207	0.6796
13-14 years	-0.174	1.216	0.8864	-0.534	1.097	0.6266
15+ years	0.660	1.465	0.6526	0.911	1.238	0.4616
Parity						
Nulliparous	Ref	Ref	Ref	Ref	Ref	Ref
1 child	0.037	1.128	0.9738	0.972	0.955	0.3087
2 children	-0.129	0.897	0.8853	0.807	0.796	0.3101
3-4 children	0.807	0.859	0.3472	0.106	0.808	0.8953
5+ children	1.388	1.207	0.2499	1.049	1.254	0.4028
Ever use of oral contraception	0.858	0.655	0.1906	0.739	0.562	0.1884
Smoking status						
Never smoking	Ref	Ref	Ref	Ref	Ref	Ref
Prior smoking	-0.749	0.651	0.2494	0.510	0.590	0.3874
Current smoking	-0.194	0.927	0.8347	-0.110	0.763	0.8851
Case status	1.706	0.689	0.0133	1.893	0.559	0.0007
Age in years	0.097	0.022	<.0001	0.064	0.020	0.0012
BMI*Race interactions						
BMI*Black	-0.007	0.023	0.7493	-0.034	0.033	0.3026
BMI*Asian	0.062	0.032	0.0546	-0.061	0.029	0.0357
BMI*Hawaiian or pacific islander	0.078	0.033	0.0201	-0.360	0.295	0.2228
BMI*Other	0.000	0.025	0.9903	0.059	0.050	0.237
BMI*Age at menarche interactions						
BMI*11-12 years	0.006	0.022	0.7923	-0.025	0.021	0.219
BMI*13-14 years	0.006	0.021	0.7851	-0.006	0.019	0.7453
BMI*15+ years	0.021	0.026	0.4217	0.014	0.022	0.5343
BMI*Parity interactions						
BMI*1 child	-0.014	0.021	0.4935	-0.013	0.017	0.4468
BMI*2 children	0.007	0.017	0.6626	-0.017	0.014	0.2411
BMI*3-4 children	-0.009	0.016	0.5828	-0.004	0.014	0.7914
BMI*5+ children	-0.014	0.020	0.4882	-0.023	0.020	0.252

Model term	Cohort studies			Case-control studies		
	Estimate	SE	Pr > |t|	Estimate	SE	Pr > |t|
BMI*Ever use of oral contraception	-0.007	0.011	0.5674	-0.020	0.009	0.0312
BMI*Smoking interactions						
BMI*Prior smoking	0.026	0.012	0.0264	-0.012	0.010	0.2257
BMI*Current smoking	0.051	0.017	0.0025	-0.008	0.014	0.5718
BMI*Case status	-0.047	0.012	<.0001	-0.078	0.009	<.0001
Black Race*Age at menarche interactions						
Black*11-12 years	-0.335	0.584	0.5659	0.921	0.820	0.2611
Black*13-14 years	-0.084	0.575	0.8835	0.492	0.778	0.5266
Black*15+years	-0.059	0.643	0.9267	0.728	0.850	0.392
Black race*Parity interactions						
Black*1 child	0.151	0.561	0.7881	0.762	0.835	0.3619
Black*2 children	0.774	0.477	0.1043	1.328	0.785	0.0905
Black*3-4 children	0.651	0.446	0.1447	1.086	0.796	0.1722
Black*5+ children	0.528	0.490	0.2815	1.280	0.839	0.1269
Black race*Ever use of oral contraception	0.484	0.308	0.1167	-0.308	0.370	0.4052
Black race*Smoking interactions						
Black race*Prior smoking	0.032	0.296	0.9144	0.229	0.361	0.5268
Black race*Current smoking	-0.028	0.370	0.9405	-0.069	0.437	0.874
Black race*Case status	0.148	0.367	0.687	-0.460	0.458	0.3156
Black race*Age in years	0.021	0.020	0.2776	-0.004	0.021	0.8538
Asian Race*Age at menarche interactions						
Asian*11-12 years	-0.139	0.582	0.8117	0.596	1.000	0.5509
Asian*13-14 years	0.224	0.571	0.6946	0.576	0.925	0.5338
Asian*15+years	-0.119	0.670	0.8588	0.251	0.942	0.7903
Asian race*Parity interactions						
Asian*1 child	0.056	0.478	0.9071	0.300	0.384	0.4351
Asian*2 children	0.031	0.412	0.9391	0.278	0.364	0.445
Asian*3-4 children	-0.183	0.401	0.6486	-0.807	0.397	0.0421
Asian*5+ children	0.045	0.598	0.9405	-1.179	0.617	0.0561
Asian race*Ever use of oral contraception	0.476	0.335	0.155	-0.386	0.245	0.1161
Asian race*Smoking interactions						
Asian race*Prior smoking	0.497	0.286	0.0823	0.344	0.361	0.3409
Asian race*Current smoking	0.057	0.498	0.9092	-0.110	0.528	0.8352
Asian race*Case status	-0.312	0.317	0.3246	-0.190	0.219	0.3858
Asian race*Age in years	-0.007	0.017	0.6711	0.032	0.016	0.0385
Hawaiian Race*Age at menarche interactions						
Hawaiian*11-12 years	-1.358	0.802	0.0905	-4.739	15.272	0.7563
Hawaiian*13-14 years	-1.712	0.931	0.0659	2.407	12.717	0.8499
Hawaiian*15+years	-0.180	0.853	0.833	3.275	12.796	0.798
Hawaiian race*Parity interactions						
Hawaiian*1 child	4.481	5.608	0.4242	3.658	18.489	0.8432
Hawaiian*2 children	4.444	5.583	0.4261	-0.054	20.058	0.9979

Model term	Cohort studies			Case-control studies		
	Estimate	SE	Pr > |t|	Estimate	SE	Pr > |t|
Hawaiian*3-4 children	4.907	5.567	0.378	5.104	15.445	0.7411
Hawaiian*5+ children	5.036	5.559	0.365	6.774	15.508	0.6623
Hawaiian race*Ever use of oral contraception	0.498	0.552	0.3667	4.614	7.287	0.5266
Hawaiian race*Smoking interactions						
Hawaiian race*Prior smoking	-0.114	0.606	0.8505	-2.474	2.111	0.2411
Hawaiian race*Current smoking	0.069	0.649	0.9157	-6.178	6.679	0.355
Hawaiian race*Case status	0.578	0.604	0.3388	1.116	1.545	0.4702
Hawaiian race*Age in years	-0.025	0.033	0.4464	-0.110	0.142	0.4412
Other Race*Age at menarche interactions						
Other*11-12 years	0.322	0.567	0.57	-0.707	1.148	0.538
Other*13-14 years	0.462	0.555	0.4056	-0.704	1.061	0.5073
Other*15+years	-0.030	0.658	0.9637	-4.796	5.509	0.384
Other race*Parity interactions						
Other*1 child	0.345	0.567	0.5424	0.382	1.569	0.8078
Other*2 children	0.250	0.497	0.6153	-0.123	1.271	0.923
Other*3-4 children	0.218	0.444	0.6229	1.207	1.231	0.3269
Other*5+ children	0.410	0.475	0.3882	0.321	1.638	0.8446
Other race*Ever use of oral contraception	0.404	0.288	0.1611	0.584	0.842	0.4881
Other race*Smoking interactions						
Other race*Prior smoking	0.055	0.276	0.8417	-0.443	0.831	0.5943
Other race*Current smoking	-0.831	0.527	0.1151	-0.559	1.360	0.6813
Other race*Case status	0.669	0.293	0.0223	0.600	0.947	0.5261
Other race*Age in years	-0.031	0.016	0.0452	-0.021	0.040	0.5981
Age at menarche*Parity interactions						
11-12 years*1 child	1.106	0.593	0.0623	-0.012	0.476	0.98
11-12 years*2 children	-0.345	0.389	0.3753	-0.068	0.379	0.8574
11-12 years*3-4 children	0.011	0.365	0.977	0.309	0.376	0.412
11-12 years*5+ children	-0.138	0.477	0.7721	-0.148	0.576	0.7973
11-12 years*Ever use of oral contraception	0.206	0.279	0.4612	-0.292	0.266	0.2721
11-12 years*Smoking interactions						
11-12 years*Prior smoking	0.461	0.283	0.1031	0.052	0.271	0.8469
11-12 years*Current smoking	0.072	0.375	0.8488	0.392	0.387	0.3105
11-12 years*Case status	0.013	0.305	0.9659	0.088	0.260	0.7354
11-12 years*Age in years	-0.015	0.017	0.3671	0.009	0.016	0.5741
Age at menarche*Parity interactions						
13-14 years*1 child	0.905	0.577	0.1169	0.176	0.429	0.6806
13-14 years*2 children	-0.096	0.370	0.7946	-0.045	0.343	0.8962
13-14 years*3-4 children	-0.103	0.350	0.7679	0.143	0.341	0.6739
13-14 years*5+ children	-0.187	0.454	0.6796	0.043	0.522	0.9341
13-14 years*Ever use of oral contraception	0.041	0.267	0.8783	-0.060	0.244	0.8041
13-14 years*Smoking interactions						
13-14 years*Prior smoking	0.390	0.271	0.1491	0.126	0.248	0.612

Model term	Cohort studies			Case-control studies		
	Estimate	SE	Pr > |t|	Estimate	SE	Pr > |t|
13-14 years*Current smoking	-0.137	0.362	0.7043	0.323	0.355	0.363
13-14 years*Case status	0.080	0.290	0.7831	0.171	0.238	0.4725
13-14 years*Age in years	-0.003	0.016	0.8731	0.011	0.015	0.4592
Age at menarche*Parity interactions						
15+ years*1 child	1.386	0.655	0.0344	-0.203	0.474	0.668
15+ years*2 children	0.477	0.454	0.2932	-0.423	0.388	0.2755
15+ years*3-4 children	0.343	0.429	0.4237	-0.336	0.391	0.3908
15+ years*5+ children	0.515	0.534	0.3353	-0.406	0.587	0.4889
15+ years*Ever use of oral contraception	0.196	0.315	0.5332	0.052	0.279	0.8528
15+ years*Smoking interactions						
15+ years*Prior smoking	0.515	0.319	0.1061	0.129	0.284	0.651
15+ years*Current smoking	0.252	0.416	0.5446	0.160	0.395	0.6866
15+ years*Case status	0.582	0.340	0.0869	0.063	0.270	0.8166
15+ years*Age in years	-0.033	0.019	0.0804	-0.017	0.016	0.2916
Parity interactions						
1 child*Ever use of oral contraception	-0.116	0.254	0.6466	-0.010	0.219	0.9647
1 child*Prior smoking	-0.694	0.232	0.0027	-0.123	0.216	0.5671
1 child*Current smoking	-0.911	0.322	0.0047	0.185	0.267	0.4887
1 child*Case status	0.006	0.246	0.9789	0.207	0.188	0.2693
1 child*Age in years	0.000	0.013	0.9777	-0.015	0.011	0.1758
2 children*Ever use of oral contraception	-0.101	0.204	0.6212	-0.029	0.184	0.8753
2 children*Prior smoking	-0.516	0.188	0.0059	-0.063	0.176	0.7188
2 children*Current smoking	-0.797	0.258	0.002	0.083	0.235	0.7249
2 children*Case status	-0.129	0.194	0.5062	-0.039	0.159	0.8043
2 children*Age in years	0.006	0.011	0.5997	-0.005	0.010	0.6053
3-4 children*Ever use of oral contraception	-0.166	0.192	0.3877	0.067	0.184	0.7147
3-4 children*Prior smoking	-0.498	0.176	0.0047	0.010	0.175	0.9529
3-4 children*Current smoking	-0.845	0.238	0.0004	0.206	0.237	0.3847
3-4 children*Case status	-0.167	0.182	0.3586	0.138	0.160	0.3894
3-4 children*Age in years	0.000	0.010	0.985	-0.009	0.010	0.3747
5+ children*Ever use of oral contraception	-0.635	0.241	0.0084	0.000	0.248	0.9996
5+ children*Prior smoking	-0.310	0.234	0.1858	-0.357	0.253	0.1572
5+ children*Current smoking	-0.764	0.315	0.0154	-0.040	0.332	0.903
5+ children*Case status	-0.017	0.256	0.9478	0.127	0.243	0.6017
5+ children*Age in years	0.001	0.014	0.9353	-0.013	0.015	0.3806
Ever use of oral contraception interactions						
Ever use of oral contraception*Prior smoking	-0.228	0.130	0.0791	0.012	0.119	0.9224
Ever use of oral contraception*Current smoking	-0.238	0.185	0.1976	0.031	0.165	0.851
Ever use of oral contraception*Case status	0.302	0.146	0.0391	-0.029	0.114	0.7983
Ever use of oral contraception*Age in years	-0.011	0.008	0.1507	-0.009	0.007	0.1865
Smoking interactions						

Model term	Cohort studies			Case-control studies		
	Estimate	SE	Pr > |t|	Estimate	SE	Pr > |t|
Prior smoking*Case status	-0.221	0.138	0.1094	-0.006	0.115	0.9557
Prior smoking*Age in years	0.010	0.008	0.1865	0.000	0.007	0.95
Current smoking*Case status	-0.246	0.218	0.259	0.052	0.162	0.7475
Current smoking*Age in years	0.006	0.011	0.5605	-0.002	0.009	0.8469
Case status*Age in years	-0.006	0.008	0.4428	0.006	0.007	0.3435

[1]Binomial logistic regression model for ever use of estrogen only hormone replacement therapy, stratified by study type, with random intercept by study site and fixed effects as appear in the table.

Table 7.9 Study 3: Diabetes bias model parameters, based on complete case analysis, by study type[1]

Model term	Cohort studies			Case-control studies		
	Estimate	**SE**	**Pr > |t|**	**Estimate**	**SE**	**Pr > |t|**
Intercept	-10.850	2.133	0.0014	-8.446	1.531	<.0001
BMI (per kg/m^2)	0.162	0.025	<.0001	0.110	0.022	<.0001
Race						
White	Ref	Ref	Ref	Ref	Ref	Ref
Black	2.039	1.253	0.1037	2.612	1.329	0.0494
Asian	-1.272	1.611	0.43	0.659	1.057	0.5331
Hawaiian or Pacific Islander	3.925	2.304	0.0885	-6.201	15.425	0.6877
Other	0.453	1.503	0.7629	-1.403	2.102	0.5044
Age at menarche						
<11 years	Ref	Ref	Ref	Ref	Ref	Ref
11-12 years	1.308	1.323	0.3228	-1.386	1.133	0.2212
13-14 years	-0.843	1.282	0.511	-1.913	1.026	0.0622
15+ years	1.220	1.594	0.4439	-2.254	1.232	0.0672
Parity						
Nulliparous	Ref	Ref	Ref	Ref	Ref	Ref
1 child	0.055	1.345	0.9675	1.150	0.992	0.2463
2 children	-0.827	1.160	0.4759	0.480	0.910	0.5978
3-4 children	0.563	1.070	0.5991	0.112	0.925	0.9034
5+ children	0.377	1.314	0.7739	1.777	1.323	0.1793
Ever use of oral contraception	2.938	0.777	0.0002	1.306	0.621	0.0355
Smoking status						
Never smoking	Ref	Ref	Ref	Ref	Ref	Ref
Prior smoking	0.575	0.773	0.4569	0.623	0.681	0.3601
Current smoking	1.100	1.207	0.3621	-0.293	0.877	0.7383
Case status	2.341	0.825	0.0046	1.836	0.620	0.0031
Age in years	0.053	0.026	0.0447	0.025	0.021	0.2358
Ever use of EHRT	1.710	1.406	0.2239	0.341	1.103	0.757
BMI*Race interactions						
BMI*Black	-0.041	0.015	0.0062	-0.049	0.018	0.0074
BMI*Asian	0.078	0.030	0.01	0.010	0.017	0.5463
BMI*Hawaiian or pacific islander	-0.050	0.029	0.0876	0.000	0.061	0.9948
BMI*Other	-0.013	0.021	0.5236	-0.041	0.030	0.1733
BMI*Age at menarche interactions						
BMI*11-12 years	0.005	0.016	0.7649	0.040	0.015	0.0075
BMI*13-14 years	0.031	0.015	0.0417	0.031	0.014	0.0215
BMI*15+ years	0.009	0.019	0.638	0.035	0.017	0.0415
BMI*Parity interactions						
BMI*1 child	0.008	0.016	0.6262	0.002	0.014	0.8979
BMI*2 children	0.004	0.014	0.7649	-0.003	0.013	0.8082
BMI*3-4 children	-0.019	0.013	0.1339	-0.005	0.012	0.6691

Model term	Cohort studies			Case-control studies		
	Estimate	SE	Pr > |t|	Estimate	SE	Pr > |t|
BMI*5+ children	-0.021	0.015	0.1625	-0.001	0.017	0.9618
BMI*Ever use of oral contraception	-0.033	0.009	0.0004	-0.014	0.008	0.0875
BMI*Smoking interactions						
BMI*Prior smoking	-0.002	0.009	0.7972	-0.008	0.009	0.3496
BMI*Current smoking	-0.002	0.016	0.8803	0.008	0.013	0.5267
BMI*Case status	-0.034	0.009	0.0003	-0.031	0.009	0.0003
BMI*Ever use of EHRT	0.005	0.019	0.7982	0.029	0.015	0.0509
Black Race*Age at menarche interactions						
Black*11-12 years	0.173	0.426	0.6854	0.652	0.554	0.2398
Black*13-14 years	-0.138	0.429	0.7472	0.864	0.503	0.0861
Black*15+years	-0.015	0.491	0.9763	0.330	0.593	0.5779
Black race*Parity interactions						
Black*1 child	-0.135	0.441	0.76	-0.844	0.503	0.0934
Black*2 children	0.087	0.411	0.8314	-0.239	0.458	0.6018
Black*3-4 children	0.274	0.369	0.4572	-0.231	0.449	0.6073
Black*5+ children	0.829	0.379	0.0286	0.487	0.468	0.2974
Black race*Ever use of oral contraception	0.098	0.235	0.6755	-0.054	0.285	0.8509
Black race*Smoking interactions						
Black race*Prior smoking	-0.008	0.230	0.9738	-0.112	0.284	0.6932
Black race*Current smoking	0.571	0.310	0.0655	-0.556	0.340	0.1019
Black race*Case status	0.050	0.275	0.8552	0.315	0.305	0.3026
Black race*Age in years	-0.007	0.016	0.6452	-0.018	0.017	0.2745
Black race*Ever use of EHRT	-0.137	0.415	0.7422	0.091	0.471	0.8466
Asian Race*Age at menarche interactions						
Asian*11-12 years	-0.770	0.502	0.1251	0.556	0.664	0.4017
Asian*13-14 years	-0.529	0.510	0.3	0.464	0.589	0.4306
Asian*15+years	0.314	0.575	0.5857	0.178	0.611	0.7704
Asian race*Parity interactions						
Asian*1 child	-1.020	0.586	0.0818	-0.190	0.339	0.5757
Asian*2 children	-0.674	0.427	0.1145	0.187	0.324	0.5643
Asian*3-4 children	-0.866	0.414	0.0366	0.347	0.325	0.2843
Asian*5+ children	-1.962	0.832	0.0183	0.562	0.430	0.1907
Asian race*Ever use of oral contraception	-1.120	0.341	0.001	-0.291	0.205	0.156
Asian race*Smoking interactions						
Asian race*Prior smoking	0.106	0.328	0.7469	-0.170	0.341	0.6185
Asian race*Current smoking	-1.096	0.830	0.187	-1.151	0.553	0.0375
Asian race*Case status	-0.812	0.403	0.0439	0.295	0.182	0.1063
Asian race*Age in years	0.049	0.020	0.0126	-0.006	0.012	0.6273
Asian race*Ever use of EHRT	-1.139	0.688	0.098	0.279	0.395	0.48
Hawaiian Race*Age at menarche interactions						
Hawaiian*11-12 years	-1.588	0.715	0.0264	-1.396	2.423	0.5646
Hawaiian*13-14 years	-0.949	0.726	0.1911	-1.968	2.187	0.3683

Model term	Cohort studies			Case-control studies		
	Estimate	SE	Pr > |t|	Estimate	SE	Pr > |t|
Hawaiian*15+years	-1.493	0.921	0.1049	-7.479	6.248	0.2313
Hawaiian race*Parity interactions						
Hawaiian*1 child	-2.062	1.256	0.1006	-1.033	2.299	0.6531
Hawaiian*2 children	-1.783	1.004	0.0757	-0.390	3.285	0.9054
Hawaiian*3-4 children	-0.948	0.763	0.2141	-3.760	2.324	0.1057
Hawaiian*5+ children	-1.160	0.770	0.1322	0.083	1.725	0.9619
Hawaiian race*Ever use of oral contraception	-0.583	0.491	0.2348	5.206	7.410	0.4823
Hawaiian race*Smoking interactions						
Hawaiian race*Prior smoking	0.108	0.541	0.8421	2.167	1.429	0.1295
Hawaiian race*Current smoking	0.437	0.628	0.4867	-0.449	1.610	0.7803
Hawaiian race*Case status	0.171	0.561	0.7606	0.702	1.086	0.5179
Hawaiian race*Age in years	0.017	0.030	0.5753	-0.015	0.081	0.856
Hawaiian race*Ever use of EHRT	0.060	0.797	0.9398	-3.404	10.889	0.7546
Other Race*Age at menarche interactions						
Other*11-12 years	-0.169	0.467	0.7168	1.543	0.964	0.1094
Other*13-14 years	0.184	0.458	0.6873	2.324	0.887	0.0088
Other*15+years	-0.145	0.555	0.7943	2.366	1.070	0.027
Other race*Parity interactions						
Other*1 child	-0.475	0.748	0.5253	1.936	0.774	0.0124
Other*2 children	0.164	0.552	0.7658	0.838	0.705	0.2347
Other*3-4 children	0.344	0.486	0.4796	-0.030	0.803	0.9702
Other*5+ children	1.061	0.482	0.0279	2.141	0.831	0.01
Other race*Ever use of oral contraception	0.169	0.290	0.56	0.323	0.493	0.5119
Other race*Smoking interactions						
Other race*Prior smoking	0.308	0.271	0.2554	0.833	0.482	0.084
Other race*Current smoking	-0.186	0.567	0.7432	0.384	0.695	0.5803
Other race*Case status	-0.240	0.336	0.4753	0.491	0.558	0.3787
Other race*Age in years	-0.001	0.017	0.9645	-0.007	0.024	0.7536
Other race*Ever use of EHRT	-0.274	0.457	0.5478	1.706	0.887	0.0544
Age at menarche*Parity interactions						
11-12 years*1 child	-0.133	0.467	0.7749	-0.658	0.409	0.1083
11-12 years*2 children	-0.270	0.406	0.5067	-0.237	0.378	0.5297
11-12 years*3-4 children	-0.339	0.377	0.3689	0.016	0.395	0.9676
11-12 years*5+ children	-0.360	0.440	0.4137	-1.236	0.487	0.0111
11-12 years*Ever use of oral contraception	-0.544	0.274	0.0474	-0.432	0.265	0.1037
11-12 years*Smoking interactions						
11-12 years*Prior smoking	0.043	0.262	0.8709	-0.385	0.267	0.149
11-12 years*Current smoking	-0.040	0.419	0.9236	0.392	0.408	0.3368
11-12 years*Case status	-0.389	0.267	0.1461	-0.628	0.278	0.0241
11-12 years*Age in years	-0.004	0.017	0.8307	0.022	0.016	0.1575
11-12 years*Ever use of EHRT	0.094	0.458	0.8368	-0.148	0.481	0.7584
Age at menarche*Parity interactions						

Model term	Cohort studies			Case-control studies		
	Estimate	SE	Pr > |t|	Estimate	SE	Pr > |t|
13-14 years*1 child	-0.199	0.459	0.6647	-0.451	0.370	0.2238
13-14 years*2 children	-0.241	0.398	0.5453	-0.070	0.352	0.8416
13-14 years*3-4 children	-0.270	0.370	0.4656	0.360	0.365	0.3246
13-14 years*5+ children	-0.311	0.427	0.4667	-1.026	0.438	0.0192
13-14 years*Ever use of oral contraception	-0.479	0.269	0.0744	-0.422	0.242	0.0808
13-14 years*Smoking interactions						
13-14 years*Prior smoking	0.068	0.256	0.7912	-0.388	0.244	0.1117
13-14 years*Current smoking	-0.139	0.412	0.7355	0.397	0.375	0.2901
13-14 years*Case status	-0.686	0.261	0.0087	-0.410	0.253	0.106
13-14 years*Age in years	0.014	0.016	0.3998	0.029	0.014	0.0429
13-14 years*Ever use of EHRT	-0.346	0.453	0.4455	0.299	0.423	0.4803
Age at menarche*Parity interactions						
15+ years*1 child	-0.623	0.578	0.2809	-0.530	0.446	0.2341
15+ years*2 children	-0.684	0.495	0.167	-0.140	0.426	0.7429
15+ years*3-4 children	-0.444	0.444	0.3178	0.292	0.438	0.5049
15+ years*5+ children	-0.245	0.512	0.6321	-1.050	0.535	0.0496
15+ years*Ever use of oral contraception	-0.221	0.333	0.5064	-0.468	0.285	0.1009
15+ years*Smoking interactions						
15+ years*Prior smoking	0.101	0.318	0.7508	-0.545	0.299	0.0685
15+ years*Current smoking	0.496	0.479	0.3003	0.364	0.427	0.3945
15+ years*Case status	-0.488	0.350	0.1637	-0.468	0.287	0.1033
15+ years*Age in years	-0.010	0.020	0.6113	0.037	0.016	0.0243
15+ years*Ever use of EHRT	-0.153	0.541	0.778	0.114	0.490	0.8157
Parity interactions						
1 child*Ever use of oral contraception	0.013	0.299	0.9662	-0.395	0.243	0.1041
1 child*Prior smoking	0.047	0.267	0.8591	0.512	0.246	0.0371
1 child*Current smoking	-0.076	0.415	0.8543	-0.001	0.313	0.9975
1 child*Case status	-0.163	0.285	0.5674	0.061	0.219	0.7807
1 child*Age in years	0.005	0.017	0.7744	-0.001	0.012	0.9456
1 child*Ever use of EHRT	-0.872	0.498	0.0802	-0.461	0.344	0.1795
2 children*Ever use of oral contraception	-0.084	0.252	0.7385	-0.373	0.223	0.0946
2 children*Prior smoking	-0.189	0.227	0.405	0.171	0.221	0.439
2 children*Current smoking	-0.036	0.353	0.9178	0.123	0.283	0.6637
2 children*Case status	-0.204	0.236	0.386	0.106	0.202	0.5985
2 children*Age in years	0.019	0.014	0.1764	0.003	0.011	0.7603
2 children*Ever use of EHRT	-0.558	0.390	0.1523	-0.537	0.297	0.0709
3-4 children*Ever use of oral contraception	-0.015	0.235	0.9491	-0.342	0.225	0.1293
3-4 children*Prior smoking	-0.432	0.209	0.0385	0.324	0.219	0.1386
3-4 children*Current smoking	-0.505	0.325	0.1202	0.071	0.289	0.8066
3-4 children*Case status	-0.390	0.219	0.0743	0.115	0.201	0.5686
3-4 children*Age in years	0.012	0.013	0.3797	0.004	0.011	0.7017
3-4 children*Ever use of EHRT	-0.360	0.342	0.2918	-0.546	0.292	0.0619

Model term	Cohort studies			Case-control studies		
	Estimate	SE	Pr > |t|	Estimate	SE	Pr > |t|
5+ children*Ever use of oral contraception	-0.087	0.270	0.746	-0.493	0.288	0.087
5+ children*Prior smoking	0.109	0.243	0.6531	0.572	0.287	0.0465
5+ children*Current smoking	-0.584	0.394	0.1386	0.208	0.385	0.5885
5+ children*Case status	-0.105	0.266	0.6948	0.369	0.275	0.1798
5+ children*Age in years	0.014	0.016	0.3904	-0.005	0.016	0.7547
5+ children*Ever use of EHRT	-0.633	0.440	0.1498	-0.393	0.429	0.3595
Ever use of oral contraception interactions						
Ever use of oral contraception*Prior smoking	0.095	0.150	0.527	0.219	0.147	0.1359
Ever use of oral contraception*Current smoking	0.108	0.241	0.6551	-0.168	0.192	0.3806
Ever use of oral contraception*Case status	0.151	0.175	0.388	0.168	0.133	0.2056
Ever use of oral contraception*Age in years	-0.021	0.010	0.0271	-0.002	0.008	0.7994
Ever use of oral contraception*Ever use of EHRT	0.197	0.272	0.4686	-0.268	0.217	0.217
Smoking interactions						
Prior smoking*Case status	-0.105	0.163	0.5199	0.132	0.142	0.3526
Prior smoking*Age in years	-0.009	0.010	0.3539	-0.010	0.008	0.2271
Prior smoking*Ever use of EHRT	0.040	0.258	0.8761	-0.139	0.223	0.5332
Current smoking*Case status	0.341	0.274	0.214	-0.313	0.196	0.1096
Current smoking*Age in years	-0.019	0.015	0.2218	0.005	0.010	0.6092
Current smoking*Ever use of EHRT	0.078	0.377	0.8361	-0.046	0.292	0.874
Case status*Age in years	-0.009	0.010	0.3804	-0.009	0.007	0.2138
Case status*Ever use of EHRT	-0.359	0.276	0.1935	-0.522	0.205	0.0108
Age in years*Ever use of EHRT	-0.024	0.016	0.1357	-0.009	0.014	0.485

[1]Binomial logistic regression model for diagnosis of diabetes, stratified by study type, with random intercept by study site and fixed effects as appear in the table.

8. REFERENCES

1. Ferlay, J., et al., *Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008.* International Journal of Cancer, 2010. **127**(12): p. 2893-2917.
2. Strinic, T., et al., *Socio-demographic characteristics of women with endometrial carcinoma.* Coll Antropol, 2003. **27 Suppl 1**: p. 55-9.
3. Purdie, D.M. and A.C. Green, *Epidemiology of endometrial cancer.* Best Pract Res Clin Obstet Gynaecol, 2001. **15**(3): p. 341-54.
4. Haidopoulos, D., et al., *Risk factors in women 40 years of age and younger with endometrial carcinoma.* Acta Obstet Gynecol Scand, 2010. **89**(10): p. 1326-30.
5. Tansavatdi, K., B. McClain, and D.M. Herrington, *The effects of smoking on estradiol metabolism.* Minerva Ginecol, 2004. **56**(1): p. 105-14.
6. Brinton, L.A. and R.N. Hoover, *Estrogen replacement therapy and endometrial cancer risk: unresolved issues. The Endometrial Cancer Collaborative Group.* Obstet Gynecol, 1993. **81**(2): p. 265-71.
7. Bokhman, J.V., *Two pathogenetic types of endometrial carcinoma.* Gynecol Oncol, 1983. **15**(1): p. 10-7.
8. SG, S., K. RJ, and N. F, *Tumours of the uterine corpus. Epithelial tumours and related lesions.* , in *Pathology and genetics. Tumours of the breast and female genital organs.*, TavassoliFA and DevileeP, Editors. 2003, IARC Press: Lyon. p. 221–32.
9. Sherman, M.E., *Theories of endometrial carcinogenesis: a multidisciplinary approach.* Mod Pathol, 2000. **13**(3): p. 295-308.

10. Kaaks, R., A. Lukanova, and M.S. Kurzer, *Obesity, endogenous hormones, and endometrial cancer risk: a synthetic review.* Cancer Epidemiol Biomarkers Prev, 2002. **11**(12): p. 1531-43.

11. Bjørge, T., et al., *Body size in relation to cancer of the uterine corpus in 1 million Norwegian women.* International Journal of Cancer, 2007. **120**(2): p. 378-383.

12. *Weight control and physical activity*, in *IARC Handbook for Cancer Prevention*, V. H and B. F, Editors. 2002, IARC Press: Lyon, France. p. 1–315.

13. Tornberg, S.A. and J.M. Carstensen, *Relationship between Quetelet's index and cancer of breast and female genital tract in 47,000 women followed for 25 years.* Br J Cancer, 1994. **69**(2): p. 358-61.

14. La Vecchia, C., et al., *Anthropometric indicators of endometrial cancer risk.* Eur J Cancer, 1991. **27**(4): p. 487-90.

15. Ioannidis, J.P.A., et al., *Commentary: Meta-analysis of Individual Participants' Data in Genetic Epidemiology.* American Journal of Epidemiology, 2002. **156**(3): p. 204-210.

16. Rothman, K.J., S. Greenland, and T.L. Lash, *Modern Epidemiology*. 3rd ed. 2008, London: Lippincott Williams & Wilkins.

17. Cornfield, J., et al., *Smoking and lung cancer: recent evidence and a discussion of some questions.* J Natl Cancer Inst., 1959. **22**(1): p. 173-203.

18. Bross, I.D.J., *Spurious effects from an extraneous variable.* Journal of Chronic Diseases, 1966. **19**(6): p. 637-647.

19. Bross, I.D.J., *Pertinency of an extraneous variable.* Journal of Chronic Diseases, 1967. **20**(487-495).

20. Hoffman, F.O. and J.S. Hammonds, *Propagation of uncertainty in risk assessments: the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability.* Risk Anal, 1994. **14**(5): p. 707-12.

21. Greenland, S., *Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment.* Risk Anal, 2001. **21**(4): p. 579-83.

22. Greenland, S., *Multiple-bias modelling for analysis of observational data.* Journal of the Royal Statistical Society: Series A (Statistics in Society), 2005. **168**(2): p. 267-306.

23. Greenland, S., *Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods.* Int J Epidemiol, 2009. **38**(6): p. 1662-73.

24. Arah, O.A., Y. Chiba, and S. Greenland, *Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders.* Ann Epidemiol, 2008. **18**(8): p. 637-46.

25. Vanderweele, T.J. and O.A. Arah, *Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders.* Epidemiology, 2011. **22**(1): p. 42-52.

26. Lash, T.L., M.P. Fox, and A.K. Flink, *Applying Quantitative Bias Analysis to Epidemiologic Data*. 2009, New York: Springer Science+Business Media.

27. Lash, T.L., et al., *CYP2D6 Inhibition and Breast Cancer Recurrence in a Population-Based Study in Denmark.* Journal of the National Cancer Institute, 2011. **103**(6): p. 489-500.

28. Fox, M.P., T.L. Lash, and S. Greenland, *A method to automate probabilistic sensitivity analyses of misclassified binary variables.* International Journal of Epidemiology, 2005. **34**(6): p. 1370-1376.

29. Rothman, K.J., S. Greenland, and T.L. Lash, *Modern Epidemiology 3rd Edition*. 2008, Philadelphia, PA: Lippincott Williams & Wilkins.

30. Greenland, S., *Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods.* International Journal of Epidemiology, 2009. **38**(6): p. 1662-1673.

31. Arah, O.A., Y. Chiba, and S. Greenland, *Bias Formulas for External Adjustment and Sensitivity Analysis of Unmeasured Confounders.* Annals of Epidemiology, 2008. **18**(8): p. 637-646.

32. PEARL, J., *Causal diagrams for empirical research.* Biometrika, 1995. **82**(4): p. 669-688.

33. Greenland, S., J. Pearl, and J.M. Robins, *Causal diagrams for epidemiologic research.* Epidemiology, 1999. **10**(1): p. 37-48.

34. Hernan, M.A., et al., *Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology.* Am J Epidemiol, 2002. **155**(2): p. 176-84.

35. Greenland, S., *Quantifying biases in causal models: classical confounding vs collider-stratification bias.* Epidemiology, 2003. **14**(3): p. 300-6.

36. Hernan, M.A., S. Hernandez-Diaz, and J.M. Robins, *A structural approach to selection bias.* Epidemiology, 2004. **15**(5): p. 615-25.

37. Walker, A.M., *Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known.* Biometrics, 1982. **38**(4): p. 1025-32.

38. Scharfstein, D.O., A. Rotnitzky, and J.M. Robins, *Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models: Rejoinder.* Journal of the American Statistical Association, 1999. **94**(448): p. 1135-1146.

39. Greenland, S., *The Impact of Prior Distributions for Uncontrolled Confounding and Response Bias.* Journal of the American Statistical Association, 2003. **98**(461): p. 47-54.

40. Robins, J.M. and D.M. Finkelstein, *Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests.* Biometrics, 2000. **56**(3): p. 779-88.

41. Bareinboim, E. and J. Pearl. *Controlling for selection bias in causal inference*. in *Fiffteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2012.

42. Horwitz, R.I. and A.R. Feinstein, *Alternative Analytic Methods for Case-Control Studies of Estrogens and Endometrial Cancer.* New England Journal of Medicine, 1978. **299**(20): p. 1089-1094.

43. Mezei, G. and L. Kheifets, *Selection bias and its implications for case–control studies: a case study of magnetic field exposure and childhood leukaemia.* International Journal of Epidemiology, 2006. **35**(2): p. 397-406.

44. Olson, S., et al., *Maximizing resources to study an uncommon cancer: E2C2—Epidemiology of Endometrial Cancer Consortium.* Cancer Causes and Control, 2009. **20**(4): p. 491-496.

45. RUBIN, D.B., *Inference and missing data.* Biometrika, 1976. **63**(3): p. 581-592.

46. Robins, J.M. and S. Greenland, *Identifiability and Exchangeability for Direct and Indirect Effects.* Epidemiology, 1992. **3**(2): p. 143-155.

47. Cole, S.R. and M.A. Hernán, *Fallibility in estimating direct effects.* International Journal of Epidemiology, 2002. **31**(1): p. 163-165.

48. Weiderpass, E., et al., *Body size in different periods of life, diabetes mellitus, hypertension, and risk of postmenopausal endometrial cancer (Sweden).* Cancer Causes & Control, 2000. **11**(2): p. 185-192.

49. Anderson, K.E., et al., *Diabetes and endometrial cancer in the Iowa women's health study.* Cancer Epidemiol Biomarkers Prev, 2001. **10**(6): p. 611-6.

50. Kaye, S.A., et al., *Increased incidence of diabetes mellitus in relation to abdominal adiposity in older women.* Journal of Clinical Epidemiology, 1991. **44**(3): p. 329-334.

51. GREENLAND, S., *THE EFFECT OF MISCLASSIFICATION IN THE PRESENCE OF COVARIATES.* American Journal of Epidemiology, 1980. **112**(4): p. 564-569.

52. Olson, S.H., et al., *Variants in hormone biosynthesis genes and risk of endometrial cancer.* Cancer Causes Control, 2008. **19**(9): p. 955-63.

53. Strom, B.L., et al., *Case-Control Study of Postmenopausal Hormone Replacement Therapy and Endometrial Cancer.* American Journal of Epidemiology, 2006. **164**(8): p. 775-786.

54. Goodman, M.T., et al., *Diet, body size, physical activity, and the risk of endometrial cancer.* Cancer Res, 1997. **57**(22): p. 5077-85.

55. Xu, W.-H., et al., *Dietary Folate Intake, MTHFR Genetic Polymorphisms, and the Risk of Endometrial Cancer among Chinese Women.* Cancer Epidemiology Biomarkers & Prevention, 2007. **16**(2): p. 281-287.

56. Brinton, L.A., et al., *Reproductive risk factors for endometrial cancer among Polish women.* Br J Cancer, 2007. **96**(9): p. 1450-6.

57. Brinton, L.A., et al., *Reproductive, menstrual, and medical risk factors for endometrial cancer: results from a case-control study.* Am J Obstet Gynecol, 1992. **167**(5): p. 1317-25.

58. Pike, M.C., et al., *Estrogen-Progestin Replacement Therapy and Endometrial Cancer.* Journal of the National Cancer Institute, 1997. **89**(15): p. 1110-1116.

59. Rowlands, I.J., et al., *Gynecological conditions and the risk of endometrial cancer.* Gynecologic Oncology, 2011. **123**(3): p. 537-541.

60. McCann, S.E., et al., *Higher regular coffee and tea consumption is associated with reduced endometrial cancer risk.* International Journal of Cancer, 2009. **124**(7): p. 1650-1653.

61. McCann, S.E., et al., *Diet in the epidemiology of endometrial cancer in western New York (United States).* Cancer Causes Control, 2000. **11**(10): p. 965-74.

62. Lu, L., et al., *Long-term overweight and weight gain in early adulthood in association with risk of endometrial cancer.* Int J Cancer, 2011. **129**(5): p. 1237-43.

63. Olesen, S.C., et al., *Personal factors influence use of cervical cancer screening services: epidemiological survey and linked administrative data address the limitations of previous research.* BMC Health Serv Res, 2012. **12**: p. 34.

64. Gu, W., C. Chen, and K.N. Zhao, *Obesity-associated endometrial and cervical cancers.* Front Biosci (Elite Ed), 2013. **5**: p. 109-18.

65. Cole, S.R. and M.A. Hernán, *Constructing Inverse Probability Weights for Marginal Structural Models.* American Journal of Epidemiology, 2008. **168**(6): p. 656-664.