Title
Applications of Multi-Bias Analysis in Studies of the Associations between Parkinson's Disease and Cancer

Author
Brendel, Paul Christian

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Applications of Multi-Bias Analysis in Studies of the Associations between Parkinson's Disease and Cancer

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Epidemiology

by

Paul Brendel

2019

ABSTRACT OF THE DISSERTATION

Applications of Multi-Bias Analysis in Studies of the Associations between Parkinson's Disease and Cancer

by

Paul Brendel

Doctor of Philosophy of Epidemiology

University of California, Los Angeles, 2019

Professor Onyebuchi A. Arah, Chair

Nonrandomized epidemiologic studies face many obstacles in attempting to quantify the effect of an exposure on an outcome. Studies of the associations between Parkinson's disease and cancer may be particularly vulnerable to uncontrolled confounding, information bias, and selection bias. This dissertation provides graphical models and descriptions of how PD-cancer studies may be affected by biases. An overview is also provided of how investigators have attempted to control for these biases. A novel multiple bias modeling method called simultaneous multi-bias adjustment is developed to address this problem within a data fusion framework. Simulation studies are used to support the validity of this method and compare it to the more traditional approach of sequentially adjusting for multiple biases. Simultaneous multi-bias adjustment is then applied to study of the effect of PD on cancer in a retrospective cohort

study using Danish population registry data combined with behavioral information collected from questionnaires and surveys. The observed effect of PD on overall cancer in this data set is approximately null. The effect estimate remains null after simultaneous multi-bias adjustment is applied, accounting for PD misclassification and selection bias related to participation and censoring. The simulation studies and Danish cohort study reveal computational and methodological challenges in performing simultaneous multi-bias adjustment. An interactive website and R package are developed to make this method more accessible to others. By developing and demonstrating how to perform simultaneous multi-bias adjustment, showing its validity in simulated data, applying it to real-world data, and making tools for its usage, this dissertation aims to make multiple bias adjustment in causal modeling more accessible to other investigators.

The dissertation of Paul Brendel is approved.

Chad Hazlett

Beate Ritz

Zuo-Feng Zhang

Onyebuchi A. Arah, Committee Chair

University of California, Los Angeles

2019

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

CI	Confidence interval
DAG	Directed acyclic graph
iPD	Idiopathic Parkinson's disease
OR	Odds ratio
PASIDA	Parkinson in Denmark Study
PD	Parkinson's disease
RMSE	Root Mean Square Error
RR	Risk ratio
SD	Standard deviation
SE	Standard error
SI	Simulation interval

ACKNOWLEDGEMENTS

I am incredibly thankful for Dr. Arah showing me the world of causal inference, believing in my abilities, and exemplifying how a scientist should live. Guidance from the rest of my dissertation committee is also greatly appreciated. Support from my friends and colleagues within UCLA Fielding School of Public Health, including Ryan Cook, Sam Wing, Brian Huang, and Tommy Gibson, was indispensable. Lastly, I would never have reached this point without the loving support of my parents and other family members.

VITA

EDUCATION

2014	MPH in Epidemiology University of Pittsburgh, Pittsburgh PA, USA
2012	BS in Neuroscience University of Pittsburgh, Pittsburgh PA, USA

RESEARCH EXPERIENCE

June 2018 – Sep. 2018	Graduate Intern Clinical Pharmacology Modeling and Simulation Amgen Thousand Oaks, CA, USA
Jan. 2016 – Sep. 2016	Research Assistant Department of Epidemiology UCLA Fielding School of Public Health Los Angeles, CA, USA
May 2013 – Aug. 2013	Epidemiology Intern Pittsburgh Summer Institute in Applied Public Health Allegheny County Health Department Pittsburgh, PA, USA
Sep. 2010 – April 2013	Research Assistant Learning Research and Development Center University of Pittsburgh Pittsburgh, PA, USA

TEACHING EXPERIENCE

Winter 2019	Statistical Modeling in Epidemiology
Fall 2018	Logic, Causation, and Probability
Fall 2015	Intro. to Functional Anatomy of the Central Nervous System

PUBLICATIONS

Moore M.W., Brendel P. C., & Fiez J. A. (2014). Reading faces: Investigating the use of a novel face-based orthography in acquired alexia. Brain & Language, 129, 7-13. DOI: https://doi.org/10.1016/j.bandl.2013.11.005

1. INTRODUCTION

1.1 PARKINSON'S DISEASE ETIOLOGY AND RELATIONSHIP WITH CANCER

Parkinson's disease (PD) is the second most common neurodegenerative disorder behind Alzheimer's disease. It is characterized by the loss of dopaminergic cells in the substantia nigra pars compacta. PD features both motor and non-motor symptoms. The cardinal symptoms of PD include: resting tremor, bradykinesia, rigidity, and loss of postural reflexes [1]. PD has an estimated overall prevalence of 571 cases per 100,000 people in the total population [2]. PD prevalence increases with age; those aged 50-59 years have 156 cases per 100,000 people, whereas those aged 80+ experience 2,498 cases per 100,000 people [2]. In the United States, PD created an estimated economic burden exceeding $14.4 billion in 2010 [3]. There is currently no therapy for the underlying neurodegeneration, but several symptomatic therapies are available to help improve patient quality of life. The therapies stimulate dopamine receptors or increase dopamine levels [4]. Levodopa, the precursor to dopamine, is the most effective treatment for the motor symptoms of PD [5].

The etiology of PD has been studied extensively but is still largely unknown. Older age is a well-established risk factor for PD. Many epidemiological studies have shown that smokers have a reduced risk of PD [6]. This trend initially gave rise to the idea that tobacco has a neuroprotective effect; however, recent evidence suggests that it may be more appropriate to explain the relationship in terms of differences in nicotinic reward/ease of quitting nicotine between those with and without PD (i.e. reverse causality) [7]. The risk of PD is about 1.5 times higher in males than in females [8]. The reduced risk in females may be attributed to a protective effect of higher circulating levels of estrogen during early reproductive life [9].

Genetics also play a role in the pathogenesis of the disease; at least 16 loci (PARK1 – PARK16) and 11 genes have been associated with inherited forms of the disease [10]. Several different occupational exposures have been studied ever since it was found that 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) can lead to pathogenesis resembling PD [11]. A meta-analysis of pesticides and PD found an increased risk of PD in those exposed to pesticides (OR=1.46), but there was substantial heterogeneity among the included studies [12]. Coffee (and presumably caffeine) may protect against PD; a meta-analysis of 13 studies found a pooled relative risk of PD of 0.69 (95% CI, 0.59-0.80) in coffee drinkers compared to non-coffee drinkers [13]. Physical activity may have a protective effect against PD, but this effect is largely dependent on age, gender, and type of activity [14]. Some other proposed risk factors needing further investigation include: heavy metals, head trauma, alcohol, and diet.

Beyond these risk factors, associations between PD and other health conditions have been studied. Inverse associations between comorbidities appear when a lower-than-expected probability of certain diseases occurs in individuals diagnosed with other health conditions [15]. An inverse comorbidity association between PD and cancer was first described by Doshay in 1954 [16], [17]. Many cohort and case-control studies have since corroborated similar associations between PD and many cancers. Meta-analyses by Bajaj et al. and Catala-Lopez et al. report estimates of the effect of PD on overall cancer as a relative risk (RR) of 0.73 (95% CI, 0.63-0.83) and pooled effect size (ES) of 0.83 (95% CI, 0.76-0.91), respectively [18], [19]. The protective effect is slightly stronger in males (RR=0.71, 0.57-0.88) compared to females (RR=0.82, 0.68-0.98) [18]. An opposite trend of increased cancer risk in PD patients has been observed in melanoma. A meta-analysis specifically examining the effect of PD on melanoma found a pooled odds ratio (OR) of 2.11 (95% CI, 1.26-3.54) [20]. While a consensus has formed

regarding the strength of the relationship between PD and different types of cancer, no single, concrete causal mechanism has been established to explain how PD might cause cancer.

1.2 CAUSAL INFERENCE

The objective of most epidemiological studies is to quantify the effect of an exposure on an outcome, measured as a risk difference, risk ratio, odds ratio, or hazard ratio. Using the counterfactual framework developed by Neyman, Rubin, and others, the individual effect of an exposure occurs when the value of an outcome under one exposure ($Y_{a=0}$) differs from the value the outcome would have under a different exposure ($Y_{a=1}$) and with all other factors identical in both scenarios besides the exposure. Since the conditions for both effects must be identical, only one of these two effects is observable, the other being hypothetical or counterfactual. It follows from this lack of observability that individual effects cannot be estimated. Using counterfactual notation and treating PD and cancer as dichotomous variables, an individual effect of PD on cancer would be written as $CANCER_{PD=1} \neq CANCER_{PD=0}$.

A population effect occurs when the proportion of subjects, all with the same exposure, who experience an outcome ($P[Y_{a=1}=1]$) differs from the proportion of subjects, all with a different exposure, who experience that outcome ($P[Y_{a=0}=1]$) and with all other factors identical in both scenarios besides the exposure. Unlike individual effects, population causal effects can be estimated by ensuring that the two subject groups are identical in their distribution of all relevant covariates. Using counterfactual notation and treating PD and cancer as dichotomous variables, the population effect of PD on cancer can be written as a risk difference ($P[CANCER_{PD=1}=1] - P[CANCER_{PD=0}=1] \neq 0$), risk ratio ($P[CANCER_{PD=1}=1] / P[CANCER_{PD=0}=1] \neq 1$), or odds ratio ($(P[CANCER_{PD=1}=1] / P[CANCER_{PD=1}=0]) /$

($P[CANCER_{PD=0}=1] / P[CANCER_{PD=0}=0]) \neq 1$). This paper henceforth refers to population effects when mentioning causal effects.

Estimates of the effect of an exposure or intervention on an outcome consist of three components: the true effect, random error, and systematic error. For effect estimates to represent true effects, random error and systematic error must be minimized. Random error is the residual variability of an effect measure that occurs because of a lack of sufficient knowledge to perfectly predict events [21]. In epidemiological studies, a major contributor of random error is sampling variability, error that results from the inability to include everyone from the target population (or a broader conceptual population of everyone with the biological experience) who could have been included in the study. Random error in epidemiological studies can be minimized by increasing the study size, efficiently apportioning subjects into study groups, and efficiently stratifying the data into covariate subcategories [21]. The random error of an effect estimate is usually quantified and reported with a standard error and confidence interval.

Systematic error or bias includes all the other (nonrandom) forces that harm the internal validity of a study. All epidemiologic biases are generally subsumed under three categories: uncontrolled confounding, selection bias, and information bias [21]. Confounding can be thought of as the distortion of an exposure-outcome relationship due to a lack of exchangeability of external variables among study groups. More precisely, confounding occurs when the conditional expectation ($E[Y|X=x]$) differs from the controlled expectation ($E[Y|do(X=x)]$) in the marginal measures setting (parallels for the conditional measures setting exist). Selection bias occurs when the observed association in those selected for analysis differs from the association in those who are eligible for analysis [22]. Lastly, information bias can occur in epidemiological studies when the exposure, outcome, or both are incorrectly classified or measured with error

[23]. This misclassification or measurement error can be either independent or dependent, depending on whether the measurement error for the exposure is related to the measurement error in the outcome [23]. Also, the misclassification can be differential if the measurement error in the exposure/outcome is affected by the outcome/exposure, or non-differential if the measurement error in the exposure/outcome is **not** affected by the outcome/exposure [23].

Besides assuming a lack of bias, including conditional exchangeability, a few other assumptions are needed to identify causal effects. Consistency is assumed, meaning that a group's potential outcome under an exposure can be observed by measuring that outcome in people who experience the exposure. Positivity is assumed, meaning that all subjects have a positive probability of receiving each possible treatment/exposure within strata of the confounders. Lastly, a lack of interference is assumed, known as the stable unit treatment value assignment. This assumption implies that a subject's potential outcome under a particular exposure should not depend on the exposure value of other subjects.

Causal assumptions made throughout this paper will be depicted using directed acyclic graphs (DAGs) [24]–[26]. DAGs consist of nodes, each representing a variable, and arrows between the nodes. An arrow from node X to node Y represents a direct, causal effect (of unknown strength) of X on Y. If X has an arrow into Z and Z has an arrow into Y, then X has an effect on Y that is causally mediated by Z (an indirect effect). Causal information travels along open paths in the DAG. A path between two nodes is considered open unless it is closed by either: 1) two arrows on the path pointing into the same variable, referred to as a collider; or 2) conditioning or selecting on a non-collider along the path, which is represented in the DAG by putting brackets around the variable (e.g. $[X]$). DAGs are acyclic, thus do not contain cycles ($X \rightarrow Y$, $Y \rightarrow X$) or self-loops ($X \rightarrow X$).

The backdoor criterion is used to assess whether direct and indirect effects of *X* on *Y* have a causal interpretation [24]. A variable *Z* satisfies the backdoor criterion for (*X*, *Y*) if: 1) no node in *Z* is a descendent of *X*; and 2) *Z* blocks every path between *X* and *Y* that contains an arrow into *X*. When a variable *Z* satisfies the backdoor criterion relative to (*X*, *Y*), then the control of *Z* allows for the identification of the causal effect of *X* on *Y*. Otherwise, open, backdoor paths create bias between variables. A biasing path between two variables will create association without causation between these variables. When a covariate *Z*, in addition to other covariates *C*, removes the confounding of an *X*-*Y* relationship after control of (*Z*, *C*), and no subset of (*Z*, *C*) can remove the confounding, *Z* is referred to as a confounding variable [27].

Bias is also present and can be visualized in DAGs when there is conditioning on a variable that is a collider or the descendant (consequence) of a collider [22], [28]. Controlling for the collider opens a biasing path that would have otherwise been closed. This type of bias is commonly seen in case-control studies, which by design select patients based on values of the outcome variable. Bias is created when this selection is also related to the exposure under study.

1.3 BIAS ANALYSIS

Conventional results from epidemiologic analyses normally quantify the random error, but fail to quantify the systematic error [29]. Researchers typically address potential sources of bias through speculative discussion. Such discussion is often limited by the creativity of the researcher and can be influenced by whether the researcher likes his or her results. A series of methods, coined bias analyses, emerged to provide a quantitative and more transparent demonstration of how effect estimates are influenced by bias.

One method, a “simple” sensitivity analysis, involves replacing the sources of uncertainty with fixed values [30]. The conventional analysis is then repeated with different values of the

uncertainty parameters [30]. This method can be expanded by replacing the fixed values with specific probability distributions for each parameter via Monte Carlo risk assessment or Bayesian methods. In Monte Carlo risk assessment, a value is drawn from each parameter and the analysis is repeated with multiple draws. A summary of the resulting distribution of effect estimates is then presented. These methods were initially developed and implemented using population-level (marginal) uncertainty parameters that failed to account for the role of measured covariates already adjusted for. However, newer methods have emerged involving record-level parameters. A record-level approach allows for more specific representation of bias by allowing for uncertainty parameters to vary with individual covariate levels [31].

A variety of external adjustment formulas are also commonly used to remove bias. Model-specific formulas are used to generate a bias factor, and this bias factor is subtracted from the exposure-outcome effect estimate [32]. The resulting effect estimate is then considered free of the suspected source of bias, based on the assumptions used to generate the bias factor [32].

When multiple biases arise in epidemiologic studies, the biases are generally analyzed sequentially [21], [33]. The order of the corrections will affect the final result; corrections should be made in the reverse of the order in which the errors occurred in the data-generating process [33]. There is no consistent order in which threats to validity arise, so determining the order of corrections may be difficult. A bias analysis with record-level uncertainty parameters could circumvent this problem by allowing for the adjustment of biases simultaneously instead of sequentially. A simultaneous adjustment would also be ideal since biases do not necessarily act independently of each other, and these interactions would not be captured in a sequential approach.

As improvements in technology lead to larger sources of data, increased study sizes should minimize the impact of sampling variation on the total uncertainty in statistical estimates. On the other hand, bigger data will inflate the chances of being precisely wrong whenever systematic error is substantial. Bias analysis is therefore most valuable in big data studies that report minimal uncertainty with narrow confidence intervals (i.e. higher precision) and are susceptible to a limited number of systematic errors [33]. Such misleading, highly precise estimates could be used as evidence for policy intervention, so it is imperative that a more accurate representation of uncertainty is provided. A study with many sources of systematic error, however, is a good candidate for a new, better designed study as opposed to bias analysis.

To make valid causal inferences, it is essential that epidemiological investigators go beyond a qualitative description of potential error sources and adopt quantitative bias analysis. Although adaptation of these methods has been slow, the importance of bias analysis and guides to good practice are gaining prominence [32], [34], [35]. Although it is important to expand on the comprehensiveness of bias analysis techniques, it is also necessary to create metrics and tools that will improve the adoption of these methods, as seen in the E-value and E-value calculator [36], [37].

1.4 SPECIFIC AIMS OF THIS DISSERTATION

The public health focus of this dissertation is to describe and add clarity to the PD-cancer relationship. This dissertation will provide an overview of the effect estimates seen in the PD-cancer literature. It will highlight several biases that may be impacting PD-cancer studies and describe how investigators have attempted to control for these biases. Lastly, a PD-cancer study will be conducted that incorporates multiple bias adjustment.

The methodological focus of this dissertation is to develop a method to adjust for multiple biases simultaneously. Simulation studies will be used to confirm the validity of this method and demonstrate its advantages over sequential multi-bias approaches. Lastly, this dissertation will develop and provide tools to make simultaneous multi-bias analysis accessible to others.

The project is divided into four studies, each with a specific aim:

Study 1: To review the causal nature of the relationship between Parkinson's disease and cancer, including a summary of the epidemiological and biological evidence and descriptions and causal graphs of potential sources of bias.

Study 2: To develop a method that simultaneously adjusts for any combination of uncontrolled confounding, exposure misclassification, and selection bias that serves to "reconstruct" the unbiased data.

Study 3: To apply simultaneous multi-bias analysis to understand the effect of PD on cancer using data from the Danish National Hospital Registry.

Study 4: To create an interactive web application and an R package to assist investigators in implementing simultaneous multi-bias analysis.

2. STUDY 1

CAUSAL REVIEW OF STUDIES OF THE ASSOCIATIONS BETWEEN PARKINSON'S DISEASE AND CANCER

2.1 ABSTRACT

Introduction: Numerous epidemiological studies have attempted to quantify and to describe the causal relationship between PD and cancer; however, there is no clear understanding of the causal mechanism relating these variables. The current study aimed to provide a causal overview of the potential biases that could explain some of the published PD-cancer associations and to give examples of how investigators have attempted to adjust for these biases to help guide future study designs.

Methods: Peer-reviewed literature on PD-cancer connections was obtained from PubMed. DAGs were used to depict the plausible causal and non-causal connections between PD and cancer based on the published literature and best available knowledge.

Results: Across a variety of study designs, PD-cancer studies show a consistent, protective effect of PD on the occurrence of most non-melanoma cancers. The inverse relationship between PD and cancer may be susceptible to a variety of biases including uncontrolled confounding, information bias, and selection bias. Differences in disease onset and survival may contribute to the heterogeneity in PD-cancer vs. cancer-PD estimates and the differences observed among specific cancer subtypes.

Conclusion: Insufficient bias adjustment in epidemiological studies and incomplete knowledge of biological processes impede the understanding of the true causal link between PD and cancer, if any.

2.2 INTRODUCTION

Numerous epidemiological studies have found PD patients to be at a decreased risk of developing many types of cancer [18]. This inverse relationship has been studied for decades because there is still no biological mechanism to explain the relationship and because most studies do not produce an adequately unbiased estimate of the effect of PD on cancer. Truly understanding this relationship could provide insight into the underlying etiologies of these two diseases. Such insight is crucial considering both diseases are highly prevalent among the elderly and lead to significant physical and economic burden.

In order to provide additional support for an exposure-outcome relationship, Epidemiologists generally conduct several studies of the relationship using different study populations, study designs, and analysis methodologies. Presumably, each additional study identifies a weakness of a previous study and attempts to improve on the weakness. To identify potential areas of improvement an investigator needs to thoroughly review the literature on the particular relationship for unaddressed sources of systematic error. What if, instead, this information was detailed in a single review?

This paper provides a causal overview of the PD-cancer relationship. It summarizes the effect estimates seen in the literature. Potential sources of bias due to uncontrolled confounding, information bias, and selection bias are described and modeled with DAGs. Examples of how investigators have attempted to control for these biases are provided. Lastly, additional factors influencing this relationship related to disease onset and survival are discussed. By providing a causal review of PD-cancer, this paper should help guide the design and analysis of future PD-cancer studies.

2.3 METHODS

The PD-cancer publication search strategy is described as follows. PD-cancer studies were initially identified from the meta-analysis by Baja et al. [18]. Additional studies from November 2009 until April 2019 were collected from PubMed searching the keywords: 'parkinson's', 'cancer', 'neoplasm', 'melanoma', and 'inverse'. The reference list of each paper was manually reviewed for any additional, relevant studies.

2.4 PD-CANCER STUDY RESULTS

Meta-analyses examining PD preceding cancer by Bajaj et al. and Catala-Lopez et al. report overall cancer risk estimates in PD patients as a relative risk (RR) of 0.73 (95% C.I. 0.63-0.83) and pooled effect size (ES) of 0.83 (95% C.I. 0.76-0.91), respectively (Table 2.1) [18], [19]. Estimates of overall cancer risk may not be very meaningful considering that the strength and direction of the effect of PD on cancer seems to vary based on the type of cancer. For example, PD had a stronger protective effect on lung cancer relative to overall cancer, whereas PD was reported to be a risk factor for melanoma. When Bajaj et al. exclusively analyzed studies where melanoma and other skin cancers could be excluded, the overall effect estimate barely changed (RR=0.71; 95% C.I. 0.63-0.79).

Table 2.1 PD-cancer meta-analysis results

Meta-analysis	Overall cancer risk	# of studies	Lung cancer risk	# of studies	Melanoma risk	# of studies
Bajaj, 2010	0.73; 0.63-0.83[a]	29	0.46; 0.41-0.51[a]	10	1.41; 0.90-2.19[a]	8
Cátala-Lopez, 2014	0.83; 0.76-0.91[b]	10	0.44; 0.35-0.55[b]	8	1.65; 1.39-1.96[b]	6

[a] Risk ratio, 95% confidence interval
[b] Pooled effect size, 95% confidence interval

Fewer studies have investigated the opposite direction of effect, cancer preceding PD. Evidence from five case-control studies is less conclusive but suggests similar results to those of PD preceding cancer (Table 2.2) [38]–[42].

Table 2.2 Estimates of the effect of cancer on PD

Study	Odds ratio; 95% CI		
Cui, 2019	Overall cancer[a]: 1.14; 0.89-1.46	Smoking-related cancer[a]: 0.75; 0.46-1.22	Non-smoking-related cancer[a]: 0.97; 0.61-1.56
D'Amelio, 2004	Malignant neoplasms[b]: 0.5; 0.1-2.1	Non-malignant neoplasms[b]: 0.3; 0.1-0.7	
Driver, 2007	Overall cancer[c]: 0.83; 0.57-1.21	Smoking-related cancer[c]: 0.74; 0.35-1.57	Non-smoking-related cancer[c]: 0.88; 0.59-1.32
Elbaz, 2002	Overall cancer[d]: 0.79; 0.49-1.27	Lung cancer[d]: 1.00; 0.14-7.10	Melanoma[d]: 1.50; 0.25-8.98
Olsen, 2006	Overall cancer[d]: 1.04; 0.96-1.12	Lung cancer[d]: 0.42; 0.22-0.80	Melanoma[d]: 1.44; 1.03-2.01

[a] Adjusted for age at PD diagnosis or index date, sex, urbanization, SES, Charleston Comorbidity Index; cases and controls matched by age and sex
[b] Adjusted for smoking, alcohol, and coffee consumption; cases and controls matched by age and sex
[c] Adjusted for smoking status, alcohol use, BMI, and exercise; cases and controls matched by age
[d] Cases and controls matched by age and sex

Two studies examined the PD-cancer relationship in both directions (Table 2.3). Both studies reported a stronger protective effect in PD preceding cancer versus cancer preceding PD. Freedman et al. used data from the Surveillance, Epidemiology, and End Results-Medicare (SEER) linked database for both a cohort (cancer $\rightarrow$ PD) and case-control study (PD $\rightarrow$ cancer) [43]. The cohort and case-control study each used the same source population, diagnostic criteria for PD/cancer, and outcome models with the same adjustments (sex, race, age, cancer registry, and frequency of physician visits). Despite these similarities in design and analysis, a lack of PD-overall cancer association was seen in the cohort study, but PD was associated with

reduced subsequent overall cancer in the case-control study. Similar estimates were seen for the cohort and case-control designs for overall cancer when automobile injuries were evaluated in place of PD. These inconsistencies led Freedman et al. to believe that the true relationship between PD and cancer is likely non-causal and that any non-null association could be explained by bias, notably detection bias (fewer screening and diagnostic tests in those with PD).

Table 2.3 Results of studies that assessed the PD-cancer relationship in both directions

Study	PD → Cancer			Cancer → PD		
	Overall	**Lung**	**Melanoma**	**Overall**	**Lung**	**Melanoma**
Fois, 2009	0.61; 0.53-0.70[a]	0.5; 0.4-0.8[a]	0; 0-1.01[a]	0.76; 0.70-0.82[a]	0.5; 0.4-0.7[a]	0.5; 0.2-0.9[a]
Freedman, 2016	0.77; 0.71-0.82[b]	0.69; 0.62-0.76[b]	1.03; 0.88-1.21[b]	0.97; 0.92-1.01[c]	0.81; 0.72-0.92[c]	1.12 0.95-1.31[c]

[a] Rate ratio, 95% confidence interval
[b] Odds ratio, 95% confidence interval
[c] Hazard ratio, 95% confidence interval

Fois et al. used the Oxford Record Linkage Study to study the occurrence of cancer in people admitted with PD before and after an admission of cancer [44]. Reference cohorts were established by including those admitted for common orthopedic, dental, ENT, and other minor disorders. Similar results were seen in cancer before and after PD; there were decreased rate ratios for overall cancer, smoking-related cancers, and some non-smoking-related cancers. These significant differences were not seen when motor neuron disease and multiple sclerosis were examined instead of PD. The authors suggest that biological processes may explain these results, pointing to a neuroprotective effect of smoking (i.e. PD patients are less likely to smoke, thus less likely to have cancer) and genetic factors that are common in both diseases, such as *parkin* mutations.

A recent cohort study by Lin et al. is the first epidemiological study to report PD-cancer estimates of effect in stark contrast to those reported in previous studies [45]. This study, which followed a Taiwanese cohort, is also one of the first cohort studies to examine this relationship in a non-Western population. The study reports the following hazard ratios for cancer in PD patients: overall (1.58; 95% CI 1.50-1.65), lung (1.56; 95% CI 1.38-1.76), melanoma (2.75; 95% CI 1.35-5.59). The estimates for overall and lung cancer are in the opposite direction of those estimates reported in other Western studies. The authors attribute their opposing results to the different genetic background and environmental exposures of the Taiwanese population.

A meta-analysis by Liu et al. looked specifically at the relationship between PD and melanoma [20]. Their analysis of 12 studies resulted in a pooled PD-melanoma OR of 2.11 (95% CI 1.26-3.54). The analysis had a couple of key concerns. First, there was significant heterogeneity across studies. Second, eight of the studies had fewer than ten cases with PD and melanoma, which brings into question issues of sufficient statistical power. Despite these issues, three other reviews of the PD-melanoma relationship each confirm the increased risk of melanoma in PD patients [46]–[48]. While the exact causal nature is unclear, the reviews agree that levodopa use does not contribute to the causal effect, as had previously been hypothesized.

2.5 BIAS IN THE PD-CANCER LITERATURE

Biological explanations may not fully explain the PD-cancer effect estimates seen in the literature. There could be various sources of systematic error contributing to a non-causal relationship. In this section, we will introduce these biases, describe how they may be distorting the PD-cancer relationship, and explain how investigators of the PD-cancer relationship have attempted to account for these biases in their studies.

2.5.1 UNCONTROLLED CONFOUNDING

Uncontrolled confounding occurs when a common cause of both exposure and outcome is not appropriately controlled for in the study design or analysis, resulting in a spurious association between the exposure and outcome (Figure 2.1). Randomized controlled trials are often considered the "gold standard" because randomization into exposure groups helps prevent confounding since the randomization scheme is the only factor influencing the exposure. Until researchers completely understand the etiology of PD and cancer, it is impossible to confirm all the common causes of both diseases. Therefore, to prevent uncontrolled confounding investigators must choose adjustment variables based on the best available knowledge and acknowledge the possibility that there may be uncontrolled, unknown common causes. At a minimum, studies of PD-cancer should adjust for age and sex.

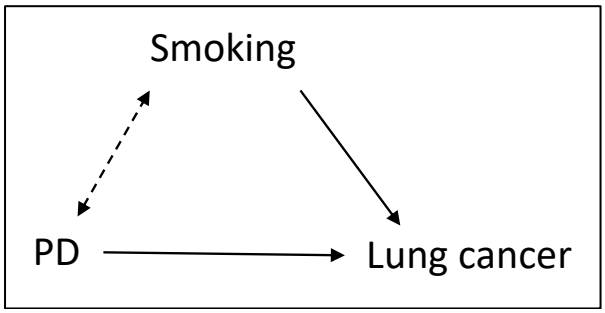

Figure 2.1 Uncontrolled confounding in the PD-cancer relationship (assuming a causal link)

Smoking is also an important confounder when examining any of the several cancer subtypes known to be affected by smoking (Figure 2.1). Many epidemiological studies have shown that smokers have a reduced risk of PD [6]. Uncertainty exists regarding the exact causal nature and direction of the smoking-PD relationship [7]. However, unless one argues that smoking acts as a mediator between PD and smoking-related cancer, smoking must be controlled for in the study design or analysis.

A common problem that contributes to uncontrolled confounding in studies of PD-cancer is the study's inability to obtain smoking or other behavioral data. Many studies use large, population-based registries or health insurance claims data, which do not contain information regarding behavioral variables. Furthermore, studies that do have smoking data typically have imprecise measurements (e.g. never smoker, past smoker, current smoker), and smoking control could still lead to residual confounding. Several studies acknowledge that their estimates may be biased from the confounding that results from the inability to obtain smoking measurements [43]–[45]. Investigators lacking smoking information may compare estimates from "smoking-related" and "not smoking-related" cancers. In their PD-cancer meta-analysis, Bajaj et al. found that smoking-related cancers had the stronger inverse relationship with PD (RR=0.61, 95% CI=0.58-0.65), but that nonsmoking-related cancers still had a significant inverse relationship (RR=0.80, 95% CI=0.77-0.84) [18]. This stratification helps bypass the need for smoking adjustment when the outcome is non-smoking related, but still results in a biased estimate for smoking-related cancer.

The strategy used by most investigators to select covariates for confounding control in PD-cancer studies is not ideal and likely results in effect estimates with residual confounding. The most common approach involves basing the adjustment set on the variables that will control confounding between PD and **overall** cancer. This same adjustment set (or no adjustment) is then used in the analyses of PD and **specific** cancers. However, different cancers have distinct risk factors, so different cancers will commonly require different adjustment sets to control for confounding. For example, estrogen level will be an important confounder on the effect of PD on breast cancer, but not a confounder for the effect of PD on lung cancer. Using the prevailing

approach for covariate selection, different PD-cancer effect measures may have different degrees of residual confounding.

Nine cohort studies and seven case-control studies were used in the Bajaj et al. PD-cancer meta-analysis (Table 2.4) [18]. Every study appropriately controlled for age and most controlled for sex. There was not a single study that used different adjustment sets for different cancer subtypes. The two studies by Driver et al. made the strongest effort to address the problem of inadequate confounding control in PD-cancer studies. These two studies, both using data from the Physician's Health Study, obtained lifestyle factors, including smoking status, alcohol use, BMI, and exercise frequency from a questionnaire mailed to study participants [40], [49]. Both studies provide crude and multivariable adjusted estimates of PD-overall cancer, but neither study uses this rich covariate data to adjust for the PD effect on specific cancers. The scarce number of outcomes within each cancer subtype probably explains why such estimates are absent.

Table 2.4 Studies incorporated in the Bajaj et al. meta-analysis and their respective adjustment sets

Cohort studies	Adjustment set
Barbeau, 1963	Age
Driver, 2007	Age, sex, smoking, alcohol, BMI, exercise
Elbaz, 2005	Age, sex, smoking
Fois, 2009	Age, sex, year of first hospital admission, region
Guttman, 2004	Age, sex
Jansson, 1985	Age
Minami, 2000	Age
Olsen, 2005	Age, sex
Pressley, 2003	Age
Case-control studies	**Adjustment set**
D'Amelio, 2004	Age, smoking, alcohol, coffee
Driver, 2007	Age, sex, smoking, alcohol, BMI, exercise
Elbaz, 2002	Age, sex
Levine, 1992	Age, sex
Olsen, 2006	Age, sex
Powers, 2006	Age, smoking, ethnicity, education
Rajput, 1987	Age, sex

2.5.2 INFORMATION BIAS

Information bias can occur in epidemiological studies when the exposure, the outcome, or both are incorrectly classified [21]. The misclassification can be either independent or dependent, depending on whether the measurement error for the exposure is related to the measurement error of the outcome [23]. Also, the misclassification can be either differential, if the measurement error of the exposure/outcome is affected by the outcome/exposure, or non-differential, if the measurement error of the exposure/outcome is **not** affected by the outcome/exposure [23]. In the case of PD-cancer, it seems likely that errors in classification are independent, since distinct diagnostic methods are used for PD and cancer, but different circumstances can arise and lead to both differential and non-differential misclassification (Figure 2.2).

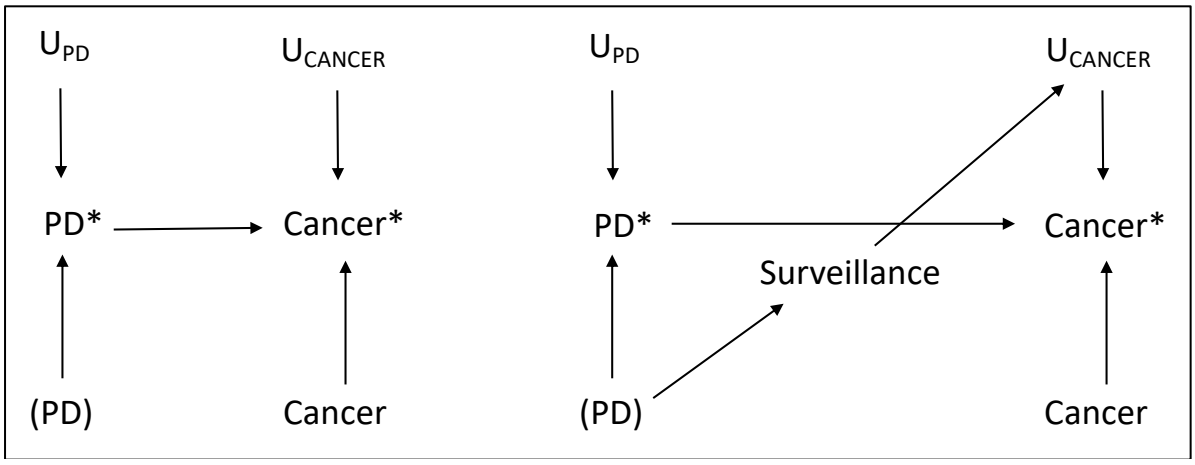

Figure 2.2 Information bias in PD-cancer (* represents misclassified variable, () represents missing variable, U_A represents all factors other than A that determine the value of A*)

PD is a particularly difficult condition to diagnose, with the only definite diagnosis requiring pathological examination post-autopsy [1]. Clinicians normally rely on the presence of multiple motor symptoms to make a diagnosis, with the cardinal symptoms including resting tremor, bradykinesia, rigidity, and loss of postural reflexes [1]. A recent meta-analysis examined 11 studies that included diagnostic parameters regarding clinical diagnosis of PD and found that these studies had a pooled diagnostic accuracy of 80.6% [50]. Having one in five subjects have an incorrect PD diagnosis due to clinical error can introduce independent, non-differential misclassification in PD-cancer studies (Figure 2.2).

Available statistics for cancer misclassification and misdiagnosis are generally lacking despite the huge financial and medical implications. A study of malpractice claims from a large risk management database found that 59% of claims related to diagnostic errors pertained to cancer diagnosis [51]. Another study analyzed cancer diagnosis error rate by comparing patient same-site cytologic and histologic specimens and found an error rate of 11.8% [52].

To address issues of independent, non-differential misclassification, Elbaz et al. assessed the reliability of their PD classification through a validation sub-study [39]. PD was initially classified as parkinsonism with three additional criteria: (1) no other cause, (2) no documentation of unresponsiveness to levodopa at doses of at least 1 gram/day in combination with carbidopa, and (3) no prominent or early signs of more extensive involvement not explained otherwise. The investigators validated the PD classification via evaluation by one of three movement disorder specialists. The specialists did not have access to patient medical records during the assessments. They found that 96.6% (57/59) of those initially classified as PD fulfilled the PD diagnostic criteria at direct examination. None of the non-PD patients were found to have PD or

parkinsonism (0/58). The results of these clinical examinations provide evidence suggesting that independent, non-differential PD misclassification was negligible.

Surveillance bias, a type of independent, differential misclassification bias, can occur when the exposure, or factors related to the exposure, influences the amount or duration of effort used in detecting the outcome (Figure 2.2) [21]. An outcome is more likely to be detected if medical professionals spend more time looking for it. A positive exposure-outcome relationship could then merely result from the exposure's ability to increase medical surveillance.

PD patients may experience increased cancer surveillance upon immediate diagnosis. The patient will present to the physician with motor and/or cognitive deficits, and the physician may order multiple tests, including cancer screenings, to find a clinical explanation for the symptoms. PD patients may later experience decreased cancer surveillance as their symptoms progress and their quality of life decreases. There is less urgency to detect/treat cancer if the patient already has a very limited quality of life due to the PD symptoms. It may seem that PD patients are experiencing less cases of cancer when, in fact, they are experiencing no difference in cancer rate and their cancer is simply not being diagnosed. In addition, PD patients are more likely to be institutionalized than age and sex-matched controls [53]. This difference could result in increased medical contacts and additional opportunities for detection of certain cancers in PD patients.

Becker et al. acknowledged the potential for surveillance bias in PD-cancer studies and attempted to account for this error in their analysis of the UK General Practice Research Database [54]. Within their longitudinal study, the investigators performed a nested case-control analysis (cases corresponded to those with incident cancer) and adjusted for the number of medical contacts prior to the index date. They found that PD patients had more physician visits

than PD-free patients (53% vs. 39% with more than 10 healthcare contacts), but that adjusting for medical contacts further moved the PD-cancer relative risk away from the null.

Freedman et al. expand on these attempts to control for surveillance bias [43]. Their PD-cancer study likewise adjusts for the frequency of physician visits. In addition, this study employs a negative comparison control, automobile injuries, to determine if a health outcome with no biological relationship to cancer could be found to affect cancer risk, presumably by influencing screening or diagnostic tests for cancer. The study found a reduced risk of cancer following both PD and automobile accidents, which is consistent with the idea that any significant medical incident could lead to under-ascertainment of cancer (i.e. an inverse exposure-outcome relationship).

2.5.3 SELECTION BIAS

Selection bias occurs when the effect estimate seen in those selected for analysis differs from that of the target population [22]. This bias occurs in case-control studies due to the inappropriate selection of controls and occurs in cohort studies due to differential censoring between the exposed and unexposed [22]. The key causal mechanism shared by all forms of selection bias is collider stratification, which involves conditioning on a variable, termed a collider, that is a shared common effect of two other variables.

In case-control studies, selection bias occurs when the outcome influences study selection (per study design) and the exposure is also related to selection (Figure 2.3). A collider stratification bias ensues due to the study inherently examining only those subjects that were selected for participation. This bias can be avoided by designing a study such that controls are sampled in a manner to ensure that their exposure distribution is comparable to that of the source population (or ensure that exposure is not related to selection other than through the outcome or a

covariate that can be adjusted for at the analytical stage). Selection bias can affect PD-cancer case-control studies due to PD causing a variety of factors known to affect study participation (Figure 2.3). Specific health problems in older adults affect recruitment and retention [55]. Some of these problems that are affected by PD include: cognitive slowing, dementia, manual dexterity difficulties, fall risk, and limited life expectancy. Failure to adjust for these health problems could lead to biasing paths through study selection.

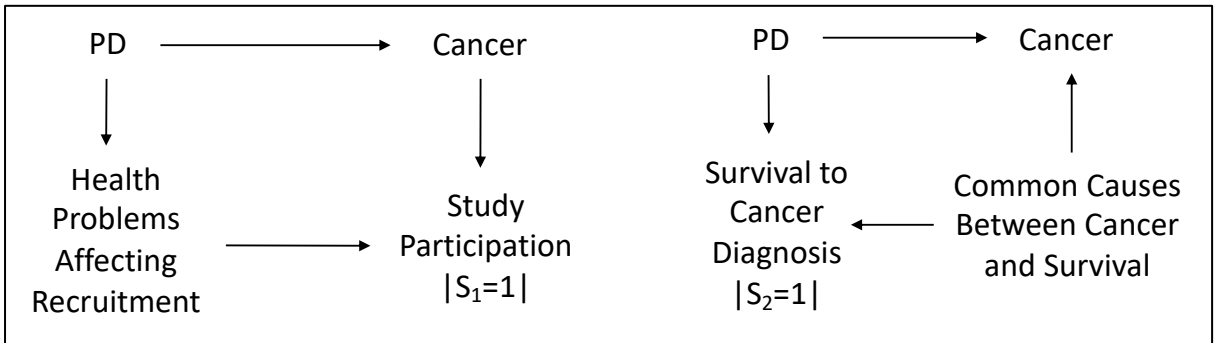

Figure 2.3 Selection bias in PD-cancer case-control studies

A specific type of selection bias, known as survivor bias, can also occur in case-control studies [21]. In addition to conditioning on selection, case-control studies inherently condition on those who have survived until an age old enough to experience the outcome. A biasing path can form when the exposure affects survival to outcome diagnosis and when survival is related to the outcome via a shared common cause (Figure 2.3). Survivor bias may influence the inverse PD-cancer relationship due to PD patients selected for studies having "survival" qualities, making them less likely to get cancer compared to those less healthy PD patients who didn't survive long enough to make it into the study. Specifically, this bias can be seen in PD-cancer case-control studies due to the lack of adjustment for factors that affect both cancer and survival

to cancer diagnosis (Figure 2.3). These factors may include: chronic inflammation, immunosuppression, diet, and obesity.

Selection bias occurs in cohort studies when those with the exposure participate or are lost to follow-up at a different rate than those without the exposure (Figure 2.4). Cohort studies inherently select on those who are not lost to follow-up. Study follow-up, or the lack of censoring, is therefore a potential point of collider stratification. In PD-cancer cohort studies, collider stratification can occur when PD and cancer affect censoring through health problems known to affect retention (Figure 2.4). Health problems caused by cancer that may affect retention include: major emotional decline, frequent hospitalizations, easy fatigability, and limited life expectancy [55].

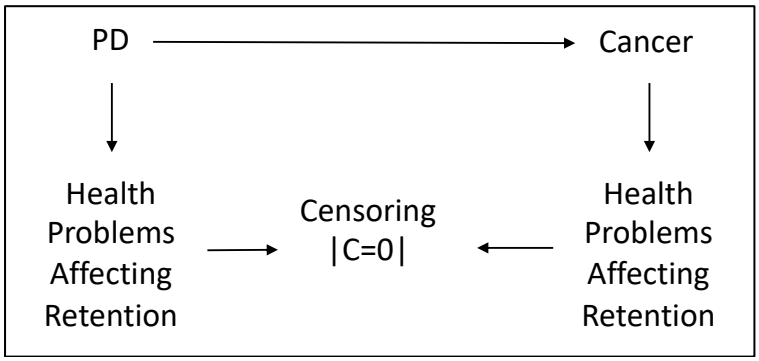

Figure 2.4 Selection bias in PD-cancer cohort studies

Distortions due to selection bias are rarely mentioned by PD-cancer investigators. However, the investigation of the effect of PD on colorectal cancer by Boursi et al. made a specific effort to control for survivor bias [56]. Their study used a nested case-control design with incidence density sampling, matching cases (those with a cancer diagnosis) to controls by age, sex, practice site, and duration of follow-up. By matching cases and controls by duration of

follow-up, the investigators attempted to remove any causal relationship between PD and survival to cancer diagnosis, thus preventing the collider stratification bias through survival.

Cui et al. used inverse probability weighting methods to adjust for potential selection bias in the PD-cancer relationship [42]. This technique involves calculating the probability of selection for each individual included in the analysis and then incorporating the inverse of this probability as a weight in the outcome regression [22], [57]. The selection model incorporated both study participation and survival. Probabilities were obtained from logistic regression models with predictors including PD, cancer, smoking, age, sex, SES, urbanization, and previous cancer history.

2.6 DISEASE ONSET AND SURVIVAL CONSIDERATIONS

Different onset and survival patterns within PD and cancer could contribute to the different effect estimates seen in PD preceding cancer versus cancer preceding PD. Onset of PD typically occurs after age 60 and is rarely diagnosed before age 50 [6]. Cancer, on the other hand, is much more likely to occur before age 60. In 2014, cancer was the leading cause of death in Americans aged 45-54 and the second leading cause of death in Americans aged 35-44 [58]. A meta-analysis found that those with PD experience an approximate decrease in survival of 5% per year of follow-up, while SEER reports an age-adjusted 5-year survival rate for overall cancer of only 66.9% [59], [60]. As a result of these differences in onset and survival, it is easier for PD patients to survive long enough to receive a cancer diagnosis compared to cancer patients surviving long enough to receive a PD diagnosis. Discrepancies will also be seen when studies allow for different follow-up periods.

Similarly, different onset and survival patterns among cancer subtypes could help explain the heterogeneity of effects of PD on specific cancers. Melanoma and lung cancer are notably

different in this regard. Melanoma has a median age of diagnosis of 57 and is one of the most common cancers in young adults [61]. It has a high 5-year survival rate: 89.5% for males, 94.0% for females [60]. On the other hand, lung cancer has a median age of diagnosis of 71 with approximately 0.2% of lung cancers diagnosed in patients aged 20-34 [62]. It has a much lower 5-year survival rate than melanoma: 14.9% for males, 20.8% for females [60].

These differences between melanoma and lung cancer could have significant implications in PD-cancer studies. Due to its younger age of onset and its high survival rate, it may be easier to find melanoma in PD patients, unless the study is restricting its classification criterion to subjects whose first cancer appeared after PD. Based on average age of onset, a person diagnosed with PD would have to wait approximately 10 years for lung cancer onset, making it difficult for lung cancer cases to be detected in a cohort study. These variations highlight the importance of varying the study design, the window between exposure and outcome, and the cancer classification (e.g. “first cancer” vs. “any cancer”) in studies of the effect of PD on cancer.

2.7 DISCUSSION

This review sought to describe the PD-cancer relationship and to explore various non-causal paths that could be contributing to the observed relationship. Investigators have explored this relationship across a variety of study designs and data sources. The consensus from observational studies suggests that most cancers appear less frequently in PD patients, with the notable exception of melanoma. Many studies focused primarily on the effect of PD on overall cancer; results from specific cancers were consequently under-powered. Most of the studies examined a Western, white population. A greater variety of populations under analysis could

help explain the role of genetics in this relationship. Lastly, this review described the importance of considering onset and survival when studying the effect of one disease on another.

PD-cancer effect or association estimates may be susceptible to a variety of biases. Potential confounders may not be controlled in the analysis due to an inability to obtain the necessary data, as occurs with smoking, or because we simply don't have a complete understanding of the etiologies of PD and cancer. Surveillance bias (independent, differential misclassification) may have a positive or negative effect; PD may lead to decreased cancer detection due to the burden of motor/cognitive symptoms or may lead to increased cancer detection through the increased medical contacts that come with institutionalization. Selection bias may occur due to the symptoms of PD and cancer that negatively affect recruitment and retention. By selecting on a study population aged 65+, a survivor bias may be observed since these subjects have characteristics that allowed them to survive until an old age and may be less likely to acquire future disease. Each of these biases was individually addressed to some extent in a PD-cancer study, but no single study did an adequate job of adjusting for all the biases.

Even if the two diseases are completely independent of each other, some instances of PD-cancer comorbidity will occur simply due to chance. For a person aged 65-69, the estimated incidence rate for PD is 134.03 per 100,000 people [63] and is 1,691.6 per 100,000 people for (overall) cancer [60]. Using these rates, about two cases of PD-cancer comorbidity would be expected to occur by chance per 1,000,000 people aged 65-69.

Methodological inconsistencies among PD-cancer studies can impede the ability to make between-study comparisons. Studies often apply different adjustment sets for confounding control. Heterogeneity is also seen in the PD classification criteria. A study may include patients with all types of parkinsonism [17] or exclusively PD [49]. A study may create a unique

diagnostic criteria [64] while others relied on the ICD standard [65]. Lastly, studies will allow for different windows of detection for cancer after PD. A narrower window should make it less likely for patients to receive a cancer diagnosis.

Future PD-cancer investigations could benefit by considering all the causal pathways that are incorporated in the effect estimates of previous studies. The causal review provided by this paper can help guide future study designs and analyses by providing a comprehensive overview of the potential for biases and examples of how investigators have attempted to control for these biases. Investigators can compare their idea of the PD-cancer causal model to those in the review to see if any biasing paths had not been considered. They may also see how their model compares to those of the investigators that have already studied this relationship, which can be particularly helpful when explaining why drastically different effect estimates are found between studies.

This paper highlights several key areas in which investigators should work to reduce bias in PD-cancer studies. However, the inaccessibility of data with rich covariate information and the number of potential biases likely make it impossible to conduct a perfectly unbiased study. Future analyses should therefore consider adjusting for these biases using quantitative bias analysis. Presenting PD-cancer estimates under a variety of plausible bias scenarios will provide more insight into this relationship than continually relying on qualitative discussion of the biases.

3. STUDY 2

SIMULTANEOUS ADJUSTMENT OF UNCONTROLLED CONFOUNDING, SELECTION BIAS, AND EXPOSURE MISCLASSIFICATION IN MULTIPLE-BIAS MODELING

3.1 ABSTRACT

Introduction: Adjusting for multiple sources of bias usually involves adjusting for one bias at a time, with careful attention to the order in which these biases are adjusted. An alternative approach to multiple-bias adjustment involves the simultaneous adjustment of all biases via imputation or regression weighting. This method corresponds to treating the sources of bias as sources of missing data and serves to "reconstruct" the unbiased data that would have been observed based on the assumed nature and degree of connections between the sources of bias and the observed, incomplete data.

Methods: Causal diagrams and probabilistic expressions are used to develop a missing data approach to bias analysis, outlining the steps to performing simultaneous multi-bias analysis. A simulation study was performed to confirm the validity of this method and assess the sensitivity of bias-adjusted effect estimates to misspecification of the bias parameters. An additional simulation study was conducted to compare the simultaneous approach to the sequential adjustment of multiple biases when the correct order of biases is unknown.

Results: Using data affected by uncontrolled confounding, exposure misclassification, and selection bias, the first simulation study revealed that a non-biased effect estimate can be obtained using simultaneous multi-bias adjustment when correct bias parameters are applied. Incorrect specification of every bias parameter by +/- 25% still produced an effect estimate with less bias than the observed, biased effect. The second simulation study resulted in non-biased effect estimates when multiple biases were adjusted simultaneously, but the sequential approach to multiple bias adjustment resulted in biased or non-biased results, depending on whether the correct sequence of adjustments was applied.

Conclusion: Simultaneous multi-bias analysis is a useful tool to help understand how multiple biases could affect an observed effect estimate.

3.2 INTRODUCTION

Although qualitative discussion of potential biases remains common in epidemiological discussions, researchers are increasingly embracing bias analysis as a key to quantifying threats to validity [33]. Methods for adjusting for single biases have grown substantially [30], [31], [66]–[69]. These methods have included simple sensitivity analysis, Monte Carlo risk analysis, Bayesian uncertainty assessment, and external adjustment formulas among others. Variations in these methods include which bias parameters are needed, whether bias parameters are fixed or probabilistic, and whether the bias parameters are applied to the data (i.e. a missing data approach) or to the observed effect estimate (i.e. external adjustment).

On the other hand, the development of methods for the adjustment of multiple biases has been comparatively stagnant [70], [71]. Instruction from Greenland and Lash et al. both describe an approach in which biases are adjusted sequentially [21], [29], [33]. To implement this method, attention must be paid to the order in which the biases are adjusted. As described by Greenland: "One can imagine each correction moving a step from the biased data back to the unbiased structure, as if hypothetically 'unwrapping the truth from the data package' [29]." Such an approach can be difficult if the true sequence of biases is hard to ascertain. In addition, these adjustments can be time-consuming and prone to error if many biases are to be evaluated.

This chapter introduces a method to adjust for multiple biases simultaneously. This method generalizes the concept of combining inverse probability of selection weighting (IPSW) with predictive value weighting, as introduced by Johnson et al. [22], [72], [73]. It relies on predicting the probability of the missing data (uncontrolled confounders, misclassified exposure, or selection) using the available data and externally obtained information or assumptions of the effect of this data on that which is missing (i.e. bias parameters). These predicted probabilities

are then incorporated as simulated values or weights in the outcome regression. This chapter provides steps for performing simultaneous multi-bias adjustment on any combination of uncontrolled confounding, exposure misclassification, and selection bias. A simulation study was performed to demonstrate the validity of this method and explore the sensitivity of effect estimates to misspecified bias parameters. Lastly, a second simulation study was conducted to compare the simultaneous and sequential approaches to multiple bias adjustment.

3.3 METHODS

3.3.1 SIMULTANEOUS MULTI-BIAS ADJUSTMENT

The following binary variables are defined: X = exposure, Y = outcome, $\boldsymbol{C}$ = vector of known confounders, X^* = misclassified exposure, S = selection. These variables are used to represent potential multi-bias scenarios in DAGs (Figure 3.1). Bias can be evaluated in DAGs using the backdoor criterion and other rules [22], [25], [26], [74], [75]. The direct effect of X on Y in these causal models is distorted by backdoor paths stemming from uncontrolled confounding ($X \leftarrow U \rightarrow Y$), information bias in the form of exposure misclassification ($X^* \leftarrow (X) \rightarrow Y$), and selection (i.e. collider stratification) bias ($X \rightarrow |S| \leftarrow Y$). Values of variables U and X are unknown and observations with S=0 are missing. An overlined variable refers to a variable whose value is assigned by the investigator. In the multi-bias analysis, $\bar{X}$ and $\bar{U}$ represent assigned values of exposure and uncontrolled confounder and X_{SIM} and U_{SIM} represent imputed values of exposure and uncontrolled confounder.

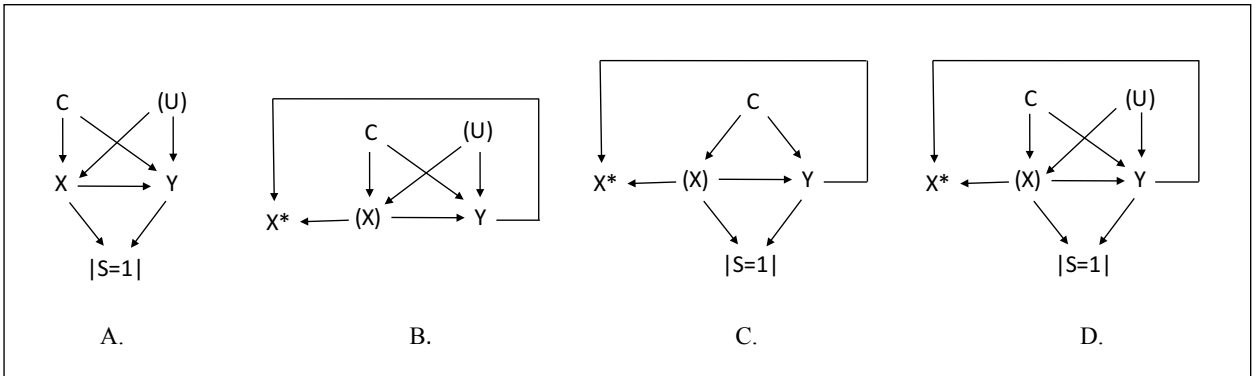

Figure 3.1 Four possible multi-bias scenarios: uncontrolled confounding and selection bias (A), uncontrolled confounding and exposure misclassification (B), exposure misclassification and selection bias (C), and uncontrolled confounding, selection bias, and exposure misclassification (D).

Simultaneous multi-bias analysis combines subject-level data with assumptions of causal strengths to recreate an unbiased dataset. There are 8 steps to performing this adjustment (Table 3.1). Steps 1-5 lead to the calculation of the predicted probabilities of the unknown variables. After these probabilities are established, the investigator can choose to incorporate these probabilities into the data through imputed values, a regression weight, or a mixture of both. Table 3.1 outlines these steps using the weighting approach under various combinations of uncontrolled confounding, exposure misclassification, and selection bias. We detail these steps below.

Table 3.1 Weighting approach to performing simultaneous multi-bias analysis

Step	Uncontrolled confounding & selection bias	Uncontrolled confounding & exposure misclassification	Selection bias & exposure misclassification	All three biases
1. Determine the observed probability and the desired joint probability.	Observed: $P(x, y, \boldsymbol{c}\|S = 1)$ Desired: $P(x, y, \boldsymbol{c}, u)$	Observed: $P(x^*, y, \boldsymbol{c})$ Desired: $P(x, y, \boldsymbol{c}, u)$	Observed: $P(x^*, y, \boldsymbol{c}\|S = 1)$ Desired: $P(x, y, \boldsymbol{c})$	Observed: $P(x^*, \boldsymbol{c}, y\|S = 1)$ Desired: $P(x, y, \boldsymbol{c}, u)$
2. Find the probability that when multiplied by the observed joint probability equals the desired joint probability.	$\frac{P(u\|x, y, \boldsymbol{c})}{P(S = 1\|x, y)}$	$P(x, u\|x^*, y, \boldsymbol{c})$	$\frac{P(x\|x^*, y, \boldsymbol{c})}{P(S = 1\|x^*, y, \boldsymbol{c})}$	$\frac{P(x, u\|x^*, y\,\boldsymbol{c})}{P(S = 1\|x^*, y, \boldsymbol{c})}$
3. Write this bias-adjusting expression into corresponding statistical models: one for exposure or confounder (the predictive value) and one for selection (the IPSW).	$logit\big(P(U = 1)\big) = \alpha_0 + \alpha_1 X + \alpha_2 Y + \boldsymbol{\alpha_3 C}$ $logit\big(P(S = 1)\big) = \delta_0 + \delta_1 X + \delta_2 Y$	$\log\left(\frac{P(X=1,U=1)}{P(X=0,U=0)}\right) = \beta_{1,0} + \beta_{1,1} X^* + \beta_{1,2} Y + \boldsymbol{\beta_{1,3} C}$ $\log\left(\frac{P(X=1,U=0)}{P(X=0,U=0)}\right) = \beta_{2,0} + \beta_{2,1} X^* + \beta_{2,2} Y + \boldsymbol{\beta_{2,3} C}$ $\log\left(\frac{P(X=0,U=1)}{P(X=0,U=0)}\right) = \beta_{3,0} + \beta_{3,1} X^* + \beta_{3,2} Y + \boldsymbol{\beta_{3,3} C}$	$logit\big(P(X = 1)\big) = \theta_0 + \theta_1 X^* + \theta_2 Y + \boldsymbol{\theta_3 C}$ $logit\big(P(S = 1)\big) = \delta_0 + \delta_1 X^* + \delta_2 Y + \boldsymbol{\delta_3 C}$	$\log\left(\frac{P(X=1,U=0)}{P(X=0,U=0)}\right) = \beta_{1,0} + \beta_{1,1} X^* + \beta_{1,2} Y + \boldsymbol{\beta_{1,3} C}$ $\log\left(\frac{P(X=0,U=1)}{P(X=0,U=0)}\right) = \beta_{2,0} + \beta_{2,1} X^* + \beta_{2,2} Y + \boldsymbol{\beta_{2,3} C}$ $\log\left(\frac{P(X=1,U=1)}{P(X=0,U=0)}\right) = \beta_{3,0} + \beta_{3,1} X^* + \beta_{3,2} Y + \boldsymbol{\beta_{3,3} C}$ $logit\big(P(S = 1)\big) = \delta_0 + \delta_1 X^* + \delta_2 Y + \boldsymbol{\delta_3 C}$
4. Externally obtain the regression coefficients for the predictive value and IPSW (i.e. the bias parameters).				

5. Using the bias parameters and individual-level data, compute the predictive value and IPSW. Save the resulting probabilities.				
6. Replicate the data and assign values for exposure and/or uncontrolled confounder.	In first replicate: $\bar{U} = 1$ In second replicate: $\bar{U} = 0$	In first replicate: $\bar{X} = 1, \bar{U} = 1$ In second replicate: $\bar{X} = 1, \bar{U} = 0$ In third replicate: $\bar{X} = 0, \bar{U} = 1$ In fourth replicate: $\bar{X} = 0, \bar{U} = 0$	In first replicate: $\bar{X} = 1$ In second replicate: $\bar{X} = 0$	In first replicate: $\bar{X} = 1, \bar{U} = 1$ In second replicate: $\bar{X} = 1, \bar{U} = 0$ In third replicate: $\bar{X} = 0, \bar{U} = 1$ In fourth replicate: $\bar{X} = 0, \bar{U} = 0$
7. Create individual weights based on the predictive value corresponding to the assigned $\bar{X}$ and/or $\bar{U}$ divided by the IPSW.	When $\bar{U} = 1$: Weight = $\frac{P(U=1)}{P(S=1)}$ When $\bar{U} = 0$: Weight = $\frac{P(U=0)}{P(S=1)}$	When $\bar{X} = 1, \bar{U} = 1$: Weight = $P(X = 1, U = 1)$ When $\bar{X} = 1, \bar{U} = 0$: Weight = $P(X = 1, U = 0)$ When $\bar{X} = 0, \bar{U} = 1$: Weight = $P(X = 0, U = 1)$ When $\bar{X} = 0, \bar{U} = 0$: Weight = $P(X = 0, U = 0)$	When $\bar{X} = 1$: Weight = $\frac{P(X=1)}{P(S=1)}$ When $\bar{X} = 0$: Weight = $\frac{P(X=0)}{P(S=1)}$	When $\bar{X} = 1, \bar{U} = 1$: Weight = $\frac{P(X=1,U=1)}{P(S=1)}$ When $\bar{X} = 1, \bar{U} = 0$: Weight = $\frac{P(X=1,U=0)}{P(S=1)}$ When $\bar{X} = 0, \bar{U} = 1$: Weight = $\frac{P(X=0,U=1)}{P(S=1)}$ When $\bar{X} = 0, \bar{U} = 0$: Weight = $\frac{P(X=0,U=0)}{P(S=1)}$
8. Perform weighted outcome regression.	$logit\big(P(Y = 1)\big) = \omega_0 + \omega_1 X + \omega_2 \bar{U} + \boldsymbol{\omega_3 C}$	$logit\big(P(Y = 1)\big) = \omega_0 + \omega_1 \bar{X} + \omega_2 \bar{U} + \boldsymbol{\omega_3 C}$	$logit\big(P(Y = 1)\big) = \omega_0 + \omega_1 \bar{X} + \boldsymbol{\omega_2 C}$	$logit\big(P(Y = 1)\big) = \omega_0 + \omega_1 \bar{X} + \omega_2 \bar{U} + \boldsymbol{\omega_3 C}$

First, the investigator must establish statements for the observed (biased) joint probability and the joint probability for the desired (non-biased) causal model. Using these two probabilities, a bias-adjusting probability is found by the following equation:

bias-adjusting weight = desired joint probability / observed joint probability

This probability derivation is not unlike that seen in inverse probability of treatment weighting, where the weight may have a numerator with the probability of the treatment conditional on previous treatments (i.e. that which is desired) and the denominator would have the probability of the treatment conditional on previous treatments and the bias-inducing covariates (i.e. that which is observed) [76]. Adjustments for selection bias will include a bias-adjusting probability whose denominator is equal to the probability of selection, as seen in single-bias IPSW [22].

The bias-adjusting weight or expression is then expressed as regression models for the bias-adjusting probability's numerator (the predictive value(s)) and for the bias-adjusting weight's denominator (the IPSW). In the case of both uncontrolled confounding and exposure misclassification, two logistic regression models or a single multinomial logistic regression can be used for the predictive values. In the latter case, the joint probability is represented by the four potential values of the X, U combination.

Regression coefficients for the predictive value and selection models (i.e. bias parameters) are then externally obtained. Plausible parameter values can derive from a variety of sources: previous studies, a validation sub-study, an expert opinion, or simply the investigator's best estimate. As with all bias analyses, it is important to clearly communicate the derivation of

each parameter and to consider assessing a range or distribution of values for each parameter [34]. Once these values are obtained, they are combined with the individual-level data to solve for the selection and the predictive value probabilities of each subject.

The investigator may choose an imputation approach or weighting approach to incorporate the predictive value probabilities of X or U. With the imputation approach, the new value (X_{SIM} or U_{SIM}) is simulated using the predictive value probabilities. In the weighting approach, subjects are first assigned each potential value of exposure and/or unobserved confounder. Thus, the data is replicated 2 times (if X **or** U is missing) or 4 times (if X **and** U are missing) and variable $\bar{X}$ and/or $\bar{U}$ is created, with each replicate containing a different value of the assigned variable(s). The weight is then assigned according to the value of $\bar{X}$ and/or $\bar{U}$; if $\bar{X}$ and/or $\bar{U} = 1$, then the weight equals the predictive value, if $\bar{X}$ and/or $\bar{U} = 0$, then the weight equals 1 – the predictive value. The selection probability is incorporated as an IPSW.

The final outcome regression therefore includes: (1) the imputed values, X_{SIM} or U_{SIM}, with the IPSW or (2) the assigned values, $\bar{X}$ and/or $\bar{U}$, with a weight equaling the predictive value weight divided by the selection probability weight. The exponential of the coefficient of X (or $\bar{X}$) represents the bias-adjusted odds ratio of the direct effect of X on Y.

Imputation and weighting serve to reconstruct the data that would have been observed in the absence of bias given the assumptions implicit in the bias parameters. In the case of selection weighting, observations with a lower probability of selection are given a greater weight in the analysis and vice versa, which serves to restore the initial variable distributions seen in the source population [22]. In the case of predictive value weighting, subjects are assigned each potential value of exposure and/or unobserved confounder and the values that are most probable are given the most weight. Using $\bar{X}$ and $\bar{U}$ along with the predictive value weights essentially

recreates a data set in which the correctly classified exposure and unobserved confounder are both included [67].

3.3.2 SIMULATION STUDY 1: PROOF OF CONCEPT

A simulation study was performed to demonstrate the validity of the simultaneous multi-bias analysis. The method was evaluated using two simulated data sets of binary variables whose causal relations were based off Figure 3.1 DAG D – the triple bias scenario. Monte Carlo simulations generated two data sets of 100,000 rows each (n_{obs} = 100,000) (Table 3.2). One has strong individual biasing paths (Simulation A) and one has weak individual biasing paths (Simulation B).

Table 3.2 Data generating mechanism for the binary variables in Simulations A and B

Variable	Description	Probability
C	Known confounder	0.5
U	Unknown confounder	0.5
X	Unknown, true exposure	$expit(-2 + log(1.5)C + \psi_1 U)$
Y	Outcome	$expit(-2.5 + log(2)X + log(1.5)C + \psi_1 U)$
S	Selection	$expit(\psi_1 X + \psi_1 Y)$
*X**	Misclassified exposure	$expit(\psi_2 + \psi_3 X + log(1.25)Y)$

In Simulation A, $\psi_1 = \log(2)$, $\psi_2 = -1$, and $\psi_3 = \log(5)$. These values were intended to create strong confounding by *U* and strong effects of *X* and *Y* on selection. The resulting misclassified exposure has P(*X** = 1) = .65 or .70 when *X* = 1 and P(*X** = 1) = .27 or .32 when *X*

= 0. Intercepts were selected to keep the probabilities of true exposure, outcome, and selection bound within (.12, .29), (.076, .33), and (.50, .80), respectively. In Simulation B, $\psi_1 = \log(1.25)$, $\psi_2 = -1.5$, and $\psi_3 = \log(15)$. These values were intended to create weak confounding by *U* and weak effects of *X* and *Y* on selection. The resulting misclassified exposure has P(*X** = 1) = .77 or .81 when *X* = 1 and P(*X** = 1) = .18 or .22 when *X* = 0. Intercepts were selected to keep the probabilities of true exposure, outcome, and selection bound within (.12, .20), (.076, .24), and (.50, .61), respectively.

The following regression model (equation 1) represents the triple-bias scenario in which values of true exposure, confounder *U*, and selection are unknown to the investigator:

$$(1)\quad logit(P(Y = 1|X^*, C)) = \alpha_Y + \alpha_{YX^*}X^* + \alpha_{YC}C$$

Here the biased $OR_{YX} = \exp(\alpha_{YX^*})$ will not equal the unbiased $OR_{YX} \approx 2$, which is known based on the simulation of *Y* in both data sets.

The analysis began by identifying the observed (biased) joint probability, (X^*=x^*, *C*=*c*, *Y*=*y* | *S*=1), and the desired (bias-free) joint probability, P(*X*=*x*, *C*=*c*, *U*=*u*, *Y*=*y*). Since $P(x, c, u, y) = \Sigma_{X^*} P(x, u \mid x^*, c, y) P(x^*, c, y \mid S=1) P(S=1) / P(S=1 \mid x^*, c, y)$, the bias-adjusting probability was determined to be $P(x, u \mid x^*, c, y) / P(S=1 \mid x^*, c, y)$. The numerator was simplified as follows: $P(x, u \mid x^*, c, y) = P(u \mid x, x^*, c, y) P(x \mid x^*, c, y) = P(u \mid x, y) P(x \mid x^*, c, y)$. The three probabilities for *U*, *X*, and *S* were rewritten as the predictive value and selection probabilities (equations 2-4):

$$(2)\quad logit(P(U = 1|X, Y)) = \alpha_U + \alpha_{UX}X + \alpha_{UY}Y$$

$$(3)\quad logit(P(X = 1|X^*, Y, C)) = \delta_S + \delta_{SX^*}X^* + \delta_{SY}Y + \delta_{SC}C$$

$$(4)\quad logit(P(S = 1|X^*, Y, C)) = \beta_S + \beta_{SX^*}X^* + \beta_{SY}Y + \beta_{SC}C$$

Having *U*, *X*, and *S* in the data allowed for the fitting of these models to obtain the correct bias parameters, which, although impossible in real-world practice, is necessary for proper evaluation of the bias-adjustment method. To get the 11 parameters, models 2-4 were fit using iteratively reweighted least squares using data from Simulations A and B. As a reminder, real-world strategies for obtaining bias parameters do not involve calculating these parameters from within the sample data but require obtaining the parameters from external sources.

The imputation approach was used for incorporating the predicted probabilities of *X* and *U*. X_{SIM} was simulated using the probabilities obtained from combining the bias parameters with the individual data for *X**, *C*, and *Y*, as in equation 3. U_{SIM} was simulated using the probabilities obtained from combining the bias parameters with the individual data for X_{SIM} and *Y*, as in equation 2. The probability of selection was calculated using the bias parameters and individual data for *X**, *C*, and *Y*, as in equation 4. Lastly, the outcome regression weighted by the inverse probability of selection was fit (equation 5):

$$(5)\quad logit\big(P(Y = 1|X_{SIM}, C, U_{SIM})\big) = \phi_Y + \phi_{YX}X_{SIM} + \phi_{YC}C + \phi_{YU}U_{SIM}$$

To evaluate if bias was appropriately corrected, $\exp(\phi_{YX}) = OR_{YX}$ should approximately equal 2, corresponding to the bias-free OR_{YX} seen in the derivation of *Y* (Table 3.2). To account for the uncertainty of Monte Carlo procedures and to obtain the sampling distribution and confidence interval for OR_{YX}, the analysis was performed on n_{sim} =1,000 bootstrap samples. The

median, 2.5th percentile and 97.5th percentile from the distribution of n_{sim} OR_{YX} estimates were used for the point estimate and 95% simulation interval. Every sampled observation had a value of S=1, thus incorporating selection bias into the data.

In real-world applications, an investigator will not know if the bias parameters he/she obtained are correct, so understanding the sensitivity and resilience of OR_{YX} to misspecification of the bias parameters is essential. The above analyses were therefore repeated using different, incorrect bias parameters to assess changes in OR_{YX} in response to changes in the bias parameters. These parameters were exponentiated so that misspecification could occur on a linear scale instead of a log scale. The effect of distortions to bias parameter(s) on the estimate of OR_{YX} were assessed based on the bias and RMSE, which incorporates both the bias and standard error of the estimate.

The analysis was run using R version 3.2.2. By default, R operates using one CPU core. To significantly increase the speed of bootstrapping, multiple CPU cores were utilized via parallel processing.

3.3.3 SIMULATION STUDY 2: SIMULTANEOUS VERSUS SEQUENTIAL ADJUSTMENT OF MULTIPLE BIASES

A second simulation study was performed to demonstrate the advantage of simultaneous multi-bias adjustment to the sequential adjustment of multiple biases. The sequential method is only valid when biases are adjusted in the correct order, whereas the simultaneous method should be valid regardless of the order in which threats to validity occurred. Using data sets affected by different bias sequences, this study aims to show that simultaneous multi-bias adjustment with correct bias parameters will always lead to a non-biased effect estimate, whereas the sequential approach only results in a non-biased effect estimate when the correct sequence of adjustments is applied.

Two data sets were simulated (n_{obs} = 500,000), each having a different order in which uncontrolled confounding, exposure misclassification, and selection bias occurred (Table 3.3). In both data sets, the first bias to occur was uncontrolled confounding; uncontrolled confounding is a population-level phenomenon that always occurs first in the data-generating process [21]. Simulation C has the following order of biases: 1. uncontrolled confounding, 2. exposure misclassification, 3. selection bias. Simulation D has the following order of biases: 1. uncontrolled confounding, 2. selection bias, 3. exposure misclassification. The binary variables in both simulations are defined as follows: X = true exposure, Y = outcome, U = uncontrolled confounder, X^* = misclassified exposure, S = selection. As seen in the generation of Y, X has a protective effect on Y (OR_{YX} = 0.5) in both simulations. In practice, variables X, U, and S would be unknown to the investigator, therefore contributing to bias. The key difference between the simulations can be seen in the generation of S. In Simulation C, S is dependent on X and Y, meaning that the selection bias preceded the exposure misclassification. In Simulation D, S is dependent on X^* and Y, meaning that the exposure misclassification preceded the selection bias. Besides this difference, the data-generating mechanisms are identical in both simulations.

Table 3.3 Data generating mechanism of binary variables for Simulations C and D

Variable	Description	Probability
U	Unknown confounder	0.5
X	Unknown, true exposure	$expit(-1 + \log(1.5)\,U)$
Y	Outcome	$expit(-1 + \log(0.5)\,X + \log(1.5)\,U)$
X^*	Misclassified exposure	$P(X^* = 1 \mid X = 1, Y = 1) = 0.8$ $P(X^* = 1 \mid X = 1, Y = 0) = 0.7$ $P(X^* = 1 \mid X = 0, Y = 1) = 0.2$ $P(X^* = 1 \mid X = 0, Y = 0) = 0.3$

S	Selection	**Simulation C:** $expit(-0.5 + \log(1.5)X + \log(1.5)Y))$	**Simulation D:** $expit(-0.5 + \log(1.5)X^{*} + \log(1.5)Y))$

The analysis will mimic the scenario in which data for X and U are missing and records are only available for subjects with S=1. Despite the missing and incomplete information, bias analysis can allow for an unbiased OR_{YX} to be obtained if the correct bias parameters are used. Knowledge of the complete data with X, U, and S, which is otherwise assumed to be missing in the analysis, will be used to obtain the correct values of the bias parameters

The weighting approach for simultaneous multi-bias analysis was applied, as explained in Section 3.3.1, with weight equal to E(X, U | X*, Y) / P(S=1 | X*, Y). This study used the simple approach to sequential multi-bias adjustment shown by Lash, Fox, and Kink [33]. To start, a 2x2 table was set up with the observed exposure-outcome frequencies. A simple bias adjustment is applied to the data, a new frequency table is obtained, and this new table is carried forth to the next simple bias adjustment until all biases are accounted for. These simple adjustments require knowledge of the probability of the uncontrolled confounder within exposure-outcome combinations, the probability of selection within exposure-outcome combinations, and the sensitivity and specificity within each outcome as the bias parameters. The adjustment for uncontrolled confounding results in multiple tables across each strata of the confounder, so a pooling method (e.g. Mantel-Haenszel) is needed to obtain the marginal effect estimate.

The value of interest was the bias-adjusted exposure-outcome odds ratio, OR_{YX}. Three values of OR_{YX} were obtained from Simulations C and D: one adjusted sequentially in the *correct* order, one adjusted sequentially in the *incorrect* order (inverting selection bias and exposure misclassification adjustment), and one adjusted simultaneously. Since both data sets were simulated such that the unbiased $OR_{YX} = 0.5$, the bias-adjusted estimate should approximate this

value. Performance was measured by the *bias* as the difference between 0.5 and the bias-adjusted odds ratio. Both methods bootstrapped over n_{sim}=1,000 samples to obtain a simulation interval. Every sampled observation had a value of S=1, incorporating selection bias into the data.

The analysis was run using R version 3.2.2. By default, R operates using one CPU core. To significantly increase the speed of bootstrapping, multiple CPU cores were utilized via parallel processing.

3.4 RESULTS

3.4.1 SIMULATION STUDY 1: PROOF OF CONCEPT

Both data sets were sampled with replacement over S=1 and fit to equation 1 to obtain OR_{YX} estimates biased from uncontrolled confounding, exposure misclassification, and selection bias: Simulation A biased OR_{YX} = 1.46 (95% CI: 1.41, 1.50), Simulation B biased OR_{YX} = 1.54 (95% CI: 1.48, 1.60).

Results from the simultaneous multi-bias analysis are provided in Table 3.4. As expected, when correct bias parameters were used, bias-adjusted $OR_{YX} \approx 2$. Modifying single bias parameters by +/- 25% while leaving the other parameters at the correct value usually resulted in $|Bias|$ less than 0.1. However, misspecification of $e^{\delta_{XY}}$ was particularly impactful in creating bias in OR_{YX}, with $|Bias|$ between 0.3 and 0.5. Larger bias was observed when multiple bias parameters were distorted compared to single parameter misspecification.

Table 3.4 Simultaneous multi-bias analysis in Simulations A and B

Simulation A				
Misspecified parameter(s)	**Degree of Misspecification (%)**	**Bias-adjusted OR_{YX}**	**Bias**	**RMSE**

none	0	2.03 (1.96, 2.11)	-0.0326	0.0491
e^{α_0}	+25	2.04 (1.97, 2.12)	-0.0433	0.0568
e^{α_0}	-25	2.03 (1.95, 2.10)	-0.0257	0.0455
$e^{\alpha_{UX}}$	+25	1.97 (1.90, 2.04)	0.0306	0.0473
$e^{\alpha_{UX}}$	-25	2.13 (2.05, 2.21)	-0.1272	0.1329
$e^{\alpha_{UY}}$	+25	1.98 (1.90, 2.05)	0.0247	0.0446
$e^{\alpha_{UY}}$	-25	2.11 (2.05, 2.19)	-0.1135	0.119
e^{δ_0}	+25	2.02 (1.95, 2.09)	-0.0178	0.0392
e^{δ_0}	-25	2.05 (1.98, 2.13)	-0.0543	0.0659
$e^{\delta_{XX^*}}$	+25	2.02 (1.95, 2.09)	-0.0185	0.0409
$e^{\delta_{XX^*}}$	-25	2.04 (1.96, 2.12)	-0.0365	0.0531
$e^{\delta_{XY}}$	+25	2.47 (2.38, 2.56)	-0.4680	0.4701
$e^{\delta_{XY}}$	-25	1.58 (1.52, 1.64)	0.4215	0.4226
$e^{\delta_{XC}}$	+25	2.02 (1.96, 2.10)	-0.0242	0.0443
$e^{\delta_{XC}}$	-25	2.04 (1.97, 2.12)	-0.0411	0.0553
e^{β_0}	+25	2.03 (1.96, 2.10)	-0.0302	0.0478
e^{β_0}	-25	2.03 (1.96, 2.11)	-0.0330	0.0499
$e^{\beta_{SX^*}}$	+25	2.06 (1.99, 2.13	-0.0608	0.0717
$e^{\beta_{SX^*}}$	-25	2.00 (1.92, 2.07)	0.0044	0.0372
$e^{\beta_{SY}}$	+25	2.04 (1.97, 2.12)	-0.0404	0.0554
$e^{\beta_{SY}}$	-25	2.03 (1.95, 2.10)	-0.0228	0.0434
$e^{\beta_{SC}}$	+25	2.03 (1.97, 2.10)	-0.0326	0.0483
$e^{\beta_{SC}}$	-25	2.03 (1.96, 2.11)	-0.0333	0.0503
e^{α_0}, $e^{\alpha_{UX}}$, $e^{\alpha_{UY}}$	+25	1.92 (1.85, 1.99)	0.0781	0.0856
e^{α_0}, $e^{\alpha_{UX}}$, $e^{\alpha_{UY}}$	-25	2.18 (2.10, 2.25)	-0.1762	0.1806
e^{δ_0}, $e^{\delta_{XX^*}}$, $e^{\delta_{XY}}$, $e^{\delta_{XC}}$	+25	2.40 (2.32, 2.48)	-0.3958	0.3978
e^{δ_0}, $e^{\delta_{XX^*}}$, $e^{\delta_{XY}}$, $e^{\delta_{XC}}$	-25	1.57 (1.50, 1.65)	0.4250	0.4266
e^{β_0}, $e^{\beta_{SX^*}}$, $e^{\beta_{SY}}$, $e^{\beta_{SC}}$	+25	2.06 (2.00, 2.14)	-0.0630	0.0734
e^{β_0}, $e^{\beta_{SX^*}}$, $e^{\beta_{SY}}$, $e^{\beta_{SC}}$	-25	2.00 (1.93, 2.07)	0.0017	0.0362
Simulation B				
Misspecified parameter(s)	**Degree of Misspecification (%)**	**Bias-adjusted OR_{YX}**	**Bias**	**RMSE**
none	n/a	2.01 (1.93, 2.10)	-0.0149	0.0466
e^{α_0}	+25	2.01 (1.93, 2.10)	-0.0108	0.0456
e^{α_0}	-25	2.01 (1.92, 2.10)	-0.0106	0.0483
$e^{\alpha_{UX}}$	+25	1.99 (1.90, 2.08)	0.0147	0.0475
$e^{\alpha_{UX}}$	-25	2.04 (1.96, 2.14)	-0.0447	0.0651
$e^{\alpha_{UY}}$	+25	1.99 (1.91, 2.08)	0.0082	0.0445
$e^{\alpha_{UY}}$	-25	2.04 (1.94, 2.12)	-0.0377	0.0589
e^{δ_0}	+25	1.98 (1.91, 2.07)	0.0182	0.0448
e^{δ_0}	-25	2.05 (1.97, 2.15)	-0.0467	0.0666
$e^{\delta_{XX^*}}$	+25	1.98 (1.90, 2.07)	0.0151	0.0460
$e^{\delta_{XX^*}}$	-25	2.03 (1.94, 2.12)	-0.0305	0.0563
$e^{\delta_{XY}}$	+25	2.35 (2.25, 2.46)	-0.3546	0.3583
$e^{\delta_{XY}}$	-25	1.63 (1.56, 1.71)	0.3704	0.3724
$e^{\delta_{XC}}$	+25	2.00 (1.91, 2.07)	0.0044	0.0412
$e^{\delta_{XC}}$	-25	2.03 (1.94, 2.12)	-0.0306	0.0555
e^{β_0}	+25	2.01 (1.93, 2.10)	-0.0093	0.0447

e^{β_0}	-25	2.01 (1.93, 2.10)	-0.0106	0.0454
$e^{\beta_{SX^*}}$	+25	2.03 (1.93, 2.12)	-0.0271	0.0539
$e^{\beta_{SX^*}}$	-25	1.99 (1.91, 2.07)	0.0139	0.0461
$e^{\beta_{SY}}$	+25	2.02 (1.93, 2.11)	-0.0181	0.0491
$e^{\beta_{SY}}$	-25	2.00 (1.92, 2.10)	-0.0023	0.0463
$e^{\beta_{SC}}$	+25	2.01 (1.92, 2.10)	-0.0094	0.0473
$e^{\beta_{SC}}$	-25	2.01 (1.92, 2.10)	-0.0143	0.0464
e^{α_0}, $e^{\alpha_{UX}}$, $e^{\alpha_{UY}}$	+25	1.95 (1.86, 2.03)	0.0521	0.0675
e^{α_0}, $e^{\alpha_{UX}}$, $e^{\alpha_{UY}}$	-25	2.03 (1.94, 2.12)	-0.0295	0.0551
e^{δ_0}, $e^{\delta_{XX^*}}$, $e^{\delta_{XY}}$, $e^{\delta_{XC}}$	+25	2.24 (2.16, 2.33)	-0.2442	0.2484
e^{δ_0}, $e^{\delta_{XX^*}}$, $e^{\delta_{XY}}$, $e^{\delta_{XC}}$	-25	1.64 (1.55, 1.72)	0.3637	0.3665
e^{β_0}, $e^{\beta_{SX^*}}$, $e^{\beta_{SY}}$, $e^{\beta_{SC}}$	+25	2.04 (1.96, 2.13)	-0.0437	0.0633
e^{β_0}, $e^{\beta_{SX^*}}$, $e^{\beta_{SY}}$, $e^{\beta_{SC}}$	-25	2.01 (1.92, 2.09)	-0.0064	0.0441

The sensitivity of OR_{YX} to changes in the bias parameters when all 11 bias parameters were misspecified by a common factor was assessed (Figure 3.2). The degree of misspecification and the amount of bias were positively related. In both simulations, it was found that the odds ratio resulting from multi-bias adjustment in which each bias parameter was misspecified by +/- 25% still produced an odds ratio estimate that better approximated the true effect when compared to the odds ratio with no bias adjustment.

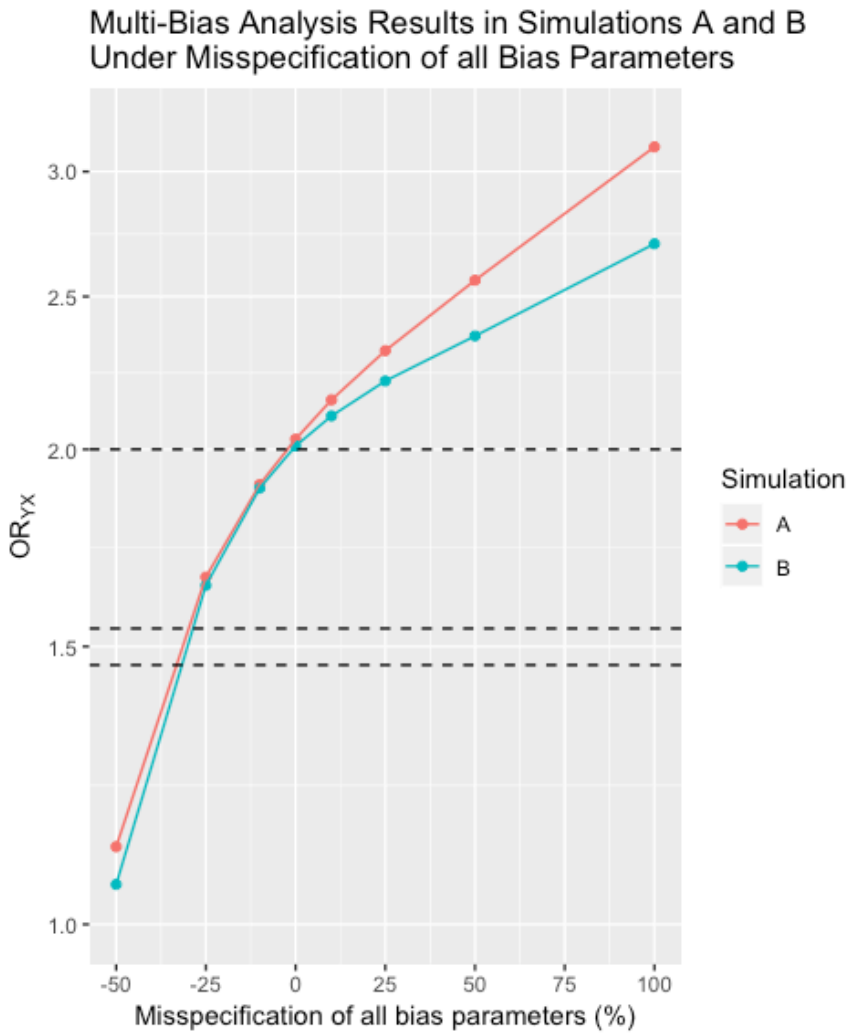

Figure 3.2 Multi-bias analysis results in Simulations A and B under misspecification of all bias parameters.
*Dotted lines represent the non-biased OR_{YX} (2.00), the biased OR_{YX} in Simulation A (1.46), and the biased OR_{YX} in Simulation B (1.54)

3.4.2 SIMULATION STUDY 2: SIMULTANEOUS VERSUS SEQUENTIAL ADJUSTMENT OF MULTIPLE BIASES

Both data sets were sampled with replacement over S=1 and fit to the logistic regression of Y on X* to obtain OR_{YX} estimates biased from uncontrolled confounding, exposure misclassification, and selection bias: Simulation C biased OR_{YX} = 0.66 (95% CI: 0.64, 0.68), Simulation D biased OR_{YX} = 0.67 (95% CI: 0.65, 0.69).

The six bias-adjusted OR_{YX} estimates are provided in Table 3.5. In Simulation C sequential multi-bias adjustment led to an unbiased OR_{YX} estimate (0.50, 95% CI: 0.48, 0.51) when the correct order of adjustments (1. exposure misclassification, 2. selection bias, 3. uncontrolled confounding) was applied but otherwise led to a biased OR_{YX} estimate (0.58, 95% CI: 0.57, 0.60). Likewise, sequential multi-bias adjustment in Simulation D led to an unbiased OR_{YX} estimate (0.51, 95% CI: 0.50, 0.52) when the correct order of adjustments (1. selection bias, 2. exposure misclassification, 3. uncontrolled confounding) was applied but otherwise led to a biased OR_{YX} estimate (0.43, 95% CI: 0.42, 0.44). Simultaneous multi-bias adjustment resulted in unbiased OR_{YX} estimates in both Simulations C (0.50, 95% CI: 0.50, 0.51) and D (0.50, 95% CI: 0.50, 0.51).

Table 3.5 Multi-bias analysis results in Simulations C and D

	Bias-adjusted OR_{YX}		
Multi-bias adjustment method	**Sequential (1) exposure misclassification (2) selection bias (3) uncontrolled confounding**	**Sequential (1) selection bias (2) exposure misclassification (3) uncontrolled confounding**	**Simultaneous**
Simulation C	0.50 (0.48, 0.51)	0.58 (0.57, 0.60)	0.50 (0.50, 0.51)
Simulation D	0.43 (0.42, 0.44)	0.51 (0.50, 0.52)	0.50 (0.50, 0.51)

3.5 DISCUSSION

This paper introduced a novel, simultaneous approach to multiple bias adjustment. A tutorial on how to perform this method on any combination of epidemiological biases was provided. A simulation study using data with an exposure-outcome relationship biased by uncontrolled confounding, selection bias, and exposure misclassification confirmed that an estimate with near-zero bias was obtained when the correct bias parameters were applied. The robustness of effect estimates to distortions in the bias parameters was assessed. Single

parameter misspecification of +/-25% generally led to $|Bias| < 0.10$. In both simulations, it was found that +/- 25% misspecification of all bias parameters produced a bias-adjusted effect estimate with less bias than the observed effect estimate with no bias adjustment. Thus, one can be confident that a biased effect estimate adjusted via simultaneous multi-bias analysis with near-accurate bias parameters is more valid than the estimate without bias adjustment, assuming biases were correctly identified in the DAG.

A second simulation study compared the sequential and simultaneous approaches to multiple bias adjustment. The analyses were performed in data sets that could both be represented by the same DAG, but with biases that occurred in different orders. The sequential multi-bias adjustment only resulted in non-biased odds ratios when adjustments were made in the reverse order in which they occurred in the data. The simultaneous approach, on the other hand, successfully produced non-biased estimates in both data sets without having to consider the order the biases. With this approach, the same steps were applied to both data sets, the only difference was the values of the bias parameters. This simulation study highlights a key advantage of the simultaneous approach; the investigator no longer must worry about obtaining substantially different results based on the order of bias adjustments. Determining the order in which biases occurred can be challenging, particularly when working with secondary data.

This advantage of simultaneous multiple-bias adjustment comes with a cost – deriving a larger number of potentially less intuitive bias parameters. The usual strategies of obtaining parameter estimates from the literature or from expert opinion may be applied [21]. A more efficient strategy would be to use data from a sub-study to inform the bias parameters. For example, a subset of the data may have information for the uncontrolled confounder or a better-classified exposure. In this case, models for U and X may be fitted to the data subset to obtain

bias parameters. A similar approach may be used to obtain the selection bias parameters if information is present for subjects who were invited but chose not to participate. Lastly, more advanced simulation strategies may be used to avoid having to reason about the bias parameters backwards (e.g. from X and C to U) [32]. For example, all the observed relationships may be combined with the hypothesized effects of U on X and Y to simulate a new dataset. This new data could then be fit to a model for U to obtain the bias parameters.

While coming up with values for all the bias parameters may seem like an arduous task, it is important to consider that every study inherently makes assumptions regarding these bias parameters [29]. Studies without bias adjustment inherently assume that if all models are correctly specified all estimates are valid and that the only source of uncertainty in these estimates is random error. Such assumptions are generally lazy and incorrect. Any attempt to improve on these implausible assumptions is worth the effort of the investigator. Multiple effect estimates that derive from different DAGs or different bias assumptions can and should be presented [34]. This transparency allows the reader to understand the resilience of the effect estimate to various bias scenarios and can also help editors identify key areas of improvement [35]. It is possible and advisable to incorporate uncertainty into each bias parameter. Bias parameters may be represented by probability distributions instead of fixed values (i.e. probabilistic bias analysis) [21]. This semi-Bayesian approach allows for the simulation interval of the effect estimate to incorporate uncertainty due to both random error and systematic error [21].

It is important to consider other applications of bias analysis besides attempting to produce the best bias-adjusted effect estimate with the best bias parameters. One can also perform the exercise of evaluating bias strengths that would lead to a null effect estimate or an

effect estimate whose confidence interval includes the null, as in the E-value [36]. By understanding the bias strengths that would "explain away" the observed effect, one gets a better idea of how likely the effect is non-null. Existing tools for these analyses focus on the case of uncontrolled confounding; corresponding tools for multiple biases have yet to be developed.

The multi-bias analysis presented here is limited to binary exposures and outcomes. Future work should expand on this method to include other variable types, model families, and measures of effect (e.g. risk differences). More complex scenarios with multiple biasing paths should also be considered. The simulation studies conducted here only evaluated the case of a single uncontrolled confounder and a single selection bias mechanism. Lastly, considering the computational complexity of these methods, additional work is needed to make simultaneous bias-adjustment more accessible to researchers across disciplines. Analytical tools should be created that can assist with performing the bias adjustment.

4. STUDY 3

QUANTIFYING THE EFFECT OF PD ON CANCER WITH SIMULTANEOUS MULTI-BIAS ANALYSIS USING DATA FROM THE DANISH NATIONAL HOSPITAL REGISTRY

4.1 ABSTRACT

Introduction: Studies of the relationship between PD and cancer have mostly showed that PD has a protective effect on cancer. There are a variety of biases that could influence this relationship including uncontrolled confounding, information bias, and selection bias. Simultaneous multi-bias analysis was used with data from PASIDA to obtain a bias-adjusted PD-cancer effect estimate.

Methods: A retrospective cohort study was designed to quantify the effect of PD on cancer within five years. The observed crude and multivariable-adjusted effect estimates were determined for overall, smoking-related, and non-smoking-related cancer. Simultaneous multi-bias adjustment for PD misclassification and collider stratification at study participation and censoring was used to obtain a bias-adjusted PD-cancer effect estimate.

Results: The observed multivariable-adjusted effect of PD on subsequent overall cancer had an OR of 1.07 (95% CI: 0.83-1.39). Estimates differed substantially when considering smoking-related versus non-smoking-related cancers. After bias adjustment for both exposure misclassification and selection bias, the PD-overall cancer effect estimate remained consistent (OR = 1.07; 95% CI: 0.80-1.47).

Conclusion: Parkinson's disease had a null effect on overall cancer within a five-year window in this Danish study population. Neither selection bias nor exposure misclassification had a major impact on effect estimates of PD on overall cancer.

4.2 INTRODUCTION

A potential inverse relationship between PD and cancer was first described by Doshay in 1954 [16], [17]. Numerous observational studies have since attempted to quantify this relationship. Meta-analyses by Bajaj et al. and Catala-Lopez et al. report estimates of the effect of PD on overall cancer as a relative risk of 0.73 (95% CI: 0.63-0.83) and pooled effect size of 0.83 (95% CI: 0.76-0.91), respectively [18], [19]. There is currently no concrete biological model to account for this relationship, but proposed mechanisms involve common genes between the two diseases, disruption of the ubiquitin-proteasome system, inflammation, and the microbiome. Both conditions have a significant public health and economic burden in developed countries, so understanding their relationship and etiologies is imperative.

PD-cancer studies commonly suffer from various potential sources of bias. These studies commonly rely solely on electronic medical records. Consequently, behavioral variables are unknown to the investigator. Uncontrolled confounding may occur due to the lack of adjustment for smoking and caffeine and alcohol consumption. Information bias can occur due to inaccurate PD diagnoses. The only gold standard diagnosis for PD involves examining the brain post-autopsy [1]. Relying on the presence or absence of motor features makes PD a difficult condition to diagnose; a recent meta-analysis found a pooled diagnostic accuracy for PD of 80.6% [50]. Selection bias may occur as a result of studies inherently conditioning on those who participate and were not lost to follow-up and if participation or loss-to-follow-up is associated with uncontrolled PD/cancer risk factors [22]. An open, biasing path between PD and cancer may exist if there is collider stratification at the points of study participation and censoring. PD and cancer are both burdensome conditions, and their presence may affect whether a patient wants to join a study or stay in the study.

A retrospective cohort study was used to determine the effect of PD on subsequent cancer within five years using data from the Danish National Hospital Registry and the Danish Cancer Registry and information from interviews or surveys of the study participants. The PD-cancer association in this data may suffer from selection bias and exposure misclassification. Simultaneous multi-bias adjustment was used to obtain an unbiased effect estimate. Medical data for non-participants and censoring information for participants were used to inform selection bias adjustment. PD diagnosis sensitivity and specificity were combined with assumptions of the impact of covariates on PD diagnosis accuracy to determine the parameters used for exposure misclassification adjustment. PD-cancer estimates under different bias adjustments showed which, if any, biases contributed most.

4.3 METHODS

PASIDA (Parkinson's in Denmark Study) is a population-based case-control study conducted in Denmark to identify environmental and genetic risk factors for PD. Study details have been described elsewhere [7], [77]. Patients over 35 years of age were identified from the Danish National Hospital Register files between 1996 and 2009 at ten major neurological departments with a PD diagnosis (ICD-10 code: G20) assigned by at least one neurologist. Reference subjects were selected from the Danish Central Population Registry matched on birth year and sex, being alive and without a PD diagnosis at the time of PD identification. Complete medical record review was conducted to verify idiopathic PD (iPD) diagnosis for all PD patients who responded to the study invitation and 378 of these patients were excluded due to non-iPD. Among 2,718 eligible subjects with PD, 1,815 were enrolled. Of 3,626 controls initially contacted, 1,885 were enrolled.

Structured telephone interviews or self-administered questionnaires were conducted between 2007 and 2009 to obtain lifestyle and behavioral information. Variables that were collected and used in this study include: education, urbanization, physical activity, smoking, and family history of PD. The degree of urbanization was based on population density of the participants' community. Physical activity was defined as any vigorous leisure time physical activity during the young, adult, or older adult periods (15–25, 25–50 or > 50 years). Cigarette smoking was condensed to a binary variable based on whether the subject smoked >= 1 cigarette. The Danish Cancer Registry is a nationwide population-based registry containing data on each primary cancer in all residents in Denmark since 1943 and had data through 2016. This registry was used to identify cancer diagnoses before and after the start of follow-up classified by Danish Cancer Society diagnosis code [78], [79].

A retrospective cohort study was implemented to assess the effect of PD on cancer within a five-year window. The start of follow-up was classified as the date of PD diagnosis for PD subjects. Non-PD subjects had the same start of follow-up as their matched PD subject. Logistic regression models were used to estimate the risk of cancer in five years in those with and without PD. Here the odds ratio approximates the risk ratio due to a low incidence of the outcome. Separate models were fit with and without confounder adjustment. The following variables were considered as potential confounders: age, sex, urbanization, physical activity, smoking, family history of PD, and cancer preceding follow-up. The analysis examined overall cancer, smoking-related cancer, and non-smoking-related cancer as outcomes.

These PD-cancer effect estimates may be distorted due to exposure misclassification and selection bias (Figure 4.1). Although diagnoses were conducted by neurological specialists, PD is a difficult disease to diagnose. The only definite diagnosis requires pathological examination

post-autopsy and diagnoses are generally made based on the presence of multiple motor symptoms [1]. Potential factors that may influence the variability of the PD diagnosis accuracy include age, PD family history, urbanization, and education. Two different processes may contribute to selection bias. Collider stratification may occur due to the study inherently conditioning on those who agreed to participate (Participation = 1) and those not lost to follow-up (Censoring = 0). The exposure (PD), outcome (cancer within five years of follow-up), or any of the confounders may affect censoring. It is assumed that PD and cancer before follow-up may affect study participation.

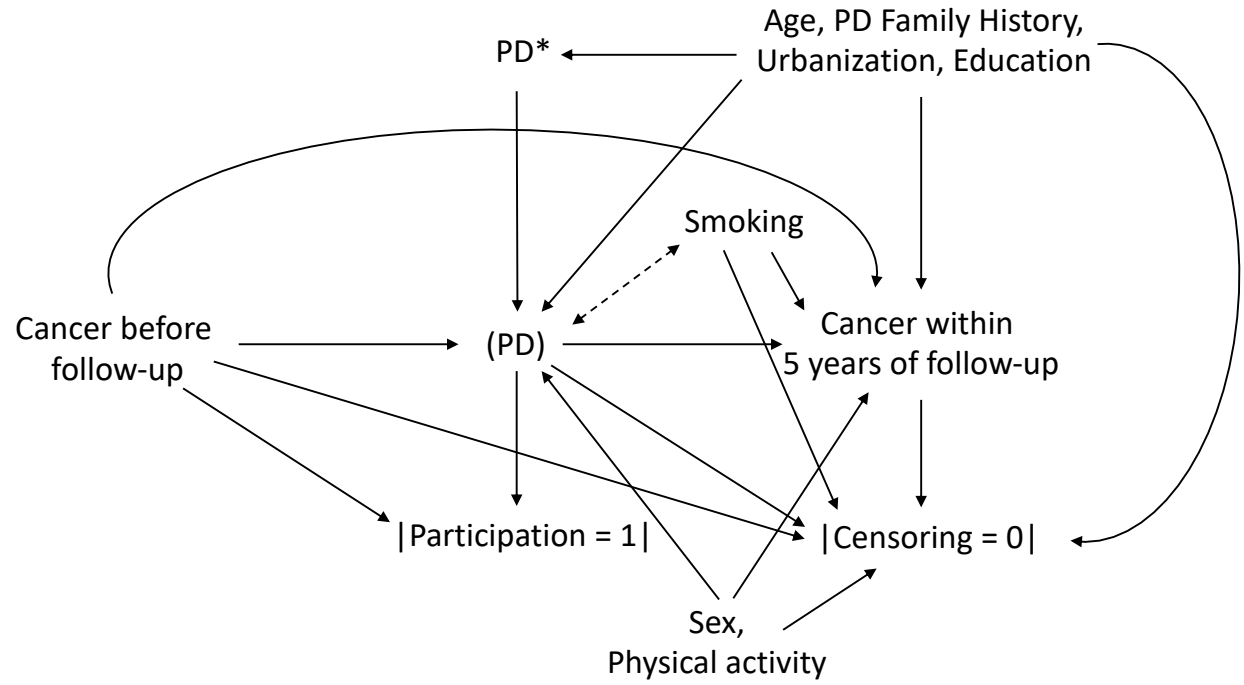

Figure 4.1 PASIDA causal model

Simultaneous multi-bias adjustment was used to adjust for exposure misclassification and selection bias. This method is described in detail in Chapter 2. The imputation approach was applied with the following model representing the probability of the true exposure:

$$logit(P(PD = 1|PD^*, Age, Sex, Smk, FH, PA, Urb, Edu, PC)) =$$

$$\alpha_0 + \alpha_{PD^*}PD^* + \alpha_{Age}Age + \alpha_{Sex}Sex + \alpha_{Smk}Smk + \alpha_{FH}FH + \alpha_{PA}PA + \alpha_{Urb}Urb + \alpha_{Edu}Edu + \alpha_{PC}PC$$

where PD = true, missing Parkinson's disease status, PD* = misclassified PD status, Sex = sex (female = 1), Smk = smoking, FH = PD family history, PA = physical activity, Urb = urbanization (1 = capital city, 0 = provincial, rural, or peripheral region), Edu = education (1 = grade 7-10, 0 = past grade 10), and PC = previous cancer.

Bias parameters for the exposure model were derived from a simulated data set. The eight confounders were simulated based on their frequency in PASIDA. Using the PASIDA data, PD was regressed on the eight confounders. The (true) PD was simulated using the regression coefficients from this model. Next, the literature was reviewed to obtain the sensitivity and specificity of PD diagnoses. A meta-analysis by Rizzo et al. found a sensitivity of 81.3% and specificity of 83.5% for PD diagnosed by experts [50]. The (misclassified) PD was simulated using these values of the sensitivity and specificity in addition to some additional assumptions: a positive family history increases the probability of a correct diagnosis by 2%; urbanization (i.e. living in a capital city) increases the probability of a correct diagnosis by 2%; higher education increases the probability of a correct diagnosis by 2%; every ten year increase in age from the average age decreases the probability of a correct diagnosis by 1%; every ten year decrease in age from the average age increases the probability of a correct diagnosis by 1%. Using these simulated values, the exposure model was fit, and the resulting coefficients represent the bias parameters.

To account for selection bias, the following IPSW was used as a weight in the outcome regression:

$$\frac{1}{P(S_1 = 1|PD^*, PC)P(S_2 = 0|PD^*, Cancer, Age, Smk, S_1 = 1)}$$

where S_1 = participation, S_2 = censoring, and Cancer = cancer that occurred after follow-up but before censoring.

The following models were used to obtain the probability of participation and censoring for each subject:

$$logit\big(P(S_1 = 1|PD^*, PC)\big) = \beta_0 + \beta_{PD^*}PD^* + \beta_{PC}PC$$

$$logit(P(S_2 = 1|PD^*, Cancer, Age, Smk, S_1 = 1) = \delta_0 + \delta_{PD^*}PD^* + \delta_{Cancer}Cancer + \delta_{Age}Age + \delta_{Smk}Smk$$

The participation model was fit using data for all the subjects who were invited to participate in PASIDA. PD and previous cancer status were known for all invitees. Data on censoring events (death, emigration, or disappearance) was available for all participants. Censoring was initially regressed against PD, cancer (after follow-up but before censoring), and all the confounders. Only four variables were found to be significant predictors of censoring: PD, cancer, age, and smoking.

Probabilistic bias parameters with a normal distribution were used to incorporate uncertainty in the bias parameters. Bootstrapping with parallel processing was used to obtain a valid simulation interval. The analysis was performed using R version 3.5.1.

4.4 RESULTS

The PASIDA subject demographics are presented in Table 4.1. PD patients and reference subjects were similar regarding most variables. There was less smoking observed in

the PD patients. There was also a greater prevalence of PD family history within the PD patients.

Table 4.1 PASIDA study population

Characteristic	PD patients (N = 1815) *n (%)*	Reference subjects (*N* = 1885) *n (%)*
Age at start of follow-up (years), mean (SD)	62.3 (9.3)	62.4 (9.5)
Sex		
Male	1072 (59.1)	1119 (59.4)
Female	743 (40.9)	766 (40.6)
Education		
7^{th}-10^{th} grade	1405 (77.4)	1474 (78.2)
Ordinary/extended technical preparation exam or higher commercial exam	86 (4.7)	98 (5.2)
High school certificate or HF exam	306 (16.9)	293 (15.5)
Unknown or unspecified	18 (1.0)	20 (1.1)
Degree of Urbanization		
Capital	444 (24.5)	563 (29.9)
Provincial cities	1119 (61.7)	964 (51.1)
Rural areas	167 (9.2)	209 (11.1)
Peripheral regions	82 (4.5)	148 (7.9)
Unknown or unspecified	3 (0.2)	1 (0.1)
Physical activity[a]		
Ever	1337 (73.7)	1453 (77.1)
Never	400 (22.0)	406 (21.5)
Unknown or unspecified	78 (4.3)	26 (1.4)
Smoking[b]		
Ever	901 (49.6)	1207 (64.0)
Never	909 (50.1)	667 (35.4)
Unknown or unspecified	5 (0.3)	11 (0.6)
Family history of PD		
Yes	251 (13.8)	99 (5.3)
No	1564 (86.2)	1786 (94.7)
Any cancer preceding follow-up		
Yes	201 (11.1)	214 (11.4)
No	1614 (88.9)	1671 (88.6)

[a] Defined as any strenuous leisure physical activity during 15-25, 25-50, or >50 age period.
[b] Defined as >= 1 cigarette smoking.

Estimates of the effect of PD on subsequent cancer are presented in Table 4.2. When analyzing the effect of PD on overall cancer, the effect is close to null (OR=1.07; 95% CI: 0.83-1.39). A strong protective effect was observed for the effect of PD on smoking-related cancer

(OR=0.41; 95% CI: 0.24-0.72), whereas PD was a moderate risk factor for non-smoking-related cancer (OR=1.37; 95% CI: 1.03-1.83). Multivariable adjustment for eight confounders (age at start of follow-up, sex, education, urbanization, physical activity, smoking, family history of PD, and previous cancer) did not have a major impact on any of the estimates.

Table 4.2 Crude and adjusted association between PD and subsequent cancer in PASIDA

	PD patients (N= 1815) *n (%)*	Reference subjects (N = 1885) *n (%)*	Crude OR (95% CI)	Multivariable adjusted OR (95% CI)[a]
Overall cancer (*n* = 303)	155 (8.5)	148 (7.8)	1.10 (0.87, 1.39)	1.07 (0.83, 1.39)
Smoking-related cancer[b] (*n* = 73)	22 (1.2)	51 (2.7)	0.44 (0.27, 0.73)	0.41 (0.24, 0.72)
Non-smoking-related cancer (*n* = 238)	135 (7.4)	103 (5.5)	1.39 (1.07, 1.81)	1.37 (1.03, 1.83)

[a] Multivariable analysis adjusted for age, sex, education, urbanization, physical activity, smoking, family history of PD, and previous cancer.
[b] Included pharynx, oesophagus, stomach, colon incl. rectosigmoideum, rectum and anus, liver, pancreas, nasal cavities, middle ear and sinuses, larynx, lung, bronchus and trachea, cervix uteri, ovary, fallopian tube and broad ligament, kidney, renal pelvis and ureter, urinary bladder, myeloid leukemia.

The bias parameters for the PD imputation model were: $\alpha_0 \sim N(-1.13, 0.34)$, $\alpha_{PD*} \sim N(2.92, 0.09)$, $\alpha_{Age} \sim N(0.0056, 0.0046)$, $\alpha_{Sex} \sim N(-0.20, 0.09)$, $\alpha_{Smk} \sim N(-0.66, 0.09)$, $\alpha_{FH} \sim N(1.26, 0.16)$, $\alpha_{PA} \sim N(-0.15, 0.10)$, $\alpha_{Urb} \sim N(-0.29, 0.10)$, $\alpha_{Edu} \sim N(-0.19, 0.11)$, $\alpha_{PC} \sim N(-0.06, 0.14)$. The bias parameters for the participation model were: $\beta_0 \sim N(0.09, 0.04)$, $\beta_{PD*} \sim N(0.61, 0.05)$, $\beta_{PC} \sim N(-0.04, 0.06)$. The bias parameters for the censoring model were $\delta_0 \sim N(-15.42, 1.29)$, $\delta_{PD*} \sim N(0.84, 0.23)$, $\delta_{Cancer} \sim N(1.07, 0.25)$, $\delta_{Age} \sim N(0.16, 0.02)$, $\delta_{Smk} \sim N(0.42, 0.23)$.

Table 4.3 shows the PD-overall cancer effect estimate under different forms of bias adjustment. The most thorough analysis, which accounts for selection bias, PD misclassification, and eight observed confounders, produced an odds ratio of 1.07 (95% SI: 0.80-1.47). The PD-overall cancer odds ratio stayed within the range of 1.00-1.15 in each of the bias adjustment scenarios.

Table 4.3 Association between PD and subsequent overall cancer in PASIDA, adjusted for different combinations of biases

Model	Conventional analysis		Probabilistic bias analysis	
	OR	95% CI	Median OR	95% SI
Unadjusted	1.10	0.87, 1.39		
Observed confounding	1.07	0.83, 1.39		
Selection bias & Observed confounding			1.14	0.85, 1.65
Exposure misclassification & Observed confounding			1.02	0.79, 1.34
Selection bias & Exposure misclassification & Observed confounding			1.07	0.80, 1.47

4.5 DISCUSSION

This study attempted to apply a novel method for multiple bias adjustment to better understand the relationship between Parkinson's disease and cancer. In PASIDA the unadjusted effect of PD on overall cancer within a five-year window was found to be OR=1.10 (95% CI: 0.87-1.39). This result differs from the general pattern in the literature of a protective effect of PD on overall cancer. The crude estimates were in opposite directions when assessing only smoking-related (OR = 0.44; 95% CI: 0.27-0.73) and non-smoking-related cancers (OR = 1.39; 95% CI: 1.07-1.81). Although past studies typically show a greater protective effect in smoking-related versus non-smoking-related cancer, a difference of this magnitude between the two estimates was not expected.

Estimates were obtained with bias adjustment for exposure misclassification, selection bias, and both. The effect estimate adjusted for both biases (OR = 1.07; 95% CI: 0.80-1.47) was nearly identical to the estimate without bias adjustment. The estimates with single bias adjustment indicate that exposure misclassification and selection bias may be acting equally in

opposite directions with exposure misclassification leading to a smaller odds ratio towards the null and selection bias leading to a larger odds ratio away from the null. This PD misclassification follows the general convention that non-differential misclassification with independent errors tends to bias effect estimates towards the null [80]. The results of selection bias adjustment indicate that more PD patients would be diagnosed with cancer if study participation and censoring was the same for those with and without PD and cancer.

The fitted model for participation showed that cancer before follow-up had a very weak effect on participation (OR = 0.96). As a result, collider stratification bias due to participation was likely non-impactful. The fitted model for censoring showed that Parkinson's disease and cancer after follow-up both had strong effects on subsequent loss to follow-up (OR = 2.31 and 2.92, respectively). A person without PD and cancer was therefore much more likely to have been followed for the entire five-year duration. This selection bias mechanism may play a role in explaining the differences observed in the effect estimates among different cancers. For example, suppose that PD increases the risk of melanoma, as is currently hypothesized [20], [46]–[48]. A patient with PD and melanoma would consequently have a lower chance of being diagnosed with a "late detection cancer" (cancer occurring after melanoma but before the end of follow-up) compared to a subject without PD since this patient is more likely to be censored following the melanoma diagnosis. In other words, the effect of PD on specific cancers may be influenced by the timing in which cancers appear, are symptomatic, and are diagnosed due to differential censoring.

The bias adjustment in this study demonstrates an important technique that is applicable to any bias analysis method: bias parameters can be derived from a simulated data set that brings together information from different sources to construct the assumed data generating process or

causal model. The PD misclassification parameters in PASIDA were obtained in this fashion. A simulated data set that was generated through assumptions of how variables are distributed in PASIDA, PD sensitivity and specificity found in the literature, and opinions of the investigators regarding how some confounders could impact PD misdiagnosis. This technique was the best approach because no available data or literature source could have provided all the parameters.

A key strength of this study was the thorough attention to systematic error via multiple bias modeling and the presentation of results under different bias adjustments. Nevertheless, a potentially impactful source of bias, surveillance bias, was not accounted for. Surveillance bias, a form of differential misclassification, can occur when subjects in one exposure group have a higher probability of having the outcome detected as a result of increased frequency or duration of attention [21]. Studies attempting to determine the relationship among diseases of aging may be vulnerable to this bias [43]. To adjust for this bias, the number of neurological examinations and/or cancer screenings would have to be obtained and balanced between the exposure groups.

Additionally, this study had rich data available for each subject on variables associated with PD, including PD family history, urbanization, education, exercise, and smoking. However, it was observed that the effect estimates changed minimally after adjustment for confounders. This result may be explained by residual confounding. In this study, smoking was defined as ≥ 1 cigarette ever smoked. This categorization serves to distinguish smokers versus non-smokers in a broad sense but does not capture the variability in the duration and quantity of cigarettes smoked among smokers. Thus, some, but not all, of the confounding by smoking is accounted for in the analysis. This residual confounding may also apply to physical activity.

This study was not able to obtain effect estimates under longer windows of examination than five years. The sensitivity of results under varying lengths of follow-up would be helpful to

include. A larger sample size would help to minimize random error. It would also provide sufficient power to detect effects of PD on specific cancers. Future studies should examine the effect of PD on specific cancers, with consideration of whether the cancer is smoking-related or not and whether the cancer is generally detected earlier or later relative to other cancers. Smoking should be obtained and balanced between PD groups by a detailed measure, such as pack years, which accounts for both the duration and intensity of smoking. Studies in non-European and less homogenous populations will assist with generalizability and may add insight to the role of genetics in the PD-cancer association.

5. STUDY 4

INTERACTIVE WEBSITE AND R PACKAGE FOR SIMULTANEOUS MULTI-BIAS ADJUSTMENT

5.1 ABSTRACT

Introduction: Simultaneous multi-bias adjustment has been described in detail, validated in a simulation study, and applied to real-world data. However, various barriers exist that prevent wide-spread adaptation of this method, including insufficient understanding of the methodology and inadequate statistical programming skills. By creating tools such as an interactive website and an R package that assist in bias analysis, the incorporation of multiple bias adjustment will become more accessible to investigators.

Methods: R version 3.5.1 was used to create two tools to assist with simultaneous multi-bias adjustment. An interactive web application was created using the Shiny package version 1.2.0 in R. The "multibias" R package for simultaneous multi-bias adjustment was built using the "devtools" package version 2.1.0 and documented using the "roxygen2" package version 6.1.1.

Results: The web application for simultaneous multi-bias adjustment (https://pcbrendel.shinyapps.io/multibias/) includes the following sections: instructions, data import, classifying variables, and quantifying bias parameters and viewing results. The "multibias" R package for simultaneous multi-bias adjustment (https://github.com/pcbrendel/multibias) includes functions that adjust for any combination of uncontrolled confounding, exposure misclassification, and selection bias. It also includes four simulated data sets with biased exposure-outcome relationships and data from the Evans County Heart Study. Documentation is provided to describe the package, functions, and data.

Conclusion: An R package and a Shiny app were created to help facilitate easier implementation of simultaneous multi-bias analysis. The Shiny app is recommended for non-R users with data available in comma, semicolon, or tab-separated values; whereas the R package is recommended for R users who would like to expand their analyses beyond the capabilities of the Shiny app.

5.2 INTRODUCTION

Quantitative bias analysis should be a key component of epidemiological investigations that attempt to quantify the effect of an exposure on an outcome. It demonstrates how the validity of an estimate can be susceptible to various sources of systematic error and how the uncertainty is affected by both random error and systematic error. Despite these benefits, bias analyses are often absent in epidemiological publications due to a lack of understanding of the methodology, an inability to program the analysis, and minimal demand from the reviewer. To overcome these first two barriers, bias-adjustment tools can be created to help facilitate the analysis [37]. Two such tools can be created with R software – an interactive web application to assist those with minimal programming skills and an R package to allow for more features and customizability beyond the web app.

As an open-source software, R code can easily be created and shared with others. In R, the fundamental unit of shared code is the package [81]. A package includes code, data, documentation, and tests. The coding functions included in packages help others automate repetitive or complex tasks. Packages can be distributed in a variety of ways, such as through the Comprehensive R Archive Network (CRAN) or GitHub.

Shiny is an R package that is used to build interactive web applications straight from R. These apps can be deployed on the web for free via the RStudio hosting service. Shiny apps are built upon a reactive programming model. There are three kinds of reactive programming objects: reactive sources, reactive conductors, and reactive endpoints. Reactive sources (typically accessed as input objects in Shiny) provide the signal to downstream objects to re-execute. Reactive endpoints (typically accessed as output objects in Shiny) are told to re-execute by the reactive environment. Reactive conductors are placed between sources and endpoints.

They are used to improve slow or computationally intensive operations and can be implemented with the reactive() function in Shiny.

5.3 METHODS

An interactive web application was created using the "Shiny" R package version 1.2.0. The "multibias" R package for simultaneous multi-bias adjustment was built using the "devtools" R package version 2.1.0. The "devtools" package also assisted in testing the functionality of the "multibias" package as it was under development and allows others to load the "multibias" package from any location. The "roxygen2" R package version 6.1.1 provided convenient features for documenting all aspects of the package, including the main description, functions, and data. All coding was done in R version 3.5.1.

5.4 RESULTS

The web application (https://pcbrendel.shinyapps.io/multibias/) is organized in four different sections that can be selected via tabs at the top of the page. The first section provides instructions regarding how to use the app and how variables are defined. It also provides a brief description of the bias adjustment methodology. The next section allows the user to upload data that is comma, semicolon, or tab-separated. The user can select whether to display all of the data or just the head. The third section provides inputs for the user to classify which variables correspond to the exposure, outcome, and confounder(s). After these variables are identified, results of the logistic regression of the outcome on the exposure and confounder(s) are displayed. Here, the user can see the exposure-outcome odds ratio and confidence interval corresponding to the estimate with no bias-adjustment. The last section provides the tools for bias adjustment. The user selects which combination of multiple biases is present: (1) uncontrolled confounding and selection bias, (2) uncontrolled confounding and exposure misclassification, (3) exposure

misclassification and selection bias, (4) or all three biases. The appropriate bias models are displayed depending on which biases are identified. The user can also choose whether to use fixed or probabilistic bias parameters. If probabilistic bias parameters are desired, the user must then select whether to use parameters with a Normal or Uniform distribution. The user then inputs the values for the bias parameters as the fixed value, the mean and standard deviation (of the Normal distribution), or the minimum and maximum (of the Uniform distribution). The bias models serve as the reference for interpreting each bias parameter. Lastly, the user selects the range of the confidence interval and how many bootstrap samples are desired. To begin the analysis, the user must press the action button. After this button is pressed, the screen will display the bias-adjusted exposure-outcome odds ratio, the confidence interval of the estimate, and a histogram of the distribution of estimates from each bootstrap sample. The amount of time needed to perform the analysis will vary depending on the size of the data and the number of bootstrap samples. If probabilistic bias parameters are used, the confidence interval will incorporate uncertainty in both the random error and the systematic error.

The "multibias" R package includes functions for adjusting for the various combinations of uncontrolled confounding, exposure misclassification, and selection bias. Each function requires the same set of arguments: the data frame, the variable in the data corresponding to the exposure, the variable in the data corresponding to the outcome, the variable(s) in the data corresponding to the confounder(s), and the bias parameters. A series of simulated data sets with each bias combination are also included in the package. Each of these data sets has a true exposure-outcome OR= 2, so the user can use this data to practice multiple bias adjustment. In addition, data is included from the Evans County cohort study in which white males were followed for seven years, with coronary heart disease as the outcome of interest [82]. The

functions and data sets are all documented, and this documentation can be viewed in the "Help" tab of RStudio. The function documentation provides the bias models that are needed to inform the bias parameter input. The package home page (https://github.com/pcbrendel/multibias) provides key details on how to install and use the package and version updates.

5.5 DISCUSSION

An R package and Shiny interactive web application were developed to help investigators perform simultaneous multi-bias analysis. While these tools help overcome some key barriers to implementing bias analysis, they are still predicated on the investigator having a firm understanding of causal reasoning. One needs appropriate causal assumptions in order to draw causal conclusions. If the investigator has data whose causal relationships are not correctly reflected in the bias models, then the estimates drawn from the bias analysis will be invalid. It is therefore recommended that anybody using these tools has read the contents of Chapter 3, which explains the derivation of simultaneous multi-bias analysis, and its papers cited within.

While the Shiny app has the significant benefit of not requiring any R programming skills, the "multibias" R package can allow for customization beyond the features of the Shiny app. The R package allows the user to work with any kind of data format that can be loaded into R. The user is also not restricted to having all bias parameters being a fixed value or Normal/Uniform probability distribution; each parameter can come from any kind of probability distribution or fixed value. Lastly, the R package user has the benefit of performing additional summaries and data visualization with the vector of bootstrap estimates.

Both tools can be developed in the future to allow for greater flexibility. The tools should be able to accommodate data represented by more complex DAGs, types of information bias beyond exposure misclassification, and data that is not restricted to binary variables and

logistic regression models. The Shiny app can be improved to allow for a detailed progress bar while the bootstrap estimates are being generated and the ability to download the results as a .csv or summary report. Future directions for the R package include expanding the warning and error messages and submitting the package to the CRAN, which "provides discoverability, ease of installation, and a stamp of authenticity [81]."

6. CONCLUSIONS

Any study that attempts to infer causation from observational data (i.e. a natural experiment) will have to account for potential sources of bias before a causal conclusion can be reached. Investigators can speculate on potential sources of bias that weren't controlled for in the design or analysis, but this isn't helpful for understanding how the estimate would change under the control of bias. In the epidemiological literature, it is becoming more commonplace to see the quantitative adjustment of a single bias source by bringing in causal assumptions related to the bias. This effort represents a significant step forward in improving the validity and interpretability of causal estimates. Building on this progress, the next step forward involves quantitative bias adjustment of every suspected source of bias. The studies in this dissertation help to achieve this goal.

Study 1 demonstrates how to dissect the biases influencing a cause-effect relationship, in this case, the PD-cancer relationship. This process involves reading the literature to understand the consensus on the strength of the effect, the hypothesized sources of bias, and which biases are generally controlled for or not controlled for. Directed acyclic graphs are an essential tool here for illustrating the causal model that incorporates each of these biases. This overall process, called a causal review, is necessary to inform multiple bias modeling. Study 1 discusses many potential biases of the PD-cancer relationship including uncontrolled confounding by smoking, PD misclassification due to inaccurate diagnoses from clinical examinations, and selection bias due to PD and cancer each affecting whether a subject will participate and stay in the study.

Study 2 introduces a new approach to multiple bias modeling. This method, called simultaneous multi-bias adjustment, involves the use of imputation and/or regression weighting to create an unbiased representation of the data. Analytical tools to assist other investigators in

performing this method are provided in Study 4. This bias adjustment technique is more generalizable than the current method of sequentially adjusting for each bias and does not require the investigator to know whether selection biases preceded any variable misclassification. The trade-off for using simultaneous multi-bias adjustment is that several less-intuitive bias parameters may be required. As was seen in the exposure misclassification adjustment in the PASIDA analysis of Study 3, some complex reasoning and data simulation may be required to derive the necessary bias parameters.

A retrospective cohort study within a Danish population was used to quantify the effect of PD on cancer in Study 3. The analysis incorporated the simultaneous multi-bias analysis method developed in the previous study. This study does not purport to have reached a decisive conclusion on the strength of the effect of PD on cancer nor to have explained why PD causes cancer. Both answers will depend on future research into the underlying biology of these diseases, including the role of genetics and the microbiome, and additional epidemiological investigation. The results of Study 3, however, provide evidence that Parkinson's disease and overall cancer may be independent of each other, in contrast to what most epidemiological studies suggest. It also shows that the relationship could be impacted by exposure misclassification and selection bias in equal and opposite directions under plausible bias models and parameters. It is suggested that future studies narrow the focus to specific cancers instead of overall cancer because different cancer types may have unique biological mechanisms, may require distinct adjustment sets for confounding, and may be impacted by censoring-related selection bias in unique ways depending on the time to diagnosis relative to other cancers.

This dissertation demonstrates the challenges associated with making causal inference tools generalizable across studies. The multi-bias adjustment in Study 3 was not able to be

performed using the bias adjustment tools created in Study 4 because these tools do not accommodate nested selection bias models or bias models with more than three confounders. Additional work is needed to make an all-encompassing tool for bias analysis that has no restrictions on the number of variables or biases and that supports several bias adjustment techniques. The ability of bias analysis to scale to wide-spread adaptation will depend on the quality of the tools available to perform the analysis.

REFERENCES

[1] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *J. Neurol. Neurosurg. Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[2] T. Pringsheim, N. Jette, A. Frolkis, and T. D. L. Steeves, "The prevalence of Parkinson's disease: A systematic review and meta-analysis," *Mov. Disord.*, vol. 29, no. 13, pp. 1583–1590, 2014.

[3] S. L. Kowal, T. M. Dall, R. Chakrabarti, M. V. Storm, and A. Jain, "The current and projected economic burden of Parkinson's disease in the United States," *Mov. Disord.*, vol. 28, no. 3, pp. 311–318, 2013.

[4] L. V. Kalia and A. E. Lang, "Parkinson's disease," *Lancet*, vol. 386, no. 9996, pp. 896–912, 2015.

[5] B. S. Connolly and A. E. Lang, "Pharmacological Treatment of Parkinson Disease," *Jama*, vol. 311, no. 16, p. 1670, 2014.

[6] L. de Lau and M. Breteler, "Epidemiology of Parkinson's disease," *Lancet*, vol. 5, pp. 525–535, 2006.

[7] B. Ritz, P. C. Lee, C. F. Lassen, and O. A. Arah, "Parkinson disease and smoking revisited: Ease of quitting is an early sign of the disease," *Neurology*, vol. 83, no. 16, pp. 1396–1402, 2014.

[8] G. F. Wooten, "Are men at greater risk for Parkinson's disease than women?," *J. Neurol. Neurosurg. Psychiatry*, vol. 75, no. 4, pp. 637–639, 2004.

[9] N. Greene, C. F. Lassen, K. Rugbjerg, and B. Ritz, "Reproductive factors and Parkinson's disease risk in Danish women," *Eur. J. Neurol.*, vol. 21, no. 9, pp. 1–11, 2014.

[10] O. Corti, S. Lesage, and A. Brice, "What Genetics Tells us About the Causes and Mechanisms of Parkinson's Disease," *Physiol. Rev.*, vol. 91, no. 4, pp. 1161–1218, 2011.

[11] J. Langston, P. Ballard, J. Tetrud, and I. Irwin, "Chronic Parkinsonism in humans due to a product of meperidine-analog synthesis," *Science (80-.).*, vol. 219, pp. 979–980, 1983.

[12] H. Ahmed, A. I. Abushouk, M. Gabr, A. Negida, and M. M. Abdel-Daim, "Parkinson's disease and pesticides: A meta-analysis of disease connection and genetic alterations," *Biomed. Pharmacother.*, vol. 90, pp. 638–649, 2017.

[13] M. A. Hernán, B. Takkouche, F. Caamaño-Isorna, and J. J. Gestal-Otero, "A meta-analysis of coffee drinking, cigarette smoking, and the risk of Parkinson's disease," *Ann. Neurol.*, vol. 52, no. 3, pp. 276–284, 2002.

[14] I. F. Shih, C. Starhof, C. F. Lassen, J. Hansen, Z. Liew, and B. Ritz, “Occupational and recreational physical activity and Parkinson’s disease in Denmark,” *Scand. J. Work. Environ. Heal.*, vol. 43, no. 3, pp. 210–216, 2017.

[15] R. Tabarés-Seisdedos and J. L. Rubenstein, “Inverse cancer comorbidity: A serendipitous opportunity to gain insight into CNS disorders,” *Nat. Rev. Neurosci.*, vol. 14, no. 4, pp. 293–304, 2013.

[16] L. Doshay, “Problem situations in the treatment of paralysis agitans,” *JAMA*, vol. 156, pp. 680–684, 1954.

[17] B. Jansson and J. Jankovic, “Low cancer rates among patients with Parkinson’s disease.,” *Ann. Neurol.*, vol. 17, no. 5, pp. 505–9, 1985.

[18] A. Bajaj, J. A. Driver, and E. S. Schernhammer, “Parkinson’s disease and cancer risk: a systematic review and meta-analysis,” *Cancer Causes Control*, vol. 21, no. 5, pp. 697–707, 2010.

[19] F. Catalá-López *et al.*, “Inverse and direct cancer comorbidity in people with central nervous system disorders: A meta-analysis of cancer incidence in 577,013 participants of 50 observational studies,” *Psychother. Psychosom.*, vol. 83, no. 2, pp. 89–105, 2014.

[20] R. Liu, X. Gao, Y. Lu, and H. Chen, “Meta-analysis of the relationship between Parkinson disease and melanoma,” *Neurology*, vol. 76, no. 23, pp. 2002–2009, 2011.

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

[22] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins, “A Structural Approach to Selection Bias,” *Epidemiology*, vol. 15, no. 5, pp. 615–625, 2004.

[23] M. A. Hernán and S. R. Cole, “Invited commentary: Causal diagrams and measurement bias,” *Am. J. Epidemiol.*, vol. 170, no. 8, pp. 959–962, 2009.

[24] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. New York, NY: Cambridge University Press, 2009.

[25] J. Pearl, “Causal Diagrams for Empirical Research,” *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.

[26] S. Greenland, J. Pearl, and J. M. Robins, “Causal Diagrams for Epidemiologic Research,” *Epidemiology*, vol. 10, no. 1. Epidemiology, pp. 37–48, 1999.

[27] T. J. Vander Weele and I. Shpitser, “On the definition of a confounder,” *Ann. Stat.*, vol. 41, no. 1, pp. 196–220, 2013.

[28] O. A. Arah, “Analyzing Selection Bias for Credible Causal Inference: When in Doubt, DAG It Out,” *Epidemiology*, vol. 30, no. 4, 2019.

[29] S. Greenland, “Multiple-bias modelling for analysis of observational data,” *J. R. Stat. Soc.*, vol. 168, pp. 267–306, 2005.

[30] S. Greenland, “Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment,” *Risk Anal.*, vol. 21, no. 4, pp. 579–83, 2001.

[31] C. A. Thompson and O. A. Arah, “Selection bias modeling using observed data augmented with imputed record-level probabilities,” *Ann. Epidemiol.*, vol. 24, no. 10, pp. 747–753, 2014.

[32] O. A. Arah, “Bias Analysis for Uncontrolled Confounding in the Health Sciences,” *Annu. Rev. Public Health*, vol. 38, no. 1, 2017.

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

[34] T. L. Lash, M. P. Fox, R. F. Maclehose, G. Maldonado, L. C. Mccandless, and S. Greenland, “Good practices for quantitative bias analysis,” *Int. J. Epidemiol.*, vol. 43, no. 6, pp. 1969–1985, 2014.

[35] M. P. Fox and T. L. Lash, “On the Need for Quantitative Bias Analysis in the Peer-Review Process,” *Am. J. Epidemiol.*, vol. 185, no. 10, pp. 865–868, 2017.

[36] T. J. Van Der Weele and P. Ding, “Sensitivity analysis in observational research: Introducing the E-Value,” *Ann. Intern. Med.*, vol. 167, no. 4, pp. 268–274, 2017.

[37] M. B. Mathur, P. Ding, C. A. Riddell, and T. J. VanderWeele, “Web Site and R Package for Computing E-values,” *Epidemiology*, vol. 29, no. 5, pp. e45–e47, 2018.

[38] J. H. Olsen, S. Friis, and K. Frederiksen, “Malignant melanoma and other types of cancer preceding Parkinson disease.,” *Epidemiology*, vol. 17, no. 5, pp. 582–587, 2006.

[39] A. Elbaz *et al.*, “Nonfatal Cancer Preceding Parkinson’s Disease: A Case-Control Study,” *Epidemiology*, vol. 13, no. 2, pp. 157–164, 2002.

[40] J. A. Driver, T. Kurth, J. E. Buring, J. M. Gaziano, and G. Logroscino, “Prospective case-control study of nonfatal cancer preceding the diagnosis of Parkinson’s disease,” *Cancer Causes Control*, vol. 18, no. 7, pp. 705–711, 2007.

[41] M. D’Amelio *et al.*, “Tumor diagnosis preceding Parkinson’s disease: A case-control study,” *Mov. Disord.*, vol. 19, no. 7, pp. 807–811, 2004.

[42] X. Cui, Z. Liew, J. Hansen, P. C. Lee, O. A. Arah, and B. Ritz, “Cancers Preceding

Parkinson's Disease after Adjustment for Bias in a Danish Population-Based Case-Control Study," *Neuroepidemiology*, pp. 136–143, 2019.

[43] D. M. Freedman *et al.*, "Associations between cancer and Parkinson's disease in U.S. elderly adults," *Int. J. Epidemiol.*, vol. 0, no. 0, pp. 1–11, 2016.

[44] A. F. Fois, C. J. Wotton, D. Yeates, M. R. Turner, and M. J. Goldacre, "Cancer in patients with motor neuron disease, multiple sclerosis and Parkinson's disease: record linkage studies.," *J. Neurol. Neurosurg. Psychiatry*, vol. 81, no. 2, pp. 215–221, 2010.

[45] P.-Y. Lin, S.-N. Chang, T.-H. Hsiao, B.-T. Huang, C.-H. Lin, and P.-C. Yang, "Association Between Parkinson Disease and Risk of Cancer in Taiwan," *JAMA Oncol.*, vol. 1, no. 5, p. 633, 2015.

[46] R. Zanetti, D. Loria, and S. Rosso, "Melanoma, Parkinson's disease and levodopa: causal or spurious link? A review of the literature," *Melanoma Res.*, vol. 16, no. 3, pp. 201–206, 2006.

[47] J. D. Vermeij, A. Winogrodzka, J. Trip, and W. E. J. Weber, "Parkinson's disease, levodopa-use and the risk of melanoma," *Park. Relat. Disord.*, vol. 15, no. 8, pp. 551–553, 2009.

[48] J. J. Ferreira *et al.*, "Skin cancer and Parkinson's disease," *Mov. Disord.*, vol. 25, no. 2, pp. 139–148, 2010.

[49] J. A. Driver, G. Logroscino, J. E. Buring, J. M. Gaziano, and T. Kurth, "A Prospective Cohort Study of Cancer Incidence Following the Diagnosis of Parkinson's Disease," *Cancer Epidemiol. Biomarkers Prev.*, vol. 16, no. 6, pp. 1260–1265, 2007.

[50] G. Rizzo, D. Martino, S. Arcuti, M. Copetti, A. Fontana, and G. Logroscino, "Accuracy of clinical diagnosis of Parkinson's disease: A systematic review and Bayesian meta-analysis," *Mov. Disord.*, vol. 30, p. S441, 2015.

[51] T. K. Gandhi *et al.*, "Missed and delayed diagnoses in the ambulatory setting: A study of closed malpractice claims," *Ann. Intern. Med.*, vol. 145, no. 7, pp. 488–496, 2006.

[52] S. S. Raab *et al.*, "Clinical impact and frequency of anatomic pathology errors in cancer diagnoses," *Cancer*, vol. 104, no. 10, pp. 2205–2213, 2005.

[53] S. Fielding, A. D. Macleod, and C. E. Counsell, "Medium-term prognosis of an incident cohort of parkinsonian patients compared to controls," *Park. Relat. Disord.*, vol. 32, pp. 36–41, 2016.

[54] C. Becker, G. P. Brobert, S. Johansson, S. S. Jick, and C. R. Meier, "Cancer risk in association with Parkinson disease: a population-based study.," *Parkinsonism Relat. Disord.*, vol. 16, no. 3, pp. 186–190, 2010.

[55] S. Mody, Lona; Miller, Douglas K., McGloin, Joanne, Freeman, Marcie; Marcantonio, Edward; Magaziner, Jay; Studenski, "Recruitment and Retention of Older Adults in Aging Research," *J. Am. Geriatr. Soc.*, vol. 56, no. 12, pp. 2340–2348, 2008.

[56] B. Boursi, R. Mamtani, K. Haynes, and Y. X. Yang, "Parkinson's disease and colorectal cancer risk-A nested case control study," *Cancer Epidemiol.*, vol. 43, pp. 9–14, 2016.

[57] J. M. Robins 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*, vol. 56, pp. 779–788, 2000.

[58] "10 Leading Causes of Death by Age Group, United States – 2014," *National Center for Injury Prevention and Control, CDC*. [Online]. Available: https://www.cdc.gov/injury/wisqars/pdf/leading_causes_of_death_by_age_group_2014-a.pdf.

[59] A. D. Macleod, K. S. M. Taylor, and C. E. Counsell, "Mortality in Parkinson's disease: A systematic review and meta-analysis," *Mov. Disord.*, vol. 29, no. 13, pp. 1615–1622, 2014.

[60] "SEER Cancer Statistics Review 1975-2013," *National Cancer Institute*. [Online]. Available: https://seer.cancer.gov/archive/csr/1975_2013/results_merged/topic_survival.pdf.

[61] Z. Ali, N. Yousaf, and J. Larkin, "Melanoma epidemiology, biology and prognosis," *Eur. J. Cancer, Suppl.*, vol. 11, no. 2, pp. 81–91, 2013.

[62] C. Dela Cruz, L. Tanoue, and R. Matthay, "Lung Cancer: Epidemiology, Etiology, and Prevention," *Clin Chest Med*, vol. 32, no. 4, p. 11, 2011.

[63] J. Driver, "Incidence and remaining lifetime risk of PD in advanced age," vol. 48, no. Suppl 2, pp. 1–6, 2010.

[64] A. Elbaz *et al.*, "Risk of cancer after the diagnosis of Parkinson's disease: A historical cohort study," *Mov. Disord.*, vol. 20, no. 6, pp. 719–725, 2005.

[65] M. Guttman, P. M. Slaughter, M. E. Theriault, D. P. DeBoer, and C. D. Naylor, "Parkinsonism in Ontario: Comorbidity associated with hospitalization in a large cohort," *Mov. Disord.*, vol. 19, no. 1, pp. 49–53, 2004.

[66] S. Greenland, "Basic Methods for Sensitivity Analysis of Biases," *Int. J. Epidemiol.*, vol. 25, no. 6, pp. 1107–1116, 1996.

[67] M. P. Fox, T. L. Lash, and S. Greenland, "A method to automate probabilistic sensitivity analyses of misclassified binary variables," *Int. J. Epidemiol.*, vol. 34, no. 6, pp. 1370–

1376, 2005.

[68] O. A. Arah, Y. Chiba, and S. Greenland, “Bias Formulas for External Adjustment and Sensitivity Analysis of Unmeasured Confounders,” *Ann. Epidemiol.*, vol. 18, no. 8, pp. 637–646, 2008.

[69] T. J. Vander Weele and O. A. Arah, “Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders,” *Epidemiology*, vol. 22, no. 1, pp. 42–52, 2011.

[70] S. Greenland and J. M. Robins, “Confounding and misclassification,” *Am. J. Epidemiol.*, vol. 122, no. 3, pp. 495–506, 1985.

[71] T. L. Lash and R. A. Silliman, “A sensitivity analysis to separate bias due to confounding from bias due to predicting misclassification by a variable that does both,” *Epidemiology*, vol. 11, no. 5, pp. 544–549, 2000.

[72] R. H. Lyles and J. Lin, “Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting,” *Stat. Med.*, vol. 29, no. 22, pp. 2297–2309, 2010.

[73] C. Y. Johnson, P. P. Howards, M. J. Strickland, D. K. Waller, and W. D. Flanders, “Multiple bias analysis using logistic regression: an example from the National Birth Defects Prevention Study,” *Ann. Epidemiol.*, vol. 28, no. 8, pp. 510–514, 2018.

[74] M. A. Hernan, “Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology,” *Am. J. Epidemiol.*, vol. 155, no. 2, pp. 176–184, 2002.

[75] S. Greenland, “Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias,” *Epidemiology*, vol. 14, pp. 300–306, 2003.

[76] J. M. Robins, M. Á. Hernán, and B. Brumback, “Marginal Structural Models and Causal Inference in Epidemiology,” *Epidemiology*, vol. 11, no. 5, pp. 550–560, 2000.

[77] L. Kenborg *et al.*, “Lifestyle, family history, and risk of idiopathic parkinson disease: A large danish case-control study,” *Am. J. Epidemiol.*, vol. 181, no. 10, pp. 808–816, 2015.

[78] H. Storm, E. Michelsen, I. Clemmensen, and J. Pihl, “The Danish cancer registry – history, content, quality and use,” *Dan Med Bull*, vol. 44, pp. 535–539, 1997.

[79] M. L. Gjerstorff, “The Danish cancer registry,” *Scand. J. Public Health*, vol. 39, no. 7, pp. 42–45, 2011.

[80] A. M. Jurek, S. Greenland, G. Maldonado, and T. R. Church, “Proper interpretation of non-differential misclassification effects: Expectations vs observations,” *Int. J.*

Epidemiol., vol. 34, no. 3, pp. 680–687, 2005.

[81] H. Wickham, *R Packages: Organize, Test, Document, and Share your Code*. Sebastopol, CA: O'Reilly Media, 2015.

[82] D. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text*, 3rd ed. Springer, 2012.