Bias Analyses in Population-Based Studies of Parkinson’s Disease

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

by

Xin Cui

2015
ABSTRACT OF THE DISSERTATION

Bias Analyses in Population-Based Studies of Parkinson’s Disease

by

Xin Cui

Doctor of Philosophy in Epidemiology

University of California, Los Angeles, 2015

Professor Beate Ritz, Chair

This dissertation work investigates two types of biases in population-based studies of Parkinson’s disease (PD), one is survivor bias in estimating the association between PD and cancer, the other is exposure misclassification due to residential mobility and time-varying exposure. Negative associations between PD and cancer have been found in epidemiological studies. Several mechanisms were proposed to explain such reported inverse associations. In the first and second part of this work we propose two similar survivor bias mechanisms that may account for the observed negative associations with cancer both prior to and after the diagnosis of PD as an alternative explanation. Using a large Danish population-based case-control study of
Parkinson’s disease as an example, we utilize causal theory, Monte Carlo methods and inverse probability-of-censoring weights techniques to quantitatively investigate how the observed negative association can be explained by the hypothesized bias. These results suggest that for cancer both before and after the diagnosis of PD, survivor bias could be an alternative explanation for the observed association between cancer and PD with reasonable bias structure and assumptions. In the last part of this work we investigate possible exposure misclassification due to residential mobility and changes in pesticide application using a California population-based case-control study of Parkinson’s disease as an illustration. We simulate scenarios where detailed residential histories were lacking and only the enrollment address was used as a proxy for all addresses to estimate long-term pesticide exposures. Results show that the exposures could be either over- or under- estimated depending on pesticide, time period of interest, as well as threshold for identifying the binary exposure variables. The exposure misclassification is not necessarily non-differential and the direction of the bias is inconsistent. When estimating long-term environmental exposures using one address only it may result in exposure misclassification and not always guarantee non-differentiality.
The dissertation of Xin Cui is approved.

Onyebuchi A. Arah
Marjan Javanbakht
Michael Jerrett
Beate Ritz, Committee Chair

University of California, Los Angeles

2015
Table of Contents

LIST OF TABLES ..................................................................................................................... vii
LIST OF FIGURES .................................................................................................................. viii
ACKNOWLEDGEMENTS ......................................................................................................... x
VITA ........................................................................................................................................ xi

1. INTRODUCTION ...................................................................................................................... 1
   1.1. PARKINSON’S DISEASE ................................................................................................. 1
   1.2. SURVIVOR BIAS IN ASSESSING THE ASSOCIATION BETWEEN PARKINSON’S DISEASE AND CANCER ......................................................................................... 1
   1.3. EXPOSURE MISCLASSIFICATION DUE TO RESIDENTIAL MOBILITY AND CHANGES IN PESTICIDE USE ...................................................................................... 2
   1.4. SPECIFIC AIMS OF THIS DISSERTATION ....................................................................... 3

2. STUDY 1 .................................................................................................................................... 5
   2.1. ABSTRACT ....................................................................................................................... 6
   2.2. INTRODUCTION .............................................................................................................. 6
   2.3. METHODS ..................................................................................................................... 9
      2.3.1. STUDY POPULATION .............................................................................................. 9
      2.3.2. DATA COLLECTION .............................................................................................. 10
      2.3.3. STANDARD PROTOCOL APPROVALS, REGISTRATIONS, AND PATIENT CONSENT ................................................................................................................ 11
      2.3.4. SURVIVOR BIAS STRUCTURE .............................................................................. 11
      2.3.5. SIMULATIONS FOR BIAS PARAMETERS .................................................................. 15
      2.3.6. STATISTICAL ANALYSIS ..................................................................................... 18
   2.4. RESULTS ....................................................................................................................... 19
   2.5. DISCUSSION ................................................................................................................. 23

3. STUDY 2 .................................................................................................................................. 30
   3.1. ABSTRACT ....................................................................................................................... 31
   3.2. INTRODUCTION .............................................................................................................. 32
   3.3. METHODS ..................................................................................................................... 34
      3.3.1. STUDY POPULATION .............................................................................................. 34
      3.3.2. DATA COLLECTION .............................................................................................. 34
      3.3.3. STANDARD PROTOCOL APPROVALS, REGISTRATIONS, AND PATIENT CONSENT ................................................................................................................ 35
      3.3.4. SURVIVOR BIAS STRUCTURE .............................................................................. 35
LIST OF TABLES

Table 2.1 Likelihood of selection by cancer and PD status based on causal assumptions. Page 16

Table 2.2 Demographic characteristics for survived individuals (S1=1) in the PASIDA Study. Page 19

Table 2.3 Adjusted odds ratios and relative odds ratios (RORs) for PD diagnosis-cancer before index date associations. Page 20

Table 2.4 Findings of prior studies investigating the association with cancer before diagnosis of PD. Page 26

Table 2.5 Cancer diagnosis before PD for unenrolled cases (by iPD verification status). Page 28

Table 3.1 Estimated associations between survival and each predictor. Page 40

Table 3.2 Demographic characteristics of the cohort. Page 43

Table 3.3 Incidence rate ratio (IR) for PD-cancer associations. Page 43

Table 3.4 Findings of prior studies investigating the association with cancer after diagnosis of PD. Page 48

Table 4.1 Demographic characteristics of study population. Page 62

Table 4.2 Correlation coefficient between the annual average pesticide exposures estimated using complete residential history and enrollment address only. Page 64

Table 4.3 Ratio of odds ratio (ROR) by comparing the odds ratios estimated from the two types of addresses. Page 71
LIST OF FIGURES

Figure 2.1 Basic survivor bias mechanism in studies of cancer before PD diagnosis. Page 12

Figure 2.2 Bias mechanism in the PASIDA Study (smoking-related cancers before PD diagnosis). Page 14

Figure 2.3 Bias mechanism in the PASIDA Study (non-smoking-related cancers before PD diagnosis). Page 15

Figure 2.4 Survivor bias and selection bias due to non-participation corrected ORs for smoking-related cancers. Page 22

Figure 2.5 Survivor bias and selection bias due to non-participation corrected ORs for non-smoking-related cancers. Page 23

Figure 3.1 Basic survivor bias mechanism in studies of cancer after PD diagnosis. Page 36

Figure 3.2 Bias mechanism in the PASIDA Study (smoking-related cancers after PD diagnosis). Page 38

Figure 3.3 Bias mechanism in the PASIDA Study (non-smoking-related cancers after PD diagnosis). Page 38

Figure 3.4 Percentage of death without developing cancer for smoking-related cancer and non-smoking-related cancer, respectively (by PD status). Page 41-42

Figure 4.1 Trends in pesticide use in the tri-County area, 1974-1999. Page 57

Figure 4.2 Annual average pesticide use in California (pound per acre), 1974-1999. Page 58

Figure 4.3 Annual average pesticide exposures estimated using actual address history and enrollment address for the 5-, 10-, 15-, 20-, and 26-year averaging periods. Page 63
Figure 4.4 Prevalence of exposures for the 5-, 10-, 15-, 20-, and 26-year averaging periods and 3 exposure definitions. Page 65-67

Figure 4.5 Sensitivity and specificity comparing exposures estimated from the enrollment address with complete residential history using 3 exposure definitions. Page 68-69

Figure 4.6 Odds ratio for the association between PD and pesticide exposures estimated for exposures using enrollment addresses or complete residential history and 3 exposure definitions. Page 70
ACKNOWLEDGEMENTS

It is my pleasure to acknowledge the roles of individuals who were instrumental for completion of my Ph.D. research. First of all, I would like to express my sincere appreciation to my dissertation committee chair and valued advisor, Dr. Beate Ritz, for her constant guidance and encouragement, without which this work would not have been possible. I couldn’t be luckier to have her as my advisor and mentor. I would like to give my special thanks to the valuable input of Dr. Onyebuchi Arah, who helped so much to shape this work. Dr. Arah made great efforts and trained me how to push myself forward. Special thanks go to Dr. Marjan Javanbakht and Dr. Michael Jerrett for their precious time and support. Without their help, I cannot move thus far. It was a great pleasure working with my supportive lab colleagues who helped me out whenever I needed them: Dr. Pei-Chen Lee, Dr. Zeyan Liew, Chenxiao Ling, Andrew Park, and Kimberly Paul. I would also like to thank Dr. Johnni Hansen at Danish Cancer Society who helped me with the Danish data, and Dr. Myles Cockburn at University of Southern California who helped me with the GIS-based exposure assessments. My deepest appreciation belongs to my family and friends for their unconditional love and support. Without them I would not be who I am today.
VITA

EDUCATION

2012 Master of Public Health in Epidemiology
University of California, Los Angeles, CA, USA

2010 Bachelor of Medicine in Preventive Medicine
Capital Medical University, Beijing, China

RESEARCH EXPERIENCE

2011 - present Graduate Student Researcher
Department of Epidemiology, Fielding School of Public Health,
University of California, Los Angeles, CA, USA

2006 - 2010 Research assistant
Department of Epidemiology and Health Statistics, School of Public
Health and Family Medicine, Capital Medical University, Beijing, China

PUBLICATIONS


PRESENTATIONS

Cui X, Arah OA, Ritz R. “Using Probabilistic Bias Analysis to Assess the Sensitivity of a Case-control Study to Possible Survivor Bias and Selection Bias due to Non-Participation”. 47th Annual Society for Epidemiologic Research Meeting. Seattle, WA. June 2015. (Oral Presentation)

1. INTRODUCTION

1.1. PARKINSON’S DISEASE
Parkinson’s disease is the second most common neurodegenerative disease after Alzheimer disease and presents with a developing tremor, rigidity, bradykinesia and postural instability. It is a complex disorder involving not only motor impairment but also deficits in behavior, cognition, and daily function. The incidence of the disease is approximately 6.3 million worldwide and rises steeply with age. The prevalence of PD is approximately 1% in people 60 years and older in the US. The etiology of PD remains elusive but is widely considered to be multi-factorial, involving a combination of exposures to neurotoxins, genetic factors, and aging. Environmental factors such as pesticides that could cause dopaminergic cells death were identified as risks factors of PD. It also has been hypothesized that nicotine protects against neuronal toxic insults and reduces disease progression. Due to the probable role of dopamine in the pathogenesis of PD, a large number of epidemiological studies, especially case-control studies, have been conducted to study the possible effects of environmental factors on PD.

1.2. SURVIVOR BIAS IN ASSESSING THE ASSOCIATION BETWEEN PARKINSON’S DISEASE AND CANCER
A growing number of epidemiological studies examined the relationship between Parkinson’s disease and cancer, and suggested an inverse association between the two diseases. The decreased cancer risk was seen both before and after the diagnosis of PD, and the only cancer type for which there appeared to be an increased risk was melanoma. This observation is intriguing, since cancer and PD have been attributed to distinct and somewhat opposing
biological mechanisms with the former being characterized by uncontrolled cell proliferation, and the latter by inappropriate cell loss that occurs progressively. Epidemiologic observations that suggested inverse associations between cancer and PD have generated interest because they promise that we might be able to gain a better understanding of cancer and PD etiology, and raise hope for new leads in developing prevention or treatment strategies.

Research showed that cancer and PD share some genetic and biological pathways and several mechanisms were proposed to explain the inverse associations reported between the two diseases in epidemiologic studies. The most obvious explanation for lower cancer rates in PD patients, however, has been the high proportion of nonsmokers and early quitters among PD patients for all smoking-related cancers, yet this leaves unexplained why non-smoking-related cancers seem also under-represented in PD patients in most studies. An alternative non-causal explanation that may account for the observed negative association between cancer and PD is survivor bias, but to date no study investigated this important hypothesis quantitatively. Before committing large intellectual and financial resources to studies to identify a presumed biology for intriguing inverse associations between cancer and neurodegenerative diseases in the elderly, it is imperative to assess simpler explanations and conduct cheap but informative quantitative bias analyses.

1.3. EXPOSURE MISCLASSIFICATION DUE TO RESIDENTIAL MOBILITY AND CHANGES IN PESTICIDE USE

A large number of animal studies and epidemiological studies have suggested that exposure to pesticides is a risk factor for Parkinson’s disease. Due to the long latency period of PD, long-
term pesticide exposure assessment is required estimating the effect of pesticide exposures on the risk of PD. The geographic information system (GIS)-based technique has increasingly been used in epidemiological studies to assess ambient agricultural pesticide exposures for not only PD but other chronic health effects, such as cancer. Since GIS-based exposure assessment relies heavily on spatial information, the availability of address information for individuals is a crucial element in determining the accuracy of this exposure assessment method, especially when the exposures relevant for the health outcome are at relatively low levels and span years which is likely to be the case for chronic disease research.

Information on residential history may be limited in studies of elder populations with memory loss or studies based on registries or medical records. Therefore, the residential location at study entry or at time of diagnosis is often used as a proxy for the complete residential history assuming residential mobility is less likely to impact the estimation of the long-term pesticide exposures over the study period. However, few studies have investigated such assumption and potential pesticide exposure misclassification due to spatial and temporal variation in residential mobility.

1.4. SPECIFIC AIMS OF THIS DISSERTATION

Study 1: To quantify a hypothesized survivor bias structure and investigate in case-control studies how the observed inverse association with cancer prior to the diagnosis of PD can be explained by the bias using Monte Carlo simulation techniques and a large Danish population-based study of Parkinson’s disease as an example.
Study 2: To investigate a similar survivor bias structure and illustrate in cohort studies how the observed inverse association with cancer after the diagnosis of PD can be explained by the bias using inverse probability-of-censoring weights technique and the same Danish population-based study of Parkinson’s disease as an example.

Study 3: To investigate possible exposure misclassification of long-term ambient pesticide exposure due to residential mobility and changes in pesticide use using a population-based case-control study of Parkinson’s disease as an illustration.
2. STUDY 1

CAN SURVIVOR BIAS EXPLAIN THE INVERSE ASSOCIATION BETWEEN CANCER AND RISK OF PARKINSON’S DISEASE?
2.1. ABSTRACT

Previous research suggests that cancer survivors are at lower risk of developing Parkinson’s disease (PD) compared with the general population. The most obvious explanation for lower cancer rates in PD patients has been the high proportion of nonsmokers and early quitters among PD patients reducing the number of smoking-related cancers, yet this leaves unexplained why non-smoking-related cancers seemed also to be under-represented among PD patients. These results have given rise to speculations about pathogenetic mechanisms that may prevent cancer in PD and raised hope for new leads in developing prevention or treatment strategies. However, spending resources on large scale investigations into this phenomenon are not justifiable if plausible bias mechanisms such as survivor bias and selection bias due to non-participation in epidemiologic studies account for these consistently observed negative associations. Using directed acyclic graphs and Monte Carlo simulation techniques, we quantify the hypothesized bias and investigate how the observed negative association can be explained by bias using a large Danish population-based case-control study of Parkinson’s disease as an example. Simulations showed that negative associations between cancer and PD not attributable to smoking can be explained by strong differential non-survival of cancer patients who later would develop PD. Our results have more general implications for a more cautious interpretation of negative associations found between comorbidities in diseases of the elderly since survivor/participation bias rather than biology might be the most likely explanation.

2.2. INTRODUCTION

A growing body of research suggests an inverse relationship between the occurrence of cancer and age-related neurodegenerative diseases [1-3]. Cancer survivors are found to be at lower risk
of developing Parkinson’s disease (PD) compared with the general population with some exceptions, mainly for melanoma [4-8]. Cancer and age-related neurodegenerative diseases, such as PD, have been attributed to distinct and somewhat opposing biological mechanisms with the former being characterized by uncontrolled cell proliferation, and the latter by abnormal cell (neurons) loss that occurs progressively. Epidemiologic observations that suggested inverse associations between cancer and PD have generated interest because they promise that we might be able to gain a better understanding of cancer and PD etiology, and raise hope for new leads in developing prevention or treatment strategies [9, 10]. For example, since PD is a male dominant disease, early on it has been suggested that endogenous estrogen may protect against PD [11], but estrogens are also linked to increased risk of estrogen responsive cancers in women [12]. Also, inhibition of the ubiquitin-proteasome system has been suggested as a potential treatment for cancer [13], while a failure of the system has been linked to pathogenesis of PD [14]. A number of possible biological mechanisms have been invoked when genes that seemed to bridge both diseases were identified. At the beginning of the modern genomics era, polymorphisms of genes encoding for detoxifying enzymes were the first to be associated with increased risk of PD [15] and decreased cancer risk [16]. Most prominent such hypotheses gained traction when known PD genes (SNCA, PARK2, DJ-1, and LRRK2) were found to also encode proteins that can enhance cell growth or reduce cell death [10, 17]; however, these genes only contribute to rare familial forms of PD that explain less than 10% of all PD cases [18]. The most obvious explanation for lower cancer rates in PD patients, however, has been the high proportion of nonsmokers and early quitters among PD patients for all smoking-related cancers [19], yet this leaves unexplained why non-smoking-related cancers seem also under-represented in PD patients in most studies [5, 7, 20-22]. An alternative non-causal explanation that may account for the
observed negative association between cancer and PD is survivor bias, but to date no study investigated this important hypothesis quantitatively. Before committing large intellectual and financial resources to studies to identify a presumed biology for intriguing inverse associations between cancer and neurodegenerative diseases in the elderly, it is imperative to assess simpler explanations and conduct cheap but informative quantitative bias analyses.

Survivor bias is a special case of selection bias introduced when studies condition on survival of participants by study design [23]. The onset of PD is rare before age 50 and a large majority of patients are diagnosed at age 60 or older [24]. Cancers, on the other hand, are a major health burden already before age 60 [25, 26]. In 2010, in the US, cancer was the number one leading cause of death among those of ages 40 to 59 years [27]. When studying risk of age-related diseases with late onset such as PD and Alzheimer’s disease (AD) among cancer survivors only, a selection mechanism may operate such that common cancers that are fatal in middle age selectively remove individuals who would be at risk of developing PD (or AD) at an older age. Also, those affected by both non-fatal cancer and later PD (or AD) may decide against participation in research studies that necessarily recruit from an elderly population with diminished health status due to cancer treatment and multi-morbidity. In addition, smoking, a major factor suspected to contribute to the reduced PD rates in cancer, is strongly related to a number of diseases, not only certain cancers but other health outcomes such as cardiovascular disease. In Europe, smoking-attributable mortality remains a major public health concern in the older segments of the population [28]. It is possible that individuals who are at higher risk for developing neurodegenerative diseases with an insidious onset and a long latency may instead die from other competing causes at a higher rate before developing PD (or AD) if they smoke.
Thus, PD (or AD) can only be observed among individuals who survive long enough to be diagnosed. By conditioning on survival, a spurious negative association between cancer and PD may appear even when there is no causal relationship between the two diseases.

Here, for the first time, we quantify this hypothesized survivor/selection bias and investigate whether it is a possible explanation for the reported inverse association between many cancers and risk of PD. First, we illustrated the possible survival/selection bias mechanism in a population-based case-control study of Parkinson’s disease in Denmark (PASIDA). We employed probabilistic bias analysis [29] in addition to a technique that adjusts for selection bias based on observed data augmented with imputed selection probabilities for each selected individual [30]. Applying external bias parameters estimated from the literature and data from the Danish Central Population Registry, we generated individual-level selection probabilities for our study participants and internally adjusted effect estimates for survivor/selection bias. We explore and describe the direction and strength of associations necessary for the presumed bias parameters to induce a sufficiently large bias to explain the size of the inverse associations reported for cancers prior to PD.

2.3. METHODS

2.3.1. STUDY POPULATION

The PASIDA Study, a population-based case-control study conducted in the Danish population, was designed to identify environmental and genetic factors that contribute to PD. Details of the study have previously been described [31]. Patients were identified from the Danish National Hospital Register files between 1996 and 2009 at 10 major neurological departments with PD
diagnosis (ICD-8 code: 342; ICD-10 code: G20) assigned by a neurologist. Population controls were selected from the Danish Central Population Registry matched on birth year and sex, being alive and without a PD diagnosis by the time of case identification. Among 3,508 cases originally identified, 362 died between diagnosis recorded in the Hospital Register and the date of invitation to participate in the study. Complete medical record review was conducted for the cases alive to confirm idiopathic Parkinson’s disease (iPD) and 428 were found to be ineligible due to no iPD. Among 2,718 eligible cases, i.e. alive at time of contact in 2007-2009 for the interview-based case-control study, we were able to enroll 1,813 with all data necessary for analysis, while 905 could not be enrolled due to 1) being too ill to participate (n=114); 2) refusal or no contact possible (n=758); 3) no medical record found or invalid disease data (n=19); and 4) dementia or cardiovascular disease before PD (n=14). Of the 3,626 controls initially contacted, 1,887 were enrolled with complete information, 1,739 were not enrolled due to 1) being unavailable for interview (n=1,717), and 2) dementia or cardiovascular disease before the index date (n=22).

2.3.2. DATA COLLECTION

In PASIDA, information on demographics and lifelong cigarette smoking history was obtained from a structured telephone interview. Pack-year was calculated for cigarette smoking up to the index date. Subjects with speech or physical difficulties were allowed to respond by mail (approximately 17%). We estimated socioeconomic status (SES) based on self-reported job positions collected for tax purposes for all adult Danish citizens during the 1980-1990s. Degree of urbanization at the index date was assigned based on population density as recorded in the
Danish Central Population Registry. The Danish Cancer Registry is a population-based registry containing data on the incidence of cancer in Denmark since 1943[32, 33]. We used records from the Danish Cancer Registry based on ICD-8 and ICD-10 codes to identify cancer diagnoses. Since the Danish Central Population Registry and Danish Cancer Registry were available for all Danish citizens, we were able to extract demographics and cancer diagnoses for all individuals regardless of participation status.

2.3.3. STANDARD PROTOCOL APPROVALS, REGISTRATIONS, AND PATIENT CONSENT

We obtained written informed consent from all subjects, and the UCLA Institutional Review Board and Danish Data Protection Agency approved the study protocol.

2.3.4. SURVIVOR BIAS STRUCTURE

We use Directed Acyclic Graphs (DAGs) to present the structural relationships between variables [34, 35] and a basic survivor bias mechanism under investigation in this study is illustrated in Figure 2.1. For the purposes of this simulation study, we assume that there is no true causal relationship between incidence of cancer and PD, rather that since cancer is a major cause of death in middle-age it reduces survival prior to PD onset and more strongly so in those who later develop PD, i.e. individuals at risk of PD are more susceptible to succumb to cancer death. Likelihood of survival by cancer and PD status based on our assumptions is illustrated in Table 2.1. Thus, death from cancer acts as a competing risk for PD diagnosis since PD diagnoses
are only made in individuals who survive long enough to develop PD at an older age. This might be the case if the long latency period before PD diagnosis is a period of heightened sensitivity either for dying from cancer or for developing cancer since participation in a research study may depend on health status such that those diagnosed and treated for cancer are less likely to participate in subsequent PD research. Conditioning on survival then induces a non-causal path that relates cancer to PD through common causes of survival or participation. This bias structure is also known as a collider-stratification bias [36] or competing mortality risks bias [37].

![Diagram](image.png)

**Figure 2.1** Basic survivor bias mechanism in studies of cancer before PD diagnosis.
Table 2.1 Likelihood of selection by cancer and PD status based on causal assumptions

<table>
<thead>
<tr>
<th>Level of selection</th>
<th>Assumptions</th>
<th>Population type</th>
<th>Likelihood of being selected (S=1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relationship between cancer and selection</td>
<td>Relationship between PD and selection</td>
<td>PD cases with cancer</td>
</tr>
<tr>
<td>Survival (S₁)</td>
<td>Cancer decreases survival</td>
<td>PD and survival are negatively associated via common causes</td>
<td>Not likely</td>
</tr>
<tr>
<td>Participation (S₂)</td>
<td>Cancer decreases participation</td>
<td>PD increases participation</td>
<td>PD cases with cancer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Likely</td>
</tr>
</tbody>
</table>

Based on previously reported associations between cancer, PD, and survival/participation, we propose two scenarios for smoking-related (Figure 2.2) and non-smoking-related (Figure 2.3) cancers, respectively, in the PASIDA Study. In the first scenario, smoking (SM) causes smoking-related cancer (SRC), which decreases the likelihood of survival (S₁) and participation in the case-control study (S₂) – the latter due to diminished health status. Since we only contacted individuals who were alive during the study period to participate in the interview-based study, those not surviving from cancer (or any other disease) or not healthy enough to be interviewed were not enrolled. The PASIDA Study is a case-control study, thus PD status may also influence participation with cases more willing to be recruited for research. In addition, some lifestyle factors such as smoking may also impact survival and participation via various health outcomes other than cancers. Variable Z represents a set of known risk factors for PD that influence survival and participation through different comorbidities, specifically age (Z₁), sex (Z₂), SES (Z₃), and urbanization (Z₄). These risk factors also influence smoking habits and the occurrence of cancer. It has been well recognized that smoking and PD are inversely associated [19, 38, 39], however, the biologic mechanism remains unclear and we recently suggested that quitting is an early symptom of PD [31]. Thus, we use a double headed arrow to represent the association.
between smoking and PD. In addition, a set of unknown or unmeasured variables (U) is presented in the causal diagram. Survival and PD are negatively related through U, which represents factors that make a person less likely to survive but more likely to subsequently have developed PD had they survived, hypothetically these factors may include high levels of stress or behavioral changes due to insidious PD that decrease survival chances, such as lowering of physical activity or autonomous nervous system dysfunction in pre-clinical PD interacting with medical treatment for cancer.

In the second scenario (Figure 2.3) we focus on cancers which are not caused by smoking (NSRC). All other covariates and relationships remain the same as in scenario 1. In both scenario 1 and 2, bias occurs by conditioning on survival and/or on participation in the study creating a spurious association between cancer and PD.

Figure 2.2 Bias mechanism in the PASIDA Study (smoking-related cancers before PD diagnosis).
2.3.5. SIMULATIONS FOR BIAS PARAMETERS

Survival was defined as a binary variable $S_1$ ($1 = \text{survived}; 0 = \text{dead}$). We used Monte Carlo techniques to generate the conditional probability of survival using the logistic equation as:

$$\logit P(S_1 = 1|\text{cancer}, \text{PD}, \text{SM}, Z_1, Z_2, Z_3, Z_4) = \beta_{S1} + \beta_{S1\text{cancer}}*\text{cancer} + \beta_{S1\text{PD}}*\text{PD} + \beta_{S1\text{SM}}*\text{SM} + \beta_{S1Z1}*Z_1 + \beta_{S1Z2}*Z_2 + \beta_{S1Z3}*Z_3 + \beta_{S1Z4}*Z_4$$

(1)
The $\beta$s represent the associations between survival and each of the predictors. For example, $\beta_{S_1}$ is the log odds of $S_1 = 1$ when all the predictors are set at the reference level; $\beta_{S_1\text{cancer}}$ is the log odds ratio (OR) relating $S_1$ and cancer when other predictors are set at the reference level. Using equation (1) we generate the selection probability for each survivor ($S_1=1$) conditioning on their cancer, PD, smoking, age, sex, SES, and urbanization values and the assumed bias parameters ($\beta$s). We refer to smoking-related cancers as SRC in scenario 1 and to non-smoking-related cancers as NSRC in scenario 2. We assume no interactions between all predictors. We specified distributions for each bias parameter based on existing knowledge and assumptions about relationships between variables as depicted in the causal structure in Figure 2.2 and 2.3.

For $\beta_{S_1}$ we assumed 0.9 as the ‘baseline prevalence’ of survival, i.e. among those who have highest survival likelihood (non-smoking cancer-free younger females living in the urban areas with high SES who were at low risk of developing PD) the probability of surviving long enough to develop PD is 0.9. To assess the impact of cancer on survival ($OR_{S_1\text{cancer}}$), we assumed the odds of survival among cancer patients to be between 0.3 to 1.0 compared with individuals without cancer, i.e. reflecting a 70 to 0% reduction in odds of survival based on reported cancer mortality rates by cancer site and stage for aging populations in European countries [40, 41]. Based on annual national mortality statistics in Denmark [42], we assumed that survival is also affected by smoking via adverse health outcomes other than smoking-related cancers ($OR_{S_1\text{SM}} = 0.9$, i.e. the odds of survival are 10% lower among smokers compared with non-smokers), as well as age ($OR_{S_1 Z_1} = 0.9$, i.e. older vs. younger), sex ($OR_{S_1 Z_2} = 0.9$, i.e. male vs. female), SES ($OR_{S_1 Z_3} = 0.9$, i.e. low vs. high), and urbanization ($OR_{S_1 Z_4} = 0.9$, i.e. rural vs. urban). For survival to PD via uncontrolled factors $U$ we assumed the $OR_{S_1\text{PD}}$ to range between 0.3 and 1.0.
based on our knowledge about possible candidates for \( U \), such as a 70 to 0\% reduction in odds of survival among individuals who later develop PD compared with those who do not.

Among those who survived (\( S_1 = 1 \)), to assess the impact of selection bias due to non-participation for PD cases and controls (cases being more motivated to participate than controls), participation was also defined as a binary variable \( S_2 \) (1 = enrolled; 0 = not enrolled). The conditional probability of participating in the interview-based case-control study was generated from the logistic equation as:

\[
\text{logit} \ P(S_2 = 1|\text{cancer, PD, SM, } Z_1, Z_2, Z_3, Z_4) = \beta_{S2} + \beta_{S2,cancer}*\text{cancer} + \beta_{S2,PD}*\text{PD} + \beta_{S2,SM}*\text{SM} + \\
\beta_{S2,Z1}*Z_1 + \beta_{S2,Z2}*Z_2 + \beta_{S2,Z3}*Z_3 + \beta_{S2,Z4}*Z_4
\]

(2)

Where the \( \beta \)s are similar to the ones in equation (1) but now represent the associations between participation (\( S_2 \)) and each of the predictors. Since demographic information is available for all individuals in Danish Central Population Registry regardless of participation status and cancer diagnosis information can be retrieved via linkage to the Danish Cancer Registry, we modeled \( S_2 \) as a function of the above listed predictors measured for both enrolled (\( S_2 = 1 \)) and unenrolled (\( S_2 = 0 \)) individuals using real data in our PASIDA case-control study. The associations between participation and each predictor were estimated for cancer (\( \text{OR}_{S2, \text{cancer}} = 0.92 \) for SRC and 0.88 for NSRC, i.e. yes vs. no), smoking (\( \text{OR}_{S2, \text{SM}} = 0.21 \), i.e. above vs. below median of controls (16.5 pack-years)), age (\( \text{OR}_{S2, Z1} = 0.54 \), i.e. above vs. below median of controls (65 years of age)), sex (\( \text{OR}_{S2, Z2} = 1.75 \), i.e. male vs. female), SES (\( \text{OR}_{S2, Z3} = 0.54 \), i.e. low vs. high), urbanization (\( \text{OR}_{S2, Z4} = 1.63 \), i.e. rural vs. urban), and PD status (\( \text{OR}_{S2, PD} = 1.47 \), i.e. yes vs. no).
For \( S_1 \) we generated normal distributions with standard deviations of 0.10 for each bias parameter. We used 95% confidence intervals (CIs) estimated from prediction models for \( S_2 \) to get the lower and upper limits for each bias parameter of \( S_2 \). Simulation of the bias parameters was repeated for 1,000 iterations.

2.3.6. STATISTICAL ANALYSIS

All variables used in the analysis were coded as binary variables. While we lacked data on smoking from non-participants, assuming missing at random (MAR) we first used multiple imputation to impute pack-years for individuals who survived but did not enroll (n=2,644) in order to estimate bias parameters for \( S_2 \). We then estimated the observed OR between cancer and PD based on the interview-based PASIDA data adjusting for smoking (from interviews), age, sex, SES, and urbanization (from Danish registries). For each type of cancer, we calculated ORs among those who survived up to the date of interview between 2007 and 2009 (both \( S_2 = 1 \) and \( S_2 = 0 \)), those who were enrolled in the interview-based case-control study (\( S_2 = 1 \)), and those who survived but did not participate in our study (\( S_2 = 0 \)). We compared these ORs using ratios of odds ratio (ROR). We calculated ORs for cancer and PD among all survivors, and compared them to the ORs for individuals who were enrolled. Confidence intervals were constructed around the logarithm of the ratios of cumulative incidence and RORs using a non-parametric bootstrap method.
By using simulated priors, selection probabilities were generated for $S_1$ and $S_2$, respectively, and we estimated the OR adjusting for survivor/participation selection biases by using $[P(S_1 = 1)P(S_2 = 1)]/ [P(S_1 = 1|\text{cancer, PD, SM, Z}_1, Z_2, Z_3, Z_4)P(S_2 = 1|\text{cancer, PD, SM, Z}_1, Z_2, Z_3, Z_4)]$ as the regression weight to estimate the association between cancer and PD adjusting for smoking, age, sex, SES, and selections. For each scenario we varied the assumed effect sizes of the bias parameters. We report medians and 95% simulation intervals (95%SI) (the 2.5th and 97.5th percentile) for the estimated bias-adjusted ORs derived from these simulations. All simulation and analyses were performed by using SAS 9.3 (SAS Institute, Cary, NC, USA).

### 2.4. RESULTS

The average age at first PD diagnosis was 62.4 for enrolled PASIDA cases and 59% were male. Being younger, male, of higher SES, and living in urban areas increased the likelihood of participation in our interview-based case-control study (Table 2.2).

<p>| Table 2.2 Demographic characteristics for survived individuals ($S_1=1$) in the PASIDA Study |
|-------------------------------------------------|-------------------------------------------------|
| Enrolled ($S_1=1$) | Not enrolled ($S_1=0$) |</p>
<table>
<thead>
<tr>
<th>Cases</th>
<th>Controls</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>1813</td>
<td>100.0</td>
<td>1887</td>
</tr>
<tr>
<td>Age (years) *</td>
<td>Mean (SD)</td>
<td>62.4±9.3</td>
<td>62.6±9.4</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1070</td>
<td>59.0</td>
<td>1121</td>
</tr>
<tr>
<td>Female</td>
<td>743</td>
<td>41.0</td>
<td>766</td>
</tr>
<tr>
<td>SES *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Academic and top self-employed</td>
<td>260</td>
<td>14.3</td>
<td>221</td>
</tr>
<tr>
<td>High self-employed and high salaried</td>
<td>275</td>
<td>15.2</td>
<td>281</td>
</tr>
<tr>
<td>Low self-employed and mid-salaried</td>
<td>341</td>
<td>18.8</td>
<td>363</td>
</tr>
<tr>
<td>Skilled worker and low salaried</td>
<td>384</td>
<td>21.2</td>
<td>422</td>
</tr>
<tr>
<td>Unskilled worker</td>
<td>149</td>
<td>8.2</td>
<td>203</td>
</tr>
<tr>
<td>Unknown or unspecified</td>
<td>404</td>
<td>22.3</td>
<td>397</td>
</tr>
<tr>
<td>Degree of urbanization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capital</td>
<td>443</td>
<td>24.4</td>
<td>564</td>
</tr>
<tr>
<td>Provincial cities</td>
<td>1118</td>
<td>61.7</td>
<td>965</td>
</tr>
<tr>
<td>Rural areas</td>
<td>167</td>
<td>9.2</td>
<td>209</td>
</tr>
<tr>
<td>Peripheral regions</td>
<td>82</td>
<td>4.5</td>
<td>148</td>
</tr>
<tr>
<td>Unknown or unspecified</td>
<td>3</td>
<td>0.2</td>
<td>1</td>
</tr>
</tbody>
</table>
Age for cases was defined as the difference between the year of first primary/secondary hospital PD diagnosis from the National Hospital Register and the year of birth. Age for controls was defined as the age at first PD diagnosis for the case that the control was matched to.

Proxy of SES was calculated using self-reported job position information given for tax information by all adult citizens from the Central Population Register from the 1980-ies and the 1990-ies.

Among enrolled participants, we found that those with smoking-related cancers had a 17% decreased risk of developing PD compared with those who did not have any cancer diagnoses, but due to the small number of cancers our confidence intervals included the null value (95%CI: 0.52-1.35). However, we did not observe an association between non-smoking-related cancers and PD risk (OR = 0.99; 95%CI: 0.63-1.55). For specific cancer sites of interest, we observed an increased prevalence of malignant melanoma and non-melanoma skin cancers prior to diagnosis of PD, with OR of 1.61 (95%CI: 0.69-3.73) and 1.46 (95%CI: 1.01-2.10), respectively. No association was found for breast cancer (OR = 0.95; 95%CI: 0.54-1.68) (Table 2.3). For those who survived up to the date of interview between 2007 and 2009 but did not participate in our study ($S_2 = 0$), the estimated effect was similar to the one for enrolled participants only for non-smoking-related cancers (OR: 0.98; 95%CI: 0.59-1.63). The RORs were close to unity except for malignant melanoma (Table 2.3).

### Table 2.3 Adjusted odds ratios and relative odds ratios (RORs) for PD diagnosis-cancer before index date associations

<table>
<thead>
<tr>
<th>Cancer</th>
<th>All survived (n=6344)</th>
<th>$S_2=1$ (n=3700)</th>
<th>Ref</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cases/controls</td>
<td>OR_{adj} (95%CI)</td>
<td>cases/controls</td>
<td>OR_{adj} (95%CI)</td>
</tr>
<tr>
<td>No any cancer</td>
<td>2450/3294</td>
<td>Ref</td>
<td>1638/1731</td>
<td>Ref</td>
</tr>
<tr>
<td>Smoking-related (SRC)</td>
<td>54/91</td>
<td>0.88 (0.62-1.25)</td>
<td>31/41</td>
<td>0.83 (0.52-1.35)</td>
</tr>
<tr>
<td>Non-smoking-related (NSRC)</td>
<td>62/87</td>
<td>0.98 (0.70-1.36)</td>
<td>38/41</td>
<td>0.99 (0.63-1.55)</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>35/61</td>
<td>0.83 (0.54-1.27)</td>
<td>24/26</td>
<td>0.95 (0.54-1.68)</td>
</tr>
<tr>
<td>Malignant melanoma</td>
<td>26/15</td>
<td>2.36 (1.25-4.48)</td>
<td>14/9</td>
<td>1.61 (0.69-3.73)</td>
</tr>
<tr>
<td>Nonmelanoma skin cancer</td>
<td>106/107</td>
<td>1.34 (1.02-1.77)</td>
<td>73/52</td>
<td>1.46 (1.01-2.10)</td>
</tr>
</tbody>
</table>

* Exclude female breast, malignant melanoma, nonmelanoma skin, and unspecified and ill-defined cancer

* For smoking-related cancers logistic regression adjusted for age at first PD diagnosis (for cases) or age at index date (for controls), sex, urbanization, SES, and smoking; for other cancers logistic regression adjusted for same variables except smoking

* Relative odds ratios (RORs) compared ORs from all survived with ORs from enrolled
Figure 2.4 presents the ORs adjusted for the association between smoking-related cancers and PD corrected for both survivor bias and selection bias due to non-participation. When adjusting for selection bias due to non-participation only ($S_2$) and assuming there is no survivor bias at the first level (both $OR_{S1_{SRC}}$ and $OR_{S1_{PD}} = 1.0$, corresponding to the most left point on the bottom line), the bias-adjusted OR was 0.75 (95%CI: 0.43-1.34) compared with the observed OR of 0.83 (95%CI: 0.52-1.35) before bias correction. However, when both $S_1$ and $S_2$ existed, the bias-adjusted OR became less strongly negative but moved towards the null value of one as the likelihood of survival decreased with both smoking-related cancers ($OR_{S1_{SRC}}$) and insidious PD ($OR_{S1_{PD}}$). Specifically, when the likelihood of survival was very low for those with both smoking-related cancers and insidious PD (through $U$), i.e. $OR_{S1_{SRC}} = 0.3$ and $OR_{S1_{PD}} = 0.3$ (corresponding to the most right point on the top line), the estimated bias-adjusted OR was 1.01 (95%CI: 0.71-1.44) indicating that the two levels of bias induced by $S_1$ and $S_2$ had opposite directions and could balance each other out when survival is strongly related to both smoking-related cancer and insidious PD. Otherwise when $OR_{S1_{SRC}}$ and $OR_{S1_{PD}}$ shifted between 0.3 and 1.0 with different combinations, the bias-adjusted ORs ranged between 0.75 and 0.94, i.e. essentially always remained negative.
Figure 2.4 Survivor bias and selection bias due to non-participation corrected ORs for smoking-related cancers.

Figure 2.5 presents the bias-adjusted ORs for non-smoking-related cancers. Similarly the bias-adjusted OR became 0.94 (95% SI: 0.54-1.64) when we only adjusted for selection bias due to non-participation compared with the observed OR of 0.99 (95% CI: 0.63-1.55). When both $S_1$ and $S_2$ were corrected, the bias-adjust ORs shifted towards the positive side from the null when both bias parameters were specified as being strong, i.e. when both $OR_{S1_{NSRC}}$ and $OR_{S1_{PD}}$ were
specified as 0.3, the bias-adjusted OR became 1.24 (95%CI: 0.87-1.78). Otherwise the bias-adjusted OR ranged between 0.90 and 1.11; i.e. remained close to the estimated null effect.

**Figure 2.5** Survivor bias and selection bias due to non-participation corrected ORs for non-smoking-related cancers.

### 2.5. DISCUSSION

In this study we hypothesized that a bias mechanism exists such that cancer prior to PD decreased the likelihood of survival and participation of those who later would become diagnosed with PD; i.e. those who died early or did not enroll were at higher risk of developing
PD producing a negative bias that might explain the observed negative associations between cancer and PD because of conditioning on survival and participation in case-control studies among multi-morbid older populations. Since our PASIDA study was an interview-based case-control study that did not entirely restrict its population to incident cases, these two levels of selection happen sequentially, i.e. the decision to participation can be made only by those who are alive during the study period when approached for interviews. These types of studies are not uncommon for neurodegenerative diseases that are relatively rare such that it takes many years for sufficient cases to accrue; furthermore for PD in particular, disease diagnosis accuracy increases over time with PD diagnoses being much more accurate at least 5 years after diagnosis. Thus, waiting for disease to progress such that the symptomatology is more informative risks survivor bias to increase especially among older cases (70 years plus) or when assessing cancer not only prior to but also after the onset of PD. Thus, it is important to take both levels of selection into consideration. The bias parameters specified in our study were estimated based on the literature, mortality statistics in Denmark, as well as individual-level information from the Danish Central Population Registry and our PASIDA interview-based study (smoking), which led to reasonable estimates for the bias parameters in our modeling of the impact of survival and participation on effect estimates. Our simulations showed that the observed negative association between cancer and risk of PD can be partially explained by the proposed bias mechanism if survival is strongly affected by cancer and at the same time strongly associated with PD risk. Therefore, it is important to consider potential survivor bias when interpreting findings that relate comorbidities causally to each other in the elderly population.
In our PASIDA study, the estimated negative effect between smoking-related cancers and subsequent PD (OR = 0.83; 95%CI: 0.52-1.35) was imprecise due to our relatively small sample size; nevertheless, it was very consistent in size with previous reports (Table 2.4). In fact, the inverse association between smoking-related cancers and PD was strengthened when we adjusted for selection bias due to non-participation (OR: 0.75; 95%SI: 0.43-1.34). This non-participation corrected effect estimate is very comparable to a much larger previous Danish record linkage study of all citizens that identified PD cases from the Danish National Hospital Register between 1986 to 1998 and reported an OR = 0.68 (95%CI: 0.58-0.81) for smoking-related cancers prior to PD index date [8]. Since this study was solely based on record linkage, selection bias from non-participation cannot occur. What our case-control study contributes and this previous study could not accomplish was 1) an extensive medical record review in which we verified iPD diagnoses; and 2) collect smoking information in interviews, data that is not available in record linkage only studies. Thus, our estimates are less likely to be affected by outcome misclassification and (negative) confounding by smoking than the previous study, yet due to its case control design it seems to have been affected by participation bias. Since PD cases stop smoking many years prior to PD diagnosis, one might say that “PD prevents smoking” and thus reduces the risk of smoking-related cancers and mortality. Negative confounding by smoking has been proposed as an explanation for the observed inverse association between cancer and PD diagnosis [19]. However, in PASIDA adjustment for pack-years of smoking did not explain the negative association between smoking-related cancers and PD, yet our estimates shifted closer to the null when we adjusted for the presumed survivor bias, indicating that these negative associations may be driven more by survival than confounding bias.
In our study we did not observe any (negative or otherwise) association between non-smoking-related cancer and risk of PD (OR=0.99; 95%CI: 0.63-1.55) but the 95%CI of our result mostly overlapped with estimates of previous studies that reported negative associations. After adjusting for selection bias due to non-participation of those who survived to interview, the bias-adjusted OR was 0.94 (95%SI: 0.54-1.64). When both survivor bias and non-participation were taken into consideration, the doubly bias-adjusted ORs became less negative as the specified bias

Table 2.4 Findings of prior studies investigating the association with cancer before diagnosis of PD

<table>
<thead>
<tr>
<th>Study</th>
<th>Publication year</th>
<th>Study area</th>
<th>Case N</th>
<th>Control/cohort N</th>
<th>Matching variable</th>
<th>Follow-up year</th>
<th>Data collection</th>
<th>Relative risk measure (95% CI)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rajput et al</td>
<td>1987</td>
<td>US</td>
<td>118</td>
<td>236</td>
<td>age, sex</td>
<td>1967-1979</td>
<td>Rochester Epidemiology Project; PD: medical record review; cancer: medical record review</td>
<td>overall: 1.3 (0.7, 2.4)</td>
<td>Prior cancer more common in PD than controls</td>
</tr>
<tr>
<td>Elbaz</td>
<td>2002</td>
<td>US</td>
<td>196</td>
<td>196</td>
<td>age, sex</td>
<td>1976-1995</td>
<td>Rochester Epidemiology Project; PD: medical record review; cancer: medical record review</td>
<td>overall: 0.79 (0.49, 1.27); smoking-related: 0.75 (0.26, 2.16); melanoma: 1.5 (0.23, 8.9)</td>
<td>Prior cancer less common in PD than controls, except melanoma</td>
</tr>
<tr>
<td>D'Amelio et al</td>
<td>2004</td>
<td>Italy</td>
<td>222</td>
<td>222</td>
<td>age, sex</td>
<td>2001-2002</td>
<td>iPD: medical record review+neurologists; cancer: self-report; medical record review</td>
<td>overall: 0.4 (0.2, 0.7); female: 0.3 (0.1, 0.7); malignant: 0.6 (0.1, 2.5); non-malignant: 0.3 (0.1, 0.7); endocrine-related: 0.3 (0.1, 0.9); non-endocrine-related: 0.4 (0.1, 0.9)</td>
<td>Prior cancer less common in PD than controls</td>
</tr>
<tr>
<td>Olsen et al</td>
<td>2006</td>
<td>Denmark</td>
<td>8090</td>
<td>32,320</td>
<td>birth year, sex</td>
<td>1986-1998</td>
<td>National Hospital Register; PD: ICD-code; cancer: Danish Cancer Registry</td>
<td>overall: 1.04 (0.96, 1.10); smoking-related: 0.68 (0.58, 0.81); melanoma: 1.4 (1.03, 2.0)</td>
<td>Overall no association; prior smoking-related cancer less common in PD than controls; melanoma prevalence increased in PD</td>
</tr>
<tr>
<td>Driver et al</td>
<td>2007</td>
<td>US</td>
<td>487</td>
<td>487</td>
<td>age, sex (male only)</td>
<td>22 yrs</td>
<td>Physician's Health Study; PD: self-report; medical record review; cancer: medical record review-report</td>
<td>overall: 0.83 (0.57, 1.21); smoking-related: 0.74 (0.35, 1.57); non-smoking-related: 0.88 (0.59, 1.32)</td>
<td>Prior cancer less common in PD than controls</td>
</tr>
<tr>
<td>Fois et al</td>
<td>2009</td>
<td>UK</td>
<td>4,355</td>
<td>26,064</td>
<td>age, sex, district of treatment, year of first recorded linkage</td>
<td>1963-1999</td>
<td>Oxford Record Linkage Study; PD: ICD-code; cancer: ICD-code</td>
<td>overall: 0.76 (0.70, 0.82); lung: 0.5 (0.4, 0.7); bladder: 0.7 (0.6, 0.9); malignant melanoma: 0.5 (0.2, 0.9)</td>
<td>Prior cancer less common in PD than reference group</td>
</tr>
<tr>
<td>Lo et al</td>
<td>2010</td>
<td>US</td>
<td>692</td>
<td>761</td>
<td>birth year; sex, respondent type</td>
<td>1994-1995; 2000-2003</td>
<td>PEAK studies; PD: medical record review; cancer: SEER recode</td>
<td>overall: 0.83 (0.54, 1.30); smoking-related: 0.57 (0.23, 1.4); non-smoking-related: 0.86 (0.54, 1.4)</td>
<td>No association</td>
</tr>
</tbody>
</table>

26
parameters were set to be stronger, and some doubly bias-adjusted ORs were even greater than null, which suggests that as the likelihood of survival decreases for non-smoking-related cancer patients with insidiously developing PD survivor bias becomes more and more negative. The magnitude of this negative bias can become a stronger force than the positive selection bias due to non-participation, leaving a negative net bias. Therefore, when correcting for both biases for non-smoking-related cancer in our data we observe an increased risk of PD, in data that starts with a negative association between cancer and PD, the association then would become null.

A major strength of our study is that individual-level information such as demographics and cancer diagnosis was available to be extracted from national registries regardless of participation status, which made it possible for us to internally adjust for the second level of selection by modeling participation as a function of cancer and measured covariates among both enrolled and un-enrolled. This makes our simulation study solidly based in real world data and distinct from other studies. Other external resources, e.g., Danish mortality statistics, made it possible for us to reasonably estimate the direction and magnitude of bias parameters and externally adjust for the first level of selection, i.e. survivor bias. In addition, we reviewed medical records and verified iPD diagnosis to limit possible misclassification of the outcome (Table 2.5). The sample size in our study was larger than for all previous interview-based case-control studies but smaller than previous linkage based studies that lacked covariates data such as for smoking and did not confirm PD diagnoses.
Table 2.5  Cancer diagnosis before PD for unenrolled cases by iPD verification status

<table>
<thead>
<tr>
<th>Cancer</th>
<th>Reviewed and verified as iPD (n=233)</th>
<th>Unable to be reviewed (n=672)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td>No any cancer</td>
<td>212</td>
<td>91.0</td>
</tr>
<tr>
<td>Smoking-related (SRC)</td>
<td>6</td>
<td>2.6</td>
</tr>
<tr>
<td>Non-smoking-related (NSRC) a</td>
<td>6</td>
<td>2.6</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>Malignant melanoma</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Non-melanoma skin cancer</td>
<td>8</td>
<td>3.4</td>
</tr>
</tbody>
</table>

a Exclude female breast, malignant melanoma, non-melanoma skin, and unspecified and ill-defined cancer.

Our simulation study has some limitations. Since our bias modeling was based on the proposed causal structure and specified bias parameters, factors that influence these assumptions may affect the bias-adjusted ORs. For example, we did not have smoking information for non-participating individuals. Thus, in order to use smoking in the prediction model of $S_2$ we had to employ multiple imputation to generate pack-years for non-participants assuming missing at random. If this assumption was not met it would have produced inaccurate estimates for parameters predicting the enrollment into the case-control study ($S_2 = 1$). Also, the association between survival and PD through $U$ is unknown; thus, we tested a range of possible values based on external resources. In addition, all variables in our analyses were binary for simplicity, which may have caused residual confounding. We assumed no interaction between the predictors in our models. If the effects on selection were modified by predictors, the selection probabilities may not be calculated accurately. Moreover, we did not take other possible biases into consideration such as additional uncontrolled confounding and measurement errors.
In this study, we propose a bias structure and illustrate that the inverse association between cancer and PD is attenuated when adjusting for both survivor bias and selection bias due to non-participation using a quantitative bias modeling approach. When survival was assumed to be strongly associated with both cancer and PD, the two biases balanced each other out but a negative association for smoking-related cancers remained; for non-smoking-related cancers the survivor bias was stronger than the selection bias due to non-participation such that bias adjusted estimates eventually became greater than the null value in our data. Since the selection mechanisms proposed in this study are common in many other age-related disease studies that have to rely on studying survivors only, careful specification of reasonable bias parameters based on prior knowledge allow estimation and correction for survivor bias and provide an alternate explanation for unexpected yet intriguing negative associations between diseases that contribute to multi-morbidity in the elderly population.
3. STUDY 2

CAN SURVIVOR BIAS EXPLAIN THE INVERSE ASSOCIATION BETWEEN PARKINSON’S DISEASE AND RISK OF CANCER?
3.1. ABSTRACT

A majority of studies that examined the relationship between Parkinson’s disease (PD) and subsequent occurrence of cancer suggested that PD patients are less likely to develop cancer. We previously proposed a survivor bias structure and illustrated that in case-control studies the inverse association with cancer prior to the diagnosis of PD is attenuated when adjusting for both survivor bias and selection bias due to non-participation when survival was assumed to be strongly associated with both cancer and PD. To quantitatively explore whether hypothesized survivor bias mechanism may also be a possible explanation for observed inverse association between PD and subsequent cancer risk, we propose a similar bias structure in cohort studies and investigate whether the observed negative association can be explained by this bias and be removed using the inverse probability-of-censoring weights technique using the information available to us from the Danish National Hospital and Cancer registry systems. Results suggested that - contrary to what has been reported previously - association with cancer occurring after a diagnosis of PD are likely underestimated, i.e. shifted towards the null or even beyond when one does not adjust for survivor bias using the inverse probability-of-censoring weights. These results suggest that for cancer both before and after the diagnosis of PD, survivor bias could be an alternative explanation for the observed association between cancer and PD with a reasonable bias structure and assumptions we were able to anchor in the Danish medical and civil registry data system. Since the bias mechanisms we described are likely common in cohort and case-control studies associating diseases with each other in elderly populations with a strong competing risk of death, survivor bias should be taken into consideration and quantitatively addressed before claiming that associations between comorbidities have biologic underpinnings.
3.2.  INTRODUCTION

A majority of studies that examined the relationship between Parkinson’s disease (PD) and subsequent occurrence of cancer suggested that PD patients are less likely to develop cancer [1, 7, 20-21, 43-44]. Research showed that cancer and PD share some genetic and biological pathways [9, 45-46] and several mechanisms were proposed to explain these observations from epidemiologic studies. However, non-causal explanations may also account for these negative associations and should be further explored quantitatively. For example, we previously proposed a bias structure to explain negative associations between cancer and subsequent PD among cancer survivors that may depend on selective survival and participation in case-control studies of elderly subjects. Cancer and PD are comorbidities with an overlapping age distribution and previous studies have investigated the cancer occurrence both before [5-8, 47-48] and after [20-22, 43, 48-51] a diagnosis of PD. Here, we will propose and quantitatively assess another bias mechanism that may explain the inverse association between PD and subsequent cancer risk.

In follow-up studies, competing risk of disease and/or death can remove subject at risk of developing the outcome of interest. Under certain conditions, removing subjects at risk from the population may introduce selection bias [35]. PD itself is not fatal; however, related complications can reduce life expectancy of patients. For example, balance problems in PD patients may result in falls, a major factor influencing the quality of life and mortality in patients with PD [52-54]. When studying the risk of a subsequent cancer diagnosis among PD patients, a censoring mechanism may exist such that PD related complications or other comorbidities selectively increase mortality in PD patients at risk of developing cancers. Since cancer can only be observed among individuals who survive long enough to be diagnosed, conditioning on
survival may generate an apparent negative association between PD and cancer even when there is no true causal relationship between the two diseases. Lifestyle factors have been suggested to be related to a number of diseases and mortality in PD patients will depend on such factors. For example, smoking is an important risk factor for many cancers as well as other fatal diseases, e.g., myocardial infarction and stroke, but negatively associated with PD, i.e. PD patients generally smoke less. However, survival in PD patients who have been smokers may even be shorter than among other smokers. There are also factors that protect against chronic diseases, and mortality in PD patients may be higher due to the lack or reduction of such protective factors. For example, physical activity has been suggested to be negatively associated with PD and cancer as well as other diseases such as cardiovascular disease and diabetes [55-60], and PD patients may slowly but steadily reduce physical activity even during the long prodromal phase of PD putting them at higher risk of developing other diseases that increase their mortality risk [61-62]. In addition, dietary habits have been suggested to be related to the etiology of many diseases. A high level of the antioxidant vitamins is associated with a decreased risk of PD [63] and cancer, and reduces the likelihood of death from other fatal diseases [64-68]. Adjustment for such lifestyle factors, therefore, is important when estimating the relationship between PD and risk of cancer.

To quantitatively explore whether this hypothesized survivor bias may be a possible explanation for the observed inverse association between PD and cancer risk, we estimated the association between PD and risk of developing cancer in a Danish population-based PD study where the unique national register systems allowed us to derive information, and applied the inverse probability-of-censoring weights technique that has been extensively discussed previously [35, 69-71] to correct for the proposed potential survivor bias. First, we model survival as a function
of the baseline risk factors including demographic characteristics, PD status, and cancer histories. Then, we generate the selection probabilities for each surviving individual and use the inverse of the probabilities of survival as weights in our model to adjust for survivor bias. We describe the direction and magnitude of the survivor bias by comparing the effect estimates before and after bias correction.

3.3. METHODS

3.3.1. STUDY POPULATION

For the present analysis, we used the PASIDA study population. Details of the study have been described previously. At baseline, the cohort included 2,718 eligible cases identified from the Danish National Hospital Register files between 1996 and 2009 at 10 major neurological departments with PD diagnosis (ICD-8 code: 342; ICD-10 code: G20) assigned by a neurologist and 3,626 population controls selected from the Danish Central Population Registry matched on birth year and sex, being alive and without a PD diagnosis by the time of case identification and initially contacted. We followed the cohort of patients and control subjects from baseline to end of 2013 for the analysis of subsequent cancer incidences. Subjects who remained free of cancer through the follow-up were censored at death or on December 31, 2013, whichever came first.

3.3.2. DATA COLLECTION

In the present analysis we used the same variables as before. Since the Danish Central Population Registry and Danish Cancer Registry were available for all Danish citizens, we were able to
extract demographics and cancer diagnoses for all individuals identified at baseline. In addition, vital status up to 2013 was extracted from the Danish Central Population Registry.

3.3.3. STANDARD PROTOCOL APPROVALS, REGISTRATIONS, AND PATIENT CONSENT

We obtained written informed consent from all subjects, and the UCLA Institutional Review Board and Danish Data Protection Agency approved the study protocol.

3.3.4. SURVIVOR BIAS STRUCTURE

We use Directed Acyclic Graphs (DAGs) to present the structural relationships between variables [34-35], and a basic survivor bias mechanism as illustrated in Supplementary Figure 3.1. For the purposes of this study, we assume that there is no true causal relationship between the incidence of PD and subsequent cancer rather that 1) PD is associated with a reduced life expectancy reducing survival prior to cancer incidence [9], and 2) survival is even shorter among those who are at higher risk of developing cancer, i.e. risk factors of cancer also decrease survival among PD patients in general. Death due to PD-related events such as falls, pneumonia etc., are competing risks shortening survival among those at higher risk of receiving a cancer diagnosis since only individuals who survive long enough to develop cancer will be observable. Specifically, PD patients with risk factors for cancers, i.e. those who smoke, are physically inactive and eat a less healthy diet, are presumed to have a shorter survival than controls with the
same risk behaviors related to cancer who do not suffer from PD. Conditioning on survival then induces a non-causal path that relates PD to cancer through common causes of non-survival.

**Figure 3.1** Basic survivor bias mechanism in studies of cancer after PD diagnosis.

Based on previously reported associations between PD and the risk of developing cancer, we propose two scenarios for smoking-related (Figure 3.2) and non-smoking-related (Figure 3.3) cancers, respectively, in the PASIDA Study. It has been well recognized that smoking and PD are inversely associated [19, 37, 72], however, the biologic mechanism remains unclear and we recently suggested that quitting is an early symptom of PD [31]. Thus, we use a double headed arrow to represent the association between smoking and PD.
In the first scenario, smoking (SM) causes smoking-related cancer (SRC), which can occur either before or after a PD diagnosis; the latter is the focus of the present analysis which complements our companion paper. The likelihood of survival (S) is decreased by both PD diagnosis and a previous cancer history. In addition, lifestyle factors such as smoking also impact survival via various other common health outcomes in the elderly such as cardiovascular disease and stroke. Variable Z represents a set of known risk factors for PD that influence survival through different comorbidities, specifically age ($Z_1$), sex ($Z_2$), SES ($Z_3$), and urbanization ($Z_4$). Information on these factors was available in the Danish national registry systems for all Danish citizens. These risk factors also influence smoking habits and the occurrence of cancer. In addition, a set of unknown or unmeasured variables (U) is presented in the causal diagram. Survival and cancer are negatively related through U, which represents factors that make a person less likely to survive but more likely to have subsequently developed cancer if they had survived.

In the second scenario (Figure 3.3) we focus on cancers which are not caused by smoking (NSRC) but might be attributable to other common risk factors that may shorten survival such as lack of physical activity and unhealthy diets. All other covariates and relationships remain the same as in scenario 1. In both scenario 1 and 2, bias occurs by conditioning on survival creating a spurious association between cancer and PD.
**Figure 3.2** Bias mechanism in the PASIDA Study (smoking-related cancers after PD diagnosis).

**Figure 3.3** Bias mechanism in the PASIDA Study (non-smoking-related cancers after PD diagnosis).
3.3.5. BIAS CORRECTION USING INVERSE PROBABILITY-OF-CENSORING WEIGHTS

Survival was defined as a binary variable \( S (1 = \text{survived}; 0 = \text{dead}) \). The conditional probability of survival was generated from the logistic equation as:

\[
\text{logit} \ P(S = 1|PD, SM, Z_1, Z_2, Z_3, Z_4, \text{cancer}_0) = \beta_S + \beta_{\text{SPD}} \times PD + \beta_{\text{SSM}} \times SM + \beta_{SZ1} \times Z_1 + \beta_{SZ2} \times Z_2 + \\
\beta_{SZ3} \times Z_3 + \beta_{SZ4} \times Z_4 + \beta_{\text{Scancer0}} \times \text{cancer}_0
\]  

(1)

The \( \beta \)s represent the associations between survival and each of the predictors. For example, \( \beta_S \) is the log odds of \( S = 1 \) when all the predictors are set at the reference level; \( \beta_{\text{SPD}} \) is the log odds ratio (OR) relating \( S \) and \( PD \) when other predictors are set at the reference level. We refer to smoking-related cancers as SRC in scenario 1 and to non-smoking-related cancers as NSRC in scenario 2. For simplicity purposes, we assume no interactions between predictors. Since demographic information and vital status are available for all individuals in the Danish Central Population Registry and cancer diagnosis information can be retrieved via linkage to the Danish Cancer Registry, we model \( S \) as a function of the above listed predictors measured for both those who survived \( (S = 1) \) and those who died \( (S = 0) \) using the data extracted from the these registries for PD cases and controls. Associations between survival and each predictor estimated from the above model are shown in Table 3.1. Using equation (1) we generate the selection probability for each survivor \( (S=1) \) conditioning on their PD, smoking, age, sex, SES, urbanization, and previous cancer history values and the bias parameters \( (\beta \)s) estimated in the above model.
3.3.6. **STATISTICAL ANALYSIS**

All variables used in the analysis were coded as binary factors. While data on lifelong cigarette smoking history was obtained from a structured telephone interview in our case-control study, these data only exist for participants who were successfully recruited (n=3,700) but not for case-control study non-participants (n=2,644). Assuming missing at random (MAR) we first used multiple imputation to impute smoking pack-year information for individuals who did not participate in our interviews before estimating bias parameters for survival.

We accumulated person-time until a subject developed an incident cancer diagnosis, died, the medical record ended or the end of the study was reached (December 31, 2013), whichever came first. We assessed incidence rates of cancer after PD diagnosis in the PD patients and in the PD-free control group and calculated incidence rate ratios with 95% confidence intervals (CIs) adjusting for smoking, age, sex, SES, urbanization, and cancer history. Survivor bias was

Table 3.1 Estimated associations between survival and each predictor

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Predictor</th>
<th>Definition</th>
<th>Bias parameter</th>
<th>Estimated association (OR(95%CI))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking-related cancer (SRC)</td>
<td>PD</td>
<td>yes vs. no</td>
<td>β_{SPD}</td>
<td>0.22 (0.19-1.26)</td>
</tr>
<tr>
<td></td>
<td>Smoking</td>
<td>above vs. below median of controls</td>
<td>β_{SSM}</td>
<td>0.72 (0.62-0.84)</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>older vs. younger</td>
<td>β_{SZ1}</td>
<td>0.31 (0.26-0.36)</td>
</tr>
<tr>
<td></td>
<td>Sex</td>
<td>male vs. female</td>
<td>β_{SZ2}</td>
<td>0.80 (0.69-0.93)</td>
</tr>
<tr>
<td></td>
<td>SES</td>
<td>low vs. high</td>
<td>β_{SZ3}</td>
<td>0.82 (0.71-0.95)</td>
</tr>
<tr>
<td></td>
<td>Urbanization</td>
<td>rural vs. urban</td>
<td>β_{SZ4}</td>
<td>1.07 (0.86-1.34)</td>
</tr>
<tr>
<td></td>
<td>Cancer history</td>
<td>yes vs. no</td>
<td>β_{scancer}</td>
<td>0.64 (0.42-0.98)</td>
</tr>
<tr>
<td>Non-smoking-related cancer (NSRC)</td>
<td>PD</td>
<td>yes vs. no</td>
<td>β_{SPD}</td>
<td>0.27 (0.23-0.31)</td>
</tr>
<tr>
<td></td>
<td>Smoking</td>
<td>above vs. below median of controls</td>
<td>β_{SSM}</td>
<td>0.67 (0.57-0.77)</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>older vs. younger</td>
<td>β_{SZ1}</td>
<td>0.32 (0.27-0.37)</td>
</tr>
<tr>
<td></td>
<td>Sex</td>
<td>male vs. female</td>
<td>β_{SZ2}</td>
<td>0.83 (0.71-0.96)</td>
</tr>
<tr>
<td></td>
<td>SES</td>
<td>low vs. high</td>
<td>β_{SZ3}</td>
<td>0.80 (0.69-0.93)</td>
</tr>
<tr>
<td></td>
<td>Urbanization</td>
<td>rural vs. urban</td>
<td>β_{SZ4}</td>
<td>1.16 (0.94-1.45)</td>
</tr>
<tr>
<td></td>
<td>Cancer history</td>
<td>yes vs. no</td>
<td>β_{scancer}</td>
<td>0.80 (0.52-1.21)</td>
</tr>
</tbody>
</table>
corrected for by using $P(S = 1)/[P(S = 1| PD, SM, Z_1, Z_2, Z_3, Z_4, cancer_0]$ as regression weights. All analyses were performed by using SAS 9.3 (SAS Institute, Cary, NC, USA).

### 3.4. RESULTS

On average, we were able to follow all PD patients and controls for as long as 9 years after the PD diagnosis date (for PD patients) or index date (for PD free controls). In terms of smoking-related cancers, more PD cases (23.5%) than PD free controls (7.5%) died without receiving a smoking-related cancer diagnosis by the end of the follow-up. For non-smoking-related cancers, the cancer free mortality was also higher for PD patients (23.7%) than PD free controls (9.0%). Follow-up year-specific death rates were much higher for PD cases than population controls and as the follow-up duration increased the gap in death rates in cases and controls widened (Figure 3.4).
The demographic characteristics are summarized in Table 3.2. For both smoking-related and non-smoking-related cancers subsequent to PD, survival to the end of the follow-up was associated with 1) younger age, 2) being female, as well as 3) higher SES level. During follow-up, 68 PD cases and 190 control subjects developed smoking-related cancers, and 90 PD cases and 160 controls developed non-smoking-related cancers. PD patients had a 14% decreased rate of developing smoking-related cancer compared with controls (95%CI: 0.68-1.10) but a 10% increased rate of developing non-smoking-related cancers compared to controls (95%CI: 0.89-1.36) (Table 3.3) yet both confidence intervals included the null. When adjusting for survivor bias using inverse probability-of-censoring weights, the bias-corrected rate ratios are 1.07
(95%CI: 0.86-1.33) for smoking-related cancer and 1.31 (95%CI: 1.07-1.59) for non-smoking-related cancer (Table 3.3).

Table 3.2 Demographic characteristics of the cohort

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>SRC</th>
<th>NSRC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Controls</td>
<td>Cases</td>
<td>Controls</td>
</tr>
<tr>
<td></td>
<td>n %</td>
<td>n %</td>
<td>n %</td>
<td>n %</td>
</tr>
<tr>
<td>Total</td>
<td>2718</td>
<td>100.0</td>
<td>3626</td>
<td>100.0</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>63.4±9.7</td>
<td>64.2±9.6</td>
<td>61.8±9.6</td>
<td>63.7±9.6</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1564</td>
<td>57.5</td>
<td>1973</td>
<td>54.4</td>
</tr>
<tr>
<td>Female</td>
<td>1154</td>
<td>42.5</td>
<td>1653</td>
<td>45.6</td>
</tr>
<tr>
<td>SES a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Academic and top self-employed</td>
<td>348</td>
<td>12.8</td>
<td>322</td>
<td>8.9</td>
</tr>
<tr>
<td>High self-employed and high salaried</td>
<td>364</td>
<td>13.4</td>
<td>417</td>
<td>11.5</td>
</tr>
<tr>
<td>Low self-employed and mid-salaried</td>
<td>495</td>
<td>18.2</td>
<td>634</td>
<td>17.5</td>
</tr>
<tr>
<td>Skilled worker and low salaried</td>
<td>598</td>
<td>22.0</td>
<td>893</td>
<td>24.6</td>
</tr>
<tr>
<td>Unskilled worker</td>
<td>279</td>
<td>10.3</td>
<td>544</td>
<td>15.0</td>
</tr>
<tr>
<td>Unknown or unspecified</td>
<td>634</td>
<td>23.3</td>
<td>816</td>
<td>22.5</td>
</tr>
<tr>
<td>Degree of urbanization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capital</td>
<td>709</td>
<td>26.1</td>
<td>1073</td>
<td>29.6</td>
</tr>
<tr>
<td>Provincial cities</td>
<td>1648</td>
<td>60.6</td>
<td>1805</td>
<td>49.8</td>
</tr>
<tr>
<td>Rural areas</td>
<td>249</td>
<td>9.2</td>
<td>378</td>
<td>10.4</td>
</tr>
<tr>
<td>Peripheral regions</td>
<td>82</td>
<td>3.0</td>
<td>148</td>
<td>4.1</td>
</tr>
<tr>
<td>Unknown or unspecified</td>
<td>30</td>
<td>1.1</td>
<td>222</td>
<td>6.1</td>
</tr>
</tbody>
</table>

Table 3.3 Incidence rate ratio (IR) for PD-cancer associations

<table>
<thead>
<tr>
<th>Cancer diagnosis after PD</th>
<th>Number of cancer cases (%)</th>
<th>IRadj (95%CI)</th>
<th>Bias-corrected IRadj (95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All (n=6344)</td>
<td>PD patients (n=2718)</td>
<td>PD free controls (n=3626)</td>
</tr>
<tr>
<td>Smoking-related (SRC) a</td>
<td>258 (4.1)</td>
<td>68 (2.5)</td>
<td>190 (5.2)</td>
</tr>
<tr>
<td>Non-smoking-related (NSRC) b</td>
<td>250 (3.9)</td>
<td>90 (3.3)</td>
<td>160 (4.4)</td>
</tr>
</tbody>
</table>

a For smoking-related cancers models adjusted for age at first PD diagnosis (for cases) or age at index date (for controls), sex, urbanization, SES, smoking, and previous cancer history.

b Non-smoking-related cancers models adjusted for age at first PD diagnosis (for cases) or age at index date (for controls), sex, urbanization, SES, smoking, and previous cancer history.
3.5. DISCUSSION

In this paper we hypothesized that a bias mechanism operates in follow-up studies of older populations with multi-morbidities that may explain inverse associations between PD and subsequent cancer diagnosis as has been observed in a number of epidemiologic studies [1, 20-22, 43, 48-51]. PD related complications or other comorbidities in PD patients may selectively decrease the likelihood of survival in those with PD who are at higher risk of developing cancer. Since a diagnosis of cancer can only be observed among surviving individuals, conditioning on survival produces a negative selection bias. In cohort studies, when PD patients and reference subjects free of PD are followed to observe the incidence of cancer, individuals may die before developing cancer and this censoring by death may depend on risk factors that also are related to PD status. Specifically, if individuals with PD die earlier are those who would also have been at higher risk of cancer had they survived; this creates a censoring mechanism that makes survivor bias unavoidable. In our PASIDA study, information about factors that influence survival during follow-up, i.e. baseline demographic characteristics, cancer histories, and vital status, are available from national registries and from interviews we conducted with participants in our large case-control study, which allowed us to model the probabilities of survival as a function of these factors in the full cohort. Using the inverse of the probabilities of survival as weights, our bias adjusted models showed that the suggested negative association between PD and risk of smoking-related cancers could be explained by the proposed survivor bias. For non-smoking-related cancers, the observed null association between PD and risk of cancer became positive when we corrected for survivor bias. These results, in addition to our previous work, suggest that for cancers not only prior to but after the onset of PD, it is important to take survival/selection
bias into consideration when interpreting results of studies conducted in comorbid elderly populations.

While in our PASIDA study the negative association between PD and subsequent smoking-related cancers (rate ratio = 0.86; 95% CI: 0.68-1.10) was not statistically significant, the point estimate was quite consistent in size with those reported in previous studies (Table 3.4). Since the average follow-up time in our study was 9 years and loss to follow-up was unavoidable, an incidence rate ratio was preferable as the effect estimate. Previous studies that also identified PD cases from the Danish National Hospital Register [21-22, 49] followed PD patients estimated standardized incidence ratios (SIR) to compare an observed with an expected number of cancer based on person-years in the PD cohort and national cancer incidence rates during the study period. However, those studies did not verify diagnoses from medical records and did not conduct interviews as done in the PASIDA study thus were able to include a much larger number of PD patients. The SIR estimates from these record linkage only studies were more precise than our incidence rate ratio estimates and the point estimates were overall more negative than ours.

In contrast, our study verified iPD diagnoses by extensive medical record review making diagnostic misclassification less likely. Also, we collected lifetime smoking information that was not available in the record linkage only studies. Since PD patients are less likely to smoke and at lower risk of developing smoking-related cancers, smoking has been proposed as a possible explanation for the negative association between PD and smoking-related cancers. However, in our PASIDA study adjustment for pack-years of smoking did not remove the negative association between the two diseases. On the other hand, addressing the proposed survivor bias
employing inverse probability-of-censoring weights moved the effect estimates towards the null, indicating that the observed negative association between PD and risk of smoking-related cancers may be due to the proposed survivor bias instead.

In our study we did not observe associations between PD and risk of non-smoking-related cancers (rate ratio = 1.10; 95%CI: 0.89-1.36). Association between PD and reported risks for non-smoking-related cancers were not consistent across previous studies with estimated risks for non-smoking-related cancers among PD patients being closer to those in the PD free reference group [9, 48] except for one study that reported an increased risk (risk ratio = 1.69; 95%CI: 1.15-2.46) [44]. For example, a large record-based case-control study nested within a cohort study reported a lower risk of non-smoking-related cancers among PD patients (OR = 0.80; 95%CI: 0.60-1.05) [20]. In studies that also identified PD cases from the Danish National Hospital Register [21-22, 50], estimated effects were negative and of similar size but the large sample size rendered them statistically significant (SIR = 0.81; 95%CI: 0.7-0.9 vs. SIR = 0.79; 95%CI: 0.71-0.86) [21-22] except for an early study (SIR = 1.01; 95%CI: 0.8-1.0) [50]. When adjusting for the proposed survivor bias, our rate ratio estimate became more strongly positively and the confidence interval did not cross the null (rate ratio = 1.31; 95%CI: 1.07-1.59), suggesting that we underestimate a positive risk for non-smoking-related cancers due to the proposed survivor bias.

Inverse probability-of-censoring weights can be used to correct for potential selection bias in cohort studies by creating a pseudo-population that would have been observed with loss to follow-up (in our case death before developing cancer) occurring but random with respect to the
measured determinants of survival [71]. For a valid estimation using inverse probability-of-
censoring weights, the exposure and censoring mechanisms must be well-defined and the
assumptions of exchangeability, positivity, and correct model specification in both the outcome
and weight model must hold [70-71]. Although the violations in conditional exchangeability are
not testable, in our analysis we measured and adjusted for the most important potential
confounders as well as common causes of survival and cancer to minimize the potential
uncontrolled confounding and violations in conditional exchangeability. With the comparatively
large sample size of our study, it is less likely to have a zero probability of survival for every
combination of observed predictors of survival, thus the assumption of positivity is likely to also
hold. However, it is possible that the functional forms of some predictors used in the correction
model were not the most appropriate ones, such that the assumption of correct model
specification cannot be guaranteed.
<table>
<thead>
<tr>
<th>Study</th>
<th>Publication year</th>
<th>Study area</th>
<th>PD case N</th>
<th>Control/cohort N</th>
<th>Matching variable</th>
<th>Follow-up year</th>
<th>Data collection</th>
<th>Relative risk measure (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jansson and Jankovic</td>
<td>1985</td>
<td>US</td>
<td>406; 6</td>
<td>No</td>
<td>No</td>
<td>Follow-up year: 1978-1984</td>
<td>PD: medical records; cancer: medical records</td>
<td>Overall malignant cancer (except non-melanoma skin cancer): relative risk 0.46 (P-value: 0.0009)</td>
</tr>
<tr>
<td>Moller et al</td>
<td>1995</td>
<td>Denmark</td>
<td>7,046</td>
<td>No</td>
<td>No</td>
<td>PD cases diagnosed in 1977-89; follow-up to end of 1990; average follow-up: 4.6 years</td>
<td>PD: Danish Hospital Register; cancer: Danish Cancer Registry; death: Danish register of deaths</td>
<td>Overall SIR: 0.88 (0.8-1.0); smoking-related: 0.49 (0.4-0.6); non-smoking-related: 1.01 (0.8-1.0)</td>
</tr>
<tr>
<td>Minami et al</td>
<td>2000</td>
<td>Japan</td>
<td>228 aged under 80</td>
<td>No</td>
<td>No</td>
<td>Follow-up year: 1984-1992</td>
<td>PD: Medical institution reports; cancer: Miyagi Cancer Registry</td>
<td>Overall SIR: 0.83 (0.46-1.37)</td>
</tr>
<tr>
<td>Olsen et al</td>
<td>2005</td>
<td>Denmark</td>
<td>14,088</td>
<td>No</td>
<td>No</td>
<td>PD cases diagnosed in 1977-98; follow-up to end of 1999</td>
<td>PD: Danish Hospital Register; cancer: Danish Cancer Registry; death: Danish National mortality files</td>
<td>Overall SIR: 0.88 (0.8-0.9); smoking-related: 0.58 (0.5-0.6); non-smoking-related: 0.81 (0.7-0.9)</td>
</tr>
<tr>
<td>Elbaz et al</td>
<td>2005</td>
<td>US</td>
<td>196</td>
<td>185</td>
<td>age, sex</td>
<td>PD cases diagnosed in 1976-95; follow-up to end of 2002</td>
<td>Rochester Epidemiology Project; PD: medical record review; cancer: medical record review</td>
<td>Overall RR: 1.64 (1.15-2.35); smoking-related: 0.77 (0.29-2.03); non-smoking-related: 1.69 (1.15-2.46)</td>
</tr>
<tr>
<td>Driver et al</td>
<td>2007</td>
<td>US</td>
<td>487</td>
<td>487</td>
<td>age</td>
<td>Median follow-up year for PD cases: 5.2 years; for reference group: 5.9 years</td>
<td>Physician's Health Study; PD: self-report; cancer: medical records and pathology reports</td>
<td>Overall RR: 0.85 (0.59-1.22); smoking-related: 0.70 (0.35-1.38); non-smoking-related: 0.91 (0.60-1.40)</td>
</tr>
<tr>
<td>Lo et al</td>
<td>2010</td>
<td>US</td>
<td>692</td>
<td>761</td>
<td>birth year; sex; respondent type</td>
<td>1994-1995; 2000-2003</td>
<td>PEAK studies; PD: medical record review; cancer: SEER code</td>
<td>Overall RR: 0.94 (0.70-1.3); smoking-related: 0.70 (0.40-1.2); non-smoking-related: 1.0 (0.71-1.4)</td>
</tr>
<tr>
<td>Becker et al</td>
<td>2010</td>
<td>UK</td>
<td>2,993</td>
<td>3,003</td>
<td>Cohort: age, sex, general practice, diagnosis date, years of history in the database prior to the diagnosis date; case-control: age, sex, calendar time</td>
<td>PD cases diagnosed in 1994-2005 (&gt; = 40 years old); follow-up to end of 2005</td>
<td>General Practice Research Database; PD: medical records; cancer: medical records</td>
<td>Cohort: overall rate ratio: 0.77 (0.64-0.92); case-control: overall OR: 0.72 (0.59-0.87); smoking-related: 0.47 (0.31-0.72); non-smoking-related: 0.80 (0.60-1.05)</td>
</tr>
<tr>
<td>Ragbajrg et al</td>
<td>2012</td>
<td>Denmark</td>
<td>20,000</td>
<td>15,020</td>
<td>No</td>
<td>PD cases diagnosed in 1977-2006; follow-up to end of 2008</td>
<td>PD: Danish Hospital Register; cancer: Danish Cancer Registry; death: Central Population Register</td>
<td>Overall SIR: 0.86 (0.83-0.90); smoking-related: 0.65 (0.60-0.70); non-smoking-related: 0.79 (0.71-0.86)</td>
</tr>
</tbody>
</table>
One strength of our study is that individual-level information such as demographics and vital status was available to be extracted from national registries, making it possible to adjust for survivor bias by modeling survival as a function of PD status and measured covariates among both those who survived and those who died; i.e. the parameters for our bias analyses were derived from actual data for the Danish population. In addition, we reviewed medical records and verified iPD diagnosis to limit possible misclassification of the outcome. The sample size in our study was larger than for all previous cohort studies with subject contact but smaller than for previous record linkage based studies that lacked covariates data such as smoking and were unable to confirm PD diagnoses.

Our study has also several limitations. We did not have smoking information for individuals not enrolled in the interview-based study. Thus, in order to use smoking in prediction models of survival we had to employ multiple imputation to generate pack-years for non-participants assuming missingness at random. If this assumption was not met it will have produced inaccurate estimates for the parameters predicting survival. Also, factors creating association between survival and cancers through unknown or unmeasured factors could not be included in the survival model. Thus there could be residual bias for which we did not adjust in our model. We assumed no interaction between the predictors in our models. If the effects on survival were modified by some of the predictors, the survival probabilities may not be calculated accurately.

In this study, we propose a bias structure and illustrate that in cohort studies the observed association with cancer after a diagnosis of PD could be shifted towards or even beyond the null value and that adjusting for survivor bias using the inverse probability-of-censoring weights
produces effect estimates that are positive rather than negative. In our companion work, a similar bias structure was proposed and we illustrated that in case-control studies the inverse association with cancer prior to the diagnosis of PD is attenuated when adjusting for both survivor bias and selection bias due to non-participation when survival was assumed to be strongly associated with both cancer and PD. For cancer both before and after the diagnosis of PD, survivor bias could be an alternative explanation for the observed association between cancer and PD with a reasonable bias structure and somewhat testable assumptions. Since the bias mechanisms we described are likely common in cohort and case-control studies in elderly populations who are at risk of death from competing causes, survivor bias should be taken into consideration and assessed quantitatively as a possible alternative explanation when associations are observed between comorbidities.
4. STUDY 3

EXPOSURE MISCLASSIFICATION USING A GEOGRAPHICAL INFORMATION SYSTEM (GIS)-BASED ASSESSMENT OF LONG-TERM AMBIENT PESTICIDE EXPOSURES: RESIDENTIAL MOBILITY AND CHANGES IN PESTICIDE USE IN A STUDY OF PARKINSON’S DISEASE
4.1. ABSTRACT

In chronic disease environmental studies, exposure data can now be more and more easily generated using geographic information system tools and methods that rely heavily on spatial information. When long-term ambient exposure measures are of interest, residential history information becomes necessary. However, many studies routinely only record or have access to residential address at study entry or at the time when a health record was created or a disease diagnosed, and it is important to evaluate whether such information can be useful under assumptions such as little residential mobility or mobility of study subjects being unlikely to impact the estimation of longer exposures more than minimally such as when exposures are temporally or spatially relatively homogeneous. Here we are using the example of long-term pesticide exposures among rural residents in California to assess how residential mobility as well as temporal and spatial patterns of exposures may contribute to exposure misclassification using a population-based case-control study of Parkinson’s disease as an illustration. We simulated scenarios for studies lacking residential histories which would use enrollment addresses only as a proxy to generate exposure estimates over different lengths of time periods prior to enrollment. We describe how estimated pesticide exposures would be misclassified and show that exposures could be either over- or under- estimated depending on the pesticide, time period of interest, as well as the cut-off points chosen to generate exposure variables. We also show that exposure misclassification is not necessarily nondifferential and that the direction of bias can be either positive or negative. Even in an elderly population with relatively little residential mobility (i.e. on average having lived 19 years at one address), temporal and spatial patterns of exposure may influence the exposures enough that comparisons based on long-term exposure estimates using
one address only may result not only in misclassification bias but one for which it is hard to predict whether the effects are under- or over- estimated.

4.2. INTRODUCTION

Recent advancements of geographic information system (GIS)-based tools have encouraged their use in environmental epidemiology studies especially those targeting health effects from agricultural pesticide exposures related to a number of diseases including cancers [73-75], autism [76-78], and Parkinson’s disease (PD) [79-81]. Since GIS-based exposure assessment relies heavily on spatial information, the availability of address information is a crucial element for exposure assessment, especially when environmental exposures relevant for the health outcome span many years or even decades such as in chronic disease research.

Given the wealth of data collected and the investments made into large cohort studies, it may seem expedient to add exposures derived with GIS-based tools to investigate environmental causes of disease. However, cohort studies usually do not collect residential history data prior to baseline and their follow-up may not cover the periods of life most relevance for disease etiology especially for cohorts comprised of elderly subjects that target diseases late in life. Also, for a growing number of studies based on registries or medical records residential history data are generally not available. Therefore, it may be tempting to use the residential location at study entry or at the time when a record was generated or a diagnosis made as a proxy for residential history under the assumption that mobility and/or exposure variation is limited over the period of interest [82-83]. Here we investigate for the first time whether and how ambient pesticide exposures are misclassified when having to make such an assumption even when spatial and
temporal variations in agricultural pesticide applications are known and the population rather stable in terms of mobility.

California is one of the most agriculturally productive states in the United States and agricultural pesticides been have extensively applied over many decades [84-85]. California’s pesticide-use reporting (PUR) program has been recognized as the most comprehensive system to record pesticide use in the world. The PUR data have been widely used to assess pesticide application patterns for populations residing in high pesticide use areas and pesticide exposures have been assessed for a number of health outcomes, including PD [86], cancers [87-90], and fetal deaths [91]. To date no study has been conducted to assess whether it is necessary to have access to address histories for an extended period of interest to assess pesticide exposures or whether and under what circumstances it might suffice to rely on one address only to derive estimates of exposure.

Pesticides increase the risk of PD [92-97] and due to its long latency period long-term pesticide exposure assessment is required. Thus, we previously collected detailed residential history data to estimate pesticide exposures in a population-based case-control study of PD in central California. Here we are using these data to simulate scenarios in which residential histories are lacking and only enrollment addresses are available as a proxy for a 26-year period. We explore and describe how estimated exposures would be misclassified under different assumptions and for different PD-relevant pesticides and furthermore assess how such misclassification may influence risk estimates we derived for PD.
4.3. METHODS

4.3.1. STUDY POPULATION

The PEG (Parkinson’s Environment and Gene) Study at the University of California, Los Angeles (UCLA) is a population-based case-control study that enrolled incident idiopathic PD patients and population-based controls from the mostly rural agricultural tri-county area (Kern, Tulare, Fresno) in central California between January 1, 2001 and early 2011. Subject recruitment methods [81, 95] and case definition criteria [98-99] have been described in details elsewhere.

Among 1,167 PD patients initially identified through neurologists, large medical groups, and public service announcements, 604 did not meet eligibility criteria: 397 had their initial PD diagnosis more than 3 years prior to recruitment, 134 lived outside the tri-county area at the time of recruitment, 51 had a diagnosis other than PD, and 22 were too ill to participate. Among 563 eligible cases, 90 could not be examined (56 declined to participate or moved away, 18 had become too ill to be examined, and 16 died prior to the scheduled appointment) leaving us with 473 subjects examined by a UCLA movement disorder specialist, of whom 94 did not meet published criteria for idiopathic PD [100-101] when examined initially 13 were reclassified during follow-up [102], and 6 subjects withdrew between examination and interview; thus in total 360 were enrolled and provided all information needed for analyses.

Population-based controls were recruited initially from Medicare lists (2001) and, after the Health Insurance Portability and Accountability Act (HIPAA) took effect, from residential tax assessor records from the tri-county area. We mailed letters of invitation to residents living at
randomly selected parcels and also attempted to identify head-of-household names and telephone numbers using marketing companies’ services and internet searches. Of the 1,212 potentially eligible controls contacted, 457 were found to be ineligible: 409 were <35 years of age, 44 were too ill to participate, and 4 primarily resided outside the study area. From 755 eligible population controls, 409 declined, became too ill to participate, or moved out of the area after screening and prior to enrollment; 346 population controls enrolled and 341 provided all information needed for analyses.

4.3.2. GIS-BASED AMBIENT PESTICIDE EXPOSURE ASSESSMENT

We obtained lifetime residential addresses from all participants who completed the interviews (N=701). Employing our GIS-based system, we combined PUR data, land use maps, and geocoded address information to produce estimates of pesticide exposure within a 500-meter radius buffer around participants’ residential addresses from 1974 to 1999 [80, 103]. A detailed description of our approach has been provided elsewhere [95, 104].

Specific pesticides have been suggested to produce dopaminergic neurodegeneration and increase the risk of PD. The suggested toxicological mechanisms involve: (1) increase of oxidative stress [105-106]; (2) inhibition of the ubiquitin proteasome system [107-108]; (3) disruption of mitochondrial function [109-110]; and (4) promotion of cell death [105, 111]. For the present analysis we selected the following pesticides of relevance to PD with different use patterns: paraquat, ziram, Maneb, chlorpyrifos, and diazinon. The temporal and spatial
application patterns for each pesticide over the past decades in the tri-county area are presented in Figure 4.1 and 4.2.

**Figure 4.1** Trends in pesticide use in the tri-County area, 1974-1999.
Figure 4.2 Annual average pesticide use in California (pound per acre), 1974-1999.

We calculated annual ambient exposures for each participant for the selected pesticides by summing the pounds of pesticides applied and weighting the total poundage by the proportion of the acreage treated in the buffer. Annual exposure estimates were generated by using (1) the complete residential address histories reported by the participant and (2) only the residential address at which we enrolled each participant into our study. Most participants resided at more than one address during the period of interest for exposure assessment (on average 10 addresses per subject). For the purpose of assessing the extent of pesticide exposure measurement error when using enrollment addresses only, we considered measures based on the complete 26-year address histories as the gold standard for ambient exposures at the residence of participants.
We also created five cumulative time windows of exposure extending backward, from 1999 when the first PD cases in our study were diagnosed to 1974 when agricultural pesticide applications in California started to be recorded in the PUR system, increasing length of the time window to allow us to describe the direction and magnitude of exposure misclassification due to residential mobility in study subjects: (1) 1974-1999; (2) 1980-1999; (3) 1985-1999; (4) 1990-1999; and (5) 1995-1999.

Thus, for each pesticide the annual pounds applied per acre were averaged across each exposure period using the complete address histories or the enrollment address only. The actual exposure a resident received may not only depend on the annual average pesticide application but also characteristics of the chemicals, such as solubility, adsorption, persistence, and volatility. Also, amount of pesticide applied does not directly reflect toxicity since chemicals with high toxicity are usually applied at lower poundage than chemicals with lower toxicity. Thus, to generate exposure variables comparable across periods, we rely on the distribution of the annual average pesticide amounts in control subjects to create binary exposure variables. To make comparisons based on the same reference values, we used the 1995-1999 median and the fourth quartile from the exposure distribution of controls to compare exposure prevalence across periods of time. We also assessed whether exposure misclassification was sensitive to how exposure was defined, i.e. we compared three definitions: (1) median average exposures in controls specific to each time period investigated; (2) median average exposures in controls for the period 1995-1999 only; and (3) 4th quartile of average exposures in controls for the period 1995-1999 only.
4.3.3. STANDARD PROTOCOL APPROVALS, REGISTRATIONS, AND PATIENT CONSENTS

All procedures described have been approved by the UCLA Institutional Review Board for human participants and informed consent was obtained from all participants.

4.3.4. STATISTICAL ANALYSIS

For each time period and pesticide we plotted the annual average ambient exposure estimated using address histories or enrollment addresses only, and calculated correlation coefficients for these two exposure estimates according to the three exposure definitions and generated and plotted exposure prevalence. We calculated sensitivity and specificity using the estimates derived from the complete residential history as the gold standard and finally conducted logistic regression analyses to estimate effect sizes for PD risk based on the two different exposure assessment methods. Additionally, we compared the odds ratio (OR) estimated for both assessment methods employing a ratio of odds ratio (ROR). All analyses were performed by using SAS 9.3 (SAS Institute, Cary, NC, USA).

4.4. RESULTS

Characteristics of study participants are summarized in Table 4.1. Study participants were primarily over 60 years of age, White, and did not report a family history of PD. Compared to controls, cases were more likely to be male, never smokers, and had fewer years of education. On average, the study population had been living at their enrollment address for 19 years (20 years for cases and 18 years for controls). Figure 4.3 summarizes for each 5-year time period the
annual average ambient pesticide exposures estimated based on the address histories and enrollment address only for cases and controls, respectively. In general, based on the address histories cases exhibited on average similar exposure trends as controls but were more highly exposed in each time period. Correlation coefficients for each type of exposure estimated according to address histories or enrollment address are presented in Table 4.2. As expected, the two exposure estimates were more highly correlated when the enrollment address was used as a proxy for the address history for a more recent time period for all five pesticides in both cases and controls and never dropped below 50% even for the longest periods we evaluated. Using address histories or solely relying on the enrollment address as a proxy for long-term exposures (back to 1974 maximum) generated difference in exposure estimates. The magnitude and direction of the differences depended on time period, type of chemical, as well as PD status. Specifically, for maneb and diazinon, using enrollment address only underestimated exposures in both cases and controls in all time periods. For chlorpyrifos, it overestimated exposures except for cases in the most recent time period (1995-1999). For paraquat and ziram, it underestimated the exposures over longer periods but overestimated exposures in most recent time periods for both cases and controls. Overall, we would have expected that annual average exposures are more accurately estimated by the enrollment address (1) in the most recent period but it was only observed for some pesticides and subgroups (i.e. diazinon in both cases and controls, chlorpyrifos in cases, and parquat in controls) but not for others; or (2) for a chemical in common, widespread, and continuous use, while we observed this for chlorpyrifos and diazinon, but did not see this equally for paraquat.
Table 4.1 Demographic characteristics of study population

<table>
<thead>
<tr>
<th></th>
<th>Cases (n=360)</th>
<th></th>
<th>Controls (n=341)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td><strong>Year of birth</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1910-1919</td>
<td>22</td>
<td>6.1</td>
<td>16</td>
<td>4.7</td>
</tr>
<tr>
<td>1920-1929</td>
<td>123</td>
<td>34.2</td>
<td>99</td>
<td>29.0</td>
</tr>
<tr>
<td>1930-1939</td>
<td>123</td>
<td>34.2</td>
<td>105</td>
<td>30.8</td>
</tr>
<tr>
<td>1940-1949</td>
<td>62</td>
<td>17.2</td>
<td>64</td>
<td>18.8</td>
</tr>
<tr>
<td>1950-1959</td>
<td>23</td>
<td>6.4</td>
<td>45</td>
<td>13.2</td>
</tr>
<tr>
<td>1960-1969</td>
<td>7</td>
<td>1.9</td>
<td>12</td>
<td>3.5</td>
</tr>
<tr>
<td><strong>Age at index year</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>68.30 (10.22)</td>
<td></td>
<td>68.20 (11.42)</td>
<td></td>
</tr>
<tr>
<td>&lt;=60</td>
<td>76</td>
<td>21.1</td>
<td>87</td>
<td>25.5</td>
</tr>
<tr>
<td>&gt;60</td>
<td>284</td>
<td>78.9</td>
<td>254</td>
<td>74.5</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>206</td>
<td>57.2</td>
<td>176</td>
<td>51.6</td>
</tr>
<tr>
<td>Female</td>
<td>154</td>
<td>42.8</td>
<td>165</td>
<td>48.4</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>290</td>
<td>80.6</td>
<td>279</td>
<td>81.8</td>
</tr>
<tr>
<td>Non-white</td>
<td>70</td>
<td>19.4</td>
<td>62</td>
<td>18.2</td>
</tr>
<tr>
<td><strong>Education (years)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>13.37 (4.21)</td>
<td></td>
<td>14.44 (3.62)</td>
<td></td>
</tr>
<tr>
<td>&lt;12</td>
<td>67</td>
<td>18.6</td>
<td>38</td>
<td>11.1</td>
</tr>
<tr>
<td>12</td>
<td>96</td>
<td>26.7</td>
<td>64</td>
<td>18.8</td>
</tr>
<tr>
<td>&gt;12</td>
<td>197</td>
<td>54.7</td>
<td>239</td>
<td>70.1</td>
</tr>
<tr>
<td><strong>Smoking</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never smoker</td>
<td>188</td>
<td>52.2</td>
<td>146</td>
<td>42.8</td>
</tr>
<tr>
<td>Former smoker</td>
<td>152</td>
<td>42.2</td>
<td>161</td>
<td>47.2</td>
</tr>
<tr>
<td>Current smoker</td>
<td>20</td>
<td>5.6</td>
<td>34</td>
<td>10.0</td>
</tr>
<tr>
<td><strong>1st degree relative with PD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>53</td>
<td>14.7</td>
<td>37</td>
<td>10.9</td>
</tr>
<tr>
<td>No</td>
<td>307</td>
<td>85.3</td>
<td>281</td>
<td>82.4</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0.0</td>
<td>23</td>
<td>6.7</td>
</tr>
</tbody>
</table>

*Index year was defined as year of PD diagnosis for cases and year of eligibility determined for controls.*
Figure 4.3 Annual average ambient pesticide exposures estimated using actual address history and enrollment address for the 5-, 10-, 15-, 20-, and 26-year averaging periods. Solid lines: exposures estimated from complete residential history; dotted lines: exposures estimated from the enrollment address only. Red: cases; green: controls.
Table 4.2  Correlation coefficient between the annual average pesticide exposures estimated using complete residential history and enrollment address only

<table>
<thead>
<tr>
<th></th>
<th>Cases and controls</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>paraquat</td>
<td>0.58909</td>
<td>0.63722</td>
<td>0.69605</td>
<td>0.77836</td>
<td>0.82634</td>
</tr>
<tr>
<td>ziram</td>
<td>0.68088</td>
<td>0.68368</td>
<td>0.75992</td>
<td>0.78197</td>
<td>0.85966</td>
</tr>
<tr>
<td>maneb</td>
<td>0.51969</td>
<td>0.53403</td>
<td>0.63779</td>
<td>0.69863</td>
<td>0.69995</td>
</tr>
<tr>
<td>chlorpyrifos</td>
<td>0.73613</td>
<td>0.75601</td>
<td>0.78197</td>
<td>0.79861</td>
<td>0.81468</td>
</tr>
<tr>
<td>diazinon</td>
<td>0.57262</td>
<td>0.65259</td>
<td>0.71639</td>
<td>0.76659</td>
<td>0.80554</td>
</tr>
</tbody>
</table>

|                     | Cases              |                   |                   |                   |                   |
| paraquat            | 0.64871            | 0.69426           | 0.74649           | 0.81201           | 0.85010           |
| ziram               | 0.71803            | 0.71520           | 0.74945           | 0.79101           | 0.87765           |
| maneb               | 0.51740            | 0.51719           | 0.60633           | 0.70932           | 0.72916           |
| chlorpyrifos        | 0.78551            | 0.79264           | 0.82350           | 0.82416           | 0.86169           |
| diazinon            | 0.59136            | 0.67385           | 0.73108           | 0.78103           | 0.83288           |

|                     | Controls            |                   |                   |                   |                   |
| paraquat            | 0.51565            | 0.56950           | 0.63749           | 0.74070           | 0.80332           |
| ziram               | 0.62091            | 0.63102           | 0.76735           | 0.76378           | 0.82893           |
| maneb               | 0.50724            | 0.55179           | 0.68706           | 0.68441           | 0.65220           |
| chlorpyrifos        | 0.67620            | 0.71011           | 0.73209           | 0.76848           | 0.76031           |
| diazinon            | 0.54775            | 0.63143           | 0.70082           | 0.75220           | 0.77724           |

Figure 4.4 shows the prevalence of exposure (by PD status) for both types of addresses used to generate different lengths of period estimates. When using the first exposure definition to define the exposure status according to the median exposure in controls in each respective time period, the exposure prevalence was higher when more and more historic time was added into the averaging time period. In order to take the overall increasing trend of pesticide use in the tri-county area into consideration (Figure 4.1), a fixed median average exposure in controls in the time period of 1995-1999 was used to compare estimates across time periods. As expected from the observed increasing trends in pesticide use over time, exposure prevalence for the averaging periods that included more years prior to enrollment was attenuated when using both actual
address history and enrollment address only. When using the 4th quartile of the exposure
distribution in controls in 1995-1999 to capture only the most extreme exposures, the estimated
exposure prevalence decreased for the averaging periods that included more previous years, with
periods. Overall for each of the three exposure definitions the prevalence of exposure differed
only slightly when basing the measures on address histories versus enrollment address only and
was quite similar in pattern for cases and controls.
Exposure prevalence estimated based on address histories and enrollment address only for each time period using exposure definition 2
Figure 4.4 Prevalence of exposures for the 5-, 10-, 15-, 20-, and 26-year averaging periods and 3 exposure definitions: (1) median average exposures in controls within each time period; (2) median average exposures in controls for the period 1995-1999; and (3) 4th quartile of average exposure in controls for the period 1995-1999. Solid lines: exposures estimated from complete residential history; dotted lines: exposures estimated from the enrollment address. Yellow/red/pink: cases; blue/green: controls.

Figure 4.5 presents sensitivity and specificity according to the three exposure definitions respectively. In general, sensitivity was consistent across the five different lengths periods with exception of 1985-1999 especially for controls, and for the 4th quartile definition sensitivity was lower for all five pesticides as expected when making the exposure cut point more stringent. Specificity, on the other hand, increased with each shortening of the averaging time interval and
was close to 1.0 when the enrollment address was used to estimate exposures in the most recent five years, i.e. 1995-1999.
Figure 4.5 Sensitivity and specificity comparing exposures estimated from the enrollment address with complete residential history using 3 exposure definitions: (1) median average exposures in controls within each time period (solid line); (2) median average exposures in controls for the period 1995-1999 (dashed line); and (3) 4th quartile of average exposures in controls for the period 1995-1999 (dotted line). Red: cases; green: controls.

In Figure 4.6 we show ORs for PD for both methods of exposure assessment, i.e. address histories and enrollment address only. For paraquat, the enrollment address only positively biased the ORs away from the null value except for the more recent time period and the strictest (4th quartile) exposure measure. For the other four pesticides, the direction of bias due to exposure misclassification depended on time period, chemical, and exposure definition. The ORs estimates using time period specific medians or the 1995-1999 median were quite consistently
biased towards the null when the enrollment address was used. Using the stricter 4\textsuperscript{th} quartile definition of exposure consistently biased the ORs towards the null over all time periods for ziram, chlorpyrifos, and diazinon but away from the null for maneb for all time periods except for 1974-1999 when using a one address proxy.

**Figure 4.6** Odds ratio for the association between PD and pesticide exposures estimated for exposures using enrollment addresses or complete residential history and 3 exposure definitions: (1) median average exposures in controls within each time period (green); (2) median average exposures in controls for the period 1995-1999 (red); and (3) 4\textsuperscript{th} quartile of average exposures in controls for the period 1995-1999 (grey). Solid lines: exposures estimated from complete residential; dashed and dotted lines: exposures estimated from enrollment address only.
The ratio of odds ratios is presented in Table 4.3. Based on a 10% change-in-estimate criterion to compare the ORs estimated from the historical and enrollment only addresses, exposure misclassification resulted in a bias of the ORs away from the null value (overestimation) for paraquat and maneb. For ziram, chlorpyrifos, and diazinon, however, the ORs where biased towards the null value. The direction of the bias was not consistent across different exposure definitions. The magnitude of ratio of odds ratio was between 0.4 and 2.0.

<table>
<thead>
<tr>
<th>Table 4.3 Ratio of odds ratio (ROR) by comparing the odds ratios estimated from the two types of addresses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paraquat</td>
</tr>
<tr>
<td>time period</td>
</tr>
<tr>
<td>1974-1999</td>
</tr>
<tr>
<td>1980-1999</td>
</tr>
<tr>
<td>1985-1999</td>
</tr>
<tr>
<td>1990-1999</td>
</tr>
<tr>
<td>1995-1999</td>
</tr>
<tr>
<td>Ziram</td>
</tr>
<tr>
<td>time period</td>
</tr>
<tr>
<td>1974-1999</td>
</tr>
<tr>
<td>1980-1999</td>
</tr>
<tr>
<td>1985-1999</td>
</tr>
<tr>
<td>1990-1999</td>
</tr>
<tr>
<td>1995-1999</td>
</tr>
<tr>
<td>Maneb</td>
</tr>
<tr>
<td>time period</td>
</tr>
<tr>
<td>1974-1999</td>
</tr>
<tr>
<td>1980-1999</td>
</tr>
<tr>
<td>1985-1999</td>
</tr>
<tr>
<td>1990-1999</td>
</tr>
<tr>
<td>1995-1999</td>
</tr>
<tr>
<td>Chlorpyrifos</td>
</tr>
<tr>
<td>time period</td>
</tr>
<tr>
<td>1974-1999</td>
</tr>
<tr>
<td>1980-1999</td>
</tr>
<tr>
<td>1985-1999</td>
</tr>
<tr>
<td>1990-1999</td>
</tr>
<tr>
<td>1995-1999</td>
</tr>
<tr>
<td>Diazinon</td>
</tr>
<tr>
<td>time period</td>
</tr>
<tr>
<td>1974-1999</td>
</tr>
<tr>
<td>1980-1999</td>
</tr>
</tbody>
</table>
4.5. DISCUSSION

Studies using one residential address such as an address recorded at study entry or time of diagnosis as a proxy for residential history during a relevant period of time assume that residential mobility impacts the estimation of environmental only minimally. However, little is known in terms of how exposure misclassification may bias the estimation of health risks derived from such estimates. Here we illustrated the impact of exposure misclassification due to assumptions about residential stability on estimates of long-term (26 years) pesticide exposure and risk of Parkinson’s disease given temporal and spatial changes in agricultural pesticide use. We simulated scenarios in which detailed residential histories are not available such as studies that use cohort enrollment or registry or medical record data that only provides a single address. Pesticide exposures here were assessed using a unique record-based database (PUR) and a GIS-based approach but our results provide scenarios that also may be informative for studies of air pollution based on long-term monitoring or modeling data [112-113], ultraviolet [114], or radon exposures [115-117], for which temporal-spatial exposure models can be generated and used to generate individual level exposures based on one residential address at the time of diagnosis or baseline assessment only. We describe how estimated pesticide exposures would be misclassified by comparing the exposure prevalence, sensitivity/specificity, and their influence on odds ratio estimates in a case control study.
Specifically, we examined how pesticide exposure misclassification due to changes in residential addresses may influence effect estimates for long term-chronic exposures averaged over a 26-year period using the enrollment address only instead of a full address history in a fairly residentially stable population. Our PEG study recruited cases and controls in a population-based manner from an elderly population, thus, participants, on average, had been living at their enrollment address for 19 years. Given so little residential mobility, the enrollment address would be expected to provide a representative proxy for the complete residential history for the years of interest in our study. However, relying on limited address information such as the address collected at baseline in a cohort study with limited coverage of the exposure period of most interest or provided in a disease registry or medical record at time of diagnosis, one address may not be representative of exposures derived for the relevant period using an accurate residential history. This may depend on how stable the study population is, as well as how exposures are distributed temporally and spatially with regard to the residential locations of subjects and whether this differs for cases and controls. Bias introduced by exposure misclassification, therefore, may have unpredictable patterns and result in biased estimates with not only different magnitudes but also directions.

Pesticide use in the tri-county area has generally been increasing between 1974 and 1999 (Figure 4.1), and average exposure patterns in controls reflect the general agricultural application patterns well as expected when controls accurately represent the exposures in the population from which the cases arose. The exposure estimates derived with the GIS-based pesticide exposure assessment method we used depended not only on mobility but also on possible variation in time and space of pesticide applications in proximity to addresses at which subjects
lived at a given point in time. We would expect that exposure estimates based on an enrollment address only will have a larger measurement error if pesticide applications vary strongly spatially and temporarily or if the subjects are highly mobile.

Our comparisons suggest that the enrollment address may underestimate exposure for some pesticides (chlorpyrifos), but overestimate for others (maneb and diazinon), and provide inconsistent estimates for some (paraquat and ziram). It is worth noting that for the most recent period of 1995-1999 when participants were most likely to be living at the enrollment address, most annual average pesticide exposures were underestimated from the enrollment address only except for chlorpyrifos among controls where the exposure was overestimated. Since our study participants were on average 68 years of age when being enrolled, it is possible that some of them moved from previously more highly exposed addresses to one historically less exposed when they became older, e.g., from farms or margins of rural towns to more urban or inner town/city areas, as has been proposed as a likely urbanization pattern in the central valley area in the past decades [118].

Agricultural use of pesticides in Central California is widespread. When we used a restrictive definition of exposure, i.e. as the 4th quartile of the pesticide distribution in controls in 1995-1999, the exposure prevalence was lower overall and the difference between the two types of exposure assessment were also generally small. Importantly, sensitivity and specificity for exposure classification was not always the same for cases and controls such that exposure misclassification could be differential or non-differential depending on pesticide exposure definitions used.
In general, sensitivity was consistent across the periods but increased slightly in the more recent time periods and for high exposures sensitivity fluctuated more. Specificity was more consistent and generally higher than sensitivity with highest level exposures having the highest specificity (close to 1.0) for all five chemicals. Thus, both sensitivity and specificity are not only influenced by the choice of exposure threshold as expected, but also depend on the exposure periods of interest, and were in general was not related to disease status.

In terms of effect estimation to determine risk of PD, the bias introduced when using one address only for exposure assessment was not always in the same direction. Some ORs were positively biased away from the null likely due to differential overestimation of some exposures in cases compared with controls. However, for most pesticides the bias was either towards the null value or varied little over time. Generally we would expect exposure misclassification nondifferential due to the use of pesticide use records and addresses. One possible reason for this to not always be the case may be that by collapsing continuous exposure measures into dichotomous ones we created differential misclassification [119-120]. For the three types of exposure definition we chose, the direction of bias was indeed sensitive to the exposure definition. Again this might suggest that nondifferential measurement error turned into differential misclassification for some of the definitions. Second, nondifferentiality alone does not guarantee bias towards the null [23]. It can produce bias away from the null if the classification error depended on errors in other variables [121-122]. For example, it is possible that residential mobility varied between PD cases and controls, e.g., sicker individuals may move more readily to urban centers or depended on socioeconomic status, therefore, in our as well as other studies of elderly subjects
nondifferential misclassification may not guarantee bias towards the null in effect estimation for these pesticides and across time periods when using only one address to approximate exposures over lengthy periods of time.

Bias analysis has been proposed to quantitatively assess the direction and magnitude of potential bias when interpreting observed results [23, 123-124]. Incorporation of bias analysis for exposure misclassification as well as other systematic biases has been encouraged [125-128] and is being applied more often in epidemiological studies [129-133]. However, to assess bias in exposure misclassification we need estimates of sensitivity and specificity that are not easily available in most studies; thus it remains unclear whether differential or nondifferential misclassification is the more appropriate assumption to be applied in formal bias analyses. It has been shown that an incorrect assumption of nondifferential misclassification [119, 134] or differential misclassification [134] in a bias analysis can lead to more biased results than using unadjusted estimate. Therefore, it is important to perform internal validation studies, whenever possible, to estimate sensitivity and specificity that will facilitate bias analysis with correct assumptions. In our study, pesticide exposures estimated from both complete residential history and enrollment address were available, which allowed us to estimate sensitivities and specificities in our own study and evaluate potential misclassification patterns. In the current analysis, a quantitative bias analysis was not conducted to correct the bias due to the inconsistent patterns we observed for different pesticides and because we had gold standard exposure assessment available. In future studies that have to rely on only one address to assess pesticide exposures over lengthy periods of time, assumptions used in the bias analysis have to be checked carefully. Our results provide some guidance for bias analyses to assess exposure
misclassification due to residential mobility and/or varying pesticide use patterns based on single addresses.

One of the strengths of our study is that the complete residential address history was available for all enrolled participants; therefore the exposures at historical addresses could be estimated and employed as “gold standard”. The detailed residential history, allows us to assess the effects of exposure misclassification for different length periods and for varying pesticide use patterns over a 26-year period. Our GIS-based method identifies the type, amount, and location of application by chemical and eliminates differential recall of pesticide exposure according to disease status, and made it possible to assess exposure misclassification due to residential mobility and/or varying pesticide use patterns without the influence of recall bias from self-reports.

One limitation of our study is that the accuracy of our GIS-based pesticide exposure estimation relies on the quality of self-reported addresses. Our previous work showed that addresses with lower geocoding accuracy tended to be assigned lower exposures than accurately geocoded addresses [95]. Also exposure misclassification for occupational addresses were not able to be assessed in this analysis because (1) 26% of the participants were missing occupational address information; and (2) exposure estimates could only be obtained for participants with an occupational address located in the tri-county area between 1974 and 1999. In the present analysis we only investigated exposure misclassification due to residential mobility, however, there could be other sources of exposure misclassification or other types of bias, such as uncontrolled confounding and selection bias.
We found that the exposures could be either over- or under- estimated depending on pesticide, time period of interest, as well as threshold for identifying dichotomous exposures. We observed that for most pesticides odds ratios were indeed biased towards the null value due to a possible nondifferential misclassification pattern. Thus, while the enrollment address used as a proxy for the complete residential history to estimate long-term exposures may not always guarantee that exposure misclassification is nondifferential especially over longer periods of time, this assumptions seems to be appropriate most of the time. Also, studies that rely on limited address information to estimate long-term environmental exposures may want to consider conducting validation studies and examine sensitivity, specificity, as well as misclassification patterns at least for a sample of participants or conduct some quantitative bias analysis to estimate the size of the bias. If it becomes more common to estimate environmental exposures based on only one address and extrapolate into the distant past, we need to consider that even if we do not rely on self-report in exposures nondifferential misclassification due to residential mobility for time- and space-varying exposures may not always guarantee non-differentiality in the bias.
5. CONCLUSIONS

The purpose of this dissertation was to investigate two types of biases in population-based studies of Parkinson’s disease, one is survivor bias in estimating the association between Parkinson’s disease and cancer, the other is exposure misclassification due to residential mobility. In studies 1 and 2 we proposed two similar survivor bias mechanisms and investigated whether survivor bias could be a possible explanation for the reported inverse association with cancer both prior to and after the diagnosis of PD. We used a Danish population-based case-control study of Parkinson’s disease as an illustration. In study 3 we used a population-based case-control study of Parkinson’s disease in central California as an example and simulated a scenario where the detailed residential histories were lacking and only the enrollment address was used as a proxy for all addresses. We explore and describe how the estimated pesticide exposures would be misclassified by estimating the exposures only from the enrollment address.

Study 1 showed that negative associations with cancer prior to the diagnosis of PD not attributable to smoking can be explained by strong differential non-survival of cancer patients who later would develop PD. Study 2 showed that the observed association with cancer after a diagnosis of PD could be positively shifted away from the null value when adjusting for survivor bias using the inverse probability-of-censoring weights. These results suggest that for cancer both before and after the diagnosis of PD, survivor bias could be an alternative explanation for the observed association between cancer and PD with a reasonable bias structure and assumptions. Since the bias mechanisms we described are likely common in cohort and case-control studies in elderly populations who are at risk of death as a competing risk, survivor bias
should be taken into consideration as a possible alternative explanation when unexpected associations are observed between comorbidities in elderly population.

Study 3 showed that the exposures could be either over- or under-estimated depending on pesticide, time period of interest, as well as threshold for identifying the binary exposure variables. The exposure misclassification is not necessarily non-differential and the direction of the bias was inconsistent. When estimating environmental exposures based on only one address and extrapolating into the distant past, we need to consider that even if we do not rely on self-report in exposures nondifferential misclassification due to residential mobility for time- and space-varying exposures may not always guarantee non-differentiality in the bias.
6. REFERENCES


4. D'Amelio, M., et al., Tumor diagnosis preceding Parkinson's disease: a case-control study. Movement Disorders, 2004(0885-3185 (Print)).


8. Olsen, J.H., K. Friis S Fau - Frederiksen, and K. Frederiksen, Malignant melanoma and other types of cancer preceding Parkinson disease. Epidemiology, 2006(1044-3983 (Print)).


10. Garber, K., Parkinson's disease and cancer: the unexplored connection. Journal of the National Cancer Institute, 2010(1460-2105 (Electronic)).


15. Checkoway, H., et al., *Genetic polymorphisms in Parkinson's disease.* Neurotoxicology, 1998(0161-813X (Print)).

16. Garte, S., *Metabolic susceptibility genes as cancer risk factors: time for a reassessment?* Cancer Epidemiology, Biomarkers, and Prevention, 2001(1055-9965 (Print)).


30. Thompson, C.A. and O.A. Arah, Selection bias modeling using observed data augmented with imputed record-level probabilities. Annals of Epidemiology, 2014(1873-2585 (Electronic)).

31. Ritz, B., et al., Parkinson disease and smoking revisited: ease of quitting is an early sign of the disease. Neurology, 2014(1526-632X (Electronic)).


33. Storm, H.H., et al., The Danish Cancer Registry--history, content, quality and use. Danish Medical Bulletin, 1997(0907-8916 (Print)).

34. Greenland, S., J.M. Pearl J Fau - Robins, and J.M. Robins, Causal diagrams for epidemiologic research. Epidemiology, 1999(1044-3983 (Print)).


37. Thompson, C.A., O.A. Zhang Zf Fau - Arah, and O.A. Arah, Competing risk bias to explain the inverse relationship between smoking and malignant melanoma. European Journal of Epidemiology, 2013(1573-7284 (Electronic)).


83


47. D'Amelio, M., et al., Tumor diagnosis preceding Parkinson's disease: a case-control study. Movement Disorders, 2004(0885-3185 (Print)).


81. Costello S, Cockburn M, Bronstein J, Zhang X, Ritz B. Parkinson's disease and


124. Lash, TL.; Fox, MP.; Fink, AK. Applying Quantitative Bias Analysis to Epidemiologic Data. New York: Springer; 2009.


