Disparities in mammography utilization patterns according to geographic access and car ownership, and long term mammography utilization among ductal carcinoma in situ survivors

by Patricia Isabelle Jewett

A dissertation submitted in partial fullfillment of
the requirements for the degree of

Doctor of Philosophy
(Population Health Sciences)

at the
UNIVERSITY OF WISCONSIN-MADISON
2017

Date of final oral examination: 07/10/2017

The dissertation is approved by the following members of the Final Oral Committee:
Ronald E. Gangnon, Professor, Population Health Sciences, Biostatistics and Medical Informatics
Amy Trentham-Dietz, Professor, Population Health Sciences
Kristen Malecki, Assistant Professor, Population Health Sciences
Elizabeth A. Jacobs, Professor, Medicine, Population Health Sciences
James LaGro Jr., Professor, Urban and Regional Planning

Contents

Acknowledgements

I would like to thank the members of my dissertation committee for mentoring this work, for their ideas and their advice, and for being open to the changes that this project underwent along the way. Especially, I would like to thank Ron Gangnon, who has mentored me and from whom I have learned so much in these past five years, who never grew impatient with my many statistics questions, and who was confident in my abilities even at times when I was not. I am deeply grateful to have been given the opportunity to be a teaching assistant in Ron Gangnon's and also Yajuan Si's classes in the Population Health Sciences department, which not only was possibly the part of my time in graduate school that I enjoyed most, but also made me understand and appreciate the material more thoroughly. I also would like to particularly thank Amy Trentham-Dietz for her mentorship in these past five years, which taught me epidemiological rigor, and gave me an idea of what it means to work as a scientific leader apart from the things they teach you in graduate school. Kristen Malecki for making me think critically about the study populations I was working with, and for making me think about the broader context of my research, Elizabeth Jacobs for bringing a clinical perspective into my work, and James LaGro for providing insights and context from an urban planning and geographic perspective.

I also would like to thank Ron Gangnon, Amy Trentham-Dietz, and all the members in Amy's lab, Julie McGregor, Kathy Peck, John Hampton, Xiaowen Zhang, Oyewale Shianbola, Natalia Arroyo, and Heidi Sahel for a wonderful five years of working together. John for many-an-idea-and-advice in SAS; Julie for always being helpful with the administrative things; Kathy for the many conversations we have had; Heidi and Wale both for brightening up everyone's mood the moment you come into the office, and also for the refreshing discussions we have had on the not-in-every-way-such-great-politics in this country; Xiao for the laughs we shared and for her focused studiousness in the adjacent cubicle that made me feel less lonely when working very early mornings or late evenings; and all of you for being extremely tolerant when I so often brought my little girls into the office to get some work done while they napped. You all made my completing this work possible by never as much as raising an eyebrow when Judith or Ruthie came in. It may have seemed that I took that for granted, but I never did, and you have no idea how much that meant to me.

I would like my other co-authors: Andy Bersch in chapter 3, and John Hampton, Elena Elkin and Polly Newcomb in chapter 4. My work as outlined in chapters 4 and 5 was supported by the National Institutes of Health grants P30 CA014520 and R01 CA067264. With regards to my work as outlined in chapter 3, I would like to acknowledge support from the University of Wisconsin Survey Center as well as SHOW participants, staff scientists and faculty for their contributions to the program. Funding for SHOW comes from the Wisconsin Division of Public Health, the Wisconsin Partnership Program (PERC) Award (223 PRJ 25DJ),

the National Institutes of Health?s Clinical and Translational Science Award (5UL 1RR025011), and the National Heart Lung and Blood Institute (1 RC2 HL101468). Investigators are also funded on a core grant to the Center for Demography and Ecology at the University of Wisconsin-Madison (P2C HD047873), the National Center for Advancing Translational Science CTSA award (UL1TR000427), and National Institute for Minority Health and Health Disparities award (1P60MD0003428).

I want to thank Quinn Fullenkamp for his good-humored way of helping me navigate the bureaucracy of graduate school. I want to thank all my instructors and peers in the Population Health Sciences Department for being an inspiration. Thanks to Michael Gleicher for his inspiring class on data visualization, which made me wish I would use data visualizations more often than I do (I am trying). Thanks to Pravleen Bajwa and Unnur Gudnadottir for sharing their SHOW work spaces with me and putting up with my frequent email 'Is there a day when you will not be here this week for me to use your desktop?'. Thanks to Andy Bersch and Tammy LeCaire for providing me with the SHOW data needed for my aim 1 analysis.

Having children during graduate school helped me stay grounded and prevented me from ever being overwhelmed. But it would have been impossible to complete this work without the help of the many people who took over part of the child care - Talia Frolkis, Kim Lemmer, Lyndsay Evans-Eder, Joie Tinberg, Sagui Lutman, my parents in law Debby and Ted Jewett, Diane Bearman, and my mom Ilse Pohl.

I want to thank Dr. Axel Bronstert and Dr. Nick Priest, my advisers for my MS thesis in Potsdam 2004/2005. They both asked me if I wanted to do a PhD, which I was not ready for at the time. I do not regret delaying my PhD, getting a glimpse at a different professional world in the years between my German Diplom and my re-entering graduate school in 2012, but the fact that they expressed that I should do a PhD was an important factor in my decision process when I finally did apply for graduate school in 2011.

I have had amazing teachers throughout my life who gave me confidence in my abilities, and helped shape the values that underly the professional and non-professional decisions I have made. Probably the most important teacher I have had was my elementary school teacher Ilse Buck who laid the foundation for everything that came after. Having little children myself has made me appreciate what difference it makes to have teachers who are passionate and enthusiastic about their work, and who are able to ignite in their students a desire to keep learning and to change this world for the better.

Writing everything I would like to thank my friends for would fill another dissertation. Therefore I just want to thank you all for letting me be in your lives, Sabine Chlosta, my cousin Johanna Schwecke, Sagui Lutman, Ruede und Marianne Reme, Judith Moering, Doerte Felsing, Kirstin Werner, Yulia Grishchenko, Andrea and Clemens Werner, Katja Gehmlich and Stefan Krause, Christiane and Frank Ehlers, Nicole Roettmer, Rebecca and Jerome Tharaud, Talia Frolkis, Angelika Breinlich and Sebastian Schaefer, Kim Lemmer, and Teresa Cousins.

Finally, I want to thank the most important people in my life, my family:

My parents for their unwavering love and trust in me ever since I can remember. My mom Ilse Pohl, for her sense of justice and her need and strength to fight for it, and for teaching me that not everyone's opinion of me matters (a lesson, I believe, not taught enough to girls in our society); my dad Peter Pohl, for the calm and humble person he was, who never sought the limelight, but who was always recognized and respected for his knowledge and sharp and accurate judgement anyway.

My sister Nina Gillmann and her husband Jan-Philipp Gillmann and their four children Jule, Justus, Fini, and Jacob for making us part of their family in Hamburg and in New York. My sister Nina, who is so intelligent, educated, funny, and warm-hearted, for being a mensch (in the Yiddish sense of that term) who I have always looked up to.

I do not know where to start to thank my husband Ethan. The last five years were only possible because of you. For your intelligence (I still think you should be the one pursuing a PhD), your interest in whatever I was doing, your critical thinking and questions, your ideas and suggestions, your way of always getting the big picture, and also your flexibility in our ever-changing childcare arrangements and workarounds in these past years, and for the stability you have created for our family. For your love and for letting me love you.

To our beautiful daughters Ruthie and Judith. Nothing has made me as happy in life as you two have, and my and your papa's universes are centered around you. You are still so small, and yet, I can already see what strong, confident, curious, intelligent, funny, and caring persons you are. It makes me so proud to see you grow up and into your own personalities. I was naive to just assume that graduate school with children would work out just fine. It did - but only because I was lucky to have such unusually uncomplicated, resilient, and bright girls as you.

Abstract

Background: In the vehicle-centered US transportation system, limited geographic and transportation access can reinforce existing social disparities in health care utilization. One objective of this work was to identify subpopulations which are likely not to own a car, and to assess the associations of car ownership and geographic access with mammography and general preventive care utilization. The second objective was to describe long-term mammography utilization among ductal carcinoma in situ (DCIS) survivors, applying a novel regression methodology that incorporated distribution assumptions on aggregated mammogram counts.

Methods: Data for this work came from the 2014-15 Survey of the Health of Wisconsin, and the 1997-2007 Wisconsin Women's Health and Wisconsin In Situ Cohort Studies. We used logistic and proportional odds regression to model the odds of car ownership, and of more frequent mammography and general care utilization, with the main predictors: driving times to, and number of, mammography facilities near women's homes, and car ownership. We used Poisson regression to estimate long-term mammography utilization among DCIS survivors, with the main predictors: treatment, age at diagnosis, and family history of breast cancer. We used data visualizations to test hypotheses that we could not evaluate statistically.

Results: Young and old age, low incomes, and racial/ethnic minority status were associated with decreased odds of car ownership. Having at least one nearby mammography facility was associated with greater mammography frequency, but effects of driving times and car ownership on health care utilization were inconsistent or non-significant. Among DCIS survivors, we detected a statistically but not clinically significant decline in mammography rates with time. Younger age and treatment with lumpectomy/radiation were associated with more frequent imaging surveillance, but without clinically meaningful effects.

Conclusion: Geographic and transportation access may not be strong utilization predictors for the health care services we assessed, especially in the general population. Future research should focus on low income groups and racial/ethnic minorities and their access to daily life activities. Future research should also confirm that DCIS survivors are stable long-term mammography users as indicated by our findings. Our modeling approach that incorporated distribution assumption on aggregated data may have applications in other fields.

List of Figures

List of Tables

1 Introduction

1.1 Research questions and general objectives

Social disparities in population health and health care utilization in the US are well known, and there are many possible explanations why socially disadvantaged groups in the US use health care less frequently and less consistently than socially advantaged groups [2–7]. Among the many predictors of health care utilization, geographic access is one. For decades, transportation systems in the United States have been vehicle-centered for almost any kind of daily living activity. The inner cities of a few metropoles are the exception, such as New York City, Chicago, Boston, Washington D.C., and San Francisco, where ample public transportation options are available. In most other places in the US, it has been assumed that people have access to a car. While that assumption may be true for the vast majority of the US population, it is probably not true for everyone, and the question arises which subgroups are less likely to own a car, and how that affects their access to basic life activities such as medical care. Diminished transportation access to health care may be reflected in less frequent utilization of health care services relative to comparable population groups that have better transportation access; and it would be an indicator of how basic life opportunities in the US are not equally distributed across the population. The first objective of this work was therefore to describe transportation access in different population groups in the US, and to assess the question: how do geographic and transportation access relate to health care utilization, specifically: utilization of mammography screening services and general preventive care, as a potential indicator of unequal population health opportunities.

The second research question was: what are long-term mammography utilization behaviors in ductal carcinoma in situ (DCIS) patients. Mammography utilization research among breast cancer survivors has received less attention than screening mammography utilization among women without a previous breast cancer diagnosis. In part, this is because there is broad consensus about mammography surveillance recommendations for breast cancer survivors who are at increased risk for a secondary cancer in the ipsilateral, or a second primary cancer in the contralateral breast, whereas the discussion about the best mammography screening regimen among women with an average risk of breast cancer is more fraught with disagreement. Other authors have written about mammography surveillance based on study populations of breast cancer survivors of multiple cancer stages, but to my knowledge, the long-term mammography utilization among DCIS patients specifically has not been previously described in detail.

The third objective of this work was to explore a creative statistical methodology that incorporated specific assumptions into a traditional regression model. In my third aim, I used interval-censored data, i.e. mammogram counts over varying periods of time for which the precise timing of events was unknown. I made specific assumptions about how event rates would be distributed over time and incorporated these

assumptions into a Poisson regression model, and compared the results with a Poisson model of average annual utilization rates, using visualization techniques to verify my assumptions.

1.2 Literature Review

1.2.1 Frameworks and terminology

At the core of the motivations for this work lie conceptual frameworks from the public health and population health literature. Kindig [8] defines *population health* as "the health outcomes of a group of individuals, including the distribution of such outcomes within the group". As such, population health is "concerned with both the definition and measurement of health outcomes and the roles of determinants". Evans and Stoddart [9] provided a framework to describe the multiple determinants of health, including the social and the physical environments people live in, and the complex interactions between them. Drawing on that framework, Kindig emphasizes that population health pays "significant attention to the multiple determinants of such health outcomes, including medical care, public health interventions, aspects of the social environment (income, education, employment, social support, culture) and of the physical environment (urban design, clean air and water), genetics, and individual behavior" [8, 10]. Therefore, population health theory points to a complex interplay between individuals, social and physical environments, health behaviors, and health outcomes.

Health disparities are "systematic, plausibly avoidable health differences that affect socially disadvantaged groups such as groups defined by race/ethnicity, skin color, religion, nationality, socioeconomic status, gender, sexual orientation, gender identity, age, geography, disability, illness, political or other affiliation, or other characteristics associated with discrimination or marginalization and are thus a measure of social injustice" [11,12]. As such, the definition of health disparities goes beyond that of health inequalities, which merely describe differences in health without paying attention to how these differences are distributed in the population. Health disparities emphasize that health inequalities are not *fairly* distributed among advantaged vs. disadvantaged populations. *Health equity* is achieved if health disparities are minimized or absent [11,13], and it is the underlying goal for a society's commitment to eliminate health disparities. There are multiple theories of how health disparities emerge within a society. Some of them are [14]:

- Fundamental causes [15]: Gradients of health keep reemerging because population groups with higher incomes, education, better access to information, and more resources can engage in better health behaviors and access better care if unhealthy than population groups that are poor, less educated, and have less access and fewer resources. Originally developed to describe social differences in health, this concept is also applicable to other forms of disadvantage, e.g. to differences in health between different racial and ethnic groups.

- Pathway models: Without contradicting the fundamental cause model, these models shift the focus away from underlying "metamechanisms" towards possible mediating processes through which the underlying causes act [14]. Examples are life-course models that emphasize the cumulative hazardous effects of stressors over people's life course [16,17], embodiment theory, according to which social disadvantages set in motion biological processes that lead to differential health outcomes [18], and complex interaction models that incorporate different pathway models while emphasizing the interactions between individuals and their environments over time [14,19].

To apply these frameworks to the objectives of this work, chapters 2-4 focus on geographic access and car ownership as indicators of people's social and physical environments, and how these interact with the health care delivery system in the US. The social and physical environments are understood as fundamental causes, and the hypothesized effects of geographic access and car ownership on health care utilization are understood as pathways by which these social causes manifest themselves as utilization disparities. The underlying motivation in chapters 2-4 is hence to provide evidence for one possible pathway by which social disparities lead to disparities in population health outcomes. Chapter 5 is focused on identifying mammography utilization patterns among DCIS survivors, rather than trying to find possible explanations of utilization patterns in the social and physical environment. In addition to their descriptive purpose, the analyses in chapter 5 were motivated by exploring a novel approach to handle aggregated count data within statistical modeling.

1.2.2 Health disparities in the US and Wisconsin

Health disparities across different socioeconomic and racial and ethnic population groups persist in the US. Although health gaps by race/ethnicity in the US have narrowed in recent years, African American and Hispanic populations still score worse than average with regards to key health indicators such as infant mortality, childhood obesity, hypertension, cancer outcomes [2, 20, 21], and also with regards to preventive care, e.g. mammography screening rates [6, 7]. Income inequalities and differences in social mobility in relation to population health outcomes in the US have received attention in the past years [3, 22].

The County Health Rankings summarize population health disparities by states and counties [23]. In Wisconsin counties, 2015 childhood poverty ranged from 6-59%, unemployment from 4.6-14.1%, 2010-2012 violent crime rates from 28-800 (Number of reported violent crime offenses per 100,000 people), and 2011 obesity prevalence from 22-39%. Estimated mammography screening uptake (having had a mammogram in the past 2 years) among eligible women ranged from 54-79% in Wisconsin in 2014 [24]. The ratio of population to number of primary care physicians in Wisconsin ranged from 560:1 to 20,220:1 in 2014 [25].

Rural and urban areas in the US each face their own obstacles to health equity. As a result of

historical policies and urban developments in the US, middle and upper class residents moved from city centers to the suburbs in the past decades, resulting in impoverished inner cities. For example, according to the 2015 American Community Survey, the percentage of children under age 18 living below poverty level was 17.8% across Wisconsin, but 31.8% in Milwaukee county (5-year averages) [1]. Racial and ethnic minorities make up a higher than average proportion in underprivileged neighborhoods which often lack essential infrastructure. A recent report on racial disparities within Wisconsin Dane county found that "about half of the area's low income black households live in approximately 15 small [...] residential concentrations [...]"' that "'typically [...] do not include a church, a full service grocery, a public school, social or civic clubs, developed open spaces, a bar, restaurant or a significant employer" and are "thinly or unevenly served by the city and county's public transit systems" [26]. Shortcomings in urban infrastructures and lack of public transportation options create dependency on private vehicle use, which can exclude populations without access to a vehicle from basic life activities [27]. The share of the US population living in urban environments has been steadily growing, such that today, more than 80% of the US population live in urban areas [28]. Rural areas suffer from sparse access to medical or other services, and few employment options, making living in rural areas unattractive for healthier working-age population groups. As a result, rural populations are older, and in the case of Wisconsin are aging faster than the US on average, with potentially increased medical care needs while having decreased access to care [29].

1.2.3 The built environment, transportation, and public health

There are many ways how population health and the built environment are intertwined [30]. Public health concerns were at the roots of early urban planning policies: zoning was introduced with the goal to keep pollution out of residential areas and to prevent the spread of infectious diseases [31]. After the second world war, collaboration between public health and urban planning declined [32]. In recent times it has become evident that the built environment can impact population health via multiple pathways, bringing the disciplines back together [33]:

Modes of transportation. How people commute impacts on their engagement in physical activity with downstream effects on chronic diseases such as obesity, diabetes and chronic heart disease. Using alternative and more active forms of transportation than driving are associated with greater levels of physical activity [27, 31, 32, 34]. Increasing obesity trends in the US in the past decades have paralleled increasing vehicle use, and a similarly parallel recent obesity trend can be seen in other countries [35].

Air pollution. The World Health Organization estimated in 2014 that air pollution is the world's largest single environmental health risk [36]. Air pollution can increase risk of and worsen asthma. From 1980 until 1996, childhood asthma prevalence in the US increased 4.3% annually [37]. Major sources of

air pollution are ozone and fine particulate matter, which can be directly linked to vehicle use for mass transportation [27].

Collisions. Transportation behaviors influence the risk of traffic injuries [27]. While the absolute risk of accidents involving pedestrians and bicyclists may increase if more people walk and bike, it has also been shown that drivers are relatively less likely to hit pedestrians and cyclists (per number of people walking and biking) when the share of non-motorized transport increases, implying that awareness among drivers increases if more people bike and walk [38]. In 2000, the US fatality rates per 100 million km traveled were 14 among pedestrians and 7.2 among cyclists. In the Netherlands, those figures were 2.5 and 2.0, with biking generating 20 percent of total traffic volume [39].

Climate change. The built environment both affects climate change and mediates how climate change affects people. Globally, transport makes up 14% of all greenhouse gas emissions [40]. For the US, the respective figure is even higher, i.e. 27% [41]. Climate change has multiple public health impacts via weather extremes, expansion of vector-, food-, and waterborne diseases, to name just a few. Urban dwellers will be more heavily affected by rising temperatures because of the heat-island effect [42].

Access to life opportunities. Access to infrastructure has an ethical dimension. Infrastructure influences how easily people can access jobs [43,44], run errands, engage in recreational activities and obtain medical and civic services. The above cited Dane County report found disproportionate distance to key civic institutions and resources, lack of infrastructure, and lower than average mobility in underprivileged neighborhoods to be possible root causes of racial disparities [26]. 'Spatial Mismatch' in Milwaukee has separated residential areas from employment opportunities, disproportionally affecting low-skill workers and racial minorities [45,46]. Access to healthy foods can also be affected by shortcomings in the infrastructure. In many places in the US, residents have to travel long distances for access to sources of healthy food [47,48]. Such 'food deserts' have been shown to be associated with unhealthy dietary habits [49–51].

1.2.4 Car ownership

Often used interchangeably, 'mobility' and 'accessibility' describe distinct albeit related concepts. Mobility describes the ability to travel, whereas accessibility describes the ease to reach destinations, which encompasses the number of different travel choices to get there [52]. Vehicle-centered transportation policies in the US have historically focused on mobility rather than accessibility. Ideally, an increase in mobility results in an increase in accessibility, but this need not be the case. Building more roads does not increase accessibility among people who do not have a car. Creating infrastructure within walking distance can increase accessibility without changing mobility. Therefore, planning for good mobility is "neither a sufficient nor a necessary condition for good accessibility" [52]. Vice versa, planning for accessibility may not result in reduced ve-

hicle use. Car ownership in the US has historically been high, and elasticity of demand in mobility seems limited [52, 53]: even among low-income groups, driving is the primary mode of transportation [54, 55]. But some population groups are vulnerable to being negatively affected by policies that assume vehicle access, e.g. racial minorities [55, 56], the elderly [33, 57], and children. Additionally, changing consumption behaviors [56, 58] in the so-called 'millenials' generation (the generation born after 1980), and an aging population may necessitate a shift in transportation paradigms.

1.2.5 Health care utilization

Health care utilization in the US has long been subject of research, and a large array of demographic and social factors beyond health insurance have been found to be associated with disparities in health care utilization, such as age, sex, income, education, race/ethnicity, cultural and language barriers, having a usual health provider, and contextual factors, e.g. urbancity, health care policies, and geographic region [13, 59–64]. One conceptual framework that has been often applied in health care utilization research is the *behavioral model*, formulated by Andersen in the late 1960s [65]. Since then it has been developed further, incorporating more concepts and ideas as the field was advancing. The behavioral model assumes that health care utilization is motivated and determined by 'predisposing', 'enabling' and 'need' factors. 'Predisposing' factors are typically demographic characteristics such as race, sex, age, and health beliefs that are hard to change. 'Enabling' factors facilitate health care utilization, e.g. health insurance, income, but also the availability of medical services. An example of a 'need' is the presence or absence of chronic conditions. In later model versions, the behavioral model further distinguished between individual and contextual (e.g. determined by physical and social environments) predisposing, enabling an need factors. The model also differentiated between specific aspects of access to care, i.e. between: potential access, defined by the presence of enabling resources; realized access, defined by the actual utilization of services; and equitable vs. inequitable access, defined as situations when need or demographic factors vs. social structure, health beliefs or enabling resources are dominant in determining health care utilization. Within the behavioral model framework, this work assesses geographic access and car ownership as enabling factors that are hypothesized to influence potential and (in)equitable access.

1.2.6 Mammography utilization among women without a breast cancer diagnosis

Mammography screening intends to detect breast cancer early to improve survival. The US Preventive Services Task Force recommends that women aged 50 to 74 years use screening mammography every two years, with individualized decision-making for women aged 40-49 according to individual context and values [66]. Uptake of mammography screening varies. Among the 72 counties in Wisconsin, adherence to biennial

mammography screening among eligible women ranged from 54% (Menominee County) to 79% (Washington County) in 2014. In Dane and Milwaukee County, respective estimated use was 75% and 69% (Percentage of female Medicare enrollees ages 67-69 who receive mammography screening) [24].

Previous studies have explored barriers to mammography screening in individual-focused behavioral models, emphasizing demographic factors such as age, education and income levels, cultural and language barriers, health beliefs, and fear of cancer [67–70]. These studies have contributed to a better understanding of individual and cultural behavioral barriers to screening, but neglected the possible role of spatial access. Recent studies found that in the 2000s, mammography use plateaued or declined [70, 71], and that mammography capacity, measured as the number of mammography machines per 10,000 women aged 40 and older, declined by 10% between 2000 and 2010 [71]. Simultaneously, women living in counties with few or no available mammography facilities were less likely to have received a mammogram recently [72].

1.2.7 Mammography surveillance after breast cancer diagnosis

Due to improvements in treatment and survival among breast cancer patients, there are now over 2 million breast cancer survivors in the US [73, 74]. Breast cancer survivors are at an increased risk of developing a secondary cancer in the ipsilateral breast (if no mastectomy was used after first diagnosis), or a second primary cancer in the contralateral breast compared with the risk of developing a primary breast cancer in the general population [75, 76]. There is broad consensus that regular mammography surveillance can help detect recurrent cancer early in order to reduce breast cancer mortality, and the current recommendation is that breast cancer survivors, including DCIS survivors, have a mammogram annually after initial treatment [77, 78]. Several studies provided evidence that use of surveillance mammography after a breast cancer diagnosis is indeed associated with reduced breast cancer mortality [79, 80].

Other studies assessed patterns and predictors of mammography surveillance after a breast cancer diagnosis [74, 81–84]. Predictors found to be associated with whether and how often breast cancer survivors utilize mammography include: age at diagnosis [74, 81–84], stage at diagnosis [81, 83], treatment regimen (surgery and radiotherapy) [81, 83, 84], a physician visit or having a usual provider [74, 82], type of health insurance [74], comorbidities [82], and in an unadjusted analysis race and marital status [84]. Some authors reported a trend towards decreasing mammography utilization with increasing time since diagnosis [74, 82]. None of the above studies focused on in situ patients specifically, but several studies found that stage 0 breast cancer was associated with more frequent mammography surveillance utilization compared with later stages [81, 83].

1.3 Specific Aims

In Chapter 2, I estimate the magnitude of delayed care due to lack of transportation in the US, and explore demographic predictors; this analysis is based on data from the 2015 National Health Interview Survey (NHIS).

In Chapter 3, I assess associations of demographic characteristics with car ownership, and relate car ownership to the utilization of preventive general medical services and mammography using a population-based cross-sectional survey. My hypotheses were:

- People who own or lease a car different demographically from people who do not own or lease a car.
- People who do not own or lease a car use general preventive care and mammography less frequently than people who own or lease a car.
- Car ownership is a mediator on the pathways of known predictors of general preventive care and mammography utilization.

In Chapter 4, I assess the associations between driving time and availability of mammography facilities near women's residences as predictors, and as outcome frequency of mammography screening among breast cancer-free women living in Wisconsin 1995-2007 who were eligible for preventive breast cancer screening. My hypotheses were:

- Longer driving time and fewer available facilities near women's place of residence are related to lower frequency of mammography screening use.
- The above suggested associations affect primarily low income and low education groups, and rural populations which are more affected by long driving distances and limited transportation options

In Chapter 5, I describe patterns in mammography surveillance utilization over time among DCIS survivors. My hypotheses were:

- The frequency of mammography surveillance decreases with increasing time since diagnosis.
- Mammography surveillance are not uniformly distributed throughout any given year, but are highest in and around the month of diagnosis in each year.
- Utilization rates differ by treatment regimen, family history of breast cancer, and age at diagnosis.

1.4 Contributions of the dissertation

Chapters 3 and 4 emphasize the social justice dimension of vehicle-centered transportation from a public health perspective: the US vehicle-centered transportation system may create barriers to basic life activities such as health care utilization for people with limited access to a car. Previous work has focused on other public health and environmental problems of vehicle-centered transportation systems, taking the perspective of people who have a choice to drive or not, or the perspective of people whose health is indirectly affected by other people's driving. Public health benefits of reduced vehicle use described in those former studies included increased physical activity and downstream positive health effects among people who drive less, reductions in greenhouse gas emissions and air pollution, and reduced traffic with potential reductions in commuting-related stress and in traffic-related morbidity and mortality, as outlined in chapter 1.2.4. This work, in contrast, takes the perspective of people for whom limited transportation is not a choice while living in a transportation culture that is extremely driving-focused. Hence, this work describes the US driving culture as one mechanism how existing social disparities in public health are reinforced through exclusion from transportation; a perspective which has been underemphasized in the literature.

Chapter 2 estimates the magnitude of delayed care due to lack of transportation using a US population-based survey, and it identifies vulnerable groups that are most likely affected by transportation limitations with regards to health care utilization. While it may not provide new insights into which groups are vulnerable beyond what has been previously reported by other authors, it estimates the problem on a US-wide scale both for the general population and for specific population groups, illustrating key differences by demographic characteristics, which sets the context for chapters 3 and 4.

Chapter 5 adds to the body of literature that monitors long term mammography utilization among breast cancer patients by specifically focusing on ductal carcinoma in situ patients: the existing literature has mostly focused on preventive mammography screening before a breast cancer diagnosis, or described mammography surveillance among survivors of breast cancer across multiple breast cancer stages. Some authors have reported that compared to later stage survivors, stage 0 breast cancer patients tend to use mammography more often, subsequently focussing their discussions on low-frequency mammography users with later stage disease. In our work, we assessed mammography surveillance behaviors among ductal carcinoma in situ patients exclusively.

This work also makes contributions in statistical data analysis methodology:

- While geographic access barriers have been explored in previous studies, possible nonlinear effects of driving times and of the spatial availability of health care facilities have not been consistently assessed. By using splines in our work to assess the effects of continuous predictors, we may have captured effects more accurately

compared with previous work.

- Chapter 5 explores a creative modeling strategy: we incorporated explicit modeling assumptions on the annual distribution pattern of mammography surveillance rates among breast cancer survivors into a Poisson regression model that was based on interval-censored data. While the data were aggregated counts over periods of time, incorporating a specific distribution assumption with time allowed for more specific estimates at specific time points.

- Chapters 4 and 5 demonstrate the use of visualizations where formal statistical testing was not possible either because of lack of statistical power or because of lack of an appropriate formal test, allowing for insights that would not otherwise have been possible.

While not one the primary foci of this dissertation, this work also contributes to the literature that monitors trends in car ownership over time. In the past decade, possible changes in car-ownership among people born after 1980 (so-called 'millenials') have been observed and discussed both in academic studies and the news [56, 58, 85, 86]. There seems to be a trend towards changing car ownership patterns in younger generations, but it is not clear where that trend is headed in the long term: do younger generations forego car ownership completely, or do they just delay getting a car? How prevalent is non-car ownership in different age groups? What role may the financial crisis in 2008/9 and the economic recovery since then have played? In addition, more car-sharing options and models have become available in recent years, e.g. Lyft, and Uber, which may be a substitute for owning a car for some. This work does not give a final answer to these questions, but it adds one cross-sectional snapshot to the body of literature in the field.

2 Description and estimated magnitude of delayed health care due to transportation barriers in the US

2.1 Abstract

Background: We wanted to describe transportation limitations as a risk factor for delaying health care, and estimate the magnitude of the problem of delayed health care because of transportation barriers in the US.
Methods: We used data from the 2015 National Health Interview Survey (NHIS) to estimate the population fraction who reported having delayed medical care due to lack of transportation in the past 12 months. Using logistic regression, we identified potentially vulnerable population groups who tended to delay care due to lack of transportation.
Results: Approximately 2% (95% CI 1.8-2.1) of all NHIS participants reported having delayed medical care in the past 12 months because of lack of transportation. Indicators of social disadvantage such as low incomes,

being of non-Hispanic other ethnicity or African American race, and low health status were associated with increased odds of reporting having delayed care due to lack of transportation in the past 12 months. We estimated that vulnerable racial/ethnic minority members with low income were more than ten times (95% CI range 7.9-23.6 times) as likely to report having delayed care for transportation reasons compared with similar non-Hispanic white individuals with high incomes and high education levels.

Conclusion: Our findings imply that social disadvantages are strongly associated with an increased risk of delaying health care due to lack of transportation. Whether transportation is likely to be a limiting factor for the utilization of health care strongly depends on the subpopulation an individual is part of.

2.2 Introduction

The analyses in Chapters 3 and 4, which assessed car ownership and geographic access in relation to health care utilization, were based on Wisconsin data. While these data may qualify to assess the hypothesized relative associations, they may not qualify to estimate the magnitude of the problem on a US-wide scale because the Wisconsin population differs demographically from the US population. For example, a greater than US average fraction of the Wisconsin population is non-Hispanic white (82.4% vs. 62.3%), a smaller fraction of Wisconsin families is poor (8.6% vs. 11.3%), and a greater Wisconsin population fraction lives in rural areas compared with the US (29.8% vs. 19.3%) [1].

The aim of this analysis was hence to provide the context for the subsequent analyses from a US-wide perspective. In order to answer the question: "Is delayed care due to transportation limitations a relevant population health problem at all?", we wanted to estimate the magnitude of delayed care due to lack of transportation in the US population. In an exploratory profile analysis, we additionally aimed to identify vulnerable groups who are more likely than others to self-report delaying care due to lack of transportation.

2.3 Methods

2.3.1 Data Source

Data for this analysis came from the 2015 National Health Interview Survey (NHIS). The NHIS is an annual cross-sectional household interview survey that collects data on health, socioeconomics, and demographics on families and individuals in the US. Subject eligibility, sampling frame, and data collection have been described in detail elsewhere [87]. In brief, the NHIS randomly samples from the current civilian noninstitutionalized US population at the time of interview, excluding patients in long-term care facilities, people on active duty with the Armed Forces, prisoners, and US nationals living abroad. The 2015 survey population included 33,673 adults, and 12,291 children. Members of the same family shared a common family identifier across

separate data files. For our analysis, we used data from the 2015 family, adult, child, and person files [88].

2.3.2 Measures

The outcome of interest in our analysis was whether or not participants self-reported having delayed medical care due to lack of transportation in the previous 12 months. Primary predictors of interest in this analysis were: education of the adult with the highest education in each family (no high school vs. high school vs. some college vs. academic associate degree vs. at least college degree), income to federal poverty level ratio (PIR) (<100% vs. 100-199% vs. 200-399% vs. ≥400%), race/ethnicity (Hispanic vs. non-Hispanic white vs. non-Hispanic African American vs. Asian vs. non-Hispanic other), self-reported health (excellent vs. very good vs. good vs. fair vs. poor), language of interview (English vs. Spanish vs. English and Spanish vs. other), having a usual place to go to when sick (yes vs. no vs. more than one place), sex, age at interview (years), and number of months without health insurance in the past 12 months.

2.3.3 Statistical Analysis

We imputed missing predictor values using the aregImpute() function from the R Hmisc package [89], generating 5 imputed datasets. We then ran a logistic regression using the svyglm() function from the R survey package [90], modeling the log odds of self-reported having delayed medical care due to lack of transportation in the past 12 months, while accounting for NHIS survey weights. The results from the 5 imputed datasets were combined using the R mitools package [91]. In addition to testing for nonlinear effects of age, we explored a potential interaction between PIR and race/ethnicity. We did not formally adjust for multiple hypotheses testing despite exploring multiple potential predictors, but incorporated the possibility of random statistical significance into our interpretation of the findings. We reported estimated odds ratios (OR), probabilities, and 95% confidence intervals (CI).

2.4 Results

The results of the profile analysis of delayed care are shown in Table 1. Population groups with an increased likelihood to self-report having delayed care due to lack of transportation were: non-Hispanic other (compared with non-Hispanic white, OR 2.20, 95% CI 1.34-3.62), non-Hispanic African American (compared with non-Hispanic white, OR 1.41, 95% CI 1.11-1.78), women (compared with men, OR 1.32, 95% CI 1.11-1.58), and people without a college degree (compared with at least a college degree, OR range 1.52-1.92, 95% CI range 1.12-2.63). People with decreasing incomes were consistently more likely to self-report having delayed care due to lack of transportation (compared with PIR >400%, OR range 2.05 - >9, 95% CI range 1.31

Table 1: Odds Ratios of reporting to have delayed care in the past 12 months due to lack of transportation, (N=45,964), National Health Interview Survey 2015

Parameter	**OR**	**95% CI**
Age, per 5 Years	1.00	0.97-1.02
Sex		
Male	0.76	0.63-0.90
Female	1.00 (Ref.)	
Race/Ethnicity		
Non-Hispanic White	1.00 (Ref.)	
Non-Hispanic African American	1.41	1.11-1.78
Hispanic	0.95	0.73-1.24
Asian	0.90	0.60-1.36
Non-Hispanic Other	2.20	1.34-3.62
Income to Federal Poverty Level Ratio		
<100%	9.10	5.84-14.19
100-199%	5.24	3.37-8.16
200-399%	2.05	1.31-3.22
≥400%	1.00 (Ref.)	
Degree of Family Member with		
Highest Level of Education		
No High School	1.92	1.40-2.63
High School	1.59	1.18-2.14
Some College	1.85	1.36-2.52
Academic Associate	1.52	1.12-2.06
At least College	1.00 (Ref.)	
Self-reported Health		
Excellent	1.00 (Ref.)	
Very Good	1.31	1.01-1.69
Good	2.20	1.72-2.83
Fair	4.31	3.23-5.75
Poor	7.28	5.14-10.33
Number of Months without		
Health Insurance (per Additional Month)	0.97	0.93-1.02
Having a Usual Place when Sick		
Yes	1.00 (Ref.)	
More than 1 Place	1.24	0.69-2.24
No	1.00	0.80-1.26
Language Used in Interview		
English	1.00 (Ref.)	
English and Spanish	1.00	0.62-1.61
Spanish	0.84	0.57-1.22
Other	0.38	0.16-0.94

Table 2: Estimated probabilities and 95% confidence intervals of having delayed care in the past 12 months due to lack of transportation for hypothetical individuals. The first 3 individuals refer to 50 year-old females with some college education, an income between 100-199% the federal poverty level, fair self-reported health, who did the interview in English, reported having a usual place to go to when sick, and who had continuous health insurance coverage in the past 12 months, from varying racial/ethnic groups. The fourth individual is a comparable white woman with a college degree and an income ≥400% of the federal poverty level. (N=45,964), National Health Interview Survey 2015

Individual	**Estimated Probability [%]**	**95% CI**
African American	10.6	7.9-13.9%
Non-Hispanic Other	15.7	10.1-23.6%
Non-Hispanic White	7.8	6.1-9.9%
Non-Hispanic White, Higher Education and Income	0.9	0.5-1.3%

- >14), and a similar consistent pattern was found with worsening self-reported health status (compared with 'excellent health', OR range 1.31- >7, 95% CI range 1.01- >10). People who were interviewed using a different language than English or Spanish were less likely to report having delayed care due to lack of transportation (compared with English, OR 0.38, 95% CI 0.16-0.94). If we had used a Bonferroni adjustment for our multiple predictors, the comparisons of no high school vs. college degree, all PIR levels vs. >400%, Non-Hispanic other vs. white, and good/fair/poor health vs. excellent health, and sex would have remained significant. Overall, 2.0% (95% CI 1.8-2.1%) of all participants reported having delayed care due to lack of transportation (aged 18 and older: 2.1%, 95% CI 1.9-2.3%; aged 17 and younger: 1.6%, 95% CI 1.4-1.9%).

Table 2 shows the estimated probabilities for individuals who represent critical subpopulations based on our estimates (numbers shown are for a female, 50 year-old African American with some college education, an income between 100-199% the federal poverty level, fair self-reported health, who did the interview in English, reported having a usual place to go to when sick, and who had continuous health insurance coverage in the past 12 months; compared with a similar Non-Hispanic other and a similar Non-Hispanic white individual, and compared with a similar middle class white individual with higher income and education levels (income above 400% of the federal poverty level, and with at least a college degree). While among middle-class Non-Hispanic whites the probability of delaying care for transportation reasons may be negligible, this is not the case among low-income African American or Non-Hispanic other subpopulations, among whom a non-negligible minority (up to 10-15%) may delay care for transportation reasons.

2.5 Discussion

Approximately 2% (95% CI 1.8-2.1) of all NHIS participants reported having delayed medical care in the past 12 months because of lack of transportation. Among children, that number was slightly smaller (1.6%, 95% CI 1.4-1.9). Low incomes, minority status (Non-Hispanic other or African American), education levels lower than a college degree, being female, and low health status were associated with increased odds of having delayed care due to lack of transportation in the past 12 months.

Our analyses indicate that the issue of having limited access to care because of transportation limitations may only affect a tiny fraction of the overall US population. This is not surprising, since previous authors have reported that the overwhelming majority of US residents own at least one car across multiple income strata [55]. However, our analyses also emphasize that there are large differences in the likelihood to delay care by population subgroups. Our estimates in Table 2 demonstrate that middle class white individuals are highly unlikely to report having delayed care due to lack of transportation; but this possibility becomes a lot more likely for low-income racial/ethnic minority groups.

This study was an exploratory analysis and has limitations. Delaying care due to lack of transportation was self-reported, and it was not specified what kind of care was delayed and for how long. There may be other predictors of whether or not people delay care due to transportation reasons which we did not account for, the most obvious factor being whether or not people had access to a car, of which we had no measure in NHIS. We also had no measure of urbanicity of participants, but there are likely differences in transportation access between rural and urban settings. Although NHIS oversamples minority populations, it is still possible that very poor populations and minorities were more likely to refuse participation. Given our finding that these population groups are more likely to delay care due to lack of transportation, our estimates may be conservative.

2.6 Conclusion

Within most well-off, middle class population groups, transportation as a limiting factor for health care utilization may not seem like a problem of population health interest. However, from the perspective of a less well-off minority subpopulation, this question may be judged differently. The large relative differences in the probability of delaying health care because of transportation limitations in different subpopulations set the framework for the subsequent analyses.

3 Differences in demographics and in preventive general care and mammography utilizations by car ownership in Wisconsin

3.1 Abstract

Background: Transportation systems in the US have been vehicle-centered for decades, creating great dependency on private vehicle use, and possibly creating unequal access to life activities among people who do not own a car. We wanted to identify population groups that are less likely to own a car, and to assess whether car ownership was associated with and potentially mediated the effects of known predictors of general preventive care and mammography utilization.

Methods: We used data from the 2014-15 Survey of the Health of Wisconsin, a population based household examination study. We used logistic regression and visualizations to describe demographic differences between car owners and non-car owners in Wisconsin. We used logistic and proportional odds regression to assess whether car ownership was associated with and mediated the effect of known predictors on the frequency of general preventive care and mammography screening utilization.

Results: Young and old age were strongly associated with lack of car ownership (35 vs. 25 years, OR of owning a car vs. no car 3.42, 95% CI 1.98-5.88; 75 vs. 50 years OR 0.35, 95% CI 0.20-0.61). Incomes above the federal poverty level increased the likelihood that people owned a car (poverty income ratio PIR 300% vs. 100% OR 4.04, 95% CI 2.35-6.93), but with higher incomes the effect plateaued (PIR 800% vs. 500% OR 1.15, 95% CI 0.70-1.88). People without a car had lower education levels, and were less likely to have health insurance. Compared with low-income non-Hispanic Whites, low-income racial/ethnic minorities were less likely to own a car. We did not find evidence that car ownership was associated with, or mediated the effect of known predictors of general care or mammography utilization.

Conclusion: The nonlinear relationships between age, income and car ownership we found confirmed previous work in the field. Our visualizations highlighted a potential interaction between low income and minority status with regards to car ownership, which lack of statistical power may have hidden otherwise. As health care services that are rarely used, general preventive care and mammography utilization may not be much affected by car ownership either in the general population or among low-income groups. Future research should focus on vulnerable population groups, i.e. low income groups, racial/ethnic minorities, and young and old age groups, and investigate access to other medical and non-medical life activities, e.g. utilization of dental care and extracurricular activities among children. Further research should also consider the role that transportation policies may play in reducing potential barriers and mitigating effects of non-car ownership in this population.

3.2 Introduction

Transportation systems in the US have been vehicle-centered for decades. There are only few cities in the US with a dense public transportation network to provide access to everyday activities, such as the inner cities of New York City, Boston, Washington D.C., or San Francisco. Most Americans live in places where public transportation is unavailable, insufficient, or unreliable [27,92]. US transportation policies in the 20th century supported increased use of vehicles and federal highway and transportation system that promoted the use of private vehicles over other modes of public transportation [93, 94]. Also, compared with other industrialized countries in Europe, the US are more spacious, and population density is low. For example, in Germany, approximately 80 million people live in an area of 357,000 square kilometers, with a population density of 231.5 person per square kilometer [95], compared with 324 million people living in an area of 9,833,000 square kilometers, with a population density of 35.4 persons per square kilometer in the US [96]. As a result, transportation is heavily dependent on private vehicle use in the US [27,97] which was reflected in increases in per capita car ownership and miles traveled throughout the 20th century [98].

Vehicle-centered transportation has documented detriments to population health. Vehicle use is at the expense of transportation related physical activity [34, 99, 100]. These changes have been occurring at the same time that jobs have become increasingly sedentary. Sedentary behavior is associated with adverse health risks including cardiovascular disease and cancer [27, 101, 102]. Beyond individual health benefits, there are other societal and public health benefits from reduced vehicle use: greenhouse gas emissions, air pollution, and injuries from car accidents would decrease [36, 41, 103]; the economy would be less dependent on gas prices, private households could reduce travel costs, and there would be less traffic congestion in and around cities [53, 104]. Mediated through all of the above, reduced vehicle use could also improve people's quality of life [105].

Vehicle-centered transportation systems also have both social and environmental justice implications. They limit people's mobility if there are no transportation alternatives, so people with or without a car do not share equal access to life opportunities, reinforcing already existing social disadvantages among vulnerable population groups. For example, car ownership has been related to greater employment and higher earnings in the US1 [43, 106, 107]. The relationship may be reciprocal: people without a job may not be able to afford a car, but people without a car may also be limited in where they can apply for work. A previous study indicated that owning at least one car per household was common in the US across most income strata [55]. Only at very low incomes did car ownership drop, indicating that in the US, car ownership was nonlinearly related to income. But since the 1990s, the percentage of driver's license holders in the population has stagnated or even decreased, as well as miles traveled and car ownership, especially among young

adults in the US [108–111]. This trend may have been reinforced by the 2008 financial crisis [112–114], but it is not yet clear whether this trend is continuing or reversed, or whether car ownership has since stabilized at a new level.

One way how social disadvantages manifest themselves is through social disparities in health care utilization. Social determinants of health car utilization that have been previously identified include income, education, cultural and language barriers, and health beliefs [13, 59–62, 65, 115–117]. Transportation barriers have also been identified as social determinants influencing access to care among low-income populations in the US [63, 64, 118–121]. In our analysis, we focused on the relationship between limited transportation and health care access, with our main research questions: which population groups are likely not to own a car, and do these groups tend to delay medical care? We hypothesized that car ownership would be associated with, and mediate the effects of known predictors of health care utilization, focusing on preventive general care and mammography utilization as two examples of care.

To identify population groups that are most likely not to own a car, we conducted a profile analysis describing demographic differences between car owners and non-car owners in Wisconsin. We then assessed whether car ownership was associated with and mediated the association between other known predictors and the use of routine primary preventive care among Wisconsin adults, and the use of mammography services among eligible adult Wisconsin women.

3.3 Methods

3.3.1 Study Population

Data for our analyses came from the 2014-2015 Survey of the Health of Wisconsin (SHOW) program. SHOW is an ongoing population based health examination survey since 2008. It gathers health-related data on population samples that are representative of Wisconsin residents on an annual (2008-2013) or tri-annual (2014-2016) basis. The SHOW protocol and its informed consent documents have been approved by the University of Wisconsin Health Sciences Institutional Review Board. The target population for the Survey of Health of Wisconsin (SHOW) is non-institutionalized, non-active duty Wisconsin residents. The selection process of participants into the 2008-2013 sample is described in detail elsewhere [122]. The sampling frame was modified in 2014 to include counties as part of the sampling frame and for annual representative samples to now become triannual samples. Briefly, participants are randomly selected in two stages: First, counties are randomly selected weighted by urbanicity and proportional to the size of the population to ensure that the sample is representative of the spatial and sociodemographic range of the Wisconsin population. Next, Census Block groups (CBGs) or clusters of CBGs are used as primary sampling units (PSUs) within counties.

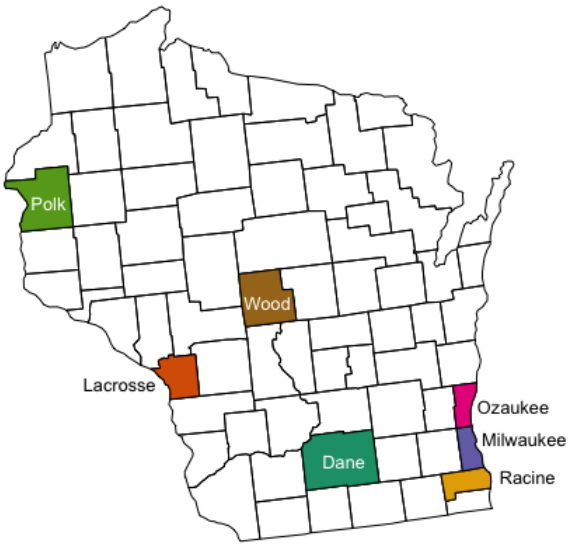

Figure 1: Surveyed counties, Survey of the Health of Wisconsin, 2014-2015

PSUs are also stratified by population size and the percentage of the population living below poverty line. Secondly, within each CBG, a list of household addresses is generated using United States Postal Service (USPS) Delivery Sequence Files purchased from MSG-Genesys (Marketing Systems Group, Fort Washington, PA). From these, household addresses are randomly selected per CBG using simple random sampling. After informing the household that they were selected for participation, study staff makes an in-person visit, with repeated attempts if there is no response. During the initial visit, all household members regardless of age are screened for study eligibility (non-institutionalized, with full cognition to independently complete consent), and subsequently invited to participate.

The 2014-2015 SHOW surveys used for this analysis included participants from Dane, Lacrosse, Milwaukee, Ozaukee, Racine, and Wood County, see Figure 1. Participation in the 2014 and 2015 surveys was 64% (N=531) and 60% (N=709), respectively. Among the total 1,240 participants aged 18 or older, three were excluded because of missing data on car ownership, resulting in a study sample of 1,237. For our analysis of routine primary preventive care utilization, we excluded 16 participants with missing data on time since utilization. For the analysis of mammography screening utilization, we restricted our study to women aged 50-74 for whom recommendations on mammography screening existed during that time, reducing the sample to 313 women. Among those, 18 women were ineligible because they reported a personal history of breast cancer, reducing our final sample size to 295 women.

3.3.2 Measures

Data collection in SHOW was divided over 3 time points: an in-home interview (Time 1), a self-administered questionnaire (Time 2), a physical exam and biospecimen collection in a mobile or fixed clinic, and more personal data collection (Time 3). Among the data collected, we used questionnaires on demographics, health history, and medical care utilization. In addition, we used contextual variables to describe the neighborhoods people lived in, i.e. at 2010 Census Tract level: median incomes ($), population fractions below the poverty line (%) and education levels (no high school degree vs. some college vs. at least a college degree) based on the 2015 American Community Survey [1]; and based on the geocoordinates of participants' residences: Transit, Walk and Bike Scores [123] where those were available.

The outcome in our profile analysis was whether or not a participant owned or leased a car (self-reported). Primary predictors in this analysis were: income to poverty ratio (PIR), calculated as the ratio of the midpoint of self-reported household income relative to the federal poverty level for a household of equivalent size in a given year, number of people living in household, years of formal education, sex, age (years), race/ethnicity (Non-Hispanic white vs. other), RUCA code (urban vs. rural), median income ($) in residential 2010 Census Tract, and having health insurance (yes/no).

To assess the potential mediating role of car ownership for use of preventive health care serves we examined self-reported time since an individual had last seen a doctor or health care provider for a routine physical exam, check-up or screening procedure. Time since utilization was reported as days, weeks, months, or years, and was subsequently transformed to time in days for each individual. Primary predictors in this model were: self-reported health status (excellent or very good vs. good vs. fair or poor), sex, age (years), race/ethnicity (Non-Hispanic White vs. Other), having health insurance (yes/no), education (no college degree vs. at least college degree), PIR, language spoken at home (English only vs. English and/or another language), and RUCA code (urban vs. suburban vs. rural). We also sought to look at potential mediating influence of car ownership on time since a woman had last had a mammogram, with time since utilization measured in years, and transformed into a numerical mammography frequency (one mammogram per n years). Primary predictors in this model were: age (years), race (Non-Hispanic white vs. other), having health insurance (yes/no), years of education, PIR, and RUCA code (urban vs. rural).

3.3.3 Missing Data Strategy

To account for missing covariate values, we used multiple imputation in the SAS 9.4 MI procedure [124]. We did not impute the main outcomes of interest of our subsequent regression analyses, i.e. car ownership, and time since last utilization of preventive general care or mammography. 16 SHOW participants were missing

data on time since last routine primary preventive care use. Among these, one person also had missing data on car ownership, and 6 people reported owning no car. Comparing the fractions of non-car owners in the entire sample (140 out of 1,237, or 11.3%) vs. among people with missing time since utilization (6 out of 16, or 37.5%) revealed a disproportionally high fraction of non-car owners among people with missing time data (χ^2 test, P-value=0.02), indicating that the time data might not be missing at random.

3.3.4 Statistical Analysis

To select our final regression models, we combined a conceptually driven modeling approach with statistical selection criteria. Based on our literature review [13,59–62,65,115–117], we identified the main outcome and covariates. Once we decided on the covariates to include in each model, we did not subsequently exclude them based on statistical criteria, but kept them in the model regardless of statistical significance. In our initial models, we included continuous variables as cubic splines. Splines allow for flexible nonlinear models, for example to represent plateau effects that are not easily captured by polynomial models. To keep the models as simple as possible, and to save degrees of freedom, we dropped nonlinear and interaction effects if these were not statistically significant (two-sided P-values greater than .05). We limited the splines to two nonlinear terms per continuous predictor (four spline knots), and we tested for nonlinear effects before testing for interaction because our degrees of freedom would not allow for testing for nonlinearity and interaction simultaneously. We also aggregated categorical predictors to save degrees of freedom. Our final models contained all initially identified potential covariates, and statistically significant cubic splines and interaction terms. We ran our regression models as multilevel models to account for potential similarities between individuals within residential neighborhoods, nesting by residential Census Tracts. We reported odds ratios (OR) and 95% confidence intervals (CI).

For our car ownership profile analysis, we tabulated demographic characteristics by car ownership, and visualized their distributions using violin and bar plots. We ran a multilevel logistic regression modeling the log odds of owning or leasing a car. In addition to testing for potential nonlinear effects of continuous variables, we explored potential interaction between PIR and 1) gender, 2) RUCA codes, and 3) race/ethnicity. We did not formally adjust for multiple hypotheses testing although we explored multiple plausible predictors of car ownership, but incorporated the possibility of random statistical significance into our interpretation of the findings.

To assess a potential mediating role of car ownership for the effects of known predictors of general preventive care, we carried out predictor selection using proportional odds regression based on the entire study population. However, we ran the main model as a logistic regression restricting to people with PIR <200% because car ownership was strongly confounded by income. For the mammography analysis, reduced

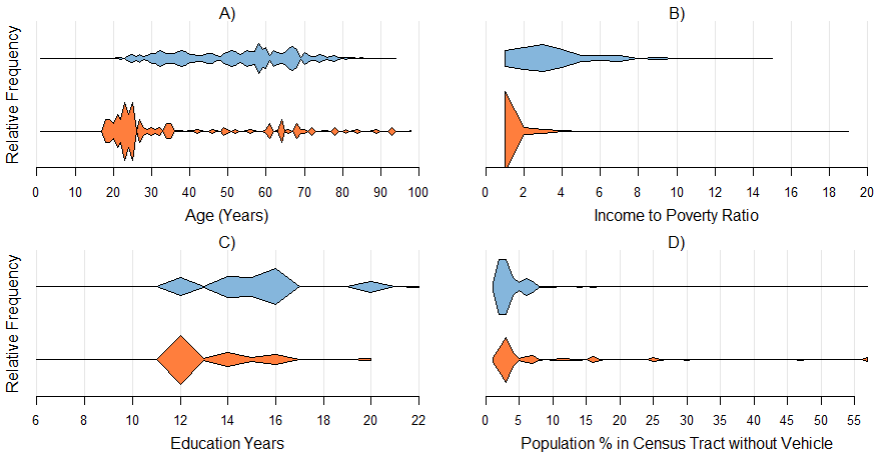

Figure 2: Distribution of baseline characteristics in car owners (blue) vs. non-car owners (orange), Survey of the Health of Wisconsin, 2014-2015. A) shows the age distribution, B) the income to poverty ratio distribution, C) distribution of education years, and D) the distribution of the population fraction without a vehicle in residential census tracts.

sample size led to convergence problems even on the full sample, and the even more reduced sample size in the model restricted to people with incomes <200% extremely limited our degrees of freedom. Therefore we used the same predictors in the mammography analysis as we did for general preventive care, after excluding sex, self-reported health, and language as predictors, and also health insurance because the latter led to complete separation problems in our model. The outcomes in the logistic models were: 1) having used primary care within the past two years, and 2) having had at least one mammogram in two years. We ran the final logistic models with and without an indicator of car ownership and evaluated changes in the effect sizes of the original predictors. As a sensitivity analysis, we also ran the final models as a proportional odds model without restricting to income for the entire study population. For our analyses across 5 imputed datasets, we used the SAS GLIMMIX and MIANALYZE procedures [125, 126], and the clmm2() and MICombine() functions from the R ordinal and mitools packages [91, 127].

3.4 Results

Table 3: Demographic baseline characteristics of participants (N=1,237) by car ownership, Survey of the Health of Wisconsin, 2014-2015

Characteristic	**Total % (N=1,237)**	**Car Owners % (N=1,097)**	**Non-Car Owners % (N=140)**
Individual Characteristics:			
Age, Years			
Min	18.0	18.0	18.0
Median	52.0	53.0	34.0

Continued from Previous Page	**Total %**	**Car Owners %**	**Non-Car Owners %**
Mean	51.3	52.3	43.4
Max	98.00	94.00	98.00
Gender			
Male	44.5	44.4	45.0
Female	55.5	55.6	55.0
Race/Ethnicity			
Non-Hispanic White	85.4	87.8	67.1
African American	4.9	3.2	18.6
Hispanic	4.0	3.4	8.6
Other	5.7	5.7	5.7
Income-to-Poverty-Line Ratio			
≤100%	8.3	5.5	30.7
>100-200%	18.8	16.9	33.6
>200-400%	31.3	32.7	20.0
>400%	41.6	44.9	15.7
Education			
No High School Degree	6.6	5.1	18.6
High School/Some College	54.8	54.1	60.7
College or Higher Degree	38.6	40.8	20.7
Type of Health Insurance			
Medicaid	11.0	9.1	25.7
Medicare/Private	84.7	87.8	60.7
None	4.3	3.1	13.6
Employment Status			
Working	59.8	61.1	50.0
Looking for Job	4.0	2.9	12.9
Not Working/Not Looking for Job	36.1	36.0	37.1
Marital Status			
Married/Living with Partner	66.7	71.6	27.9
Divorced/Separated	11.2	11.0	12.9
Widowed	6.5	5.9	10.7
Never Married	15.6	11.4	48.6
Household Size, N			
Min	1	1	1
Median	2	2	2
Mean	2.6	2.6	2.4
Max	11	11	8
Home Owner	76.7	81.3	40.7
Time since Last Use of Primary Care, Months	35.9	35.6	38.9
Time since Last Mammogram, years			
0 to 2	79.7	79.6	80.0
>2	13.9	14.2	10.0
Never	6.4	6.2	10.0
Neighborhood Characteristics:			
Urbanicity			
Urban	50.0	48.1	65.0
Suburban	17.4	18.0	12.1
Rural	32.6	33.8	22.9
Median income, $	57,334	58,342	49,434
Population Fraction in Census Tract below Poverty Line, %	7.8	6.9	14.5
Education			
No High School Degree	7.9	7.5	11

Continued from Previous Page	Total %	Car Owners %	Non-Car Owners %
High School/Some College	51.6	51.7	50.3
College or Higher Degree	40.5	40.7	38.6
Average Walk Score, 1-100	27.8	25.7	44.3
Average Bike Score, 1-100	71.8	70.1	82.4
Average Transit Score, 1-100	39.6	37.2	49.4

Table 3. shows selected characteristics at baseline of the study population, and Figure 2 shows violin plots of the distributions of age, education, PIR, and population fractions without a vehicle in residential Census Tracts, comparing car owners with non-car owners in our study population. Non-car owners were more likely to be categorized as having low socioeconomic status: compared with car-owners, non-car owners tended to be younger than 35 years or older than 60 years, while the majority of car owners was between 30 and 70 years old. Non-car-owners were more likely to be Hispanic (8.6% vs. 3.4%) or African American (18.6% vs. 3.2%), and with some exceptions tended to have low incomes (fraction with PIR <200%=64.3%), whereas car owners tended to have higher incomes (fraction with PIR <200%=22.4%). Fewer non-car owners than car-owners had more than 16 years of education, and a large group of non-car owners had fewer than 14 years of education, while among car-owners, this group was small. Non-car owners were more likely to be looking for a job (12.9% vs. 2.9%), more likely to have no health insurance (13.6% vs. 3.1%) or to be on Medicaid (25.7% vs. 9.1%), less likely to be married or living with a partner (27.9% vs. 71.6%) and more likely to be widowed (10.7% vs. 5.9%), and less likely to be home owners (40.7% vs. 81.3%). Non-car owners were also more likely to live in urban Census Tracts (65.0% vs. 48.1%), and the Census Tracts they lived in had lower median incomes ($49,434 vs. $58,342), larger population fractions below the poverty line (14.5% vs. 6.9%), slightly lower education levels (population fraction aged 25 and older without a high school degree 11.0% vs. 7.5%), and their immediate residential neighborhoods had higher Walk Scores (44.3 vs. 25.7), higher Bike Scores (82.4 vs. 70.1), and higher Transit Scores (49.4 vs. 37.2).

Figure 3 visualizes differences in car ownership among low income (<200% PIR) and higher income (≥200% PIR) populations by race/ethnicity. Regardless of income, car-ownership tended to be lower among Hispanics and African Americans, but the magnitude of the difference was largest when comparing low-income population groups. Among low-income non-Hispanic white and participants of other race/ethnicity, car ownership was still relatively common (78% out of N=240, and 80% out of N=24, respectively). Among low-income African Americans, only 49% (out of N=44) owned or leased a car, and among low-income Hispanics only 64% (out of N=25). This indicates that the relationship between income and car ownership may not be consistent across different subgroups.

The results of the logistic regression with car ownership as outcome (odds ratios (OR) and confidence

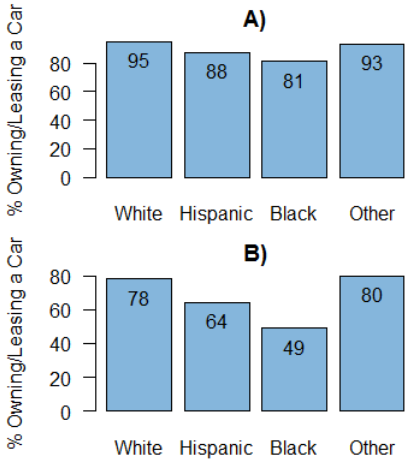

Figure 3: Car ownership among A) normal-high income groups (≥200% federal poverty level), and B) low income groups (<200% federal poverty level) by race/ethnicity, Survey of the Health of Wisconsin, 2014-2015

Table 4: Results of the profile analysis of car ownership: odds ratios (OR) of car ownership, all model covariates shown, (N=1,237), Survey of the Health of Wisconsin, 2014-2015

Parameter	**OR**	**95% CI**
Age (Years)		
35 vs. 25	3.42	1.98-5.88
50 vs. 35	3.38	2.04-5.58
75 vs. 50	0.35	0.20-0.61
Gender		
Male	0.97	0.62-1.53
Female	1 (Ref.)	
Race/Ethnicity		
Non-Hispanic White	1.50	0.80-2.81
Other	1 (Ref.)	
Income/Poverty Ratio		
200 vs. 100%	3.07	1.98-4.75
300 vs. 100%	4.04	2.35-6.93
500% vs. 200%	1.74	0.86-3.54
500% vs. 300%	1.33	0.92-1.90
800 vs. 500%	1.15	0.70-1.88
Years of Education (per 4 Years)	1.67	1.10-2.55
Health Insurance		
Yes	2.55	1.09-5.99
No	1 (Ref.)	
Urbanicity		
Urban	0.60	0.27-1.35
Rural	1 (Ref.)	
Median Census Tract income 2010, $		
25,000 vs. 15,000	2.09	0.84-5.18
50,000 vs. 25,000	3.20	1.53-6.72
75,000 vs. 50,000	0.86	0.50-1.47

intervals) are shown in Table 4. Age was significantly and nonlinearly associated with car ownership (overall significance, χ^2 test, P-value<0.0001): up to age 50, increasing age was associated with being more likely to own or lease a car (35 vs. 25 years, OR 3.42, 95% CI 1.98-5.88, and 50 vs. 35 years OR 3.38, 95% CI 2.04-5.58). At older ages, the trend was reversed (75 vs. 50 years OR 0.35, 95% CI 0.20-0.61). PIR was nonlinearly associated with car ownership (overall significance, χ^2 test, P-value<0.0001): being above the federal poverty level increased the likelihood that people owned or leased a car (PIR 200% vs. 100% OR 3.07, 95% CI 1.98-4.75, and PIR 300% vs. 100% OR 4.04, 95% CI 2.35-6.93), but with higher incomes the effect plateaued (PIR 500% vs. 300% OR 1.33, 95% CI 0.92-1.90, and PIR 800% vs. 500% OR 1.15, 95% CI 0.70-1.88). There was no evidence of a nonlinear effect of education, but linear effects of education years were significant (per 4 years, OR 1.67, 95% CI 1.10-2.55). Health insurance was also significantly associated with car ownership (insured vs. uninsured, OR 2.55, 95% CI 1.09-5.99). Increasing median income in 2010 Census Tracts were nonlinearly associated with increasing odds of car ownerships, but with plateauing effects similar to what we saw with PIR (overall significance, χ^2 test, P-value=0.007; \$50,000 vs. \$25,000, OR 3.20, 95% CI 1.53-6.72; \$75,000 vs. \$50,000, OR 0.86, 95% CI 0.50-1.47). Sex, race/ethnicity, and urbanicity were not significantly associated with car ownership in the multivariate analysis, but odds ratios tended to be above 1 when comparing Non-Hispanic Whites to other races/ethnicities, and rural residents with urban residents. Interaction terms between PIR and 1) RUCA codes (P-value=0.63) and 2) gender (P-value=0.74) and 3) race/ethnicity (P-value=0.41) were not statistically significant. Since we treated all covariates as main predictors, it is possible that some associations were significant by chance. If we had used a Bonferroni adjustment, the associations of age and PIR with car ownership would still have been significant.

The results of the logistic regressions of having used primary preventive care or mammography within the past 2 years are shown in Tables 5 and 6. Car ownership was not itself significantly associated with having used general preventive care in the past 2 years (which is a condition for potential mediation), and including car-ownership as potential mediator in the model for general preventive care did not change any of the relationships significantly, neither with regards to their statistical significance, nor with regards to effect size. In the logistic regression of having used mammography in the past 2 years, the association of car ownership was borderline significant (P=0.06), and there were some small changes in the effect sizes of most predictors. However, the results were counterintuitive: Owning a car reduced the estimated odds of having had a mammogram in the past 2 years (OR 0.19, 95% CI 0.03-1.07), and the estimated effects of the other predictors grew stronger after including car ownership in the model. Given these non-intuitive results, we looked in detail what caused these results in our data. We found that among 13 low-income women who did not own a car, 11 reported having had an annual mammogram. 7 out of these non-car owners with high mammography utilization lived in Milwaukee, 6 of them were African American, and one of them was

Table 5: Odds ratios (OR) of having used primary care within the past 2 years among low-income people (below 200% Federal Poverty Level), with and without car ownership as potential mediator, all model covariates shown, (N=328), Survey of the Health of Wisconsin, 2014-2015

	Model without Car Ownership		**Model with Car Ownership**	
Parameter	**OR**	**95% CI**	**OR**	**95% CI**
Age, per 5 Years	1.08	0.99-1.18	1.08	0.99-1.18
Gender				
Male	0.73	0.38-1.39	0.75	0.39-1.43
Female	1 (Ref.)		1 (Ref.)	
Race/Ethnicity				
Non-Hispanic White	0.93	0.41-2.13	0.93	0.41-2.13
Other	1 (Ref.)		1 (Ref.)	
Income/Poverty Ratio, per 100%	1.00	0.48-2.08	1.06	0.49-2.33
Education, per 4 Years	1.31	0.70-2.44	1.36	0.72-2.58
Self-reported Health				
Excellent/Very Good	1 (Ref.)		1 (Ref.)	
Good	0.72	0.34-1.53	0.70	0.33-1.50
Fair/Poor	1.31	0.43-4.00	1.33	0.42-4.17
Health Insurance				
Yes	5.39	2.18-13.35	6.09	2.38-15.60
No	1 (Ref.)		1 (Ref.)	
Primary Language Spoken at Home				
English	1.21	0.34-4.27	1.20	0.33-4.29
English and/or other	1 (Ref.)		1 (Ref.)	
Urbanicity				
Urban	1.22	0.57-2.61	1.20	0.54-2.65
Suburban	1.67	0.45-6.14	1.67	0.44-6.30
Rural	1 (Ref.)		1 (Ref.)	
Car Ownership				
Yes			1.15	0.5-2.66
No			1 (Ref.)	

Table 6: Odds ratios (OR) of having had at least one mammogram in two years among low-income women (below 200% Federal Poverty Level) aged 50-74 without personal history of breast cancer, with and without car ownership as potential mediator, all model covariates shown, (N=64), Survey of the Health of Wisconsin, 2014-2015

	Model without Car Ownership		**Model with Car Ownership**	
Parameter	**OR**	**95% CI**	**OR**	**95% CI**
Age, per 5 Years	1.32	0.86-2.05	1.36	0.87-2.14
Race/Ethnicity				
Non-Hispanic White	0.32	0.08-1.25	0.41	0.10-1.69
Other	1 (Ref.)		1 (Ref.)	
Income/Poverty Ratio, per 100%	1.13	0.34-3.78	1.29	0.36-4.71
Education, per 4 Years	0.97	0.37-2.54	1.20	0.42-3.38
Urbanicity				
Urban	0.72	0.24-2.17	0.60	0.19-1.87
Rural	1 (Ref.)		1 (Ref.)	
Car Ownership				
Yes			0.19	0.03-1.07
No			1 (Ref.)	

Hispanic.

As sensitivity analyses, we reran the models as proportional odds models for the entire study population (N=1,222 for the preventive care analysis, and N=295 for the mammography analysis). In these models, car ownership effects were not-significant and did not change any of the other predictors' effect sizes, see Tables 7 and 8.

Table 7: Odds ratios (OR) of longer time since primary care use (smaller OR signify a shorter amount of time since last utilization of care), with and without car ownership as potential mediator, all model covariates shown, (N=1,222), Survey of the Health of Wisconsin, 2014-2015

	Model without Car Ownership		**Model with Car Ownership**	
Parameter	**OR**	**95% CI**	**OR**	**95% CI**
Age, per 5 Years	0.90	0.87-1.00	0.80	0.64-1.00
Gender				
Male	1.14	0.94-1.39	1.14	0.93-1.38
Female	1 (Ref.)		1 (Ref.)	
Race/Ethnicity				
Non-Hispanic White	1.07	0.79-1.45	1.08	0.79-1.46
Other	1 (Ref.)		1 (Ref.)	
Income/Poverty Ratio				
200 vs. 100%	1.46 1.18-1.80	1.45	1.16-1.81	
300 vs. 100%	1.56	1.22-2.00	1.55	1.20-2.01
500% vs. 200%	1.13	0.87-1.45	1.13	0.87-1.45
500% vs. 300%	1.05	0.93-1.19	1.05	0.93-1.19
800 vs. 500%	0.98	0.83-1.17	0.98	0.83-1.17
Education, per 4 Years	1.12	0.98-1.29	1.12	0.97-1.29
Self-reported Health				
Excellent/Very Good	1 (Ref.)		1 (Ref.)	
Good	0.80	0.64-1.00	0.80	0.64-1.00
Fair/Poor	0.75	0.54-1.05	0.76	0.54-1.06
Health Insurance				
Yes	0.08	0.05-0.14	0.08	0.05-0.14
No	1 (Ref.)		1 (Ref.)	
Primary Language Spoken at Home				
English	1.05	0.69-1.58	1.05	0.69-1.59
English and/or other	1 (Ref.)		1 (Ref.)	
Urbanicity				
Urban	0.76	0.61-0.96	0.76	0.61-0.96
Suburban	0.94	0.70-1.26	0.94	0.70-1.26
Rural	1 (Ref.)		1 (Ref.)	
Car Ownership				
Yes			1.00	0.71-1.42
No			1 (Ref.)	

Table 8: Odds ratios (OR) of greater mammography frequency among women aged 50-74 without personal history of breast cancer, with and without car ownership as potential mediator, all model covariates shown, (N=295), Survey of the Health of Wisconsin, 2014-2015

	Model without Car Ownership		**Model with Car Ownership**	
Parameter	**OR**	**95% CI**	**OR**	**95% CI**
Age, per 5 Years	1.17	0.95-1.43	1.03	0.99-1.07
Race/Ethnicity				
Non-Hispanic White	0.74	0.29-1.87	0.85	0.33-2.17
Other	1 (Ref.)		1 (Ref.)	
Income/Poverty Ratio				
200 vs. 100%	1.70	0.88-3.28	1.83	0.94-3.51
300 vs. 100%	2.94	1.51-5.73	3.25	1.66-6.26
500% vs. 200%	3.01	1.42-6.39	3.16	1.50-6.51
500% vs. 300%	1.74	1.19-2.55	1.78	1.22-2.57
800 vs. 500%	1.02	0.65-1.60	1.02	0.66-1.58
Education, per 4 Years	1.25	0.77-2.02	1.30	0.80-2.11
Health Insurance				
Yes	9.53	3.12-29.16	10.08	3.30-30.77
No	1 (Ref.)		1 (Ref.)	
Urbanicity				
Urban	1 (Ref.)		1 (Ref.)	
Suburban	0.51	0.25-1.06	0.55	0.27-1.12
Rural	0.94	0.49-1.84	1.00	0.52-1.91
Car Ownership				
Yes			0.32	0.09-1.12
No			1 (Ref.)	

Table 9: Comparing characteristics in the study population to a) sampled counties, b) Wisconsin, and c) the US (based on American Community Survey 2015 [1])

Characteristic	**Study Population**	**US 2015 (%)**	**Wisconsin 2015 (%)**	**Weighted Average across Sampled Counties**
Total Population, N	1,240	316,515,000	5,742,000	1,983,000
Urbanicity				
Urban	67.4	80.7	70.2	90.2
Rural	32.6	19.3	29.8	9.8
Age Group				
<35	22.6	46.7	45.3	49.2
35-54	31.1	26.8	26.6	25.8
>54	46.2	26.4	28.1	25.0
Median Age	52.0	37.6	37.8	35.9
Race/Ethnicity				
Non-Hispanic White	85.4	62.3	82.4	68.6
Non-Hispanic Black	4.9	12.3	6.2	15.1
Non-Hispanic Other	4.0	8.4	5.1	6.4
Hispanic	5.7	17.1	6.3	9.9
Families with Incomes below Poverty Level in past 12 months	8.1	11.3	8.6	11.8
Median Income	67,500	66,011	68,064	67,487
Unemployment Rate	4.0	8.3	6.3	7.5
Education				
High School Degree or Higher	93.4	86.7	91	90.1
Bachelor's Degree or Higher	38.6	29.8	27.8	33.7
Transportation Mode for Work				
Car, Truck, Van, Alone		76.4	80.6	76.9
Car, Truck, Van, Carpooled		9.5	8.4	8.8
Public Transit		5.1	1.9	4.6
Walking		2.8	3.3	4.1
Other		1.8	1.7	2.1
Worked at Home		4.4	4.1	3.6

In order to capture demographic differences in our study population compared with the Wisconsin and US populations, Table 9 compares characteristics between the Wisconsin and US 2015 populations, and the population in the sampled counties for the 2014/15 SHOW survey and our study population. Compared with the US, the Wisconsin population was more rural, more likely to be Non-Hispanic White, slightly less likely to be poor, had a higher median income, was less likely to be unemployed, and less likely to use public transportation to commute to work. The population in the sampled counties from the SHOW 2014/15 surveys, however, included Dane and Milwaukee counties as the most urban counties in the state, with demographic distributions closer to the national average. Compared with Wisconsin, the sampled counties were highly urban, less likely to be white, younger, more likely to be poor and unemployed, but more likely to have at least a Bachelor's degree. Finally, the comparison between the study population and the expected population in the sampled counties indicated that not all population groups were equally likely to participate: People aged >54 were more likely to participate than younger age groups, non-white minorities were underrepresented in the sample, and the sample population had higher than average education levels and was less likely to be unemployed.

3.5 Discussion

Young and old age, low income, low education, and not having health insurance were associated with not owning/leasing a car. In an unadjusted comparison of low-income groups, African American and Hispanic participants were less likely to own or lease a car than low-income white or other participants. We did, however, not find evidence to support our hypothesis that owning a car was associated with or mediated the effects of known predictors of increased preventive primary care or mammography utilization.

Pucher [55] found that in the US, car ownership was common over a broad income spectrum and only declined at very low incomes. He also found that racial and ethnic minorities were more likely to use public transportation than whites, and that miles traveled continuously declined at ages older than 65. Our findings paralleled his: we found evidence of decreased odds of owning a car when comparing low to medium incomes, but non-significant odds ratios when comparing medium to high incomes; and we found that non-car ownership was associated with being Hispanic or African American in an unadjusted comparison, although race was not significantly associated with car ownership in a multivariate model. The reason why we found no association in the adjusted model may be due to lack of statistical power, or due to confounding effects, e.g. lower incomes among racial/ethnic minorities. We also found that being young or elderly was associated with not owning a car.

Our finding of a nonlinear relationship between age and car ownership (increasing odds of car

ownership up to age 50, but decreasing odds with older ages) may not be surprising, but may have relevant public health implications. Relative to the US, Wisconsin has an older population, and rural populations tend to be older than urban populations [29, 128]. Wisconsin is also disproportionally white, and white populations tend to have fewer children compared with other populations [129]. This means that in addition to being older already, the Wisconsin population will age more rapidly than the US [29]. Although car ownership is more common in rural than in urban areas, we saw that about 20% of non-car owners in our study population came from rural areas. Additionally, at older ages, some people may own a car but be reluctant to drive long distances to use health care services. Low rates of car ownership and use among older populations may aggravate the effect of already existing shortages of medical services in rural areas, increasingly limiting people's access to care as the population gets older.

We were not able to detect a measurable association between car ownership and time since general preventive, both when restricting to low-income groups and across our entire study population, which contradicts previous reports that cited transportation issues as a barrier to health care utilization [63,64,118–121]. One reason may be the outcome: General preventive care may be less likely to be delayed than other medical services, e.g. dental care, because it is not used regularly unless there is a specific medial reason for more frequent utilization. Another reason may be participation bias: racial/ethnic minorities and people with lower education levels and unemployed people were underrepresented in our sample. If the associations estimated in our propensity score model are correct, these groups are less likely than others to own a car. If they are also more likely to delay care, non-participation might have prevented us from estimating the true relationship between car ownership and health care utilization. There was also answer bias in our study: among 16 excluded people with missing responses on time since last general care utilization, a disproportionally high number reported owning no car. If the response missingness could be explained by very long time since utilization, we would have underestimated the effect of car ownership on general care utilization.

We had the counterintuitive finding that car ownership was associated with reduced odds of having had a mammogram in the past 2 years among low-income women. Given our small sample size (N=66), this finding needs to be interpreted with caution. The estimated relationship could mostly be traced to 7 women from Milwaukee, 6 of them African American, who owned no car but reported having had annual mammography. When looking at some of the other questionnaires answered by these women, they were disproportionally likely to have reported that access to public transportation was a very important consideration in their choice of residence, and they seemed to have strong ties with their residential neighborhoods. When comparing the unadjusted effect of car ownership on having had a mammogram in the past 2 years (owning vs. not owning a car OR 0.19, 95% CI 0.04-0.96) with a model that adjusted for an indicator variable for Milwaukee residence and an indicator of having reported that public transit was at least an 'important'

consideration when choosing the place of residence, the effect size of car ownership shrank considerably, and the confidence interval became much wider (owning vs. not owning a car OR 0.39, 95% CI 0.06-2.40), see appendix A. Our sample size is too small to allow for any statistical deductions based on these data, but it is possible that the low-income women who did not own a car but used mammography frequently either had transportation access through public transit or through their networks of friends, neighbors, or family. This would mean that our estimated association of car ownership in the mammography analysis would be confounded by transportation alternatives that are potentially more present in Milwaukee than in other places in Wisconsin. Another possibility is that there are more mammography facilities close to these women's homes in Milwaukee. Also, the National Breast and Cervical Cancer Early Detection Program provides free mammography screening for women who meet age, low-income, and health under-insurance criteria which could partially explain the high utilization frequency among these women [130].

Our study had strengths. Our analysis was based on a population-based sample. We included nonlinear effects of our continuous predictors, which allowed us to measure complex relationships more accurately, e.g. between age and car ownership, and income and car ownership. Our visualizations highlighted some findings that we would not have captured otherwise because of lack of statistical power: our violin plots emphasized the strong association of non-car ownership with incomes below 200% PIR, and our bar plots emphasized that among low income populations, racial and ethnic minorities may be less likely to own a car, making them more vulnerable to geographic access barriers. To classify urbanicity, we used RUCA codes whose definitions incorporate commuting behaviors, and were therefore more appropriate for our transportation-centered analyses than urban/rural Census classification measures that are based on population density and land use.

Our study also had limitations. Not owning a car was rare in our study population (140 people or 11.3% did not own or lease a car), and being non-White was also rare (1,056 or 85.4% of our study population were Non-Hispanic White). This forced us to run parsimonious regression models, and may have limited our ability to detect interaction effects in our multivariate analyses even though our visualizations indicated that car ownership by income was not consistent across racial and ethnical groups. We used car ownership in our analyses, but car ownership does not necessarily measure car access. With regards to health care utilization, car ownership may misclassify people who borrow friends' or family's cars to drive to medical appointments, or use ride services. This may have limited our ability to detect an effect of car access. It is likely that self-reported time since last general checkup / since last mammogram was subject to recall bias, leading to greater misclassification among people whose last utilization was longer ago. For example, we found that participants with missing data on time since last routine primary preventive care use were disproportionally likely to report owning no car. We did not include any formal adjustment to

account for multiple hypothesis testing in our multivariate models. For the multivariate logistic regression of car ownership, this is a valid concern. However, it is unlikely that all estimated significant relationships were simultaneously randomly significant; and age and PIR would have remained significant even with a Bonferroni adjustment. For the mediation analyses, chance significance was a lesser concern, as our primary interest was not to explore statistical significance, but differences in measured effects of known predictors of care utilization when including car ownership as potential mediator.

3.6 Conclusion

Our findings confirm previous work, indicating strong nonlinear relationships between age, income and car ownership. We may not have had the statistical power to detect significant effect modification, but our visualizations highlighted that low-income minorities may be less likely to own a car than low-income whites, making them potentially more vulnerable to transportation barriers. Future research should therefore focus on vulnerable groups, i.e. low income groups, racial/ethnic minorities, young and old age groups, and assess whether other types of medical care are likely to be delayed, e.g. dental care. Future research should also investigate potential associations of car access with utilization of non-medical life opportunities, for example extracurricular education activities among children, as these are also related to downstream population health outcomes. For Wisconsin policy makers, low levels of car ownership among the elderly should be incorporated into considerations that address aging populations in rural Wisconsin with growing medical needs. Lastly, the possibility that our estimated association of car ownership with mammography utilization among low-income women was confounded by better spatial availability of mammography services in metropolitan areas, or by transportation alternatives either through public transportation or individual networks should be subject of future research.

4 Geographic access to mammography facilities and frequency of mammography screening

4.1 Abstract

Background: Limited geographic access to mammography facilities may reduce women's mammography screening utilization. Women with low incomes and low education levels, and women living in rural areas may be most severely affected by geographic barriers.

Methods: We used interview data from the population based Wisconsin Women's Health study in 1995-2007 to assess variation of self-reported mammography use over the past 5 years. We used proportional

odds and logistic regression to test whether driving times to mammography facilities and the number of mammography facilities within a 10 km radius around women's homes were associated with mammography screening frequency among women aged 50-74, and whether associations differed between Rural Urban Commuting Areas and income and education groups.

Results: We found evidence for nonlinear relationships between facility density and driving times and mammography utilization (nonlinear effects of driving times and facility density, P-values 0.01 and 0.005, respectively). Having at least one nearby mammography facility was associated with greater mammography frequency among urban women (1 vs. 0 facilities, OR 1.26, 95% CI 1.09-1.47), with similar estimated effects among rural women. Adding more facilities had decreasing effects. Long driving times tended to be associated with lower mammography frequency, but when comparing moderate to short driving times, estimated effects were inconsistent. We found no effect modification by income, education, or urbanicity. In rural settings, non-use of mammography was higher, facility density was smaller, and driving times to facilities were longer.

Conclusion: Having at least one mammography facility near one's home may indeed increase mammography utilization, while adding more facilities may have decreasing marginal effects. While we could find no evidence of effect modification by income, education, or rural vs. urban settings, geographic access to mammography facilities was more restricted in rural areas.

4.2 Introduction

While there has been a recent controversy about benefits and harms, and optimal starting age and frequency of mammography screening [131–133], there is broad consensus that mammography screening can detect breast cancer early and reduce mortality [134]. Specific breast cancer screening recommendations have changed over time, but for women aged 50-74 mammography screening every 1-2 years has been consistently recommended in the US since the 1980s [66, 135, 136]. Screening uptake varies; for example racial minorities and low-socioeconomic status women use mammography screening less often than white, middle-class women [137, 138].

Since the 1990s, mammography utilization appears to have plateaued or even decreased [139, 140]. At the same time, there was a decrease in the number of mammography facilities, potentially reducing geographic access to mammography services [71]. Limited access to a car while having to travel large distances to few geographically available mammography facilities could restrict people's access to care. Urban sprawl and limited public transportation make many Americans dependent on personal automobile use [27, 141]. Among lowest income groups in the US, there is a disproportionate number of households without a car [55],

and those population groups may additionally have greater time restrictions, making them vulnerable to potential geographic barriers such as long driving times and few health care options.

In this analysis, we aimed to assess how driving times and geographic availability of mammography facilities relate to frequency of mammography screening among women without breast cancer, and whether these associations differ between rural and urban settings. We hypothesized that decreased geographic access would be related to lower frequency of mammography screening use, and that this relationship would be strongest for low income and low education groups. We additionally hypothesized that long driving times and fewer geographically available mammography facilities would impact rural populations more than urban populations.

4.3 Methods

4.3.1 Study Population

The Wisconsin Women's Health Study was a series of population based case control studies at the University of Wisconsin conducted between 1987 and 2007. Women with a personal history of breast cancer were the cases and women without the controls [142]. We only included controls in this analysis because our outcome was mammography screening frequency before any breast cancer diagnosis. We obtained informed consent from study participants, and the University of Wisconsin Health Sciences Institutional Review Board approved the study.

Subject eligibility has been described in detail elsewhere [142–145]. In total, 70.2% eligible Wisconsin residents enrolled in the study. Briefly, women randomly selected from lists of licensed drivers were eligible to participate as controls if they had no personal history of breast cancer, had a listed telephone number, and if they were able to complete a standardized telephone interview. We furthermore restricted our analysis to women aged 50-74 at baseline who participated between 1995-2007 because screening recommendations existed for this age group consistently during the study period, and because data on mammography facilities were available since 1995. During ongoing recruitment, not all questionnaire versions contained questions on income. Income is likely a confounder of geographic access in relation to mammography use. Therefore, we excluded women who were not asked about income at baseline.

In total, 6,075 women were eligible for our analysis. Of these, we excluded 145 women (2.4%) because of missing data on mammography use. Among the remaining 5,930 participants, we imputed main exposures and residential confounders of 38 women who could not be geocoded because their addresses were ambivalent, for example, P.O. boxes. Comparing baseline characteristics between the 145 women whom we excluded due to missing data and the women in the final sample, excluded women were older, more likely

Table 10: Baseline Characteristics among eligible vs. excluded women (N=6,075), Wisconsin Women's Health Study, 1995-2007

Characteristics	**Eligible Participants (N=6,075)** %	**Excluded Participants Due to Missing Mammography Data (N=145)** %
Age at Diagnosis, Years		
50-59	49.7	39.3
60-69	47.2	54.5
70-74	3.1	6.2
Race Ethnicity		
Non-Hispanic White	95.8	93.3
Other	4.2	6.7
Menopausal Status		
Premenopausal	13.1	4.7
Postmenopausal	86.9	95.3
Family History of Breast Cancer		
Yes	16.2	13.2
No	83.8	86.8
Postmenopausal Hormone Use		
Never	53.5	76.3
Ever	46.5	23.7
Income, $		
$\leq$30,000	32.9	34.7
30,001-50,000	29.7	34.7
50,001-100,000	28.3	26.3
>100,000	9.0	4.2
Education		
No High School Degree	7.4	11.8
High School Degree	42.9	48.4
Some College	26.6	28.0
At least College Degree	23.1	11.8
Urbanicity		
Urban	64.0	56.9
Rural	36.0	43.1

to have gone through menopause (95% vs. 87%), less likely to have used postmenopausal hormones (24% vs. 47%), less likely to have a family history of breast cancer (13% vs. 16%), less likely to be white (93% vs. 96%), less likely to have a college degree (12% vs. 23%), and more likely to have missing data across all variables (Table 10).

4.3.2 Data Collection

In telephone interviews, we collected information on family history of breast cancer, frequency of mammography screening, postmenopausal hormone use, race, education, income, and household size. Information on the number and locations of mammography facilities between 1995 and 2007 was obtained from the FDA which has maintained administrative records on certified mammography facilities in the US since 1994 [72]. Although all participants lived in Wisconsin, we accounted for mammography facilities in adjacent states close to the Wisconsin border, because some participants may have traveled to surrounding states for screening.

4.3.3 Geocoding

Methods for geocoding have been described elsewhere [72, 146]. After geocoding, each participant was assigned a corresponding 2000 Census Tracts, county and Rural Urban Commuting Area (RUCA) code [147]. Mammography facilities were geocoded by street address, using ArcGIS software (Version 9.2, ESRI, Redlands, CA). Mobile mammography facilities were assigned to the county of their mailing address [72].

4.3.4 Measures

The outcome in our analytical models was mammography frequency, measured as the self-reported number of screening mammograms in the past 5 years translated into an annual frequency. Primary exposures were driving times to nearby mammography facilities and the number of mammography facilities near a woman's home. Driving times were measured as the shortest driving time to a mammography facility near a woman's home and were determined in two steps: Using ArcGIS 10.0 Generate Near Table functionality, for each woman we identified the two closest (Euclidean distance) certified mammography facilities in the year of her study participation. Afterwards, we used the Googlemaps Distance Matrix API via the R-package httr [148] to determine for which of the two facilities driving time was the shortest. Googlemaps Distance Matrix API limits the number of requests to 2,500 per 24 hours. Therefore, driving times were calculated over 10 week days, each day around the same time to make traffic conditions comparable. Mammography facility density was defined as the number of certified mammography facilities within a 10 km radius around each

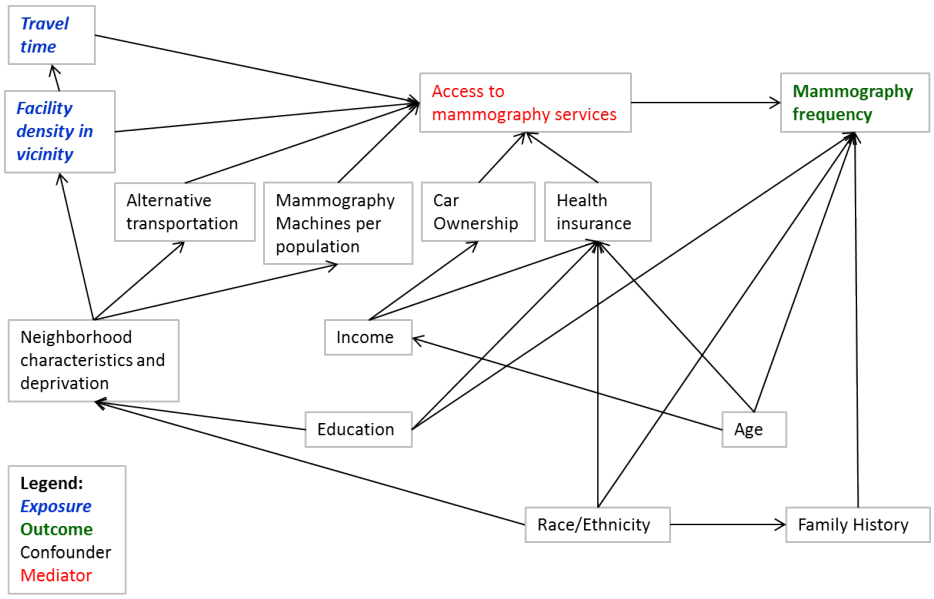

Figure 4: Conceptual model of the associations between diving times and mammography facility density with frequency of mammography screening utilization

woman's residential address in the year of her participation, and estimated using ArcGIS 10.0 buffer/intersect functionalities.

We used DAGitty [149] to draw our conceptual model as a directed acyclic graph (DAG) (Figure 4) in order to decide which other covariates to include in our analysis. DAGs are depictions of researchers' beliefs to qualitatively explain outcomes by potential determinants [150]. If a DAG correctly depicts the relationships, it identifies confounders and mediators, i.e. which variables to include and exclude from statistical models. Based on our DAG, we chose two confounder models. One included the minimal sufficient adjustment set for estimating the total effects of our exposures, which would sufficiently control for confounding assuming that our DAG correctly identified all relationships. The minimal sufficient adjustment set included as confounders: income ($), number of people per household, education (years), race (white vs. other), and mammography capacity, i.e. the number of mammography machines available per 10,000 women aged 40 and older in a county. The second model additionally adjusted for age (years), family history (yes/no), and indicators of neighborhood deprivation by Census Tract: population fraction below the poverty line (%), median 1999 income ($), population fraction without a vehicle (%), population fraction without a high school degree and with at least a college degree (%); and by County: the population fraction without health insurance (%).

4.3.5 Geospatial Analysis

To describe regional variation in mammography frequency throughout Wisconsin during the study period, we assessed the geospatial distribution of mammography screening frequencies across Wisconsin, using the R packages sp, gstat, spatstat, maps, and geoR [151–155]. Of the 5,930 participants in our analytical sample, 48 participants had coordinate duplicates because of rounding in the original geocoordinates. Of these, we randomly selected one of each duplicate pair (24 participants) and jittered their coordinates using the jitter() function in R [156]. We fit a semivariogram model to participants' self-reported mammography screening frequency, and used the fitted semivariogram in a kriging interpolation throughout Wisconsin.

4.3.6 Statistical Analysis

We used multiple imputation in the SAS MI procedure [124] to impute missing driving time and facility density for 38 women who could not be geocoded, and their Census Tract and County level confounders. We also imputed missing individual-level confounders, i.e. missing values of education, race, family history, and income.

To test the associations of driving time and mammography facility density with mammography frequency, we used ordered multinomial regression (proportional odds model) in the SAS 9.3 GLIMMIX and MIANALYZE procedures [125, 126]. We ran the regressions as multilevel models, grouping participants by residential 2000 Census Tracts to account for potential neighborhood clustering. We tested for nonlinear relationships between our main exposures and the outcome by including cubic splines of driving time and facility density. We tested for interaction between our main exposures and RUCA codes, income, and education levels, and re-ran the final model stratified by these same variables. Additionally, we ran a logistic regression, modeling the odds of having a mammography frequency greater or equal 0.5 (equivalent to at least one mammogram in two years) compared to a frequency below 0.5. Model selection was carried out on the proportional odds model; only the final model was also run as logistic regression.

4.4 Results

4.4.1 Descriptive Statistics

Table 11: Baseline characteristics of participants by mammography screening frequency, (N=5,930), Wisconsin Women's Health Study, 1995-2007

	Annual Frequency of Mammography Screening				
Characteristics	**Total** % **N**= 5930	**0 Times** % **N**= 596	**<0.5 Times** % **N**= 968	**0.5-0.9 Times** % **N**= 894	**≥1 Times** % **N**= 3472
Age at Diagnosis, Years					
50-59	49.9	46.8	50.5	55.6	48.8
60-69	47.0	48.2	45.4	40.2	49.0
70-74	3.1	5.0	4.1	4.3	2.1
Driving Time to Closest Mammography Facility, Minutes					
<5	27.4	27.5	27.5	25.6	27.8
5-10	31.7	30.2	31.2	31.8	32.1
11-20	29.7	29.7	28.8	30.8	29.6
>20	11.2	12.6	12.5	11.9	10.4
Number of Mammography Facilities within 10km Radius, N					
0	28.1	31.0	31.6	28.6	26.4
1-2	29.8	33.4	28.5	29.6	29.5
3	34.4	29.0	32.6	35.5	35.5
Number of Mammography Machines per 10,000 Women Aged 40+ in County, N					
0-1	24.6	29.2	26.0	26.0	23.0
2-3	42.9	43.3	42.4	42.8	43.0
>3	32.5	27.5	31.6	31.2	34.0
Race Ethnicity					
Non-Hispanic White	95.8	94.3	95.1	94.7	96.5
Other	4.2	5.7	4.9	5.3	3.5
Menopausal Status					
Premenopausal	14.2	16.4	18.2	18.6	11.6
Postmenopausal	85.8	83.6	81.8	81.4	88.4
Family History of Breast Cancer					
Yes	16.4	12.2	13.8	15.0	18.1
No	83.6	87.8	86.2	85.0	81.9
Postmenopausal Hormone Use					
Never	53.1	84.9	70.7	57.6	41.6

Continued from Previous Page	**Total** %	**0 Times** %	**<0.5 Times** %	**0.5-0.9 Times** %	**≥1 Times** %
Ever	46.9	15.1	29.3	42.4	58.4
Income, $					
≤30,000	33.0	53.5	41.4	32.0	27.4
30,001-50,000	30.2	25.8	29.2	29.1	31.5
50,001-100,000	27.9	16.9	24.4	27.3	30.8
>100,000	8.9	3.7	5.0	11.6	10.2
Education					
No High School Degree	7.3	15.4	8.4	7.5	5.6
High School Degree	42.8	49.8	42.7	38.9	42.6
Some College	26.6	23.3	29.0	28.4	26.0
College Degree	23.3	11.4	19.9	25.2	25.8
Urbanicity					
Urban	64.2	56.4	62.1	65.2	65.8
Rural	35.8	43.6	37.9	34.8	34.2

Table 12: Distributions of driving times and facility density, urban vs. rural (N=5,930), Wisconsin Women's Health Study, 1995-2007

Main Exposures	**Minimum**	**1st Quartile**	**Mean**	**Median**	**3rd Quartile**	**Maximum**
Number of Mammography Facilities within 10km Radius, N						
Urban	0	1	4	6.9	10	39
Rural	0	0	1	0.8	1	14
Driving Time to Closest Mammography Facility, Minutes						
Urban	0.02	4.4	6.9	8.4	11.4	37.4
Rural	0.2	5.8	12.2	13.7	19.7	138.8

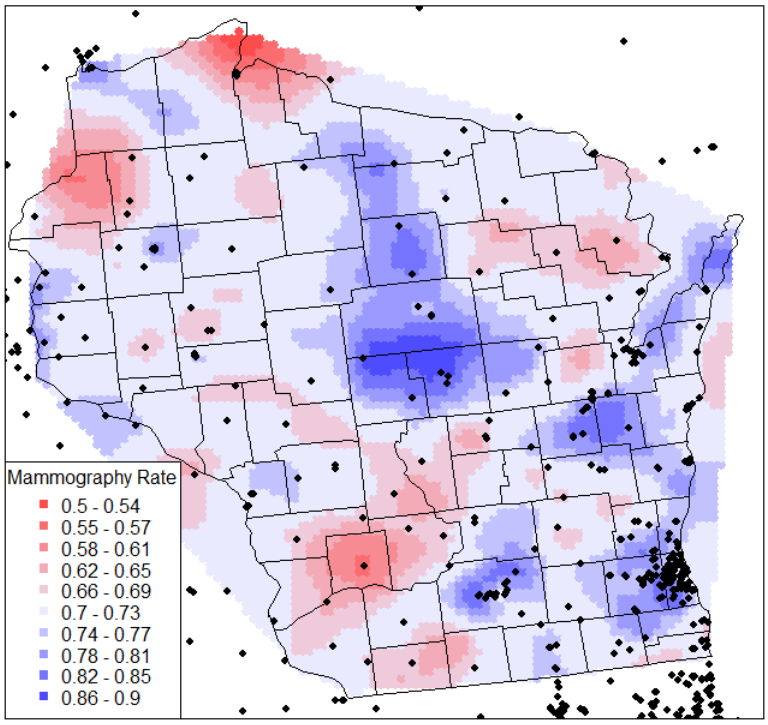

Figure 5: Location of mammography facilities and estimated spatial patterning of annual mammography utilization in Wisconsin, (N=5,930), Wisconsin Women's Health Study, 1995-2007

Table 11 shows selected characteristics at baseline of the study population. Most participants were non-Hispanic white. When comparing women who reported no screening in the past 5 years (non-users) to women who reported at least annual mammography screening, non-users had fewer mammography facilities within a 10km radius around their homes, were less likely to be white, to have a family history of breast cancer, to have gone through menopause, and to have used postmenopausal hormones, and they had lower income and education levels. The proportion of women from rural areas was similar in all utilization groups, except for non-users, which was disproportionally rural (χ^2 square test, P<0.0001). Table 12 shows a distribution summary of driving times and facility densities in rural vs. urban settings. Urban women tended to have more facilities near their homes and shorter driving times. Almost half (48%) of rural women had no facility within a 10km radius around their home.

4.4.2 Geospatial Analysis

The heat map from our kriging analysis in Figure 5 visualizes the locations of mammography facilities and utilization patterns in Wisconsin relative to mammography facility locations. Utilization above average is shaded in blue, and below average in red, and facilities are shown are marked as black dots. Mammography are clustered in and around major demographic centers (Chicago, Milwaukee, Madison, Greenbay/Appleton

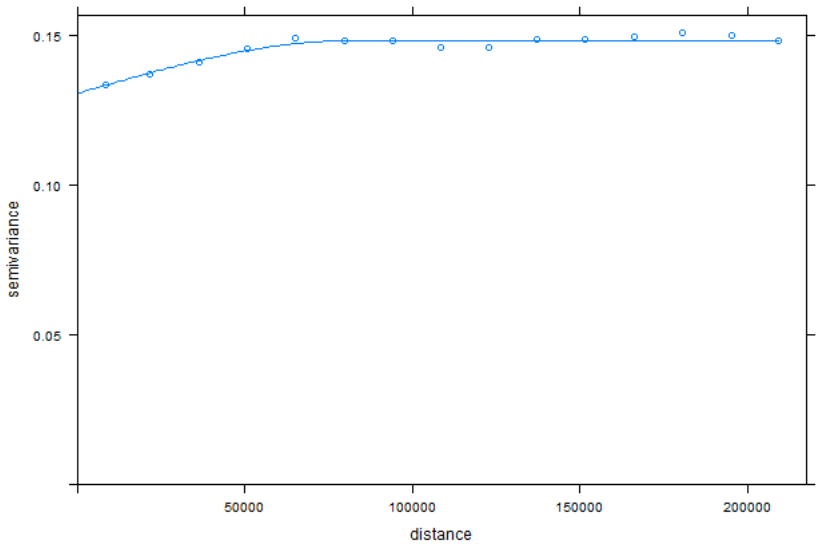

Figure 6: Semivariogram of annual mammography utilization in Wisconsin (N=5,930), Wisconsin Women's Health Study, 1995-2007

area, and near the Twin cities and Duluth), and otherwise scattered across more rural areas.

Utilization is consistently above average where many facilities are regionally clustered, especially in or near large urban areas (Milwaukee, Madison, Greenbay, Appleton, and the rural regions near Duluth and the Twin Cities), and below average where facility density is sparse. The blue above-average utilization area in central Wisconsin is centered around the I-39 and the Wisconsin Marshfield clinics. The semivariogram of mammography frequency (Figure 6) displays a large nugget effect, implying that the greatest share (88%) in utilization variability is local, whereas the increase in variance with distance that could be attributable to regional differences makes up 12% of the total variance.

4.4.3 Statistical Analysis

Odds ratios (OR) of more frequent mammography utilization from the proportional odds models are shown in Table 13 , showing results from both the minimal adjustment set model and the model with more potential confounders. Nonlinear effects of driving time and facility density were significant in both models (nonlinear effects of driving time, P-values 0.003 and 0.01, respectively; nonlinear effects of facility density, P-values 0.003 and 0.005, respectively). After stratifying the model with more confounders by urban vs. rural RUCA codes, driving time remained significant among rural women (nonlinear effects of driving time, P-value 0.02), and facility density remained significant among urban women (nonlinear effects of facility density, P-value 0.005). Interaction terms of the main exposures with urban vs. rural RUCA codes, and with income and education were not significant.

Table 13: Odds ratios of more frequent mammography screening use by geographic access (N=5,930), Wisconsin Women's Health Study, 1995-2007

Main Exposures	**Minimal Model*, Urban**		**Minimal Model*, Rural**		**Full Model**, Urban**		**Full Model**, Rural**	
	OR	**95% CI**	**OR**	**95% CI**	**OR**	**95% CI**	**OR**	**95% CI**
Driving Time								
20 vs. 10 Minutes	1.09	0.77-1.54	1.39	1.03-1.87	1.09	0.77-1.55	1.46	1.09-1.97
30 vs. 20 Minutes	0.64	0.37-1.08	0.89	0.76-1.04	0.64	0.38-1.09	0.93	0.79-1.08
40 vs. 20 Minutes	0.31	0.08-1.23	0.61	0.41-0.90	0.31	0.08-1.26	0.65	0.44-0.97
Mammography Facilities within 10km Radius, N								
1 vs. 0 Facilities	1.22	1.05-1.42	1.17	0.92-1.49	1.26	1.08-1.47	1.21	0.95-1.53
2 vs. 1 Facilities	1.15	1.03-1.29	1.10	0.94-1.30	1.18	1.05-1.33	1.12	0.96-1.32
3 vs. 2 Facilities	1.05	0.99-1.11	1.00	0.87-1.15	1.06	1.00-1.12	1.00	0.87-1.15
4 vs. 3 Facilities	0.98	0.93-1.02	0.92	0.76-1.13	0.97	0.92-1.02	0.92	0.75-1.12

* Model adjusts for education, income, number of household members, race, and mammography capacity.
** Model adjusts for education, income, number of household members, race, and mammography capacity, age, family history of breast cancer, and indicators of neighborhood deprivation by Census Tract: population fraction below the poverty line, median 1999 income, population fraction without a vehicle, population education levels, population fraction without health insurance (by county).

Table 14: Odds ratios of at least one mammogram in two years by geographic access (N=5,930) Wisconsin Women's Health Study, 1995-2007

	Urban		**Rural**	
Main Exposures	OR	95% CI	OR	95% CI
Driving Time				
20 vs. 10 Minutes	1.07	0.71-1.64	1.56	1.10-2.23
30 vs. 20 Minutes	0.68	0.35-1.31	0.94	0.78-1.14
40 vs. 20 Minutes	0.37	0.07-2.03	0.66	0.41-1.07
Mammography Facilities within 10km Radius, N				
1 vs. 0 Facilities	1.29	1.06-1.55	1.26	0.95-1.69
2 vs. 1 Facilities	1.20	1.04-1.38	1.19	0.98-1.45
3 vs. 2 Facilities	1.06	0.98-1.14	1.08	0.90-1.30
4 vs. 3 Facilities	0.96	0.91-1.02	1.00	0.78-1.29

Models adjust for education, income, number of household members, race, and mammography capacity, age, family history of breast cancer, and indicators of neighborhood deprivation by Census Tract: population fraction below the poverty line, median 1999 income, population fraction without a vehicle, population education levels, population fraction without health insurance (by county).

Table 15: Odds ratios of more frequent mammography screening use by geographic access, stratified by income and education (N=5,930), Wisconsin Women's Health Study, 1995-2007

	Annual Per Capita Income ≤11,250		**Annual Per Capita Income >11,250**		**At Least College Degree**		**No College Degree**	
Main Exposures	**OR**	**95% CI**	**OR**	**95% CI**	**OR**	**95% CI**	**OR**	**95% CI**
Driving Time								
20 vs. 10 Minutes	1.28	0.87-1.90	1.25	1.04-1.62	1.04	0.65-1.66	1.27	1.00-1.62
30 vs. 20 Minutes	0.76	0.57-1.03	0.91	0.77-1.07	0.78	0.60-1.01	0.91	0.76-1.08
40 vs. 20 Minutes	0.44	0.20-0.98	0.69	0.47-1.03	0.53	0.29-0.98	0.69	0.43-1.10
Mammography Facilities within 10km Radius, N								
1 vs. 0 Facilities	1.15	0.90-1.48	1.23	1.06-1.42	1.15	0.90-1.48	1.21	1.06-1.38
2 vs. 1 Facilities	1.10	0.91-1.33	1.16	1.05-1.29	1.11	0.92-1.33	1.15	1.04-1.26
3 vs. 2 Facilities	1.02	0.93-1.11	1.06	1.00-1.12	1.03	0.94-1.13	1.05	1.00-1.10
4 vs. 3 Facilities	0.96	0.89-1.04	0.98	0.93-1.03	0.98	0.90-1.06	0.97	0.93-1.02

Models adjust for education, income, number of household members, race, and mammography capacity, age, family history of breast cancer, and indicators of neighborhood deprivation by Census Tract: population fraction below the poverty line, median 1999 income, population fraction without a vehicle, population education levels, population fraction without health insurance (by county).

Table 16: Odds ratios of at least one mammogram in two years by geographic access, stratified by income and education (N=5,930), Wisconsin Women's Health Study, 1995-2007

	Annual Per Capita Income ≤11,250		**Annual Per Capita Income >11,250**		**At Least College Degree**		**No College Degree**	
Main Exposures	**OR**	**95% CI**	**OR**	**95% CI**	**OR**	**95% CI**	**OR**	**95% CI**
Driving Time								
20 vs. 10 Minutes	1.33	0.85-2.09	1.27	0.92-1.75	0.96	0.52-1.76	1.33	1.01-1.77
30 vs. 20 Minutes	0.76	0.55-1.05	0.96	0.80-1.16	0.79	0.57-1.11	0.93	0.77-1.14
40 vs. 20 Minutes	0.43	0.18-1.01	0.79	0.51-1.24	0.57	0.26-1.26	0.72	0.42-1.21
Mammography Facilities within 10km Radius, N								
1 vs. 0 Facilities	1.16	0.87-1.54	1.32	1.10-1.59	1.24	0.89-1.74	1.26	1.08-1.48
2 vs. 1 Facilities	1.11	0.90-1.37	1.22	1.07-1.40	1.17	0.91-1.51	1.18	1.05-1.33
3 vs. 2 Facilities	1.03	0.94-1.14	1.07	1.00-1.15	1.06	0.94-1.21	1.06	1.00-1.12
4 vs. 3 Facilities	0.98	0.90-1.07	0.97	0.91-1.03	0.98	0.88-1.11	0.97	0.92-1.02

Models adjust for education, income, number of household members, race, and mammography capacity, age, family history of breast cancer, and indicators of neighborhood deprivation by Census Tract: population fraction below the poverty line, median 1999 income, population fraction without a vehicle, population education levels, population fraction without health insurance (by county).

The final model included more potential confounders beyond the minimal adjustment set, nonlinear effects of the main exposures, no interaction terms, and was stratified by urban vs. rural RUCA codes. Among urban women, having one mammography facility within a 10km radius was associated with more frequent mammography utilization (1 vs. 0 facilities, OR 1.26, 95% CI 1.0-1.47). More facilities increased the odds of more frequent screening, but the effect decreased with each added facility and became insignificant with 2-3 nearby facilities (2 vs. 1 facilities, OR 1.18, 95% CI 1.05-1.33; 3 vs. 2 facilities, OR 1.06, 95% CI 1.00-1.12). In the rural model, effects of facility density were similar, albeit not statistically significant. Contrary to our hypothesis, driving times of 20 vs. 10 minutes were associated with increased odds of more frequent screening among rural women (OR 1.46, 95% CI 1.09-1.97). There was generally a trend towards decreased ORs with long driving times, but odds ratios for moderate vs. short driving times were inconsistent. OR from the logistic model (Table 14) that modeled the odds of having had at least 1 mammogram in two years were close to the OR from the proportional odds model. Estimated effects stratified by income and education were also similar (Tables 15 and 16).

We conducted sensitivity analyses on our final model (see appendix C). Although postmenopausal hormone use was not a confounder according to our DAG, we reran the final model including an indicator of ever use of postmenopausal hormones because hormone use was associated with more frequent mammography use in our study population. We also reran the final model excluding mammography facilities outside Wisconsin, and we reran the final model stratifying by more nuanced (4 categories) RUCA codes. Neither of these altered models resulted in substantial effect changes of our main exposures. Data on health insurance was not part of all questionnaire versions in the study, but we reran the final model for a reduced sample (N=759) of women for whom we had collected data on insurance, with and without an indicator of having health insurance. Having health insurance slightly attenuated the effect of facility density among urban women (1 vs. 0 facilities while controlling for insurance, OR 1.18, 95% CI 0.74-1.88, vs. without controlling for health insurance, OR 1.24, 95% CI 0.78-1.99), but in both model versions, the pattern still indicated increased odds of screening use with higher facility density, with decreasing effects per added facility. Finally, in a model with only linear effects, effects of driving time were not significant, and the OR for driving time and facility density were close to 1.

4.5 Discussion

We found that having at least one mammography facility within a 10 km radius may increase mammography frequency among women without breast cancer, increasing the odds of more frequent utilization by 15-25%, but adding more facilities had smaller estimated effects. With long driving times, there was a tendency

towards decreased odds of screening. We found no evidence of effect modification of geographic access by income or education, or by urbanicity. However, rural women had fewer mammography facilities near their homes and longer driving times to facilities, and were more likely to have reported no mammogram in the past 5 years.

Findings from similar studies have been inconsistent. Khan-Gates [157] reviewed articles relating geographic access variables to mammography use. Among those, Meersman et al [158] found reduced odds of a recent mammogram if only 0-1 mammography facilities were located near a woman's home, but the odds of screening hardly changed for 2-10 nearby facilities. We similarly found that higher facility density within a 10km radius was nonlinearly associated with greater odds of more frequent utilization in urban settings, without measurable marginal benefits when more than 2-3 facilities were available.

Other studies examined the relationships between driving distance to mammography facilities and utilization. Meersman et al [158] found no association between distance to facilities and mammography utilization; Engleman et al [159] found that with each additional 5 miles of distance, odds of screening use decreased by 3%. Some [160] but not all studies [161–163] have suggested that longer driving distances or driving times may be related to later stage of breast cancer at diagnosis. These inconsistent findings parallel our inconclusive results with regards to driving time. We found some evidence of a nonlinear effect of driving times on mammography utilization, and a tendency towards reduced mammography utilization with long driving times, but inconsistent odds ratios when comparing moderate to short driving times.

Other studies evaluated the number of mammography machines per population as main exposure. Elkin et al [72] found that in counties with inadequate availability of mammography screening machines, odds of screening use was reduced by 13-15%. Elting et al [164] found that the presence of a mammography facility in the county was associated with increased odds of mammography utilization. Other studies found no association [165–167]. Included only as a confounder in our study, mammography capacity per population was not significantly associated with mammography utilization in our models.

A strength of our study is that we allowed for nonlinear effects of our main exposures, which was not done in previous studies we reviewed. We found evidence that driving times and facility density were nonlinearly related to mammography utilization, while a strictly linear model did not detect any statistically or clinically significant effects. Our study also has limitations. All women in our study had a driver's license because of the sampling frame, but we had no data on individual car access. Pucher et al [55] showed that owning at least one car per household is common in the US, but becomes less common for lowest income groups. These lowest income groups may be less likely to have participated in our study, which could have prevented us from detecting effect modification by income. Furthermore, 50-74 year-old women may be more likely to have access to a car compared with younger age groups. Mammography is used infrequently,

which may make geographic access less relevant compared with other considerations regarding mammography utilization, e.g. time off work and child care. Our findings cannot be generalized to other medical services. Another limitation in our study is a historic discrepancy between driving times and outcome measures. While data on mammography use was collected 1995-2007, driving times were estimated in 2015 under potentially different road and traffic conditions. This could have created exposure misclassification, which could be one reason why we did not detect a clear effect pattern for driving times, with a trend towards decreasing odds of utilization with long driving times, but inconsistent odds ratios when comparing moderate to short driving times.

4.6 Conclusion

Our analysis emphasizes that relationships can be nonlinear. According to our findings, the availability of at least one mammography facilities near women's homes may indeed increase mammography screening utilization, with decreasing marginal effects per each added facility. While we found no evidence of effect modification by urbanicity, we did find that geographic access to mammography facilities was more restricted in rural than in urban areas. Identifying areas without any nearby mammography facilities may be one means to address under- and non-use of mammography services among eligible women, especially in rural areas.

5 Long term mammography utilization among ductal carcinoma in situ patients

5.1 Abstract

Background: The National Comprehensive Cancer Network and the American Society of Clinical Oncology recommend that women diagnosed with ductal carcinoma in situ (DCIS) who have not been treated with bilateral mammography have one surveillance mammogram each year. Whether a woman adheres to these guidelines may depend on the treatment she has had, her age, family history, and other factors. Considerations with regards to mammography surveillance may also change over time. In this study, we wanted to describe long-term mammography surveillance utilization and identify potential factors that may influence utilization trajectories among DCIS patients.

Methods: Data for this analysis came from the Wisconsin In Situ Cohort Study, conducted at the University of Wisconsin, Madison, since 1997. Information on mammography utilization was available as interval-censored data, i.e. as aggregated mammogram counts over varying periods of time. We made two Poisson regression approaches: model 1 estimated average annual mammography rates predicted by treatment reg-

imen, age at diagnosis, and family history of breast cancer. Model 2 used the same predictors, but incorporated more assumptions into the regression model: we assumed that utilization would not be uniformly distributed in each year, but that utilization would be highest in and around the month of diagnosis in each year, resulting in an oscillating pattern of utilization rates. Furthermore, we hypothesized and tested for a time trend in the underlying utilization base rate. We used data visualization techniques to validate our modeling assumptions that we could not evaluate with formal statistical tests.

Results: Mammography utilization in the first year after diagnosis exceeded utilization rates in subsequent years. After excluding observations that included person time in year 1, we detected a statistically significant, but not clinically meaningful decline in mammography utilization rates with time (annual change ranging from -0.008 to -0.013 mammograms p.a.). Long term utilization rates fluctuated around 1 mammogram per year. Younger age and treatment with lumpectomy and radiation were associated with increased utilization rates, but the effects were not clinically meaningful. Visualizing monthly distributions of women's most recent mammograms confirmed our assumption that utilization rates were not uniformly distributed over time but were highest in and around the month of diagnosis.

Conclusion: According to our findings, DCIS survivors tend to adhere to mammography surveillance guidelines with stable long-term mammography utilization rates, and without meaningful utilization differences by treatment, age or family history of breast cancer. In our model 2 approach, we exemplified how assumptions of a specific utilization pattern over time can be incorporated into a Poisson regression. This may have other statistical modeling applications when estimating specific event distributions is relevant, while the precise location of events in time or space is unknown.

5.2 Introduction

Patients of Ductal Carcinoma in situ (DCIS) have survival rates similar to the general population [131]. However, DCIS can be a precursor for invasive breast cancer [168]. After a DCIS diagnosis, the current recommendation for women who have had breast-conserving surgery or unilateral mastectomy is to have a surveillance mammogram 6-12 months after therapy and annually after that [77,78]. Mammography screening recommendations before any kind of breast cancer diagnosis have been disputed with regards to starting and ending age, and with regards to frequency [132,133,169]. Surveillance mammography recommendations after a breast cancer diagnosis have received less attention and have been less controversial than mammography screening recommendations before any kind of breast cancer diagnosis. In this work, when using the term 'mammography surveillance', we refer to the utilization of mammography after a DCIS diagnosis, as opposed to preventive mammography screening before a diagnosis.

A woman may have her own considerations that inform her mammography surveillance behavior. Previously reported factors that may impact a woman's decision process on mammography utilization are the treatment she underwent, a family history of breast cancer, age, life expectancy, and competing risks for death at a given time, and whether or not a physician recommended mammography use [81–83, 170–174]. Current surveillance guidelines have no explicit age limitations. But attitudes with regards to the above listed decision criteria and circumstances can change over time and with age, and may result in changes of surveillance behaviors with time.

It was our aim to to describe long-term mammography surveillance utilization among DCIS patients. We wanted to understand changes in mammography surveillance patterns over time, and what factors may influence utilization trajectories. We hypothesized that women would tend to adhere to surveillance guidelines more rigorously in the initial years after a DCIS diagnosis, but that screening rates would then decrease with increasing time since diagnosis. We furthermore hypothesized that rates would not be uniformly distributed throughout any given year, but that women would be more likely to have a surveillance mammogram in and around the month of diagnosis in the years following her diagnosis. We believed this annual pattern would be especially evident in the first years after diagnosis, and be less pronounced in later years. Lastly, we hypothesized that treatment regimen, family history of breast cancer, and age would be associated with mammography utilization rates.

5.3 Methods

5.3.1 Study Population

Data for this analysis come from the Wisconsin In Situ Cohort (WISC) Study, conducted at the University of Wisconsin, Madison, since 1997. WISC participants were the cases from a series of case control studies (Wisconsin Women's Health Study, WWHS) conducted between 1997 and 2007 that compared breast cancer in situ (BCIS) cases to controls without a personal history of breast cancer. The goal of the WISC study was to study BCIS survival, risk factors, and health behaviors over time. The WWHS and WISC studies were approved by the University of Wisconsin Institutional Review Board, and informed consent was obtained from each participant.

WISC participants were eligible to be included in our main analysis if they had not undergone bilateral mastectomy, and had not had any kind of breast cancer recurrence until a given follow-up interview. For example, if a participant had two follow-up interviews in total, and had no bilateral mastectomy after her initial diagnosis, and no breast cancer recurrence of any kind before her first interview, but did have any kind of breast cancer recurrence before her second follow-up interview, then the first interview was part of

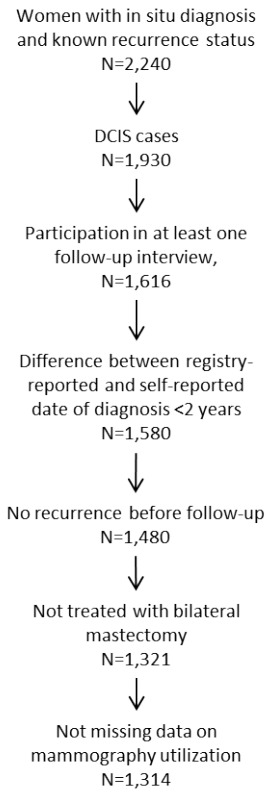

Figure 7: Study eligibility and exclusion

the analysis, but not the second.

2,240 BCIS cases (78%) enrolled in the original WWHS study. Of these, 1,930 women (86.2%) had a DCIS diagnosis, and 310 (13.8%) a different in situ diagnosis. Among the enrolled DCIS cases, 1,076 (79% of eligible) participated in the first follow-up interview, 506 (85% of eligibe) in the second, 1,232 (73% of eligible) in the third, and 1,180 (73% of eligible) in the fourth follow-up interview. In total, 1,616 (83.7% of enrolled) DCIS cases had at least one follow-up interview. In the first follow-up interview, women were asked about the date of their initial diagnosis. We excluded 36 women whose self-reported date differed from the cancer registry-reported date of diagnosis by more than 2 years. Of the remaining 1,580 participants, 100 were ineligible because they had a recurrence before follow-up, and another 159 because they were treated with bilateral mastectomy for their initial diagnosis, resulting in a final eligible sample of 1,321 women. Of 3,244 total follow-up observations before any kind of breast cancer recurrence across eligible DCIS participants, 45 observations (1.4%) had missing information on mammography utilization and were not used in this analysis, including all observations on 7 eligible women, which reduced our final analytical sample to 1,314 women. Study eligibility and exclusion due to missingness are summarized in Figure 7.

5.3.2 Data Collection

Baseline interviews with BCIS cases were conducted approximately 1.3 years after diagnosis (interquartile range: 1.0-1.5). After baseline, two follow-up telephone interviews took place in 2003-2007 and 2005-2007. Two mailed follow-up surveys took place in 2010/2011 and 2012/2013. Women enrolled continuously throughout the recruitment period 1997-2007. Therefore, not all women were eligible for every interview. In order to be eligible for an interview, at least two years had to have passed since a woman's previous interview.

In the baseline and follow-up interviews, we collected the following information: cancer-related information: mode of detection at initial diagnosis, breast cancer recurrence, treatment after initial diagnosis and subsequent recurrences; health behaviors and indicators: mammography utilization, and family history of breast cancer. Women were additionally asked to report the year and month of their last mammogram at each follow-up.

From the Wisconsin Cancer Reporting System, we obtained information on cancer pathology at first diagnosis, including information on tumor grade, stage, histology, and date of diagnosis. Pathology reports were obtained from diagnosing hospitals and clinics to confirm self-reported breast cancer recurrence.

5.3.3 Measures

Mammography surveillance utilization was the main outcome of interest for this analysis. In each follow-up interview, the questionnaire on mammography utilization was phrased slightly differently:

- In the first interview, the question was phrased: "Since your [most recent] breast diagnosis, how many mammograms have you had?" Upon that, participants could enter any number ≥ 0.
  In the first interview women were also asked to state the date of their initial diagnosis. If a woman provided this information, this resulted in a 'subjective' date of diagnosis which was potentially subject to recall bias, and an 'objective' date of diagnosis as confirmed by the Wisconsin Cancer Registry. If provided, we used the 'subjective' date of diagnosis when analyzing the first interview, and otherwise the 'objective' date of diagnosis. We excluded women if the 'subjective' date differed from the 'objective' date of diagnosis by more than 2 years.

- In the second interview, the question was phrased: "Since the year preceding last interview [formulated as calendar date], how many mammograms have you had?" Upon that, participants could enter any number ≥ 0. For most participants in the second interview, the year preceding last interview referred to the date of their initial diagnosis (if they had not participated in the first follow-up round), for others it referred to the year preceding their first follow-up interview.

- In the third interview, the question was phrased: "How many mammograms have you had in the past 5 years?" With the answer key: "None, 1, 2, 3, 4, 5, more than 5".

- In the fourth follow-up, the question was phrased: "How many mammograms have you had in the past 4 years?" With the answer key: "None, 1, 2, 3, 4, more than 4". The fourth interview also included a questionnaire on magnetic resonance imaging (MRI) utilization. The question was phrased: "How many MRIs have you had in the past 4 years?" With the answer key: "None, 1, 2, 3, 4, more than 4". We calculated the sum of the self-reported number of mammograms and MRIs; so even though we use the term mammography utilization in this analysis, this includes self-reported MRIs at the fourth interview. The reason why we included MRI in this count is that women might have adapted to using MRI as a new technology instead of using mammography.

As primary predictors of mammography utilization, we included age (years) at diagnosis, an indicator of whether a woman had a family history of breast cancer, treatment regimen for the initial diagnosis (biopsy/none vs. lumpectomy without radiation vs. lumpectomy with radiation vs. unilateral mastectomy), and a combined indicator (yes/no) of any kind of endocrine therapy (Tamoxifen, Raloxifen, aromatase inhibitors). At each follow-up interview, we updated recurrence status. If a woman had several rounds of

treatment for her initial diagnosis but before any recurrence, she was assigned the most aggressive treatment she received. For example, if a woman initially underwent lumpectomy and radiation, but then had a second treatment round with unilateral mastectomy for her initial diagnosis, she was assigned unilateral mastectomy at her first follow-up interview.

5.3.4 Statistical Analysis

The structure of our data was interval-censored: women participated in follow-up interviews at similar but not the same times since DCIS diagnosis, and self-reported the cumulative number of mammograms since varying starting times t_i at each interview. In order to estimate mammography utilization rates from the aggregated mammogram counts, we ran two Poisson models, henceforth referred to as model 1 and model 2. In both models, the number of mammograms (and in the case of the fourth interview the sum of mammograms and MRIs) over time was the primary outcome. In addition to our main analyses, we ran two sensitivity analyses:

- Sensitivity analysis 1: including observations from the first follow-up interview that were excluded in the man analysis if the 'subjective' (self-reported at first follow-up) and 'objective' dates of initial diagnosis differed from each other by more than 2 years.
- Sensitivity analysis 2: Excluding MRI from our outcome count at the 4th follow-up interview.

To estimate standard errors while accounting for nesting of multiple observations within women, we used bootstrapping by creating 1,000 subsamples of participating women with replacement. Based on the 1,000 simulations, we then estimated the 95% highest density credible intervals using the HPDinterval() function from the R coda package [175]. To impute missing values in our predictors within each bootstrap sample, we used multiple imputation using the aregImpute() function from the R Hmisc package [89], imputing missing values for family history (N missing=68), surgical treatment at initial diagnosis (N missing=46), radiation (N missing=44), and endocrine therapy (N missing=93). We generated 5 datasets with imputed values within each bootstrap. The exact number of missing values depended on the bootstrap sample.

Model 1: average annual mammography utilization

In Poisson model 1, we estimated average annual mammography utilization rates up to 18 years after diagnosis. The expected number of mammograms $E[Y]$ was estimated as:

$$E[Y] = r_1 t_1 + r_2 t_2 + r_3 t_3 ...$$

Where r_1, r_2, r_3 etc were the annual average utilization rates in year 1, 2, 3 etc after diagnosis, and t_1, t_2, t_3 etc the person time a woman spent in each respective year. The annual rates r_i were estimated as:

$$
\begin{aligned}
ln(r_1) &= ln(a1) + \beta X \\
&= ln(a1) + \beta_1 * age\ at\ diagnosis + \beta_2 * family\ history + \beta_3 * biopsy/no\ treatment \\
&+ \beta_4 * lumpectomy\ (no\ radiation) + \beta_5 * lumpectomy\ (with\ radiation) \\
&+ \beta_6 * endocrine\ therapy \\
ln(r_2) &= ln(a2) + \beta X \\
&= ln(a2) + \beta_1 * age\ at\ diagnosis + \beta_2 * family\ history + \beta_3 * biopsy/no\ treatment \\
&+ \beta_4 * lumpectomy\ (no\ radiation) + \beta_5 * lumpectomy\ (with\ radiation) \\
&+ \beta_6 * endocrine\ therapy \\
ln(r_3) &= ln(a3) + \beta X \\
&etc
\end{aligned}
$$

With unilateral mastectomy as the reference treatment. The estimation was run iteratively in two steps. Firstly, $E[Y]$ was calculated as:

$$E[Y] = a_1 * exp(\beta X)t_1 + a_2 * exp(\beta X)t_2 + a_3 * exp(\beta X)t_3 ... + a_{18} * exp(\beta X)t_{18} \quad (1)$$

With the $exp(\beta X)t_i$, as the covariates, and the a_i as the coefficients in a Poisson regression with identity link. In the first iteration, $\beta_1 ... \beta_6$ were set to zero. Subsequently, the estimates of $\beta_1 ... \beta_6$ were updated in a Poisson regression with log link by using the estimated $a_1 ... a_{18}$ from (1):

$$
\begin{aligned}
E[Y] &= \beta_0 + \beta_1 * age\ at\ diagnosis + \beta_2 * family\ history + \beta_3 * biopsy/no\ treatment \quad (2) \\
&+ \beta_4 * lumpectomy\ (no\ radiation) + \beta_5 * lumpectomy\ (with\ radiation) \\
&+ \beta_6 * endocrine\ therapy + offset
\end{aligned}
$$

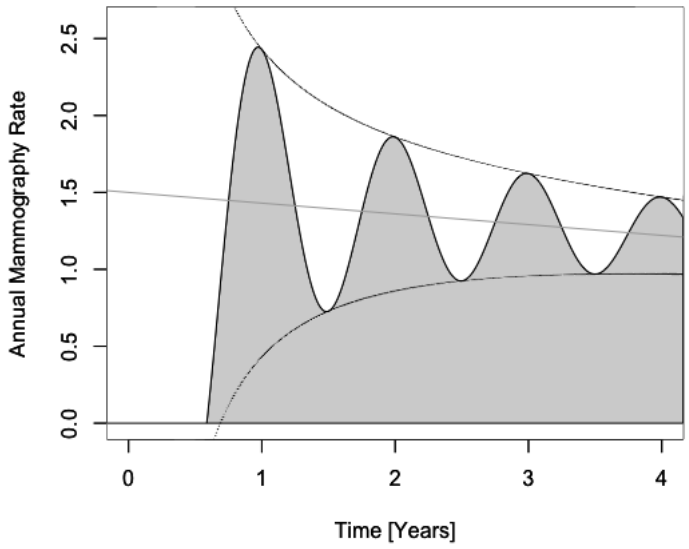

Figure 8: Model 2 approach: Conceptualizing mammography rates as a damped cosine wave. The shaded area represents the expected number of mammograms up to a given time point.

where the offset was calculated as:

$$offset = t_1 + (\frac{a_2}{a_1}t_2 + \frac{a_3}{a_1}t_3 + ... + \frac{a_{18}}{a_1}t_{18})$$

The updated β_i estimates were then inserted in (1). The process was repeated until the parameter estimates of $a_1...a_{18}$, and $\beta_1...\beta_6$ converged.

Model 2: mammography utilization approximated as oscillating linear splines

Here, we give a summary of the model 2 estimation approach. The approach is described in more detail in appendix D. We assumed that mammography surveillance rates were not uniformly distributed in each year, but that more women used mammography in and around the month of their initial diagnosis than in other months, such that mammography utilization rates would oscillate over the course of each year after diagnosis. We secondly assumed that the oscillation would flatten out with time, i.e. that utilization rates would gradually approach an arbitrarily/uniformly distributed pattern throughout each year. To formulate these assumptions mathematically, we conceptualized mammography rates as a damped cosine wave, as visualized in Figure 8. The number of mammograms up to a specific time point would then equal the shaded area under the cosine curve. In physics, a damped cosine wave is enveloped along its maxima by an exponential decay curve (see equation (11) in appendix D), fluctuating around a base rate $C(t)$. If the base

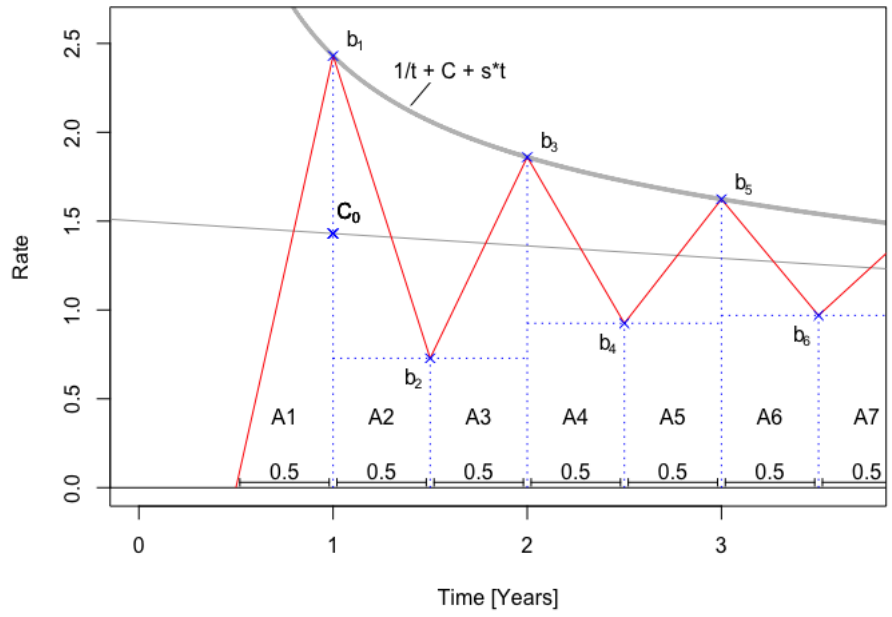

Figure 9: Altered approach for model 2: linear spline approximation of a damped cosine wave.

rate remained the same over time, $C(t)$ would be a constant. With this approach, the expected number of mammograms between two time points t_0 and t_1 would then be estimated by the integral of the damped cosine wave between t_0 and t_1 (see equations (12) and (13) in appendix D).

We made several alterations to this approach: the unknown parameters that we needed to estimate entered the solution of the integral of the damped cosine wave nonlinearly in a way that would prevent us from estimating them with generalized linear regression approaches. Instead, we approximated the oscillation curve using linear splines. Secondly, an exponential decay would make the oscillation flatten out fast for small values of t. Since there was no specific reason to assume that the oscillation decayed exponentially, we chose a hyperbolic decay curve $1/t$ instead, which allowed the oscillation to flatten out more gradually. Thirdly, we hypothesized that women had mammograms less often with time, i.e. that the base mammography rate $C(t)$ decreased with time. The altered approach is visualized in Figure 9. The maxima of the linear splines were assumed to be exactly one year apart, and the minima were assumed to occur half a year after each maximum. We hence estimated the envelope functions that framed the maxima and minima of the splines as:

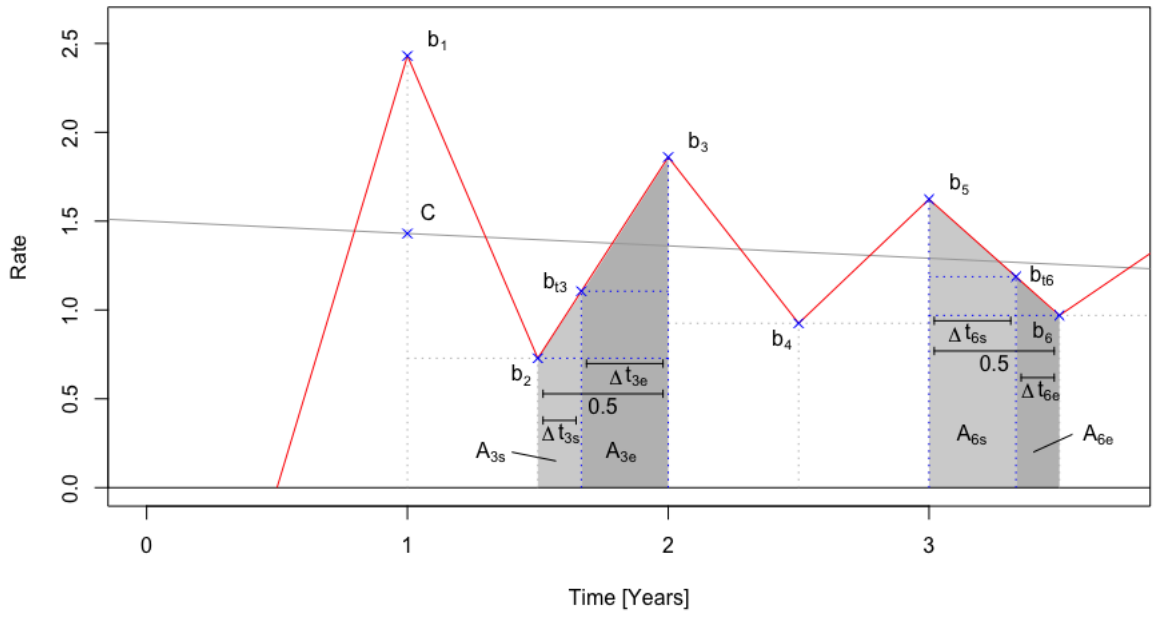

Figure 10: Calculation of partial areas in model 2

$$1/t + C + \alpha t \ (upper\ envelope)$$

$$-1/t + C + \alpha t \ (lower\ envelope)$$

Where C described the initial base rate, and α the (linear) change of the base rate with time. With these adjustments, and if women had been observed for full half-years at a time, the expected number of mammograms between t_0 = time of diagnosis and an end point t_i would be estimated by the sum of the areas A_i under the linear splines:

$$E[Y] = A_1 + A_2 + A_3 + ...A_i \tag{3}$$

Where $A_1...A_i$ represented the number of mammograms during half-year intervals (Figure 9). The areas A_i could be calculated using trigonometric properties. In our study, however, women were not observed for full half-year intervals at a time. Instead, women in our study reported the cumulative number of mammograms between varying starting time points t_0 and varying end times t_i and may only have been observed in parts of the A_i. To be able to calculate partial areas under the splines, we divided each half-year interval t_i into a

starting interval t_{is} and an ending interval t_{ie}. For example, and using the notation in Figure 10, the entire area of A_3 was calculated as the sum of the 'starting' area A_{3s}, and the 'ending' area A_{3e}:

$$\begin{aligned} A_{3s} &= b_2 * \Delta t_{3s} + (b_{t3} - b_2) * \Delta t_{3s} * 0.5 \\ A_{3e} &= b_{t3} * \Delta t_{3e} + (b_3 - b_{t3}) * \Delta t_{3e} * 0.5 \\ A_3 &= A_{3s} + A_{3e} \\ &= \frac{b_2 \Delta t_{3s}}{2} + \frac{b_3 \Delta t_{3e}}{2} + \frac{b_{t3}}{2}(\Delta t_{3s} + \Delta t_{3e}) \end{aligned} \tag{4}$$

Where Δt_{3s} represented the person time a woman spent in the 'starting' interval of the interval between 1.5 and 2 years after diagnosis, and Δt_{3e} the time a woman spent in the 'ending' interval between 1.5 and 2 years after diagnosis. For example, if a woman's observed person time started at 1.6 years after diagnosis, her value of Δt_{3s} would be 0, and her value of Δt_{3e} would be 0.4. Similarly, the starting and ending intervals in A_6 were calculated as:

$$\begin{aligned} A_{6s} &= b_{t6} * \Delta t_{6s} + (b_5 - b_{t6}) * \Delta t_{6s} * 0.5 \\ A_{6e} &= b_6 * \Delta t_{6e} + (b_{t6} - b_6) * \Delta t_{6e} * 0.5 \\ A_6 &= A_{6s} + A_{6e} \\ &= \frac{b_5 \Delta t_{6s}}{2} + \frac{b_6 \Delta t_{6e}}{2} + \frac{b_{t6}}{2}(\Delta t_{6s} + \Delta t_{6e}) \end{aligned} \tag{5}$$

Where Δt_{6s} represented the person time a woman spent in the 'starting' interval between 3 and 3.5 years after diagnosis, and Δt_{6e} the time a woman spent in the 'ending' interval between 3 and 3.5 years after diagnosis. For example, if a woman's observed person time ended at 3.3 years after diagnosis, her value of Δt_{6s} would be 0.3, and her value of Δt_{6e} would be 0.

The b_i in (4) and (5) were the maxima and minima of the splines that fell on the hyperbolic envelope curves and could hence be calculated as:

$$\begin{aligned} b_1 &= 1/1 + C + 1\alpha \\ b_3 &= 1/2 + C + 2\alpha \\ b_5 &= 1/3 + C + 3\alpha \\ &etc. \end{aligned} \tag{6}$$

and

$$
\begin{aligned}
b_2 &= -1/1.5 + C + 1.5\alpha \\
b_4 &= -1/2.5 + C + 2.5\alpha \\
b_6 &= -1/3.5 + C + 3.5\alpha \\
&etc.
\end{aligned} \tag{7}
$$

The b_{ti} in (4) and (5) could be derived from the trigonometric intersect theorem (see equation (20) in appendix D), such that:

$$
\begin{aligned}
b_{t2} &= b_1 - 2b_1\Delta t_{2s} + 2b_2\Delta t_{2s} \\
b_{t3} &= b_3 - 2b_3\Delta t_{3e} + 2b_2\Delta t_{3e} \\
b_{t4} &= b_3 - 2b_3\Delta t_{4s} + 2b_4\Delta t_{4s} \\
b_{t5} &= b_5 - 2b_5\Delta t_{5e} + 2b_4\Delta t_{5e} \\
&etc.
\end{aligned} \tag{8}
$$

For the area A_1 in the first year, we assumed that women would start surveillance at 0.5 years after diagnosis at the ealiest. Entering the relationships from (4), (5), (6), (7), and (8) into (3) led to:

$$
\begin{aligned}
E[Y] &= A_1 + A_{2s} + A_{2e} + A_{3s} + A_{3e} + ... + A_{35s} \\
&= C * [expression\ 1] - \alpha * [expression\ 2] + offset
\end{aligned} \tag{9}
$$

For the calculations of *expression* 1, *expression* 2, and the offset, see equations (21) to (23) in appendix D. Lastly, to take into account treatment regimen, age and family history similarly to model 1, we assumed that the initial base rate C differed by these risk factors, i.e.:

$$
\begin{aligned}
E[C] = C_0 &+ \gamma_1 * age\ at\ diagnosis + \gamma_2 * family\ history \\
&+ \gamma_3 * biopsy/no\ treatment + \gamma_4 * lumpectomy\ (no\ radiation) \\
&+ \gamma_5 * lumpectomy\ (with\ radiation) + \gamma_6 * endocrine\ therapy
\end{aligned}
$$

With unilateral mastectomy as the reference treatment. Including the risk factors in (9), the final Poisson model was:

$$
\begin{aligned}
E[Y] = [&C_0 + \gamma_1 * age\ at\ diagnosis + \gamma_2 * family\ history + \gamma_3 * biopsy/no\ treatment \\
&+ \gamma_4 * lumpectomy\ (no\ radiation) + \gamma_5 * lumpectomy\ (with\ radiation) \\
&+ \gamma_6 * endocrine\ therapy] * expression\ 1 \\
&- \alpha * expression\ 2 \\
&+ offset \\
\\
&= C_0 * expression\ 1 + \gamma_1 * age\ at\ diagnosis * expression\ 1 + \gamma_2 * family\ history * expression\ 1 \\
&+ \gamma_3 * biopsy/no\ treatment * expression\ 1 + \gamma_4 * lumpectomy\ (no\ radiation) * expression\ 1 \\
&+ \gamma_5 * lumpectomy\ (with\ radiation) * expression\ 1 + \gamma_6 * endocrine\ therapy * expression\ 1 \\
&- \alpha * expression\ 2 \\
&+ offset
\end{aligned}
\tag{10}
$$

Equation (10) was estimated in a Poisson regression with identity link, with C_0, α, and $\gamma_1 ... \gamma_6$ as the coefficients, *expression* 1 and *expression* 2 as the covariates, and an offset.

Person Time Calculation

In both modeling approaches we needed to calculate the observed person time women spent in different time intervals since diagnosis. For model 1 (average annual mammography utilization rates), we calculated the person time a woman spent in each year after diagnosis. For model 2 (utilization rates approximated as oscillating linear splines), we calculated more granular person time intervals, i.e. the time a woman spent in each half-year, split into a 'starting' and an 'ending' subinterval. If a woman's person time covered a whole half-year period, we set the 'starting' and 'ending' subintervals of that half-year to 0.25 each.

Person time calculations corresponded to the way the questionnaire on mammography utilization was phrased in each follow-up interview. In the first interview, women were asked about the number of mammograms since initial diagnosis. For example, if a woman participated in the first follow-up interview at 3.8 years after her initial diagnosis, her person time was calculated as one full year during year 1 after diagnosis. For model 2, the remaining 2.8 years of her observed person time were distributed assigning 0.25

years to all starting and ending time subintervals between 1 and and 3.5 years after diagnosis, and 0.3 years to the starting interval of 3.5-4 years after diagnosis. For model 1, the time subintervals were aggregated to the observed person time spent in each year.

In the second interview, women were asked about the number of mammograms since the year preceding their last interview. The last interview referred to the baseline interview if women did not participate in the first follow-up, or to their first follow-up interview otherwise. In the first case, person time was calculated the way it was done for the first follow-up interview as described above. If the latter was the case, person time was calculated as the time between the two follow-up interviews and distributed across the appropriate time subintervals. The starting and ending time points were calculated as:

$$person\ time\ starting\ point = time\ since\ diagnosis\ at\ previous\ interview\ -\ 1\ year$$

$$person\ time\ end\ point = time\ since\ diagnosis\ at\ second\ interview$$

In the 3^{rd} (4^{th}) interview, women were asked about the number of mammograms in the past 5 (4) years. The starting and ending time points were calculated as:

$$person\ time\ starting\ point = time\ since\ diagnosis\ at\ 3^{rd}\ (4^{th})\ interview\ -5\ (4)\ years$$

$$person\ time\ endpoint = time\ since\ diagnosis\ at\ 3^{rd}\ (4^{th})\ interview$$

Validating Model 2 assumptions

The assumptions we made for the distribution of mammograms over time in model 2 were:

- A trend towards less frequent mammography surveillance use with time.
- Women would tend to schedule mammography appointments in and around the month of their initial diagnosis in the years following their diagnosis.
- The oscillation of surveillance rates would become more arbitrary with time.

While we could formally evaluate the first of these assumptions by testing for a time trend in our regression models, we could not use statistical testing to find evidence for the other assumptions because

we did not have data on the precise timing of reported mammograms, but only the cumulative number of mammograms over periods of time. Instead, we tested our assumptions visually. In each follow-up interview, women were asked about the year and month of their last mammogram. Assuming that the dates of women's most recent mammogram constituted a random sample of all past mammograms, we compared the month of women's last mammogram with the month of their diagnosis. For example, if a woman had been diagnosed in January, and in a given follow-up interview reported having had her last mammogram in January, then the month difference was 0. If the same woman reported having had her last mammogram in February, the month difference would be 1, or equivalently -11. A mammogram in March would mean a month difference of 2, or equivalently -10 etc. When also accounting for the self-reported year of the most recent mammogram, we were able to estimate month difference distributions in each year since diagnosis. We displayed these distributions by year since diagnosis to visually check if there was indeed an oscillation pattern and if it flattened out with time.

5.4 Results

Baseline characteristics among participating women are shown in Table 17. The majority of our study participants was white (97%), and more than two thirds were older than 50 years when their original DCIS was detected. About 23% of the women had a family history of breast cancer, and most women had their original DCIS detected by a mammogram (87%). Most women were treated with lumpectomy and radiation (47%) or unilateral mastectomy (37%) at their initial diagnosis, and almost half of all patients underwent some form of endocrine therapy.

5.4.1 Results of the Regression Models 1 and 2

Table 17: Baseline characteristics of study participants, N=1,580, Wisconsin In Situ Cohort study, 2000-2013

Characterstic	%
Age (Years)	
<50	29.2
50-60	36.3
>60	34.6
Race	
White	96.6
Other	3.4
Family History of Breast Cancer	
Yes	23.5
No	76.5
Highest Educational Degree	
High School or Less	42.3
Some College	27.5
At Least College	30.3
Annual Income, $	
$\leq$ 30,000	20.0
30,001-50,000	28.3
50,001-100,000	37.6
>100,000	14.1
Surgical Treatment	
Biopsy/None	2.3
Lumpectomy without Radiation	9.7
Lumpectomy with Radiation	47.0
Unilateral Mastectomy	36.7
Bilateral Mastectomy	4.3
Endocrine Therapy	
Yes	46.3
No	53.7
Mode of Detection at Diagnosis	
Mammography	86.9
Other	13.1

Table 18: Model 1: Estimated average annual mammography utilization rates and 95% credible intervals (CI) for high-frequency user: a woman without family history of breast cancer, diagnosed with BCIS at age 30, treated with lumpectomy with radiation, and using endocrine therapy; low frequency user: a woman with family history of breast cancer, diagnosed with BCIS at age 70, treated with unilateral mastectomy; medium frequency user: a woman without family history of breast cancer, diagnosed with BCIS at age 55, treated with lumpectomy without radiation

Year after Diagnosis	**High Frequency User Rate**	**95% CI**	**Medium Frequency User Rate**	**95% CI**	**Low Frequency User Rate**	**95% CI**
Year 1	3.02	2.56-3.49	2.47	2.11-2.85	2.10	1.76-2.37
Year 2	1.05	0.50-1.54	0.86	0.40-1.27	0.73	0.36-1.09
Year 3	1.46	1.01-1.95	1.20	0.81-1.60	1.01	0.69-1.34
Year 4	1.34	0.94-1.74	1.10	0.78-1.42	0.93	0.66-1.20
Year 5	0.83	0.48-1.17	0.68	0.40-0.97	0.58	0.34-0.82
Year 6	1.50	1.13-1.94	1.22	0.92-1.60	1.04	0.78-1.33
Year 7	1.47	1.08-1.81	1.20	0.89-1.50	1.02	0.76-1.25
Year 8	1.05	0.72-1.39	0.86	0.58-1.14	0.73	0.49-0.96
Year 9	0.90	0.58-1.25	0.74	0.47-1.02	0.63	0.40-0.86
Year 10	1.56	1.20-1.89	1.27	0.97-1.55	1.08	0.84-1.31
Year 11	1.02	0.68-1.41	0.83	0.54-1.14	0.71	0.46-0.97
Year 12	1.58	1.12-2.01	1.29	0.89-1.65	1.10	0.74-1.37
Year 13	0.77	0.31-1.22	0.63	0.23-0.98	0.54	0.22-0.84
Year 14	1.09	0.57-1.69	0.89	0.49-1.41	0.76	0.41-1.17
Year 15	1.68	1.05-2.31	1.38	0.86-1.91	1.17	0.72-1.60
Year 16	1.27	0.45-2.07	1.04	0.37-1.69	0.88	0.34-1.46
Estimated time trend (change in mammograms per annum):						
	-0.041		-0.037		-0.034	

Estimated annual average mammography rates and rate ratios for the risk factors from model 1 are shown in Table 18. The estimates are for three women, which given our models represent low-frequency and high-frequency users in order to present the upper and lower extremes of our estimates, and a medium-frequency user:

- low-frequency user: a woman who was 70 years old at diagnosis, with a family history of breast cancer, and treated with unilateral mastectomy after diagnosis, who did not use endocrine therapy.
- high-frequency user: a woman who was 30 years old at diagnosis, without a family history of breast cancer, treated with lumpectomy and radiation after diagnosis, who used endocrine therapy.
- medium-frequency user: a woman who was 55 years old at diagnosis, without a family history of breast cancer, treated with lumpectomy without radiation after diagnosis, who did not use endocrine therapy.

The year 1 annual rate estimates in Table 18 differed substantially from the remaining annual rate estimates across all women, ranging from about 2 to 3 mammograms in year 1, while in the remaining years, utilization rates fluctuated around 1 mammogram per year for all women. The estimated annual change in the average rates was calculated as the slope of the regression line through the estimated annual rates, weighted by their inverse variance, and ranged between -0.03 and -0.04 mammograms per year.

Figure 11 compares models 1 and 2 visually: the black lines are the regression lines through the model 1 average rates over time for the above described high-frequency and low-frequency users. The regression line for the medium-frequency user is not shown to avoid visual cluttering, but would fall between the displayed regression lines. The blue lines depict the estimated base rates C for the same women according to model 2 and the estimated trend α over time . While these example estimates do not represent everyone in our study population but apply only for these specific combinations of age, family history, and treatment, the estimated patterns and time trends were approximately the same when comparing different users, see Table 18. As in Table 18, it is easily visible in Figure 11 that the estimated average utilization rates in year 1 exceeded the estimated average rates in the remaining years. As an outlier, the year 1 estimate may have unduly influenced the estimated time trends.

Table 19 shows the estimated risk factor effects from both models. The interpretation of the effect estimates in model 1 is rate ratios (RR); the interpretation of the effect estimates in model 2 is the difference in the number of mammograms per year at beginning of follow-up (base rate). The estimated effects of the risk factors were consistent across the two models with regards to significance, effect sizes and directions of effects. None of the risk factors appeared to have a strong impact on utilization rates in either model. Greater age at diagnosis was associated with lower utilization rates, but there was only a small effect

Table 19: A) Model 1: Risk factor rate ratios of mammography utilization, and B) Model 2: Risk factor coefficients (base rate change in number of mammograms per year)

A)

Risk Factor	**Rate Ratio**	**95% CI**
Age (per 5 Years)	0.98	0.97-0.99
Family History		
Yes	0.98	0.95-1.02
No	1 (Ref.)	
Surgical Treatment		
Biopsy/None	1.05	0.94-1.17
Lumpectomy, no Radiation	1.09	1.04-1.15
Lumpectomy, with Radiation	1.13	1.11-1.16
Unilateral Mastectomy	1 (Ref.)	
Endocrine Therapy		
Yes	1.06	1.03-1.09
No	1 (Ref.)	

B)

Risk Factor	**Base Rate Difference**	**95% CI**
Age (per 5 Years)	-0.03	-0.04 - -0.02
Family History		
Yes	-0.01	-0.05-0.03
No	0 (Ref.)	
Surgical Treatment		
Biopsy/None	0.05	-0.08-0.16
Lumpectomy, no Radiation	0.08	0.02-0.15
Lumpectomy, with Radiation	0.10	0.07-0.13
Unilateral Mastectomy	0 (Ref.)	
Endocrine Therapy		
Yes	0.03	0.00-0.06
No	0 (Ref.)	

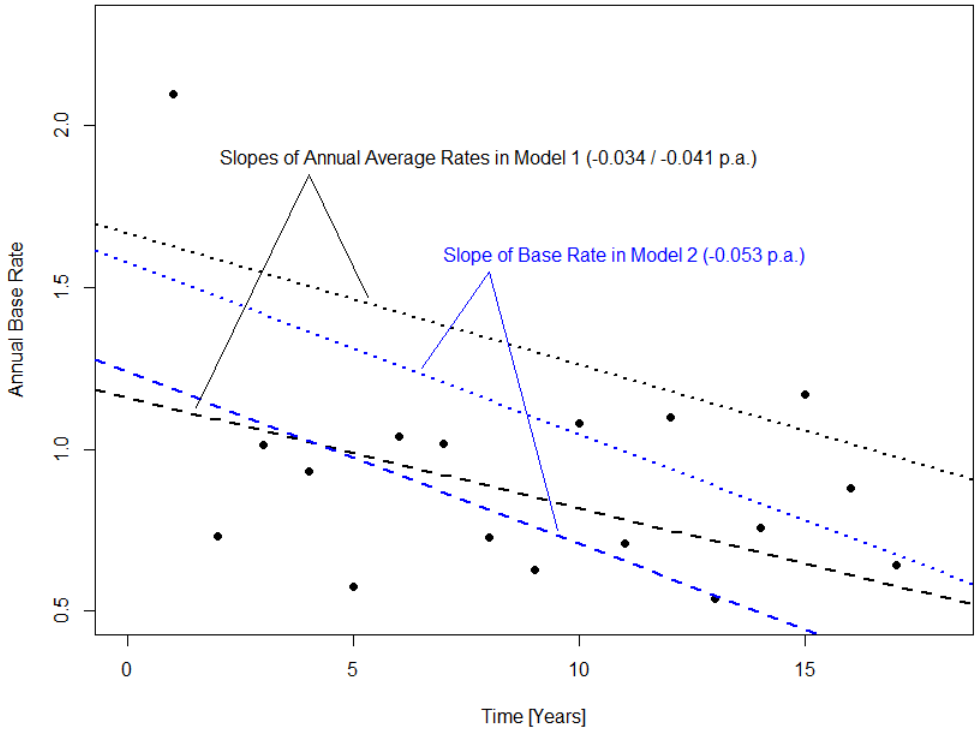

Figure 11: Model 1 and 2 in comparison. Model 1 (black points and lines): Estimated average annual mammography utilization rates (in mammograms per year) between year 1 and year 17 since diagnosis for a) a woman with family history of breast cancer, diagnosed with DCIS at age 70, treated with unilateral mastectomy (points and dashed line) and b) a woman without family history of breast cancer, diagnosed with DCIS at age 30, treated with lumpectomy, radiation, and endocrine therapy (dotted line). Model 2 (blue lines): Estimated base rate and time trend for the analogous women.

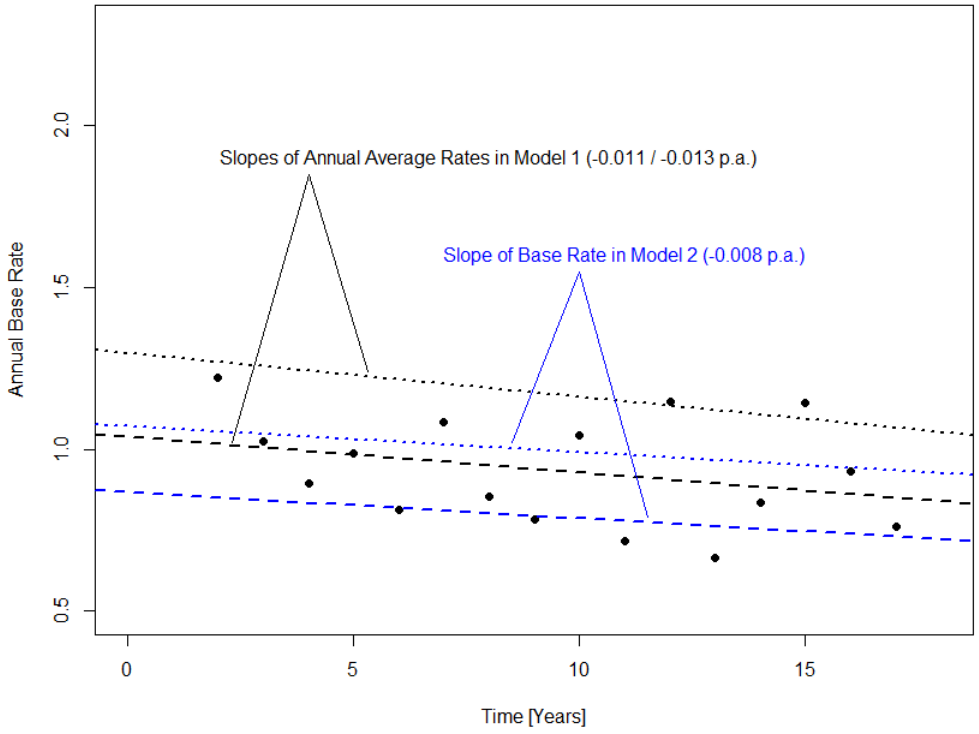

Figure 12: Model 1 and 2 in comparison after excluding observations with person time in year 1 after diagnosis: Model 1 (black points and lines): Estimated average annual mammography utilization rates (in mammograms per year) between year 1 and year 17 since diagnosis for a) a woman with family history of breast cancer, diagnosed with DCIS at age 70, treated with unilateral mastectomy (points and dashed line) and b) a woman without family history of breast cancer, diagnosed with DCIS at age 30, treated with lumpectomy, radiation, and endocrine therapy (dotted line). Model 2 (blue lines): Estimated base rate and time trend for the analogous women.

(model 1: per 5 additional years of age at baseline , RR 0.98, 95% credible interval (CI) 0.97-0.99; model 2: per 5 years -0.03 mammograms per year, 95%CI -0.04-0.02). Having been treated by lumpectomy was associated with increased utilization rates (model 1 estimates: lumpectomy without radiation vs. unilateral mastectomy RR 1.09, 95% CI 1.04-1.15; and lumpectomy with radiation vs. unilateral mastectomy RR 1.13, 95% CI 1.11-1.16; model 2 estimate: lumpectomy with radiation vs. unilateral mastectomy 0.10 additional mammograms per year, 95%CI 0.07-0.13). Use of any kind of endocrine therapy was also associated with increased utilization rates (model 1: RR 1.06, 95% CI 1.03-1.09; model 2: 0.03 additional mammograms per year, 95% CI 0.00-0.06).

Table 20: Model 1 after excluding observations with person time in year 1 after diagnosis: Estimated average annual mammography utilization rates and 95% credible intervals for high frequency user: a woman without family history of breast cancer, diagnosed with BCIS at age 30, treated with lumpectomy with radiation, and using endocrine therapy; low frequency user: a woman with family history of breast cancer, diagnosed with BCIS at age 70, treated with unilateral mastectomy; medium frequency user: a woman without family history of breast cancer, diagnosed with BCIS at age 55, treated with lumpectomy without radiation

Year after Diagnosis	**High Frequency User Rate**	**95% CI**	**Medium Frequency User Rate**	**95% CI**	**Low Frequency User Rate**	**95% CI**
Year 2	1.52	1.02-2.05	1.31	0.87-1.74	1.22	0.80-1.62
Year 3	1.28	0.85-1.77	1.10	0.74-1.55	1.02	0.69-1.43
Year 4	1.12	0.66-1.51	0.96	0.57-1.31	0.90	0.54-1.22
Year 5	1.23	0.88-1.58	1.06	0.77-1.37	0.99	0.70-1.27
Year 6	1.01	0.61-1.39	0.87	0.55-1.23	0.81	0.48-1.09
Year 7	1.35	0.98-1.66	1.16	0.84-1.43	1.08	0.80-1.33
Year 8	1.07	0.75-1.39	0.92	0.67-1.23	0.85	0.61-1.13
Year 9	0.98	0.69-1.34	0.84	0.59-1.14	0.78	0.53-1.05
Year 10	1.30	0.95-1.58	1.12	0.80-1.37	1.04	0.76-1.29
Year 11	0.89	0.54-1.22	0.77	0.44-1.03	0.72	0.45-1.00
Year 12	1.43	1.01-1.83	1.23	0.84-1.56	1.15	0.83-1.49
Year 13	0.83	0.39-1.23	0.71	0.33-1.05	0.66	0.31-0.98
Year 14	1.04	0.59-1.58	0.89	0.51-1.35	0.83	0.48-1.25
Year 15	1.42	0.88-2.01	1.22	0.70-1.71	1.14	0.72-1.64
Year 16		0.47-1.92	1.22	0.40-1.66	1.14	0.34-1.51
Estimated time trend (change in mammograms per annum):						
	-0.011		-0.012		-0.013	

Since mammography utilization in year 1 exceeded utilization rates in subsequent years, we reran both models excluding observations with person time in the first year after diagnosis. The results of these model runs are shown in Tables 20 and 21, and in Figure 12, using the same low-, high-, and medium-frequency user as above. After excluding year 1 after diagnosis, there were no more outliers in the estimated average annual mammography rates in model 1, and the estimated time trend was smaller, ranging between -0.013 and -0.008 mammograms per year for both models, see Figure 12. The estimated risk factor effects were even smaller compared to the model versions that included year 1, see Table 21. An additional 5 years of age at diagnosis resulted in an rate ratio of 0.99, 95% CI 0.98-0.99 in model 1; and -0.01 mammograms p.a., 95%CI -0.02-0.00in model 2. Having been treated by lumpectomy remained associated with increased utilization rates, but with minimal effects (model 1: lumpectomy with radiation vs. unilateral mastectomy RR 1.04, 95% CI 1.01-1.07; model 2: lumpectomy with radiation vs. unilateral mastectomy 0.04 additional mammograms p.a., 95%CI 0.01-0.06). The estimated effect of endocrine therapy did not change noticeably when year 1 was excluded. We also compared the AIC values of the two models after excluding year 1. The model 2 AIC value was slightly lower, but there was hardly a difference between them (AIC=7820.7 for model 1, and AIC=7812.2 for model 2).

Table 21: Models 1 and 2 after excluding observations with person time in year 1 after diagnosis: A) Model 1: Risk factor rate ratios of mammography utilization, and B) Model 2: Risk factor coefficients (change in number of mammograms per year)

A)

Risk Factor	**Rate Ratio**	**95% CI**
Age (per 5 Years)	0.99	0.98-0.99
Family History		
Yes	0.99	0.96-1.03
No	1 (Ref.)	
Surgical Treatment		
Biopsy/None	1.05	0.92-1.18
Lumpectomy, no Radiation	1.02	0.97-1.08
Lumpectomy, with Radiation	1.04	1.01-1.07
Unilateral Mastectomy	1 (Ref.)	
Endocrine Therapy		
Yes	1.06	1.03-1.08
No	1 (Ref.)	

B)

Risk Factor	**Base Rate Difference**	**95% CI**
Age (per 5 Years)	-0.01	-0.02-0.00
Family History		
Yes	0.00	-0.04-0.03
No	0 (Ref.)	
Surgical Treatment		
Biopsy/None	0.05	-0.07-0.17
Lumpectomy, no Radiation	0.02	-0.03 0.07
Lumpectomy, with Radiation	0.04	0.01-0.06
Unilateral Mastectomy	0 (Ref.)	
Endocrine Therapy		
Yes	0.05	0.03-0.08
No	0 (Ref.)	

Table 22: Results from the sensitivity analyses in comparison with the original results. A) Model 1: Risk factor rate ratios of mammography utilization, and B) Model 2: Risk factor coefficients (change in number of mammograms per year)

	Original Model	**Sensitivity Analysis 1**	**Sensitivity Analysis 2**
A)			
Risk Factor	**Estimate (RR)**	**Estimate (RR)**	**Estimate (RR)**
Intercept			
(Reference Rate, Mammograms p.a.)	1.510	1.526	1.390
Age (per 5 Years)	0.997	0.985	0.987
Family History			
Yes	0.992	0.996	0.993
No	1 (Ref.)		
Surgical Treatment			
Biopsy/None	1.050	1.058	1.048
Lumpectomy, no Radiation	1.018	1.015	1.015
Lumpectomy, with Radiation	1.041	1.045	1.045
Unilateral Mastectomy	1 (Ref.)		
Endocrine Therapy			
Yes	1.057	1.055	1.053
No	1 (Ref.)		
B)			
Risk Factor	**Estimate**	**Estimate**	**Estimate**
C	1.060	1.058	1.043
α		-0.008	-0.009
Age (per 5 Years)	-0.014	-0.014	-0.012
Family History			
Yes	-0.005	0.000	-0.003
No	1 (Ref.)		
Surgical Treatment			
Biopsy/None	0.049	0.056	0.046
Lumpectomy, no Radiation	0.016	0.014	0.013
Lumpectomy, with Radiation	0.036	0.040	0.039
Unilateral Mastectomy	1 (Ref.)		
Endocrine Therapy			
Yes	0.054	0.053	0.050
No	1 (Ref.)		

The results from our sensitivity analyses (sensitivity analysis1: including observations with diverging 'subjective' and 'objective' dates of initial diagnosis; sensitivity analysis 2: excluding MRI counts from the outcome counts) in comparison with the estimates from the original analysis are shown in Table 22. There were hardly any differences in the estimates in any of these model versions compared with the main models..

5.4.2 Annual Mammography Distributions

The monthly rate distributions of women's most recent self-reported mammogram in comparison with their month of diagnosis are shown in Figures 13 and 14. The distributions until year 10 supported our hypothesis that mammography utilization rates were highest in the month of diagnosis each year, i.e. when the difference between the month of diagnosis and month of last mammogram was 0, or equivalently 12 months, and that the amplitude of that annual maximum shrank with time. Monthly rates were also above average for the month difference 1/-11 (e.g. if month of diagnosis was January, and the month of last mammogram was February, i.e. one month after the diagnosis months or 11 months before), and the month differences 2/-10, and 3/-9, and 11/-1. These month differences all represented months that were close to the month of diagnosis in each year. The rates at month differences 4/-8 until 10/-2 tended to be below average, which furthermore supported our hypothesis that there was an oscillation pattern each year. The oscillation was not as symmetrical as in a damped cosine wave that we assumed in model 2, but the general trend supported our assumptions. The visualization of rates in subsequent years through the end of year 16 still confirmed the general oscillating trend, but the trend was less clearly visible, see Figure 14. Rates were still highest at month differences 0/12, 1/-11, and 2/-10, but at the remaining month differences, the rates were arbitrarily scattered around the uniformly distributed average rate line. The fact that the oscillation pattern was less prominent in later years supported our hypothesis that the oscillation trend disappeared with time.

5.5 Discussion

Both models 1 and 2 gave consistent results: Estimated mammography utilization rates in year 1 were substantially larger than utilization rates in later years. As an outlier, observed person time in year 1 may have unduly influenced time trend estimates in both models. After excluding observations with person time in year 1, the estimated utilization decline with time was reduced; and while statistically significant, it was minimal in both models (ranging between -0.008 and -0.013 mammograms p.a. in both models after excluding year 1). Estimated annual average utilization rates in model 1 fluctuated around 1 mammogram per year. Lumpectomy with and without radiation, and endocrine therapy were associated with greater, and increasing

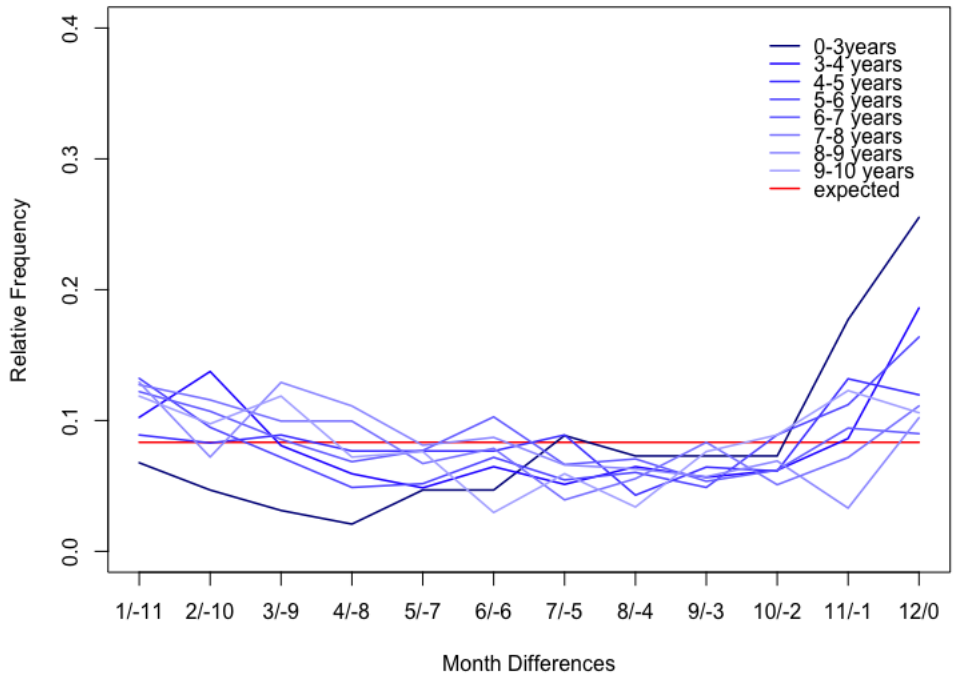

Figure 13: Distribution of differences between month of diagnosis and month of self-reported last mammogram by year since diagnosis (first 10 years after diagnosis). The red line marks the theoretical expected distribution if rates were uniformly distributed.

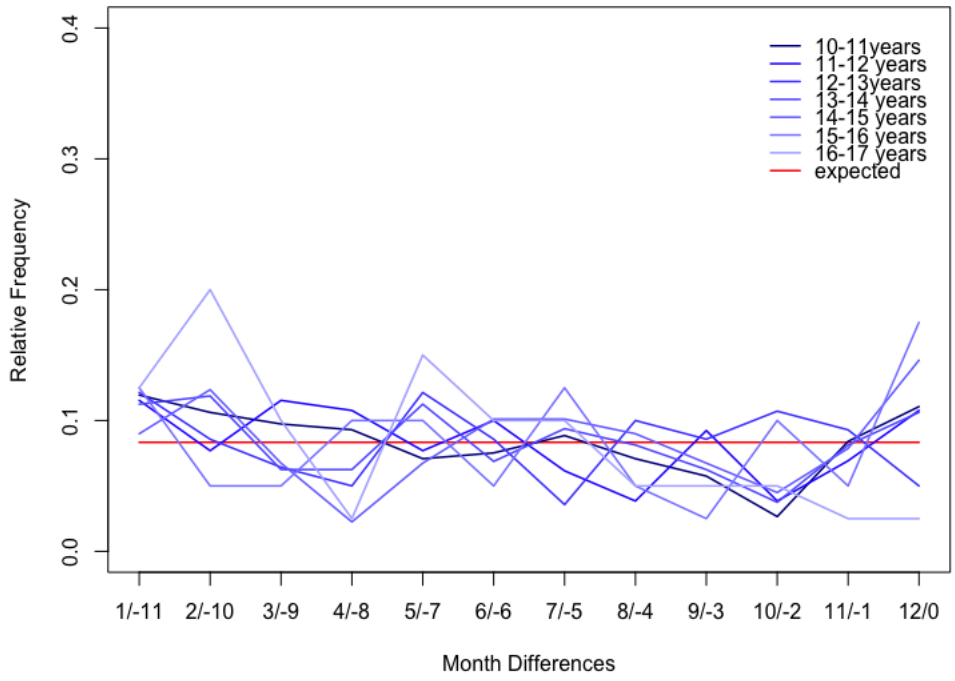

Figure 14: Distribution of differences between month of diagnosis and month of self-reported last mammogram by year since diagnosis (years 10-16). The red line marks the theoretical expected distribution if rates were uniformly distributed.

age with lower utilization rates, but the effects were not clinically significant. Our visual comparisons of monthly utilization rate distributions relative to the month of diagnosis confirmed our hypotheses that annual utilization rates oscillated around a base rate each year, but became more uniformly distributed with time.

Increased mammography utilization in year 1 is likely explained by additional monitoring mammograms that patients received during and immediately after treatment. To avoid capturing diagnostic mammograms that were part of a more frequent monitoring regime in the context of treatment, other authors excluded the first months after treatment in their evaluations [81,83]. Our findings are consistent with the findings by other authors who assessed mammography utilization among breast cancer survivors across multiple cancer stages, and who reported that stage 0 cancer patients tended to be more likely to have used mammography surveillance compared with later stage cancer patients [81–83]. However, unlike other authors, we did not detect a meaningful decline in utilization with time: although we saw a statistically significant decline in the rates, the decline was not clinically significant, especially after year 1 was excluded. Patients in our study tended to have approximately 1 mammogram each year as recommended. The minimal decline in utilization with time that we did detect may have been attributable to a gradual delay in mammography utilization, e.g. patients tending to have a mammogram 13 or 14 months instead of 12 months after their last mammogram in later years. Compared with later stage breast cancer, DCIS is disproportionally likely to have been diagnosed by mammography, and in our study, the fraction of patients who reported that their initial diagnosis was detected by a screening mammogram was 87%. Therefore, compared with other breast cancer survivors, DCIS survivors may represent a population with generally strong beliefs in the effectiveness of mammography, and may therefore use mammography more regularly both before and after a diagnosis, which could explain why we found evidence for guideline-adherent mammography surveillance utilization and no meaningful decline in utilization rates in our study. Consistently with other authors [81–83,176], we found significant increases in utilization rates with treatment by lumpectomy and radiation, and a decrease in utilization rats with increasing age, but the estimated effect sizes made no clinically meaningful difference.

When comparing our two models, there was no substantial advantage of our more complicated model 2 which made more assumptions on the distribution of mammography utilization over time compared with our more simple average rate model 1. In part, that was because the interval-censored data in our study was aggregated over several years at each follow-up. If we had had more granular, for example annual, data on each patient, model 2 might have provided greater insight into specific utilization trends over time. Especially utilization in the first years when the original DCIS diagnosis is still fresh in memory may be different from utilization in later years. If the time trend that we detected was in fact caused by people's tendency to gradually delay mammography by a few months, our model 2 might have allowed us to detect this more clearly with more granular data. Given that DCIS patients are at lower breast cancer mortality

risk than survivors of later breast cancer stages, delaying one's annual mammogram by a few months may not be considered harmful among DCIS patients who have already survived without a recurrence for a long time. If we were able find evidence that a minor decline in mammography utilization with time among DCIS patients is primarily driven by gradual utilization delays among long-term survivors, this might lessen clinical concerns about such a time trend in this specific population.

Our study had strengths: We integrated complex data that was available in different formats at different interviews: at each follow-up, women reported the number of mammograms they had had in the reference time frame differently (counts entered freely from memory vs. pre-specified count categories from a multiple choice questionnaire). Also, observed person-time was measured differently at each follow-up. Our study demonstrates how different formats of the same measure can be integrated into a coherent analysis. Our model 2 is an example of how specific assumptions on the structure of the data can be incorporated into statistical modeling. In our case, the conclusions from our average model 1 and our more complex model 2 were similar, but there may be other applications, in which learning about the specific distribution patterns of events may be more important. An example might be an infectious disease with a seasonal pattern, but with aggregated case counts over long periods of time. In such a case, it may be important to manage resources according to the seasonality of the disease, without being able to analyze time series data where the historical timing of events is precisely known. Integrating a seasonality assumption into the interval-censored data structure might help allocate resources appropriately in such a case. A different example of potential modeling applications is the use of spatial distribution assumptions, e.g. the distribution of cases by geographic altitude or urbanicity when case counts are only available in aggregated form across large areas. Another strength of our study was the use of visualization techniques to evaluate the model 2 distribution assumptions that we could not test with formal statistical tests, making creative use of additional information that was collected on women's most recent mammogram at each interview.

Our study also has limitations: There is a possibility of differential recall bias comparing participants who did or did not answer the mammography questionnaire in the follow-up interviews. However, there was only a small number of patients (N=7) who were excluded from our analysis because they did not provide information on mammography utilization, see Figure 7. According to our findings, DCIS survivors tend to stably adhere to surveillance guidelines over time, but our estimates are averaged across people who utilize mammography frequently and people who do not. Even among DCIS survivors as a potentially guideline-adherent population, there are some people who rarely utilize mammography, and identifying and describing these people in more detail should be the subject of future research.

Our models could be improved: we could include more risk factors (e.g. socio-economic predictors) of long term mammography utilization. Risk factors could be included differently in our model 2: we used

risk factors to predict the base rate at start of follow-up, assuming that a hypothetical time trend was the same across everyone, but the time trend might differ by risk factor profiles. We mostly used categorical predictors, but for our one continuous predictor (age) we did not test for potential nonlinear effects, and we did not test for possible interactions between our predictors. Finally, we excluded observations after a breast cancer recurrence, and all our study participants had DCIS, so our findings cannot be generalized to patients with recurrent or later stage breast cancer. Future research should analyze changes in mammography behaviors and treatment decisions among DCIS patients with recurrences.

5.6 Conclusion

Although we found a statistically significant decline in mammography utilization with time among DCIS survivors, the estimated downward trend was too small to make a meaningful impact on mammography utilization. Similarly, we found statistically significant differences in utilization by treatment and age, but the effects were not clinically significant. Our findings need to be replicated in other longitudinal studies of DCIS survivors, but if true, our findings indicate that DCIS survivors represent a generally guideline-adherent population with regards to mammography utilization. Our model 2 exemplifies how assumptions of a specific utilization pattern over time can be incorporated into a Poisson regression that was based on aggregated event counts over periods of time. In our case that did not lead to more insights compared with a more simple model of annual average mammography utilization rates, but there may be situations when estimating more specific event patterns may be relevant, e.g. to manage resources more effectively in time and space.

6 Conclusion

6.1 Summary of results

In chapter 2, we estimated that approximately 2% of the US population delay health care utilization because of lack of transportation. Estimated probabilities of delaying care because of transportation barriers were not equally distributed in the population: while among middle class and high income populations, delayed care because of transportation was negligible, among low income African American and Non-Hispanic other populations, there was a non-negligible minority (10-15%) who reported having delayed care because of transportation.

In our profile analysis of car ownership in chapter 3, we found that low incomes were strongly associated with not owning a car; and in an exploratory visual analysis we saw that low-income African

Americans and Hispanics were less likely to own a car than low-income whites in our study population, even though we did not have the power to confirm these differences statistically. Another strong predictor of non-car ownership was young and old age. In other words, people who reported not owning or leasing a car were demographically essentially the same as the low socioeconomic status groups that we found to be likely to report having delayed care because of lack of transportation in chapter 2. We did not, however, find evidence in chapter 3 that car ownership was associated with or mediated the effects of known predictors on frequency of general preventive care or mammography utilization, either among low-income or in the general population. We had the counterintuitive finding that owning a car was associated with being less likely to have had at least one mammogram in the past 2 years among low-income women. However, given our very small sample size, these estimates need to be interpreted with caution, and we demonstrated that there may have been confounding by the place that women lived in (appendix A), with most low-income women without a car but high mammography frequency living in Milwaukee, with likely more alternative transit options and more nearby mammography facilities close to people's homes than in other places in Wisconsin.

In chapter 4, we found evidence that increased spatial density of mammography facilities near women's homes was associated with greater frequency of mammography utilization, with decreasing marginal effects with each added facility within a 10km radius around a woman's home, and no detectable effects beyond 2-3 nearby facilities. We also found some evidence for a nonlinear relationship between driving times and frequency of mammography utilization. Long driving times tended to be associated with less frequent mammography utilization, but the effects of increasing driving times were not consistent when comparing moderate to short driving times.

With regards to long-term mammography utilization among DCIS survivors, in chapter 5 we found evidence that mammography utilization rates in the first year after diagnosis exceeded rates in later years and strongly influenced the estimated trend of utilization rates with time. After excluding the first year after diagnosis, we found that participants tended to have one annual mammogram in later years, without a clinically significant time trend or clinically meaningful differences in utilization rates by age at diagnosis, treatment, or family history. When comparing our two Poisson models, there was no noticeable advantage of using a more complex model 2 that incorporated assumptions about how utilization rates were distributed in each year compared with a more simple model 1 that calculated annual average rates. However, there may be other modeling applications when it would be more relevant to incorporate assumptions which meaningfully structure interval-censored or otherwise aggregated data. Using visualization techniques, we were able to confirm the assumptions that we made in our model 2 which we could not evaluate using formal statistical tests.

In addition to these findings that were directly relevant to our main research questions, we had

some additional findings that may be relevant for guiding future research:

-Across chapters 2-4, we found that income was nonlinearly associated with all outcomes that we were looking at: with reporting having delayed care (chapter 2), with owning a car, and with the time since last utilization of general care and mammography (chapter 3). The pattern we found was the same in all of these analyses: having incomes above the poverty level increased the likelihood of owning a car, and decreased the likelihood of having delayed care, but high incomes compared with medium incomes had decreasing or non-significant effects.

- In an exploratory analysis (appendix B), we attempted to find evidence for a 'new type' of 'non-car owner' that has been described by the media as symptomatic of the 'millenials' generation: a young, urban, highly educated and relatively wealthy non-car owner who unlike low-income populations makes a conscious choice against owning a car. However, among people with incomes above 200% the federal poverty level, we were unable to predict for individuals whether they did or did not own a car. Nevertheless, when comparing certain demographic characteristics between car owners and non-car owners aged $\leq$30 with incomes above 200% the federal poverty level, some general patterns emerged, and the logistic regression with car ownership as outcome assigned lower probabilities of car ownership to non-car owners than to car owners. Non-car owners displayed characteristics of lower socioeconomic status with regards to income, education, and home ownership, and they were younger. The most interesting finding may be that relatively wealthy non-car owners had higher walk scores and tended to live in Dane County. Dane County has been lauded as bike- and walking friendly [177]. These findings may imply that if infrastructure is in fact provided, more people decide against buying a car.

6.2 Limitations

The specific limitations for each analysis were discussed in the respective chapters. Here, overarching limitations across our analyses are reiterated by theme.

Limitations in our measures. In chapter 3, car ownership was our central measure. However, car ownership is not the same as access to a car. People who do not own a car may be able to borrow a car from friends, family and neighbors e.g. to use health care services. For the purpose of this work, non-car ownership would misclassify people who get rides or can borrow a car, as they would have transportation access that is comparable to people who do own a car. Conversely, we also did not measure the number of cars per household. Owning a car includes households with multiple adults but only one car. Households with only one car may still face access limitations since household members have to share the one car available for competing needs. None of our geographic and transportation access measures accounted for

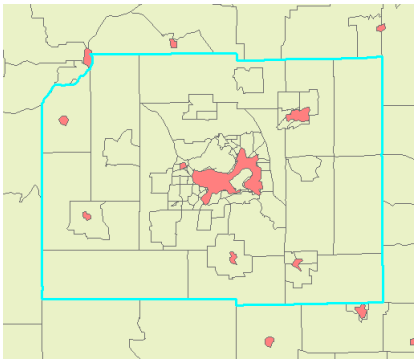

Figure 15: 2000 Census Tracts within Dane County, Wisconsin, with the city of Madison at the center

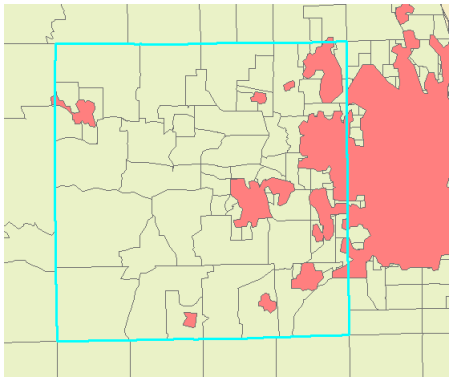

Figure 16: 2000 Census Tracts within Waukesha County, Wisconsin, West of the city of Milwaukee

quality of care, which may vary for example by location, e.g. in rural vs. urban areas, and by type of health insurance [178,179]. The availability of public transit can lead to different implications of non-car ownership with regards to access. We did not account for public transportation options as an alternative to driving in chapters 3 and 4. It was our original intention to include transit scores [123] in these analyses. However, we ended up not using them because as a measure of transportation alternatives, transit scores have their own problems:

- Within Wisconsin, transit scores are only available for Madison, Milwaukee, Greenbay.

- There are no historical transit scores available, which made them incompatible with the data we used in chapter 4.

- Transit scores measure how easy it is to leave from a given place using public transit, by counting the number of different alternative transportation options leaving from that place, and their frequency. However, transit scores do not measure how well two specific places are connected to one another, e.g. a home to a health care provider.

In chapters 3 and 4, we used rural urban commuting area (RUCA) codes to classify individuals as urban, rural, or suburban. RUCA code definitions incorporate commuting behavior criteria. An alternative

would have been to use Census measures of urbanicity whose definition is based on population size and land use patterns. The two classification systems overlap when the classification is obvious (individuals living in large cities, or individuals living in low-populated areas far from cities); but there are differences when the definitions are not as clear-cut. There is no clear consensus as to how urbanicity should be classified in public health research [180]. For the transportation-access focussed aims of this work, RUCA codes seemed like the appropriate choice as compared with Census measures of urbanicity. However, RUCA codes have their shortcomings, as demonstrated in Figures 15 and 16 as two examples. The figures show 2000 Census Tracts in Dane County, with the city of Madison a its center, and in Waukesha County, which is located just West of the city of Milwaukee. In these two counties, according to the 2-category RUCA classification, all 2000 Census Tracts are urban, and not rural. However, Census Tracts that are located along the edges of those counties have rather rural characteristics with regards to population density and access to services and life opportunities. The reason why they are classified as urban is the dominant commuting patterns in these Census Tracts (to the cities of Madison and Milwaukee) for work, i.e. these Census Tracts are influenced by the general urban characteristics of the county that they are located in. For the purposes of our analyses, these Census Tracts were likely misclassified and would have been better classified as rural. In order to improve upon the limitations of RUCA codes, it may be relevant to consider distance traveled and distance to a major highway to measure convenience of transportation, rather than accounting only for the direction of commuting flows as RUCA codes do. Additionally, Census Tract shapes and sizes are rather arbitrary. Depending on the population density, a Census Tract can be rather large or small. This can lead to distortions in underlying public health estimates due to issues with zoning and scale [181]. For the purpose of our work, likely the most important implication arising from problems of RUCA codes is that in Wisconsin, some Census Tracts with rural characteristics were wrongly classified as urban.

Sources of potential selection bias and confounding in our estimates. There is a possibility of participation bias in our studies. When we compared our SHOW study population in chapter 3 with the population that we would have expected from sampled counties we found that participants were older, less likely to come from a non-white minority, had higher than average education levels and were less likely to be unemployed. In other words, participation might have created bias by under-representing socially disadvantaged groups, which would likely be most affected by the relationships that we hypothesized, assuming our hypotheses were correct. Similarly, in chapter 4 we found that women with missing data on mammography utilization tended to have lower education levels and slightly lower incomes, and were more likely to be non-white than the people in our eligible sample, creating potential answer bias.

Car ownership is strongly confounded by income: only 50 out of 140 non-car owners in our study population in chapter 2 had incomes that were above 200% the federal poverty level. The tight correlation

with income makes it hard to separate the individual effects. We addressed this by running our mediation analyses in chapter 3 restricting to incomes lower than 200% the federal poverty line. However, this greatly reduced our sample size and hence our power to detect an effect. In addition to confounding by income, in chapter 3 we showed how the estimated effect of car ownership may be confounded by existing public transit options and better general availability of services in a place like Milwaukee that we did not control for in our analyses.

Potential misclassification: In our analyses, participants may not have accurately remembered when they last used general preventive care or mammography, or how many mammograms they had had in the past years. The recall bias may be greater the longer ago the last time of utilization was, which would cause differential bias with an unpredictable impact on our estimates. It is also possible that some of the people whose outcome measures were missing were more likely to be on the lower end of utilization spectrum: they might not have had a mammogram, or used preventive care in a long time, which may mean that missing outcome variables were not entirely missing at random.

There is also a possibility that some of our predictors were misclassified: the driving times that we used in chapter 4 were estimated using 2015 road infrastructure and driving times, whereas the data on mammography utilization was collected between 1995 and 2007. Infrastructure and traffic conditions may not have changed much since then; but if they did, the misclassification was likely non-differential, which would likely mean that our estimates were biased towards the null.

Generalizability. Our analyses are specific to the types of care that we chose as outcomes (preventive general care and mammography services), and are not generalizable to other types of health care. We did not find consistent evidence of an association between car ownership or driving time with preventive care or mammography utilization, but that does not allow for the conclusion that limited transportation has no impact on participation in other life activities. Also, the small sample sizes of the potentially most affected subpopulations may have prevented us from measuring significant effects.

6.3 Implications and Future Research

Based on our analyses, geographical and transportation access may not be the most important among other considerations that determine an individual's uptake of general preventive care or mammography services. However, we found some evidence that the availability of mammography facilities near women's homes may increase mammography utilization. Identifying places without mammography facilities within a reasonably close distance may help find strategies to ensure that women in such places still have easy access to mammography screening. We also found that among low-income groups, a non-negligible minority reported

having delayed care because of lack of transportation. We do not know which specific health car services that was referring to, and our choice of health care services, especially utilization of general preventive care may not have been representative of services that people tend to forego because of transportation barriers. One reason may that both preventive general care and mammography services rarely need to be used even when adhering to guidelines, and people may be able to find transportation alternatives when they do need to use these health care services. For example, in chapter 3 we found that people had not used preventive general care in 35-39 months on average. Transportation effects on utilization may be stronger for other medical services, e.g. dental care whose utilization is recommended 1-2 times per year.

It may be true that US residents tend to buy a car as soon as their resources allow them to, but it is wrong to assume that everybody owns a car. Statements about car ownership and mobility should be made with specific populations in mind. To give an example, our findings indicate that among African Americans with incomes below 200% the federal poverty level, not owning a car may be more common than among Non-Hispanic whites with similar incomes (22 out 44 African Americans with incomes below 200% the federal poverty level did not own a car in our study population, compared with 188 out of 240 Non-Hispanic whites). These numbers are based on small sample sizes and should be replicated in future studies. But if there are indeed subpopulations among which non-car ownership is as common as our numbers imply, that raises the question which population groups policy makers and planners should pay special attention to in the design of transportation and other infrastructure. Certainly infrastructure has to address the needs of the general population, but should it not also address the needs of already disadvantaged and vulnerable groups? For health care providers that implies that there may be a need for creative solutions in order to lower transportation barriers among patients with limited transportation options.

Given the association of non-car ownership with low incomes, and since low-income African American and Hispanic populations may have less access to a car, future research surveys should explicitly focus on the most vulnerable populations rather than the general population. The Wisconsin population is disproportionally Non-Hispanic white, and non-white racial/ethnic groups may be less likely to participate in research surveys. For example, in chapter 3 we saw that SHOW participants were more educated, older, and more likely to be white than the average population that we would have expected from the sampled counties, which biased our study population away from the low socioeconomic status and minority populations that we are most concerned about with regards to access to transportation and health care.

In addition to focussing on the most vulnerable population groups, future research should investigate other kinds of life opportunities that vulnerable groups may have limited access to. This includes other kinds of medical care (e.g. specialist care, dental care), but also activities unrelated to health care, e.g. access to extracurricular activities among children. In order to avoid confounding by income, there should possibly be

a focus on activities for which money is not a limiting factor, e.g. public library use, children's museum visits, and affordable sports activities. Although such activities may not seem like population health outcomes at first sight, they are related to downstream and long-term population health outcomes such as BMI and literacy [182], social mobility and related health outcomes [22]. As opposed to health care that is only used occasionally, effects of limited access and transportation may be stronger for life activities such as these which take place more frequently and regularly.

Future research should set one focus on elderly residents in rural areas: rural populations tend to be older, and in Wisconsin, they are ageing faster than the US on average [29]. In rural areas in the US, health care facilities are sparse, and the next available health provider may be far away. For example, in chapter 4, we found that rural women were highly likely to have no mammography facility near their residence, and we found that never-users of mammography were disproportionally rural. Car ownership decreases with age, while medical needs increase, raising the question how future rural populations with greater health care needs will have access to the appropriate health care services.

A related field of research is to deepen the understanding of underlying factors that lead to the spatial distribution of health service facilities, e.g. hospitals and HMO policies, and cost/benefit considerations that determine where new facilities are located. As was shown in Figure 5 with mammography facilities as an example, health services clinics are predictably clustered in and around larger cities in Wisconsin, and otherwise more sparsely scattered across rural areas. Understanding why health care facilities are or are not located in specific places where health care tends to be sparse is relevant in order to improve health services in underserved areas in the US. For example, if certain sites are prohibitive because they would not be financially viable, financial incentives or mobile health services might be made available in order to provide the services necessary in affected areas.

A question worth asking is whether the fact that car ownership is common across large income strata in the US is desirable. As pointed out in chapter 1, there are multiple public health, environmental, and other societal co-benefits from reduced vehicle use; but the US transportation model does not encourage such a reduction. Low-income populations have competing needs to allocate limited financial resources to. Should a car be among the first investments people make as soon as their finances allow them to? What are the opportunity costs at the individual level if low-income populations spend money on cars? These are some of the questions that future research on vehicle ownership and utilization should answer.

An exploratory analysis of a potentially confounded effect of car ownership on mammography utilization (appendix A), and an exploratory analysis of car ownership among young adults with income >200% the federal poverty level (appendix B) point towards the possibility that people may use transportation alternatives and reduce car ownership if other transportation options are available. An important future research

field is the improvement of instruments that measure access to transportation. That includes improved measurement of car access (vs. car ownership), and it includes improved measures of the availability of public transportation that do not only evaluate how easy it is to get away from a specific place as transit scores do, but rather, how well specific places are connected to one another by public transportation. An example of a potentially improved measurement of vehicle access was used in the 2015-16 National Health and Nutrition Examination Survey (NHANES) [183]. The NHANES 2015-16 included a one-time Flexible Consumer Behavior Survey Module with a questionnaire on participating households' access to a vehicle for food shopping: "How do (you/you or anyone who lives in the household) usually get to the store (or stores) where you do most of your grocery shopping?", with answer keys (here aggregated):

- in my car / in a car that belongs to someone I live with / to someone who lives elsewhere
- walk, ride bicycle, bus, subway, or other public transit
- taxi or other paid driver / someone else delivers groceries / other
- no usual mode of traveling to store.

Shopping for food may be one of the first activities people carry out by car as soon as they have access to a car because transporting groceries in anything but a vehicle is highly inconvenient, making this instrument potentially useful to measure vehicle access in contrast to measuring car ownership. The data from the 2016-15 NHANES were not available when the analyses for this dissertation were carried out, which is why we did not include an NHANES-based study in this work.

Appendix B compares car ownership by age groups. With consumption behaviors of the so-called 'millenials' generation being widely discussed by the media, a hypothesis has been stated that millenials do not forego buying cars entirely [56,58,85,86], but just delay buying a car compared with previous generations. Future research should monitor car ownership by age groups to understand potentially changing patterns of car ownership and use over time, accounting also for alternatives to car ownership, such as car sharing services. It is possible that younger generations are generally willing to forego buying a car and only purchase a vehicle eventually because available alternative transportation infrastructure does not correspond to their changing needs any more as they get older, e.g. to the transportation needs of young families. Even if they buy a car eventually, younger generations may be more inclined to reduce vehicle use unless there is no alternative. The potential willingness to avoid owning or using a vehicle would imply that there is a demand for better and denser infrastructure in cities in the US, which would be an opportunity for urban planners to design more sustainable cities with decreased negative impacts on the environment and public health.

With regards to mammography surveillance utilization among DCS survivors, we did not observe

a clinically significant downward trend of mammography utilization with time, nor clinically significant differences in base rate utilization by age, family history, or treatment, which differs from observations that other authors have previously made among patient populations with mixed cancer stage profiles.

Future studies should assess costs and benefits of long-term annual mammography surveillance among DCIS patients, and focus on non-guideline adherent patients, and include more potential predictors of mammography surveillance utilization among DCIS survivorss. As predictors, we included age at diagnosis, treatment regimen, and family history, but we did not, for example, include any potential socioeconomic determinants such as income or education, which have been reported to be associated with mammography screening utilization before any kind of breast cancer diagnosis, and may also be related to mammography use after a breast cancer diagnosis. Furthermore, our model 2 could be improved by incorporating our predictors in different ways: in our model 2, we estimated differences in baseline mammography utilization immediately after diagnosis by the included predictors, but assumed that any existing time trend would be the same across women of different risk factor profiles, which is a questionable assumption.

A separate topic for future work is the application of our model 2 ideas to other modeling scenarios and other scientific fields. If event counts or concentrations are measured aggregated over space or time, as was the case for our interval-censored mammogram counts over time, it may be useful to make specific distribution assumptions if these can be realistically made. This may lead to more accurate estimates of the location of events or concentrations in space and time.

Our investigation of mammography surveillance behaviors among DCIS patients was centered around women without a breast cancer recurrence. Women's surveillance or treatment responses to subsequent recurrences were not the focus of our study. We ran an exploratory analysis comparing treatment decisions among non-recurrent patients vs. patients with one or two breast cancer recurrences (appendix E). It can be seen that with each recurrence, more women tended to have unilateral or bilateral mastectomy, while the fraction of patients with only lumpectomy decreased, and the fraction of patients without any surgical treatment became essentially zero. How those treatment decisions after a breast cancer recurrence might influence ongoing utilization of mammography should be the topic of future research.

A Potential confounding of the car ownership effect in chapter 3

Table 23: Potential confounding in the effect of car ownership on mammography utilization among low-income women (N=66), Survey of the Health of Wisconsin, 2014-2015

Parameter	Odds Ratio	95 % CI
Car Ownership		
Unadjusted	0.19	0.04-0.96
Adjusting for Milwaukee Location	0.37	0.06-2.27
Adjusting for Milwaukee Location and Importance of Public Transit	0.39	0.06-2.40

Table 23 demonstrates potential confounding of our estimated effect of car ownership on mammography utilization among low-income women. The unadjusted effect implies that low-income women without a car are less likely to use mammography screening (owning a car vs. not owning a car, OR 0.19, 95% CI 0.04-0.96). However, looking at the data in detail revealed that the effect estimate could mostly be traced back to 7 women who did not own a car but reported using mammography annually, of which 6 were non-white women living in Milwaukee, and reported that public transit was an important or very important criterion when moving to their place of residence. After including an indicator variable for Milwaukee as a location, the effect size of car ownership considerably shrank towards the null, and became statistically non-significant (OR 0.37, 95% CI 0.06-2.27). It is likely that Milwaukee a) offers more public transit options than other places in Wisconsin, and b) that the spatial density of mammography facilities in Milwaukee is greater than in other places in Wisconsin, such that Milwaukee residents have better geographic access to mammography services even if they do not have a car, which would imply confounding of our estimated effect of car ownership by other transportation options and better overall availability of services.

B Supplemental tables for profile analysis of car ownership in chapter 3

Table 24: Comparison of car owners with non-car owners aged ≤30 (so-called 'millenials') with incomes ≤200% the federal poverty level (N=87), Survey of the Health of Wisconsin, 2014-2015

Characterstic	car owners	non-car owners
Mean Propensity score of owning a car according to our final logistic model	0.89	0.60
Mean PIR	4.82	3.77
Fraction of Home Owners (%)	54%	30%
Mean Years of Education	15.6	14.2
Mean Age (Years)	26.5	23.8
Fraction living in Dane County (%)	42%	62%
Mean Walk score (1-100)	32.6	55.4

Table 25: Car-ownership by age-group (N=1,237), Survey of the Health of Wisconsin, 2014-2015

Age (Years	% Car Owners	% Non-Car Owners
≤20	44.4	55.6
21-30	71.5	28.5
31-40	91.8	8.2
41-50	93.1	6.9
51-60	96.6	3.4
61-70	90.7	9.3
71-80	93.6	6.4
>80	81.2	18.8

In an attempt to assess whether it would be possible to identify the typical 'millenial' that has been described by the media [85, 86], i.e. a young, relatively wealthy and well educated urban resident who does not own a car, Table 24 compares car owners and non-car owners aged ≤30, and with incomes ≥200% the federal poverty level. Based on our data and logistic regression, we were not able to predict at the individual level who did or did not own a car. However, there exist distributional differences between people in this group who did or did not own a car: non-car owners were assigned lower probabilities of car-ownership by the logistic regression (on average 0.60 compared with 0.89 among car-owners), and they were younger (23.8 vs. 26.5 years), had lower incomes (3.77 PIR vs. 4.82 PIR) and lower education levels (14.2 vs. 15.6 years), and were less likely to be home-owners (30 vs. 54 %). The socioeconomic differences might, however, be a direct consequence of their younger age. The most interesting finding may be in the differences in average walk scores (55.4 vs. 32.6) and in where non-car owners tended to live: 62% of non-car owners vs. 42% of car-owners lived in Dane County.

Table 25 compares population fractions of car owners vs. non-car owners by age group, which confirms the association of young an old age with non-car ownership as estimated in our logistic regression

model. While not providing any additional insight by itself, this provides one snapshot to add to the literature that monitors car ownership by age group over time. In the context of similar future studies, this may eventually help answer the question whether younger generations are more likely to forego purchasing a car, or whether they delay purchasing a car compared with older generations.

C Supplemental tables for chapter 4

Table 26: Sensitivity Analysis 1: excluding mammography facilities outside Wisconsin. Odds ratios of greater mammography frequency by geographic access (N=5,930) Wisconsin Women's Health Study, 1995-2007

	Urban		**Rural**	
Main Exposures	OR	95% CI	OR	95% CI
Driving Time				
20 vs. 10 Minutes	1.06	0.75-1.49	1.43	1.06-1.92
30 vs. 20 Minutes	0.64	0.38-1.10	0.92	0.79-1.08
40 vs. 20 Minutes	0.32	0.08-1.29	0.65	0.44-0.97
Mammography Facilities within 10km Radius, N				
1 vs. 0 Facilities	1.24	1.06-1.44	1.17	0.92-1.49
2 vs. 1 Facilities	1.16	1.04-1.31	1.10	0.94-1.30
3 vs. 2 Facilities	1.05	0.99-1.12	1.00	0.86-1.15
4 vs. 3 Facilities	0.97	0.92-1.02	0.93	0.76-1.14

Models adjust for education, income, number of household members, race, and mammography capacity, age, family history of breast cancer, and indicators of neighborhood deprivation by Census Tract: population fraction below the poverty line, median 1999 income, population fraction without a vehicle, population education levels, population fraction without health insurance (by county).

Table 27: Sensitivity Analysis 2: including an indicator of health insurance (A: with insurance, B: without insurance). Odds ratios of greater mammography frequency by geographic access (N=759) Wisconsin Women's Health Study, 1995-2007

	Urban		**Rural**	
Main Exposures	OR	95% CI	OR	95% CI
A)				
Driving Time				
20 vs. 10 Minutes	0.80	0.28-2.31	1.03	0.40-2.67
30 vs. 20 Minutes	0.56	0.12-2.60	0.78	0.45-1.35
40 vs. 20 Minutes	0.26	0.00-16.75	0.50	0.11-2.28
Mammography Facilities within 10km Radius, N				
1 vs. 0 Facilities	1.18	0.74-1.88	0.88	0.40-1.95
2 vs. 1 Facilities	1.11	0.78-1.58	0.87	0.53-1.44
3 vs. 2 Facilities	1.00	0.83-1.21	0.88	0.54-1.43
4 vs. 3 Facilities	0.93	0.79-1.10	0.95	0.55-1.63
B)				
Driving Time				
20 vs. 10 Minutes	0.89	0.31-2.56	1.13	0.44-2.90
30 vs. 20 Minutes	0.44	0.09-2.00	0.78	0.45-1.35
40 vs. 20 Minutes	0.13	0.00-8.14	0.48	0.11-2.15
Mammography Facilities within 10km Radius, N				
1 vs. 0 Facilities	1.24	0.78-1.99	0.86	0.39-1.91
2 vs. 1 Facilities	1.16	0.81-1.64	0.87	0.53-1.43
3 vs. 2 Facilities	1.02	0.85-1.23	0.89	0.55-1.44
4 vs. 3 Facilities	0.93	0.79-1.09	0.96	0.56-1.65

Models adjust for education, income, number of household members, race, and mammography capacity, age, family history of breast cancer, and indicators of neighborhood deprivation by Census Tract: population fraction below the poverty line, median 1999 income, population fraction without a vehicle, population education levels, population fraction without health insurance (by county).

Table 28: Sensitivity Analysis 3: including an indicator of ever use of postmenopausal hormones. Odds ratios of greater mammography frequency by geographic access (N=5,930) Wisconsin Women's Health Study, 1995-2007

	Urban		**Rural**	
Main Exposures	OR	95% CI	OR	95% CI
Driving Time				
20 vs. 10 Minutes	1.06	0.74-1.51	1.45	1.06-1.97
30 vs. 20 Minutes	0.57	0.34-0.98	0.91	0.77-1.07
40 vs. 20 Minutes	0.24	0.06-0.99	0.63	0.42-0.96
Mammography Facilities within 10km Radius, N				
1 vs. 0 Facilities	1.25	1.07-1.46	1.19	0.93-1.52
2 vs. 1 Facilities	1.17	1.04-1.32	1.12	0.95-1.32
3 vs. 2 Facilities	1.05	0.99-1.12	1.01	0.86-1.17
4 vs. 3 Facilities	0.96	0.92-1.01	0.94	0.75-1.16

Models adjust for education, income, number of household members, race, and mammography capacity, age, family history of breast cancer, and indicators of neighborhood deprivation by Census Tract: population fraction below the poverty line, median 1999 income, population fraction without a vehicle, population education levels, population fraction without health insurance (by county).

Table 29: Sensitivity Analysis 4: including more granular Rural Urban Commuting Area Codes. More than 75% of women in isolated rural towns have no mammography facility within a 10km radius around their homes, which results in unrealistic estimates in this category. Odds ratios of greater mammography frequency by geographic access (N=5,930) Wisconsin Women's Health Study, 1995-2007

	Urban		**Large Rural City/Town**		**Small Rural Town**		**Isolated Small Rural Town**	
Main Exposures	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
Driving Time								
20 vs. 10 Minutes	1.10	0.77-1.56	2.60	1.28-5.28	1.63	0.80-3.32	1.19	0.78-1.84
30 vs. 20 Minutes	0.63	0.37-1.07	0.93	0.36-2.37	0.70	0.38-1.29	0.89	0.75-1.07
40 vs. 20 Minutes	3.38	0.07-1.19	0.46	0.04-5.51	0.31	0.06-1.59	0.66	0.45-0.98
Mammography Facilities within 10km Radius, N								
1 vs. 0 Facilities	1.22	1.05-1.42	1.27	0.79-2.02	1.29	0.70-2.41	1.02	0.58-1.81
2 vs. 1 Facilities	1.16	1.03-1.30	1.06	0.80-1.40	1.33	0.89-1.99	1.60	0.72-3.56
3 vs. 2 Facilities	1.05	0.99-1.12	0.80	0.62-1.05	1.19	0.43-3.24	3.92	0.06-261.41
4 vs. 3 Facilities	0.98	0.94-1.02	1.32	0.56-1.04	0.47	0.09-2.51	15.36	0.00-∞

Models adjust for education, income, number of household members, race, and mammography capacity, age, family history of breast cancer, and indicators of neighborhood deprivation by Census Tract: population fraction below the poverty line, median 1999 income, population fraction without a vehicle, population education levels, population fraction without health insurance (by county).

Tables 26-29 show the results of the sensitivity analyses in chapter 4. The final model with more confounders beyond the minimal adjustment set was rerun a) excluding mammography facilities outside Wisconsin (Table 26), b) with and without an indicator of health insurance (Table 27), c) including an indicator of ever-use of postmenopausal hormones (Table 28), and d) using a more detailed categorization of Rural Urban Commuting Area Codes (Table 29).

D Detailed model 2 approach chapter 5

In our model 2 approach, we assumed that women tended to use mammography in and around the month of their initial diagnosis, such that mammography utilization rates would oscillate over the course of each year after diagnosis. We furthermore assumed that the oscillation would flatten out with time. Mathematically, we conceptualized mammography rates to have the shape of a damped cosine wave, as visualized in Figure 8. The number of mammograms up to a specific time point would then equal the shaded area under the cosine curve. In physics, a damped cosine wave is enveloped by an exponential decay curve, and fluctuates around a base rate $C(t)$ at a given time t. If the base rate did not change with time, $C(t)$ would be a constant. With this approach, the expected number of mammograms between two time points t_0 and t_1 would then be estimated by the integral of the damped cosine wave between t_0 and t_1. The exponential decay curve that frames the maxima of a damped cosine curve is formulated as:

$$A(t) = A_0 * \exp(-\lambda t) \tag{11}$$

And the damped cosine curve approach to describe mammography utilization rates would be:

$$Rate(t) = A_0 * \exp(-\lambda t) * \cos(\omega t + \phi) + C(t) \tag{12}$$

Where:

$$\begin{aligned}
A(t) &: utilization\ rate\ at\ a\ specific\ time\ t \\
A_0 &: the\ initial\ maximum\ amplitude \\
\lambda &: decay\ constant\ describing\ how\ quickly\ the\ oscillation\ flattens\ out \\
\omega &: frequency\ of\ the\ oscillation\ = 1 * year^{-1} \\
\phi &: phase\ (\pi/2)\ to\ place\ maxima\ of\ the\ oscillation\ curve\ at\ full\ years \\
C(t) &: base\ rate\ at\ time\ t\ around\ which\ the\ rate\ oscillates
\end{aligned}$$

The expected number Y of mammograms between a starting point t_0 and and end point t_1 would then be estimated as the integral of Rate(t) between those time points:

$$E[Y] = \int_{t_0}^{t_1} (A_0 * \exp(-\lambda t) * \cos(\omega t + \phi) + C(t))dt \tag{13}$$

In order to avoid having to apply nonlinear regression methods to estimate the unknown parameters A_0, λ, and C, and to allow for a more gradual decay of the amplitudes of the oscillation curve, we made several alterations in our modeling approach. Firstly, we approached the cosine curve using linear splines. Secondly, we used the hyperbolic curve $1/t$ to describe a more gardual decline of the amplitudes with time. Thirdly, we assumed that $C(t)$ would change (decrease) with time. This altered approach is visualized in Figure 9. With these adjustments, and assuming that women were observed for full half-years at a time, the expected number of mammograms up to an end point t_1 would then be estimated by the discrete areas A_i under the spline curve:

$$E[Y] = A_1 + A_2 + A_3 + ...A_i$$

The maxima of the linear splines were assumed to be exactly one year apart, and the minima were assumed to occur half a year after each maximum. We hence estimated the envelope functions framing the maxima and minima of the splines as hyperbolic decay curves :

$$\begin{aligned} 1/t + C + \alpha t \ (upper\ envelope) \\ -1/t + C + \alpha t \ (lower\ envelope) \end{aligned} \tag{14}$$

Where C described the initial base rate, and α the change in the base rate with time. Referring to the notation in Figure 9, the areas A_1, A_2, A_3 etc. represented the number of mammograms during half-year intervals, and would be calculated as:

$$\begin{aligned} A_1 &= b_1 * 0.5^2 = b_1/4 \\ A_2 &= b_2 * 0.5 + (b_1 - b_2) * 0.5^2 = b_1/4 + b_2/4 \\ A_3 &= b_2 * 0.5 + (b_3 - b_2) * 0.5^2 = b_2/4 + b_3/4 \\ &etc. \end{aligned} \tag{15}$$

Given the hyperbolic parametrizations in (14) of the envelope curves framing the maxima and minima of the linear splines, the b_i could be calculated as:

$$\begin{aligned} b_1 &= 1/1 + C + 1\alpha \\ b_3 &= 1/2 + C + 2\alpha \\ b_5 &= 1/3 + C + 3\alpha \\ &etc. \end{aligned} \tag{16}$$

and

$$\begin{aligned} b_2 &= -1/1.5 + C + 1.5\alpha \\ b_4 &= -1/2.5 + C + 2.5\alpha \\ b_6 &= -1/3.5 + C + 3.5\alpha \\ &etc. \end{aligned} \tag{17}$$

The area calculations in (15) assume that women were observed for whole half-year intervals at a time. However, women were not observed for full half-year intervals at a time. Instead, women in our study reported mammograms between varying starting time points t_0 and varying end times t_i and may only have been observed in parts of the A_i. To be able to calculate partial areas under the splines, we divided each half-year interval t_i into a starting interval t_{is} and an ending interval t_{ie}. For example, and using the notation in Figure 10, the entire area of A_3 was calculated as the sum of the 'starting' area A_{3s}, and the 'ending' area A_{3e}:

$$\begin{aligned} A_{3s} &= b_2 * \Delta t_{3s} + (b_{t3} - b_2) * \Delta t_{3s} * 0.5 \\ A_{3e} &= b_{t3} * \Delta t_{3e} + (b_3 - b_{t3}) * \Delta t_{3e} * 0.5 \\ A_3 &= A_{3s} + A_{3e} \\ &= \frac{b_2 \Delta t_{3s}}{2} + \frac{b_3 \Delta t_{3e}}{2} + \frac{b_{t3}}{2}(\Delta t_{3s} + \Delta t_{3e}) \end{aligned} \tag{18}$$

Where Δt_{3s} represented the person time a woman spent in the 'starting' interval of the interval between 1.5 and 2 years after diagnosis, and Δt_{3e} the time a woman spent in the 'ending' interval between 1.5 and 2 years. For example, if a woman's observed person time started at 1.6 years after diagnosis, her value of Δt_{3s} would be 0, and her value of Δt_{3e} would be 0.4. Similarly, the starting and ending intervals in A_6 were

calculated as:

$$\begin{aligned} A_{6s} &= b_{t6} * \Delta t_{6s} + (b_5 - b_{t6}) * \Delta t_{6s} * 0.5 \\ A_{6e} &= b_6 * \Delta t_{6e} + (b_{t6} - b_6) * \Delta t_{6e} * 0.5 \\ A_6 &= A_{6s} + A_{6e} \\ &= \frac{b_5 \Delta t_{6s}}{2} + \frac{b_6 \Delta t_{6e}}{2} + \frac{b_{t6}}{2}(\Delta t_{6s} + \Delta t_{6e}) \end{aligned} \tag{19}$$

Where Δt_{6s} represented the person time a woman spent in the 'starting' interval between 3 and 3.5 years after diagnosis, and Δt_{6e} the time a woman spent in the 'ending' interval between 3 and 3.5 years. For example, if a woman's observed person time ended at 3.3 years after diagnosis, her value of Δt_{6s} would be 0.3, and her value of Δt_{6e} would be 0.

The values of b_i in (18) and (19) were defined in (16) and (17), and the values of the b_{ti} could be derived from the trigonometric intersect theorem, see also Figure 10:

$$\begin{aligned} \frac{\Delta t_{3e}}{0.5} &= \frac{(b_3 - b_{t3})}{(b_3 - b_2)} \Leftrightarrow b_{t3} = b_3 - 2b_3 \Delta t_{3e} + 2b_2 \Delta t_{3e} \\ \frac{\Delta t_{6s}}{0.5} &= \frac{(b_5 - b_{t6})}{(b_5 - b_6)} \Leftrightarrow b_{t6} = b_5 - 2b_5 \Delta t_{6s} + 2b_6 \Delta t_{6s} \end{aligned} \tag{20}$$

Using all these relationships, the expected number of mammograms were approximated as the sum of the half-year areas under the linear spline curves up to year 18 (the maximum interview time after initial diagnosis in our study):

$$\begin{aligned} E[Y] &= A_1 + A_2 + A_3 + ... + A_{35} \\ &= \frac{b_1}{4} \quad [A_1, see\ (15)] \\ &+ \frac{b_1 \Delta t_{2s}}{2} + \frac{b_2 \Delta t_{2e}}{2} + \frac{b_{t2}}{2}(\Delta t_{2s} + \Delta t_{2e}) \quad [A_2] \\ &+ \frac{b_2 \Delta t_{3s}}{2} + \frac{b_3 \Delta t_{3e}}{2} + \frac{b_{t3}}{2}(\Delta t_{3s} + \Delta t_{3e}) \quad [A_3] \\ &+ [A_4 + A_5 + ... + A_{34}] \\ &+ \frac{b_{34} \Delta t_{35s}}{2} + \frac{b_{t35} \Delta t_{35s}}{2} \quad [A_{35s}] \end{aligned}$$

Substituting (16) and (17) for the b_i, and (20) for the b_{tis} and b_{tie}, this results in:

$$
\begin{aligned}
E[Y] = {} & \frac{1+C+\alpha}{4} \quad [A_1] \\
& + [\ C(\Delta t_{2s} + \Delta t_{2e}) + \frac{1}{1}\Delta t_{2s} - \Delta t_{2s}^2(\frac{1}{1}+\frac{2}{3}) + \Delta t_{2e}(-\frac{1}{3}+\frac{1}{2}) - \Delta t_{2s}\Delta t_{2e}(\frac{1}{1}+\frac{2}{3}) \\
& + -\alpha(\Delta t_{2s} + \frac{\Delta t_{2s}^2}{2} + \frac{\Delta t_{2s}\Delta t_{2e}}{2} + \Delta t_{2e}(\frac{3}{4}+\frac{1}{2}))] \quad [A_2] \\
& + [\ C(\Delta t_{4s} + \Delta t_{4e}) + \frac{1}{2}\Delta t_{4s} - \Delta t_{4s}^2(\frac{1}{2}+\frac{2}{5}) + \Delta t_{4e}(-\frac{1}{5}+\frac{1}{4}) - \Delta t_{4s}\Delta t_{4e}(\frac{1}{2}+\frac{2}{5}) \\
& + -\alpha(2\Delta t_{4s} + \frac{\Delta t_{4s}^2}{2} + \frac{\Delta t_{4s}\Delta t_{4e}}{2} + \Delta t_{4e}(\frac{5}{4}+\frac{2}{2}))] \quad [A_4] \\
& + [\ C(\Delta t_{6s} + \Delta t_{6e}) + \frac{1}{3}\Delta t_{6s} - \Delta t_{6s}^2(\frac{1}{3}+\frac{2}{7}) + \Delta t_{6e}(-\frac{1}{7}+\frac{1}{6}) - \Delta t_{6s}\Delta t_{6e}(\frac{1}{3}+\frac{2}{7}) \\
& + -\alpha(3\Delta t_{6s} + \frac{\Delta t_{6s}^2}{2} + \frac{\Delta t_{6s}\Delta t_{6e}}{2} + \Delta t_{6e}(\frac{7}{4}+\frac{3}{2}))] \quad [A_6] \\
& + [A_8 + A_{10} + ... + A_{32}] \\
& + [\ C(\Delta t_{34s} + \Delta t_{34e}) + \frac{1}{17}\Delta t_{34s} - \Delta t_{34s}^2(\frac{1}{17}+\frac{2}{35}) + \Delta t_{6e}(-\frac{1}{35}+\frac{1}{34}) - \Delta t_{34s}\Delta t_{34e}(\frac{1}{17}+\frac{2}{35}) \\
& + -\alpha(17\Delta t_{34s} + \frac{\Delta t_{34s}^2}{2} + \frac{\Delta t_{34s}\Delta t_{34e}}{2} + \Delta t_{34e}(\frac{35}{4}+\frac{17}{2})) \quad [A_{34}]
\end{aligned}
\tag{21}
$$

$$
\begin{aligned}
& + [\ C(\Delta t_{3s} + \Delta t_{3e}) - \frac{2}{3}\Delta t_{3s} + \Delta t_{3s}^2(\frac{2}{3}+\frac{1}{2}) + \Delta t_{3e}(\frac{1}{4}-\frac{1}{3}) - \Delta t_{3s}\Delta t_{3e}(\frac{2}{3}+\frac{1}{2}) \\
& + -\alpha(\frac{3}{2}\Delta t_{3s} + \frac{\Delta t_{3s}^2}{2} + \frac{\Delta t_{3s}\Delta t_{3e}}{2} + \Delta t_{3e}(\frac{2}{2}+\frac{3}{4}))] \quad [A_3] \\
& + [\ C(\Delta t_{5s} + \Delta t_{5e}) - \frac{2}{5}\Delta t_{5s} + \Delta t_{5s}^2(\frac{2}{5}+\frac{1}{3}) + \Delta t_{5e}(\frac{1}{6}-\frac{1}{5}) - \Delta t_{5s}\Delta t_{5e}(\frac{2}{5}+\frac{1}{3}) \\
& + -\alpha(\frac{5}{2}\Delta t_{5s} + \frac{\Delta t_{5s}^2}{2} + \frac{\Delta t_{5s}\Delta t_{5e}}{2} + \Delta t_{5e}(\frac{3}{2}+\frac{5}{4}))] \quad [A_5] \\
& + [\ C(\Delta t_{7s} + \Delta t_{7e}) - \frac{2}{7}\Delta t_{7s} + \Delta t_{7s}^2(\frac{2}{7}+\frac{1}{4}) + \Delta t_{7e}(\frac{1}{8}-\frac{1}{7}) - \Delta t_{7s}\Delta t_{7e}(\frac{2}{7}+\frac{1}{4}) \\
& + -\alpha(\frac{7}{2}\Delta t_{7s} + \frac{\Delta t_{7s}^2}{2} + \frac{\Delta t_{7s}\Delta t_{7e}}{2} + \Delta t_{7e}(\frac{4}{2}+\frac{7}{4}))] \quad [A_7] \\
& + [A_9 + A_{11} + ... + A_{33}] \\
& + [\ C(\Delta t_{35s}) - \frac{2}{35}\Delta t_{35s} + \Delta t_{35s}^2(\frac{2}{35}+\frac{1}{18}) \\
& + -\alpha(\frac{35}{2}\Delta t_{35s} + \frac{\Delta t_{35s}^2}{2})] \quad [A_{35}]
\end{aligned}
$$

Rearranging this formula results in:

$$
\begin{aligned}
E[Y] = \mathbf{C}\ [\ 0.25 + \Delta t_{2s} + \Delta t_{2e} + \Delta t_{3s} + \Delta t_{3e} + ... + \Delta t_{35s}] \\
- \boldsymbol{\alpha}\ [\ 0.25 \\
+ (1\Delta t_{2s} + \frac{\Delta t_{2s}^2}{2} + \frac{\Delta t_{2s}\Delta t_{2e}}{2} + \Delta t_{2e}(\frac{3}{4} + \frac{1}{2})) \\
+ (2\Delta t_{4s} + \frac{\Delta t_{4s}^2}{2} + \frac{\Delta t_{4s}\Delta t_{4e}}{2} + \Delta t_{4e}(\frac{5}{4} + \frac{2}{2})) \\
+ (3\Delta t_{6s} + \frac{\Delta t_{6s}^2}{2} + \frac{\Delta t_{6s}\Delta t_{6e}}{2} + \Delta t_{6e}(\frac{7}{4} + \frac{3}{2})) \\
+ ... \\
+ (17\Delta t_{34s} + \frac{\Delta t_{34s}^2}{2} + \frac{\Delta t_{34s}\Delta t_{34e}}{2} + \Delta t_{34e}(\frac{35}{4} + \frac{17}{2})) \\
+ (\frac{3}{2}\Delta t_{3s} + \frac{\Delta t_{3s}^2}{2} + \frac{\Delta t_{3s}\Delta t_{3e}}{2} + \Delta t_{3e}(\frac{2}{2} + \frac{3}{4})]) \\
+ (\frac{5}{2}\Delta t_{5s} + \frac{\Delta t_{5s}^2}{2} + \frac{\Delta t_{5s}\Delta t_{5e}}{2} + \Delta t_{5e}(\frac{3}{2} + \frac{5}{4})) \\
+ (\frac{7}{2}\Delta t_{7s} + \frac{\Delta t_{7s}^2}{2} + \frac{\Delta t_{7s}\Delta t_{7e}}{2} + \Delta t_{7e}(\frac{4}{2} + \frac{7}{4})) \\
+ ... \\
+ (\frac{35}{2}\Delta t_{35s} + \frac{\Delta t_{35s}^2}{2})] \\
+ \textbf{offset}
\end{aligned}
\tag{22}
$$

Which can then be run as Poisson regression with identity link, with C and α as the coefficients, and with

the offset:

$$
\begin{aligned}
offset = \ & 0.25 \\
& + [\frac{1}{1}\Delta t_{2s} - \Delta t_{2s}^2(\frac{1}{1}+\frac{2}{3}) + \Delta t_{2e}(-\frac{1}{3}+\frac{1}{2}) - \Delta t_{2s}\Delta t_{2e}(\frac{1}{1}+\frac{2}{3})] \\
& + [\frac{1}{2}\Delta t_{4s} - \Delta t_{4s}^2(\frac{1}{2}+\frac{2}{5}) + \Delta t_{4e}(-\frac{1}{5}+\frac{1}{4}) - \Delta t_{4s}\Delta t_{4e}(\frac{1}{2}+\frac{2}{5})] \\
& + [\frac{1}{3}\Delta t_{6s} - \Delta t_{6s}^2(\frac{1}{3}+\frac{2}{7}) + \Delta t_{6e}(-\frac{1}{7}+\frac{1}{6}) - \Delta t_{6s}\Delta t_{6e}(\frac{1}{3}+\frac{2}{7})] \\
& + ... \\
& + [\frac{1}{17}\Delta t_{34s} - \Delta t_{34s}^2(\frac{1}{17}+\frac{2}{35}) + \Delta t_{34e}(-\frac{1}{35}+\frac{1}{34}) - \Delta t_{34s}\Delta t_{34e}(\frac{1}{17}+\frac{2}{35})] \\
& + [-\frac{2}{3}\Delta t_{3s} + \Delta t_{3s}^2(\frac{2}{3}+\frac{1}{2}) + \Delta t_{3e}(\frac{1}{4}-\frac{1}{3}) - \Delta t_{3s}\Delta t_{3e}(\frac{2}{3}+\frac{1}{2})] \\
& + [-\frac{2}{5}\Delta t_{5s} + \Delta t_{5s}^2(\frac{2}{5}+\frac{1}{3}) + \Delta t_{5e}(\frac{1}{6}-\frac{1}{5}) - \Delta t_{5s}\Delta t_{5e}(\frac{2}{5}+\frac{1}{3})] \\
& + [-\frac{2}{7}\Delta t_{7s} + \Delta t_{7s}^2(\frac{2}{7}+\frac{1}{4}) + \Delta t_{7e}(\frac{1}{8}-\frac{1}{7}) - \Delta t_{7s}\Delta t_{7e}(\frac{2}{7}+\frac{1}{4})] \\
& + ... \\
& + [-\frac{2}{35}\Delta t_{35s} + \Delta t_{35s}^2(\frac{2}{35}+\frac{1}{18})]
\end{aligned}
\tag{23}
$$

abbreviated as:

$$
\begin{aligned}
E[Y] = & \ \mathbf{C} * expression\ 1 \\
& - \boldsymbol{\alpha} * expression\ 2 \\
& + \mathbf{offset}
\end{aligned}
\tag{24}
$$

Lastly, to take into account treatment regimen, age and family history similarly to model 1, we assumed that the initial base rate C differed by these risk factors, i.e.:

$$
\begin{aligned}
E[C] = & \ C_0 + \gamma_1 * age\ at\ diagnosis + \gamma_2 * family\ history \\
& + \gamma_3 * biopsy/no\ treatment + \gamma_4 * lumpectomy\ (no\ radiation) \\
& + \gamma_5 * lumpectomy\ (with\ radiation) + \gamma_6 * endocrine\ therapy
\end{aligned}
$$

With unilateral mastectomy as the reference treatment. Including these risk factors in (22), the final Poisson

model was:

$$
\begin{aligned}
E[Y] = [&C_0 + \gamma_1 * age\ at\ diagnosis + \gamma_2 * family\ history + \gamma_3 * biopsy/no\ treatment \\
&+ \gamma_4 * lumpectomy\ (no\ radiation) + \gamma_5 * lumpectomy\ (with\ radiation) \\
&+ \gamma_6 * endocrine\ therapy] * expression\ 1 \\
&- \alpha * expression\ 2 \\
&+ offset \\
\\
&= C_0 * expression\ 1 + \gamma_1 * age\ at\ diagnosis * expression\ 1 + \gamma_2 * family\ history * expression\ 1 \\
&+ \gamma_3 * biopsy/no\ treatment * expression\ 1 + \gamma_4 * lumpectomy\ (no\ radiation) * expression\ 1 \\
&+ \gamma_5 * lumpectomy\ (with\ radiation) * expression\ 1 + \gamma_6 * endocrine\ therapy * expression\ 1 \\
&- \alpha * expression\ 2 \\
&+ offset
\end{aligned}
\tag{25}
$$

Equation (25) was estimated in a Poisson regression with identity link, with C_0, α, and $\gamma_1 ... \gamma_6$ as the coefficients, *expression* 1 and *expression* 2 as the covariates, and an offset.

E Exploratory analysis of treatment distributions after recurrences chapter 5

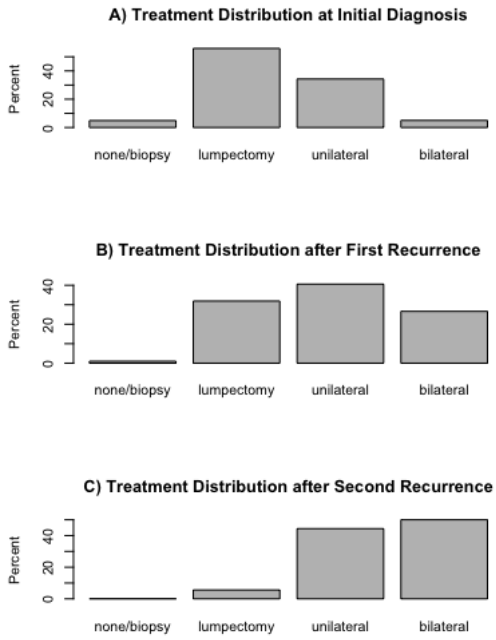

Figure 17: Surgical treatment distributions after initial diagnosis (A), a first breast cancer recurrence (B), and a second breast cancer recurrence (C),

Figure 17 visualizes the distributions of surgical treatment options (none/biopsy, lumpectomy, unilateral and bilateral mastectomy) among people without, with 1, and with 2 breast cancer recurrences. It can be seen that with each recurrence, more women tend to have more aggressive treatment, as the fraction of patients who had mastectomy increases while the fraction of patients with only lumpectomy decreases, and the fraction of patients without any surgical treatment becomes essentially zero.

F R and SAS codes

F.0.1 R Code used for analysis in chapter 2 (NHIS analysis)

```
library(gmodels)
library(multcomp)
library(survey)
library(Hmisc)

adult = read.csv(file="NHIS_samadult.csv", header=TRUE)
family = read.csv(file="NHIS_family.csv",header=TRUE)
person = read.csv(file="NHIS_person.csv",header=TRUE)
child = read.csv(file="NHIS_samchild.csv",header=TRUE)

###### step 1: merge and stack data files ################
# select variables from above files, adult file:
myvars1 <- c("HHX", "FMX", "FPX","AUSUALPL","AHCDLYR5")
adultsam <- adult[myvars1]
colnames(adultsam)[4] <- "USUALPL"
colnames(adultsam)[5] <- "HCDLYR5"

# child file:
myvars4<- c("HHX", "FMX","FPX","CUSUALPL","CHCDLYR5")
childsam <- child[myvars4]
colnames(childsam)[4] <- "USUALPL"
colnames(childsam)[5] <- "HCDLYR5"

# family file
myvars2 <- c("HHX", "FMX","FM_SIZE","FM_KIDS","FM_ELDR",
             "FM_EDUC1","FDMEDYN","FHICOVCT","FDGLWCT2",
             "FDGLWCT1","RAT_CAT5","FLNGINTV")
famsam <- family[myvars2]

#person file:
myvars3<- c("HHX", "FMX","FPX","WTFA","PSU_P","STRAT_P",
            "REGION","SEX","ORIGIN_I","HISCODI3","AGE_P",
            "HINOTMYR","PHSTAT")
personsam <- person[myvars3]

# merge adult data with persondata
merged1 <- merge(adultsam,personsam,by=c("HHX","FMX","FPX"))
# merge with family data
merged3 <- merge(merged1,famsam,by=c("HHX","FMX"))

# merge child data with person data
merged2 <- merge(childsam,personsam,by=c("HHX","FMX","FPX"))
# merge with family data
merged4 <- merge(merged2,famsam,by=c("HHX","FMX"))

# stack child and adult df into 1 file
alldf=rbind(merged3, merged4)

################ Step 2 data clean up ####################
```

```
dfclean=alldf
# clean up  HCDLYR5 (no transportation): 7,8,9 meaningless
table(dfclean$HCDLYR5)
dfclean[, 5][dfclean[, 5] == 7] <- NA
dfclean[, 5][dfclean[, 5] == 8] <- NA
dfclean[, 5][dfclean[, 5] == 9] <- NA
table(dfclean$HCDLYR5)

# income:poverty ratio: 96, 99 meaningless
table(dfclean$RAT_CAT5)
dfclean[, 24][dfclean[, 24] == 96] <- NA
dfclean[, 24][dfclean[, 24] == 99] <- NA
table(dfclean$RAT_CAT5)

# health insurance: HINOTMYR 97, 98,99 meaningless
table(dfclean$HINOTMYR)
dfclean$HINOTMYR[is.na(dfclean$HINOTMYR)] <- 0
dfclean[, 14][dfclean[, 14] == 99] <- NA
table(dfclean$HINOTMYR)

# race (HISCODI3): ok: 1:Hispanic, 2: Non-Hisp White,
#3: Non-Hisp Black, 4: Non-Hisp Asian, 5: Non-Hisp all other

# education: FM_EDUC1: 97,98,99 not meaningful
table(dfclean$FM_EDUC1)
dfclean[, 19][dfclean[, 19] == 97] <- NA
dfclean[, 19][dfclean[, 19] == 99] <- NA
table(dfclean$FM_EDUC1)

# health status (PHSTAT): 7,8,9 not meaningful
table(dfclean$PHSTAT)
dfclean[, 15][dfclean[, 15] == 7] <- NA
dfclean[, 15][dfclean[, 15] == 9] <- NA
table(dfclean$PHSTAT)

# usual place when sick (USUALPL): 7,8,9 not meaningful
table(dfclean$USUALPL)
dfclean[, 4][dfclean[, 4] == 7] <- NA
dfclean[, 4][dfclean[, 4] == 8] <- NA
dfclean[, 4][dfclean[, 4] == 9] <- NA
table(dfclean$USUALPL)

# Language spoken at Interview: 8 not meaningful;
#1: English, 2: Spanish, 3: English and Spanish, 4: Other
table(dfclean$FLNGINTV)
dfclean[, 25][dfclean[, 25] == 8] <- NA
table(dfclean$FLNGINTV)
summary(dfclean)

# Step 3: Logistic regression
dfreg=dfclean

# Recode  HCDLYR5 to 1 vs. 0 (2 means 'no' at the moment)
table(dfreg$HCDLYR5)
```

```
dfreg[,5][dfreg[,5] == 2] <- 0
table(dfreg$HCDLYR5)

table(dfreg$SEX)
dfreg[,10][dfreg[,10] == 2] <- 0
table(dfreg$SEX) # 1= male, 0=female

# Create factor variables: education, income/povert ratio,
# race, health status, language spoken at interview
# keep continuous: months without insurance, age, sex

dfreg$edufac=as.factor(dfreg$FM_EDUC1)
dfreg$incpovfac=as.factor(dfreg$RAT_CAT5)
dfreg$racefac=as.factor(dfreg$HISCODI3)
dfreg$hlthfac=as.factor(dfreg$PHSTAT)
dfreg$langfac=as.factor(dfreg$FLNGINTV)
dfreg$usualplfac=as.factor(dfreg$USUALPL)

# set reference levels
dfreg <- within(dfreg, racefac <- relevel(racefac, ref = 2))

# set up splines and months without insurance
summary(dfreg$AGE_P) # min, Q1, median, Q3, and max
# 0, 17, 38, 59, 85

for (i in 1:nrow(dfreg)){
  dfreg$cub1[i]=max(0,((dfreg$AGE_P[i]-0)**3))
  dfreg$cub2[i]=max(0,((dfreg$AGE_P[i]-15)**3))
  dfreg$cub3[i]=max(0,((dfreg$AGE_P[i]-35)**3))
  dfreg$cub4[i]=max(0,((dfreg$AGE_P[i]-65)**3))
  dfreg$cub5[i]=max(0,((dfreg$AGE_P[i]-85)**3))

  dfreg$d1[i]=(dfreg$cub1[i]-dfreg$cub5[i])/(85-0)
  dfreg$d2[i]=(dfreg$cub2[i]-dfreg$cub5[i])/(85-15)
  dfreg$d3[i]=(dfreg$cub3[i]-dfreg$cub5[i])/(85-35)
  dfreg$d4[i]=(dfreg$cub4[i]-dfreg$cub5[i])/(85-65)

  dfreg$N1[i]=dfreg$d1[i]-dfreg$d4[i]
  dfreg$N2[i]=dfreg$d2[i]-dfreg$d4[i]
  dfreg$N3[i]=dfreg$d3[i]-dfreg$d4[i]
}

# reduce poverty levels: <1, 1-1.99, 2-3.99, >4
dfreg$povagg=dfreg$RAT_CAT5
table(dfreg$povagg)
dfreg[, 44][dfreg[, 44] == 2] <- 1
dfreg[, 44][dfreg[, 44] == 3] <- 1
dfreg[, 44][dfreg[, 44] == 4] <- 2
dfreg[, 44][dfreg[, 44] == 5] <- 2
dfreg[, 44][dfreg[, 44] == 6] <- 2
dfreg[, 44][dfreg[, 44] == 7] <- 2
dfreg[, 44][dfreg[, 44] == 8] <- 3
dfreg[, 44][dfreg[, 44] == 9] <- 3
dfreg[, 44][dfreg[, 44] == 10] <- 3
```

```
dfreg[, 44][dfreg[, 44] == 11] <- 3
dfreg[, 44][dfreg[, 44] == 12] <- 4
dfreg[, 44][dfreg[, 44] == 13] <- 4
dfreg[, 44][dfreg[, 44] == 14] <- 4
dfreg[, 44][dfreg[, 44] == 15] <- 1
dfreg[, 44][dfreg[, 44] == 16] <- 2
dfreg[, 44][dfreg[, 44] == 17] <- 3
dfreg[, 44][dfreg[, 44] == 18] <- 4
table(dfreg$povagg)

dfreg$povaggfac=as.factor(dfreg$povagg)

# simplify race: aggregate White and Asian)
dfreg$race2=dfreg$HISCODI3
table(dfreg$race2)
dfreg[, 46][dfreg[, 46] == 4] <- 2
table(dfreg$race2)
dfreg$race2fac=as.factor(dfreg$race2)
dfreg <- within(dfreg, race2fac <- relevel(race2fac, ref = 2))

# Simplify education: no highschool,
# high school, some college, AA degree, at least college
dfreg$edu2=dfreg$FM_EDUC1
table(dfreg$edu2)
dfreg[, 48][dfreg[, 48] == 2] <- 1
dfreg[, 48][dfreg[, 48] == 3] <- 1
dfreg[, 48][dfreg[, 48] == 4] <- 2
dfreg[, 48][dfreg[, 48] == 5] <- 3
dfreg[, 48][dfreg[, 48] == 6] <- 4
dfreg[, 48][dfreg[, 48] == 7] <- 4
dfreg[, 48][dfreg[, 48] == 8] <- 5
dfreg[, 48][dfreg[, 48] == 9] <- 5
table(dfreg$edu2)

dfreg$edu2fac=as.factor(dfreg$edu2)
dfreg <- within(dfreg, edu2fac <- relevel(edu2fac, ref = 5))

# logistic regression, not using survey weights
# Model 1
model1=glm(HCDLYR5 ~ edufac + incpovfac + racefac + hlthfac
                     + langfac + usualplfac + SEX + AGE_P + HINOTMYR,
                     family = binomial, data=dfreg)
summary(model1)
coeffs=coef(model1)
SEs=coef(summary(model1))[, 2]
coefexp=cbind(coeffs,SEs)

# Model 2: With age splines in model
model2=glm(HCDLYR5 ~ edufac + incpovfac + racefac + hlthfac
           + langfac + usualplfac + SEX + AGE_P + N1 + N2 + N3 + HINOTMYR,
           family = binomial, data=dfreg)

# Age nonlinear?
length(coef(model2))
```

```
coef(model2)[42:44]
zerovec=rep(0,41)

# nonlinear significance of age:
K1 <-c(zerovec, 1, 0, 0, 0)
K2 <-c(zerovec, 0, 1, 0, 0)
K3 <-c(zerovec, 0, 0, 1, 0)
Km=rbind(K1, K2, K3)
Km
dim(Km)
G = glht(model2, linfct = Km)
summary(G,test = Ftest()) # P=0.327

# Model 3: Test for interaction between race and poverty
model3=glm(HCDLYR5 ~ edufac + povaggfac + race2fac + hlthfac
           + langfac + usualplfac + SEX + AGE_P + HINOTMYR
           + race2fac*povaggfac,
           family = binomial, data=dfreg)

# interaction significant?
length(coef(model3))
coef(model3)[28:36]
zerovec2=rep(0,27)

# nonlinear significance of age:
K1 <-c(zerovec2, 1,0,0,0,0,0,0,0,0)
K2 <-c(zerovec2, 0,1,0,0,0,0,0,0,0)
K3 <-c(zerovec2, 0,0,1,0,0,0,0,0,0)
K4 <-c(zerovec2, 0,0,0,1,0,0,0,0,0)
K5 <-c(zerovec2, 0,0,0,0,1,0,0,0,0)
K6 <-c(zerovec2, 0,0,0,0,0,1,0,0,0)
K7 <-c(zerovec2, 0,0,0,0,0,0,1,0,0)
K8 <-c(zerovec2, 0,0,0,0,0,0,0,1,0)
K9 <-c(zerovec2, 0,0,0,0,0,0,0,0,1)

Km=rbind(K1,K2,K3,K4,K5,K6,K7,K8,K9)
Km
dim(Km)
G = glht(model3, linfct = Km)
summary(G,test = Ftest()) # P=0.24

########  Model 4:  Final model #########
dfreg <- within(dfreg, povaggfac <- relevel(povaggfac, ref = 4))
dfreg <- within(dfreg, edu2fac <- relevel(edu2fac, ref = 5))
dfreg <- within(dfreg, racefac <- relevel(racefac, ref = 2))


# Impute missing variables
imptd <- aregImpute(~ AGE_P + SEX + racefac + hlthfac + langfac
                    + HINOTMYR + edu2fac + usualplfac + povaggfac,
                    data=dfreg, n.impute=5, nk=3)

compimp1=compimp2=compimp3=compimp4=compimp5=dfreg
```

```
imp1 <- impute.transcan(imptd, imputation=1, data=dfreg,
                        list.out=TRUE,pr=FALSE, check=FALSE)
compimp1[names(imp1)] <- imp1
imp2 <- impute.transcan(imptd, imputation=2, data=dfreg,
                        list.out=TRUE,pr=FALSE, check=FALSE)
compimp2[names(imp2)] <- imp2
imp3 <- impute.transcan(imptd, imputation=3, data=dfreg,
                        list.out=TRUE,pr=FALSE, check=FALSE)
compimp3[names(imp3)] <- imp3
imp4 <- impute.transcan(imptd, imputation=4, data=dfreg,
                        list.out=TRUE,pr=FALSE, check=FALSE)
compimp4[names(imp4)] <- imp4
imp5 <- impute.transcan(imptd, imputation=5, data=dfreg,
                        list.out=TRUE,pr=FALSE, check=FALSE)
compimp5[names(imp5)] <- imp5


# Create 5 survey design objects
imp1 <- svydesign(id=~PSU_P, strata=~STRAT_P,nest = TRUE,weights=~WTFA,data=compimp1)
imp2 <- svydesign(id=~PSU_P, strata=~STRAT_P,nest = TRUE,weights=~WTFA,data=compimp2)
imp3 <- svydesign(id=~PSU_P, strata=~STRAT_P,nest = TRUE,weights=~WTFA,data=compimp3)
imp4 <- svydesign(id=~PSU_P, strata=~STRAT_P,nest = TRUE,weights=~WTFA,data=compimp4)
imp5 <- svydesign(id=~PSU_P, strata=~STRAT_P,nest = TRUE,weights=~WTFA,data=compimp5)

# Run final model fo each imputed dataset
# accounting for survey weights

mimp1<-svyglm(HCDLYR5 ~ edu2fac + povaggfac + racefac + hlthfac
              + langfac + usualplfac + SEX + AGE_P + HINOTMYR,
              family = quasibinomial, design=imp1)

mimp2<-svyglm(HCDLYR5 ~ edu2fac + povaggfac + racefac + hlthfac
              + langfac + usualplfac + SEX + AGE_P + HINOTMYR,
              family = quasibinomial, design=imp2)

mimp3<-svyglm(HCDLYR5 ~ edu2fac + povaggfac + racefac + hlthfac
              + langfac + usualplfac + SEX + AGE_P + HINOTMYR,
              family = quasibinomial, design=imp3)

mimp4<-svyglm(HCDLYR5 ~ edu2fac + povaggfac + racefac + hlthfac
              + langfac + usualplfac + SEX + AGE_P + HINOTMYR,
              family = quasibinomial, design=imp4)

mimp5<-svyglm(HCDLYR5 ~ edu2fac + povaggfac + racefac + hlthfac
              + langfac + usualplfac + SEX + AGE_P + HINOTMYR,
              family = quasibinomial, design=imp5)

# Combine results across data sets
library(mitools)
models=list(mimp1,mimp2,mimp3,mimp4,mimp5)
betas<-MIextract(models,fun=coef)
vars<-MIextract(models, fun=vcov)
combined=MIcombine(betas,vars)
results=summary(combined)
```

```
write.csv(results,"combinedv2.csv")

# Get combined covariance matrix
covmatrix=vcov(combined)
# Get betas
coeffs=coefficients(combined)
parnum=length(coeffs)
coeffvec=as.matrix(coeffs)

######## specific comparisons ####################
# black vs. white
coeffs[9:12]
one=9
two=10

betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=-1
betavec[two]=1
betavec=t(betavec)
(b_vs_w=betavec %*% coeffvec) # lnOR =0.340165
(SEb_vs_w=sqrt(betavec %*% covmatrix %*% t(betavec)))
# SE ln OR : 0.121698

# Non Hisp Other vs. white
coeffs[9:12]
one=9
two=12

betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=-1
betavec[two]=1
betavec=t(betavec)
(b_vs_w=betavec %*% coeffvec) # lnOR =0.7902993
(SEb_vs_w=sqrt(betavec %*% covmatrix %*% t(betavec)))
# SE ln OR : 0.2534282

# Hispanoc vs. white
coeffs[9:12]
one=9

betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=-1
betavec=t(betavec)
(b_vs_w=betavec %*% coeffvec) # lnOR =-0.05255297
(SEb_vs_w=sqrt(betavec %*% covmatrix %*% t(betavec)))
# SE ln OR: 0.1370342

# Asian vs. white
coeffs[9:12]
one=9
two=11
```

```
betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=-1
betavec[two]=1
betavec=t(betavec)
(b_vs_w=betavec %*% coeffvec) # lnOR other vs w=-0.1038106
(SEb_vs_w=sqrt(betavec %*% covmatrix %*% t(betavec))) # SE ln OR : 0.2079997

# Education levels vs at least college
# no highschool vs college
coeffs[2:5]
one=2
two=3
betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=-1
betavec[two]=1
betavec=t(betavec)
(b_vs_w=betavec %*% coeffvec) # lnOR =0.6529995
(SEb_vs_w=sqrt(betavec %*% covmatrix %*% t(betavec)))
# SE ln OR : 0.1601217

# high school vs college
one=2
two=4
betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=-1
betavec[two]=1
betavec=t(betavec)
(b_vs_w=betavec %*% coeffvec) # lnOR =0.4613184
(SEb_vs_w=sqrt(betavec %*% covmatrix %*% t(betavec)))
# SE ln OR : 0.1526725

# some colleg vs college
one=2
two=5
betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=-1
betavec[two]=1
betavec=t(betavec)
(b_vs_w=betavec %*% coeffvec) # lnOR =0.6164677
(SEb_vs_w=sqrt(betavec %*% covmatrix %*% t(betavec)))
# SE ln OR : 0.1566269

# aa vs college
one=2
betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=-1
betavec=t(betavec)
(b_vs_w=betavec %*% coeffvec) # lnOR =0.4178515
```

```
(SEb_vs_w=sqrt(betavec %*% covmatrix %*% t(betavec)))
# SE ln OR : 0.1556368

## Descriptive: How many people delayed Care?
overall=svymean(~HCDLYR5,na.rm=TRUE,design=imp1) # 0.01982
confint(overall,level = 0.95)
#             2.5 %     97.5 %
#  HCDLYR5 0.01816772 0.02147262

# children vs. adults
subimp1=compimp1[which(compimp1$AGE_P<18),]
subimp2=compimp1[which(compimp1$AGE_P>=18),]
subchild <- svydesign(id=~PSU_P, strata=~STRAT_P,
                      nest = TRUE,weights=~WTFA,data=subimp1)
subadult <- svydesign(id=~PSU_P, strata=~STRAT_P,
                      nest = TRUE,weights=~WTFA,data=subimp2)

overallchild=svymean(~HCDLYR5,na.rm=TRUE,design=subchild) # 0.016312
confint(overallchild,level = 0.95)
#               2.5 %     97.5 %
#  HCDLYR5 0.01363624 0.01898876

(overallad=svymean(~HCDLYR5,na.rm=TRUE,design=subadult)) # 0.021033
confint(overallad,level = 0.95)
#               2.5 %    97.5 %
#  HCDLYR5 0.01914775 0.0229189
```

F.0.2 SAS Code used for analysis in chapter 3 (car ownership and preventive and mammography care)

```
option ls=76 ps=500 mergenoby=ERROR ;

/* read in original dataset */
%include 'T:\Core\Core_main\Source File Development\Source File Create\Formats\SHOW_MACROS.sas';
libname pjewett 'T:\CORE\CORE_Main\Analytical\Data requests\2015\206_Jewett_Car_Ownership\data' ;

data original;
set pjewett.dr206_20161108;
run;

/*make HHID and SPID numeric*/
data original_1;
set original;
HHID_num = HHID*1;
SPID_num = SPID*1;
run;

/* add contextual SES data by CTRACT2000 and county to dataset: */
/********** SES indicators per 2000 census tract **************/
proc import datafile="T:\CORE\CORE_Main\Analytical\Data requests\2015\
206_Jewett_Car_Ownership\data\Non-SAS data\census10_SES.csv"
```

```
     out=census10_SES
     dbms=csv
     replace;
     getnames=yes;
run;

proc sort data=census10_SES;
by CTRACT10;
run;

/*********** % of households w/o a vehicle in 2000 census tract *****************/
proc import datafile="T:\CORE\CORE_Main\Analytical\Data requests\2015\
206_Jewett_Car_Ownership\data\Non-SAS data\noveh_ctract10.csv"
     out=NOVEH
     dbms=csv
     replace;
     getnames=yes;
run;

proc sort data=NOVEH;
by CTRACT10;
run;

/***** uninsurance rate per county --> don't do this, instead include county indicator *********
proc import datafile="T:\CORE\CORE_Main\Analytical\Data requests\2015\
206_Jewett_Car_Ownership\data\Non-SAS data\SAHIE05_uninsured_per_county.csv"
     out=CNTY_UNINSURED
     dbms=csv
     replace;
     getnames=yes;
run;

proc sort data=CNTY_UNINSURED;
by COUNTYFIPS;
run;

/* Merge contextual SES data into original data*/
/* convert CENSUS_TRACT and COUNTY_FIPS in original data to numerical (as they are in csv files)*/
data original2;
set original_1;
CTRACT00 = CENSUS_TRACT*1;
CTRACT10=CENSUS_TRACT_2010*1;
COUNTYFIPS = COUNTY_FIPS*1;
run;

proc sort data=original2;
by CTRACT10;
run;

data merge1;
MERGE original2 (in=a) census10_SES  NOVEH;
by CTRACT10;
if a=1;
run;
```

```
/* Data clean-up*/
data tmp2;
set merge1;
if ANT070=".D" or ANT070=".R" then ANT070=.;
if ANT060_1=".D" or ANT060_1=".R" OR ANT060_1<100 then ANT060_1=.;
if ANT060_2=".D" or ANT060_2=".R" OR ANT060_2<100 then ANT060_2=.;
if ANT060_3=".D" or ANT060_3=".R" OR ANT060_3<100 then ANT060_3=.;
/*implausible BMI because of implausible height*/
if ANT_MEAS_HEIGHT_CM<100 then do; ANT_MEAS_HEIGHT_CM=.; ANT_BMI=.; end;
if CURRENT_INSURANCE=".D" then CURRENT_INSURANCE=.;
if DIQ100=".D" then DIQ100=.;
if DMQ080_1=2 then DMQ080_1=0;
if DMQ080_8=2 then DMQ080_8=0;
if DMQ080_9=2 then DMQ080_9=0;
if DMQ100=2 then DMQ100=0;
if INCOME_HH_MID=".D" OR INCOME_HH_MID=".R" then INCOME_HH_MID=.;
if IUQ010=".D" then IUQ010=.;
if IUQ015=".D" then IUQ015=.;
if IUQ050=".D" then IUQ050=.;
if IUQ120=".D" then IUQ120=.;
if IUQ130=".D" then IUQ130=.;
if IUQ220_N=".D" then IUQ220_N=.;
if IUQ260_R2=".D" then IUQ260_R2=.;
if IUQ275=".D" then IUQ275=.;
if HHQ480=".D" then HHQ480=.;
if OCQ100=".D" OR OCQ100=".R" then OCQ100=.;
if IUQ020_EMPL=2 then IUQ020_EMPL=0;
if IUQ020_HIRSP_R2=2 then IUQ020_HIRSP_R2=0;
if IUQ020_IHS_R2=2 then IUQ020_IHS_R2=0;
if IUQ020_INDIV=2 then IUQ020_INDIV=0;
if IUQ020_MEDICAID_R2=2 then IUQ020_MEDICAID_R2=0;
if IUQ020_MEDICARE=2 then IUQ020_MEDICARE=0;
if IUQ020_MIL_R2=2 then IUQ020_MIL_R2=0;
if IUQ020_OTHERIND=2 then IUQ020_OTHERIND=0;
run;


/*integrate NBRHD006 and NBRHD006_R2 variables from different surveys*/
data tmp3;
set tmp2;
NBRHD006_B_tmp=NBRHD006_B_R2;
NBRHD006_D_tmp=NBRHD006_D_R2;
NBRHD006_G_tmp=NBRHD006_G_R2;
NBRHD006_H_tmp=NBRHD006_H_R2;
NBRHD006_I_tmp=NBRHD006_I_R2;
NBRHD006_L_tmp=NBRHD006_L_R2;
NBRHD006_N_tmp=NBRHD006_N_R2;
NBRHD006_O_tmp=NBRHD006_O_R2;

if NBRHD006_B_R2=3 then NBRHD006_B_tmp=4;
else if NBRHD006_B_R2=4 then NBRHD006_B_tmp=5;

if NBRHD006_D_R2=3 then NBRHD006_D_tmp=4;
```

```
else if NBRHD006_D_R2=4 then NBRHD006_D_tmp=5;

if NBRHD006_G_R2=3 then NBRHD006_G_tmp=4;
else if NBRHD006_G_R2=4 then NBRHD006_G_tmp=5;

if NBRHD006_H_R2=3 then NBRHD006_H_tmp=4;
else if NBRHD006_H_R2=4 then NBRHD006_H_tmp=5;

if NBRHD006_I_R2=3 then NBRHD006_I_tmp=4;
else if NBRHD006_I_R2=4 then NBRHD006_I_tmp=5;

if NBRHD006_L_R2=3 then NBRHD006_L_tmp=4;
else if NBRHD006_L_R2=4 then NBRHD006_L_tmp=5;

if NBRHD006_N_R2=3 then NBRHD006_N_tmp=4;
else if NBRHD006_N_R2=4 then NBRHD006_N_tmp=5;

if NBRHD006_O_R2=3 then NBRHD006_O_tmp=4;
else if NBRHD006_O_R2=4 then NBRHD006_O_tmp=5;

run;

data tmp4;
set tmp3;
NBRHD006_B_fin=.;
NBRHD006_D_fin=.;
NBRHD006_G_fin=.;
NBRHD006_H_fin=.;
NBRHD006_I_fin=.;
NBRHD006_L_fin=.;
NBRHD006_N_fin=.;
NBRHD006_O_fin=.;

if NBRHD006_B=. then NBRHD006_B_fin=NBRHD006_B_tmp;
else if NBRHD006_B_tmp=. then NBRHD006_B_fin=NBRHD006_B;

if NBRHD006_D=. then NBRHD006_D_fin=NBRHD006_D_tmp;
else if NBRHD006_D_tmp=. then NBRHD006_D_fin=NBRHD006_D;

if NBRHD006_G=. then NBRHD006_G_fin=NBRHD006_G_tmp;
else if NBRHD006_G_tmp=. then NBRHD006_G_fin=NBRHD006_G;

if NBRHD006_H=. then NBRHD006_H_fin=NBRHD006_H_tmp;
else if NBRHD006_H_tmp=. then NBRHD006_H_fin=NBRHD006_H;

if NBRHD006_I=. then NBRHD006_I_fin=NBRHD006_I_tmp;
else if NBRHD006_I_tmp=. then NBRHD006_I_fin=NBRHD006_I;

if NBRHD006_L=. then NBRHD006_L_fin=NBRHD006_L_tmp;
else if NBRHD006_L_tmp=. then NBRHD006_L_fin=NBRHD006_L;

if NBRHD006_N=. then NBRHD006_N_fin=NBRHD006_N_tmp;
else if NBRHD006_N_tmp=. then NBRHD006_N_fin=NBRHD006_N;
```

```
if NBRHD006_O=. then NBRHD006_O_fin=NBRHD006_O_tmp;
else if NBRHD006_O_tmp=. then NBRHD006_O_fin=NBRHD006_O;

run;

/* imputation of missing data: */
/*Take log of POVERTY_IR DRINKS PREDFVL PREDSUG and impute those
so no negative values are imputed*/
data tmp5;
set tmp4;
lnpovir=log(POVERTY_IR);
run;

/* imputation step 0*/
proc mi nimpute=0 data=tmp5;
/*variables without missing values:*/
var age_consent BELOW_POVLN gender AT_LEAST_COLL MED_INC_13
DMQ080_1 DMQ080_8 DMQ080_9
No_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000
/*variables with missing values:*/
/*DMQ010 DMQ040 IUQ270 CHRONIC_COND IUQ120 CURRENT_INSURANCE
DIQ100 RACE_ETHNICITY_4CAT SITTING_TIME_GPAQ INSURANCE_TYPE_R2
IUQ010 OCQ100 HOQ065_R2 INQ201
ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI INCOME_HH_MID lnpovir IUQ050
HOQ060_R3 SF12010 NBRHD001_H NBRHD001_M NBRHD003 NBRHD004
NBRHD001_E NBRHD001_F NBRHD001_N NBRHD002 NBRHD001_P NBRHD001_I
NBRHD001_O NBRHD005_C
SMQ_DER_FORMER_NEVER_CURRENT_R2 NBRHD005_A PSH050 NBRHD001_J
NBRHD006_L_FIN NBRHD006_O_FIN COMMUNITY160 NBRHD006_B_FIN NBRHD006_D_FIN
COMMUNITY100 NBRHD006_H_FIN NBRHD006_N_FIN IUQ130 COMMUNITY110
COMMUNITY130 COMMUNITY150 NBRHD006_G_FIN NBRHD006_I_FIN COMMUNITY090
COMMUNITY140 PREDFVL*/

DMQ010 DMQ040 IUQ270 CHRONIC_COND IUQ120 CURRENT_INSURANCE DIQ100
RACE_ETHNICITY_4CAT SITTING_TIME_GPAQ INSURANCE_TYPE_R2 IUQ010
OCQ100 HOQ065_R2 INQ201
ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI INCOME_HH_MID lnpovir IUQ050
HOQ060_R3 SF12010 NBRHD001_H NBRHD001_M NBRHD003 NBRHD004
NBRHD001_E NBRHD001_F NBRHD001_N NBRHD002 NBRHD001_P NBRHD001_I
NBRHD001_O NBRHD005_C
SMQ_DER_FORMER_NEVER_CURRENT_R2 NBRHD005_A PSH050 NBRHD001_J
NBRHD006_L_FIN NBRHD006_O_FIN COMMUNITY160 NBRHD006_B_FIN NBRHD006_D_FIN IUQ130
NBRHD006_H_fin NBRHD006_N_fin;
run;


/* imputation step 1*/
proc mi nimpute=5 data=tmp5 out=Pattern seed=100
round = 1 0.1 1 0.1 1 1 1 1 0.1 0.1 0.1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0.1 0.1 1
0.01 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1
min = 18 0 1 10.3 11003 0 0 0 0 0 0 1 5 1
1 1 1 1 1 1 10 0 0 1 1 1 110 44.2 16.6 7500 -1.97 1
```

```
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1
max = 98 63.6 2 87.8 157778 1 1 1 46.7 56.7 10.6 21 6 2
2 6 2 6 4 6900 2 12 4 4 11 199 186 57.7 225000 2.96 3
4 5 6 6 4 4 6 6 6 4 6 6 6 4 3 4 6
6 5 5 5 5 5 2 5 5
;
em maxiter=500;
mcmc impute = monotone;

var age_consent BELOW_POVLN gender AT_LEAST_COLL MED_INC_13
DMQ080_1 DMQ080_8 DMQ080_9
No_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000

DMQ010 DMQ040 IUQ270 CHRONIC_COND IUQ120 CURRENT_INSURANCE DIQ100
RACE_ETHNICITY_4CAT SITTING_TIME_GPAQ INSURANCE_TYPE_R2 IUQ010 OCQ100
HOQ065_R2 INQ201
ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI INCOME_HH_MID lnpovir IUQ050
HOQ060_R3 SF12010 NBRHD001_H NBRHD001_M NBRHD003 NBRHD004
NBRHD001_E NBRHD001_F NBRHD001_N NBRHD002 NBRHD001_P NBRHD001_I
NBRHD001_O NBRHD005_C
SMQ_DER_FORMER_NEVER_CURRENT_R2 NBRHD005_A PSH050 NBRHD001_J
NBRHD006_L_FIN NBRHD006_O_FIN COMMUNITY160 NBRHD006_B_FIN
NBRHD006_D_FIN IUQ130
NBRHD006_H_fin NBRHD006_N_fin;
run;

proc mi nimpute=1 data=Pattern seed=10 out=Imputed;
class DMQ010 CURRENT_INSURANCE gender DMQ080_1
DMQ080_8 DMQ080_9
  RUCA_CODE_2000 DMQ040 IUQ270 CHRONIC_COND IUQ120
  DIQ100 INSURANCE_TYPE_R2 OCQ100 HOQ065_R2 RACE_ETHNICITY_4CAT
  INCOME_HH_MID IUQ050
  HOQ060_R3 SF12010 NBRHD001_E NBRHD001_H NBRHD001_M
  NBRHD003 NBRHD001_F NBRHD001_N NBRHD004 NBRHD001_P NBRHD002
  NBRHD001_I NBRHD001_O NBRHD005_C SMQ_DER_FORMER_NEVER_CURRENT_R2
  NBRHD005_A PSH050 NBRHD001_J
  NBRHD006_L_FIN NBRHD006_O_FIN COMMUNITY160 NBRHD006_B_FIN
  NBRHD006_D_FIN IUQ130 NBRHD006_H_FIN NBRHD006_N_FIN ;

var age_consent BELOW_POVLN gender AT_LEAST_COLL MED_INC_13 DMQ080_1 DMQ080_8 DMQ080_9
No_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000

DMQ010 DMQ040 IUQ270 CHRONIC_COND IUQ120 CURRENT_INSURANCE DIQ100
RACE_ETHNICITY_4CAT SITTING_TIME_GPAQ INSURANCE_TYPE_R2 IUQ010 OCQ100
HOQ065_R2 INQ201
ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI INCOME_HH_MID lnpovir IUQ050
HOQ060_R3 SF12010 NBRHD001_H NBRHD001_M NBRHD003 NBRHD004
NBRHD001_E NBRHD001_F NBRHD001_N NBRHD002 NBRHD001_P NBRHD001_I
NBRHD001_O NBRHD005_C
SMQ_DER_FORMER_NEVER_CURRENT_R2 NBRHD005_A PSH050 NBRHD001_J
NBRHD006_L_FIN NBRHD006_O_FIN COMMUNITY160 NBRHD006_B_FIN NBRHD006_D_FIN IUQ130
NBRHD006_H_fin NBRHD006_N_fin;
```

```
by _imputation_;

monotone logistic (DMQ010=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000);
monotone logistic (DMQ040=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000);
monotone discrim (IUQ270=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL);
monotone discrim (CHRONIC_COND=age_consent BELOW_POVLN AT_LEAST_COLL
MED_INC_13 NO_HIGHSCHOOL PERCNOVEH_OVERALL);
monotone logistic (IUQ120=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000);
monotone logistic (CURRENT_INSURANCE=age_consent BELOW_POVLN AT_LEAST_COLL
MED_INC_13 NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000);
monotone logistic (DIQ100=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000);
monotone logistic (RACE_ETHNICITY_4CAT=age_consent BELOW_POVLN AT_LEAST_COLL
MED_INC_13 NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000);
monotone regression (SITTING_TIME_GPAQ=age_consent BELOW_POVLN AT_LEAST_COLL
MED_INC_13 NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 DIQ100 RACE_ETHNICITY_4CAT);
monotone logistic (INSURANCE_TYPE_R2=age_consent BELOW_POVLN AT_LEAST_COLL
MED_INC_13 NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000);
monotone regression (IUQ010= age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 DMQ010 CURRENT_INSURANCE DIQ100 RACE_ETHNICITY_4CAT SITT
monotone logistic (OCQ100=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 IUQ010);
monotone logistic (HOQ065_R2=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010);
monotone regression (INQ201=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 DMQ010 CURRENT_INSURANCE
DIQ100 RACE_ETHNICITY_4CAT SITTING_TIME_GPAQ );
monotone regression (ANT_MEAS_HEIGHT_CM=age_consent BELOW_POVLN AT_LEAST_COLL
MED_INC_13 NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 DIQ100 HOQ065_R2
RACE_ETHNICITY_4CAT INQ201 IUQ010 SITTING_TIME_GPAQ IUQ010 INQ201);
monotone regression (ANT070=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 DIQ100 HOQ065_R2
RACE_ETHNICITY_4CAT INQ201 IUQ010 SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM);
monotone regression (WHQ_SELF_BMI=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 DIQ100 HOQ065_R2 RACE_ETHNICITY_4CAT
INQ201 IUQ010 SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM ANT070);
monotone logistic (INCOME_HH_MID=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010 INQ201
ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI);
monotone regression (lnpovir=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 DIQ100 HOQ065_R2 RACE_ETHNICITY_4CAT
INQ201 IUQ010 SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI);
monotone logistic (IUQ050=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13 NO_HIGHSCHOOL
PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM
ANT070 WHQ_SELF_BMI);
monotone logistic (HOQ060_R3=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13 NO_HIGHSCHOOL
PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM
ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (SF12010=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13 NO_HIGHSCHOOL
 PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM ANT070
```

```
 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD001_H=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD001_M=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD003=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD004=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD001_E=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD001_F=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD001_N=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD002=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD001_P=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD001_I=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD001_O=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD005_C=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (SMQ_DER_FORMER_NEVER_CURRENT_R2=age_consent BELOW_POVLN
AT_LEAST_COLL MED_INC_13 NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000
SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD005_A=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (PSH050=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD001_J=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD006_L_FIN=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010 INQ201
ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD006_O_FIN=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010 INQ201
```

```
ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (COMMUNITY160=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010 INQ201
ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD006_B_FIN=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010 INQ201
ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD006_D_FIN=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010 INQ201
ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
/*monotone logistic (COMMUNITY100=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
MET_MIN_WEEK_GPAQ NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000
SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);*/
monotone logistic (IUQ130=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010
INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD006_H_FIN=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010 INQ201
ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD006_N_FIN=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010 INQ201
ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
/*monotone logistic (NBRHD006_G_FIN=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
MET_MIN_WEEK_GPAQ NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000
SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (NBRHD006_I_FIN=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
MET_MIN_WEEK_GPAQ NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000
SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (COMMUNITY110=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
 MET_MIN_WEEK_GPAQ NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000 SITTING_TIME_GPAQ IUQ010 INQ201 ANT_M
monotone logistic (COMMUNITY130=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
MET_MIN_WEEK_GPAQ NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000
SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (COMMUNITY150=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
MET_MIN_WEEK_GPAQ NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000
SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);
monotone logistic (COMMUNITY090=age_consent BELOW_POVLN AT_LEAST_COLL MED_INC_13
MET_MIN_WEEK_GPAQ NO_HIGHSCHOOL PERCNOVEH_OVERALL RUCA_CODE_2000
SITTING_TIME_GPAQ IUQ010 INQ201 ANT_MEAS_HEIGHT_CM ANT070 WHQ_SELF_BMI lnpovir);*/
run;

/*back-transformation of the logs*/
data Imputed_2;
set Imputed;
POVERTY_IR2=exp(lnpovir);
run;

*data transformations;
data tmp6;
set Imputed_2;
/*education years*/
eduyrs=.;
educat=.;
```

```
if DMQ010=5 then eduyrs=6;
else if DMQ010=6 then eduyrs=6;
else if DMQ010=7 then eduyrs=7;
else if DMQ010=8 then eduyrs=8;
else if DMQ010=9 then eduyrs=9;
else if DMQ010=10 then eduyrs=10;
else if DMQ010=11 then eduyrs=11;
else if DMQ010=12 then eduyrs=11.5;
else if DMQ010=13 then eduyrs=12;
else if DMQ010=14 then eduyrs=13;
else if DMQ010=15 then eduyrs=14;
else if DMQ010=16 then eduyrs=15;
else if DMQ010=17 then eduyrs=15;
else if DMQ010=18 then eduyrs=16;
else if DMQ010=19 then eduyrs=20;
else if DMQ010=20 then eduyrs=20;
else if DMQ010=21 then eduyrs=22;

/*education as more aggregated variable*/
if DMQ010<=12 then educat=1; /*no highschool degree*/
else if DMQ010>12 and DMQ010<=15 then educat=2; /*High School degree; no College degree*/
else if DMQ010>15 and DMQ010<=17 then educat=3; /*associate degree;*/
else if DMQ010>17 then educat=4;/*At least College degree*/

/*calculate missing BMI values*/
if ANT_BMI=. then ANT_BMI=ANT070/((ANT_MEAS_HEIGHT_CM/100)**2);
run;

/* make permanent copy of dataset for analysis*/
data pjewett.for_pscore;
set tmp6;
run;

/*Analysis code:*/
option ls=76 ps=500 mergenoby=ERROR ;

%include 'T:\Core\Core_main\Source File Development\Source File Create\Formats\SHOW_MACROS.sas';
libname pjewett 'T:\CORE\CORE_Main\Analytical\Data requests\2015\206_Jewett_Car_Ownership\data' ;

data checkup;
set pjewett.for_pscore;
run;

/*Scaling and recoding of variables*/
data checkup2;
set checkup;
own_car=.;
homeowner=.;

if DMQ100=1 then own_car=1;
else if DMQ100=0 then own_car=0;

inc_percapita = INCOME_HH_MID/INQ201; /* per capita income*/
inc_percapita_5K = inc_percapita/5000; /* per capita income per $5,000*/
```

```
* homewonwership Y/N;
if HOQ065_R2=1 or HOQ065_R2=2 then homeowner=1;
else if HOQ065_R2=3 or HOQ065_R2=4 then homeowner=0; /*homeowner or not*/

if gender=1 then male=1;
else if gender=2 then male=0;

MED_INC10K = MED_INC_13/10000; /* median income per $10,000*/

eduyrs4=eduyrs/4; /*education years per 4 additional years*/

perc_noveh5=percnoveh_overall/5; /*pop fraction w/o a car per 5%*/
run;

/*Code outcome variables IUQ220_N and IUQ220_U into one variable into one number of days variable*/
data checkup3;
set checkup2;
dayssince=.;
check=.;

if IUQ220_N=0 then check=.; /*Presumably never had a checkup, but that's obviously not true*/

if IUQ220_N=. then do; /*missing*/
dayssince=.;
check=.;
end;

else if IUQ220_U = 1 then do; /*Days*/
dayssince=IUQ220_N;
check=1;
end;

else if IUQ220_U = 2 then do; /*Weeks*/
dayssince=IUQ220_N*7;
check=1;
end;

else if IUQ220_U = 3 then do; /*Months*/
dayssince=IUQ220_N*30.4;
check=1;
end;

else if IUQ220_U = 4 then do; /*Years*/
dayssince=IUQ220_N*365.25;
check=1;
end;

/*mammography recoded*/
if SIQ180=2 then SIQ180=0;
run;

data checkup4;
set checkup3;
```

```
language=.; /*1: only english, 2: english and another 3: another*/
marital=.; /*1: married/with partner, 2: separated/divorced, 3: widowed, 4: never married*/
usualplace=.; /*1: clinic/community center/outpatient, 2: other, 3: none*/
usualdoc=.;
chronic=.;

if (DMQ080_1=0 and DMQ080_8=0 and DMQ080_9=0) then language=3;
else if (DMQ080_1=0 and (DMQ080_8=1 OR DMQ080_9=1)) then language=3;
else if (DMQ080_1=1 and (DMQ080_8=1 OR DMQ080_9=1)) then language=2;
else if (DMQ080_1=1 and DMQ080_8=0 and DMQ080_9=0) then language=1;

if DMQ040= 1 or DMQ040=6 then marital=1;
else if DMQ040= 3 or DMQ040=4 then marital=2;
else if DMQ040= 2 then marital=3;
else if DMQ040= 5 then marital=4;

if IUQ120=2 or IUQ120=3 or IUQ120=4 then usualplace=1;
else if IUQ120=1 or IUQ120=5 then usualplace=2;
else if IUQ120=6 then usualplace=3;

if IUQ130=1 then usualdoc=1;
else if IUQ130=2 then usualdoc=0;

if CHRONIC_COND=1 then chronic=1;
else if CHRONIC_COND=2 then chronic=0;
run;

/*set up splines for education years, age, and income/poverty ratio*/
proc univariate data=checkup4 noprint;
    var POVERTY_IR2;
histogram POVERTY_IR2;
    output out=knots pctlpts=0 25 50 75 100 pctlpre=pct;
run;
proc print data=knots;
run;

DATA checkup5_1;
SET checkup4;

/*education*/
    ecub1 = max(0,(eduyrs4-1.5)**3);
    ecub2 = max(0,(eduyrs4-3.25)**3);
    ecub3 = max(0,(eduyrs4-3.75)**3);
    ecub4 = max(0,(eduyrs4-4)**3);
    ecub5 = max(0,(eduyrs4-5.5)**3);

    ed1 = (ecub1-ecub5)/(5.5-1.5);
    ed2 = (ecub2-ecub5)/(5.5-3.25);
    ed3 = (ecub3-ecub5)/(5.5-3.75);
    ed4 = (ecub4-ecub5)/(5.5-4);

    eN1 = ed1-ed4;
    eN2 = ed2-ed4;
    eN3 = ed3-ed4;
```

```
/*education: fewer knots to reduce degrees of freedom*/
    eecub1 = max(0,(eduyrs4-1.5)**3);
    eecub2 = max(0,(eduyrs4-3.25)**3);
    eecub3 = max(0,(eduyrs4-4)**3);
    eecub4 = max(0,(eduyrs4-5.5)**3);

    eed1 = (eecub1-eecub4)/(5.5-1.5);
    eed2 = (eecub2-eecub4)/(5.5-3.25);
    eed3 = (eecub4-eecub4)/(5.5-4);

    eeN1 = eed1-eed3;
    eeN2 = eed2-eed3;

/*income/poverty ratio*/
pcub1 = max(0,(POVERTY_IR2-0.15)**3);
    pcub2 = max(0,(POVERTY_IR2-1)**3);
    pcub3 = max(0,(POVERTY_IR2-2)**3);
    pcub4 = max(0,(POVERTY_IR2-5)**3);
    pcub5 = max(0,(POVERTY_IR2-11)**3);

    pd1 = (pcub1-pcub5)/(11-0.15);
    pd2 = (pcub2-pcub5)/(11-1);
    pd3 = (pcub3-pcub5)/(11-2);
    pd4 = (pcub4-pcub5)/(11-5);

    pN1 = pd1-pd4;
    pN2 = pd2-pd4;
    pN3 = pd3-pd4;

/*poverty: fewer knots to reduce degrees of freedom:*/
ppcub1 = max(0,(POVERTY_IR2-0.5)**3);
    ppcub2 = max(0,(POVERTY_IR2-1.5)**3);
    ppcub3 = max(0,(POVERTY_IR2-2.5)**3);
    ppcub4 = max(0,(POVERTY_IR2-8)**3);

    ppd1 = (ppcub1-ppcub4)/(8-0.5);
    ppd2 = (ppcub2-ppcub4)/(8-1.5);
    ppd3 = (ppcub3-ppcub4)/(8-2.5);

    ppN1 = ppd1-ppd3;
    ppN2 = ppd2-ppd3;

/*age*/
acub1 = max(0,(age_consent-18)**3);
    acub2 = max(0,(age_consent-36)**3);
    acub3 = max(0,(age_consent-52)**3);
    acub4 = max(0,(age_consent-65)**3);
    acub5 = max(0,(age_consent-98)**3);

    ad1 = (acub1-acub5)/(98-18);
    ad2 = (acub2-acub5)/(98-36);
    ad3 = (acub3-acub5)/(98-52);
    ad4 = (acub4-acub5)/(98-65);
```

```
    aN1 = ad1-ad4;
    aN2 = ad2-ad4;
    aN3 = ad3-ad4;

/*age: fewer knots to reduce degrees of freedom:*/
aacub1 = max(0,(age_consent-25)**3);
    aacub2 = max(0,(age_consent-40)**3);
    aacub3 = max(0,(age_consent-50)**3);
    aacub4 = max(0,(age_consent-80)**3);

    aad1 = (aacub1-aacub4)/(80-25);
    aad2 = (aacub2-aacub4)/(80-40);
    aad3 = (aacub3-aacub4)/(80-50);

    aaN1 = aad1-aad3;
    aaN2 = aad2-aad3;
run;

proc sort data=checkup5_1;
by _imputation_;
run;

/*create categorical and dichotomous variables*/
data checkup5;
set checkup5_1;
daycat=.;
insured=.;
white=.;
urban=.;
single=.;
otherlanguage=.;
srhealth=.;
edubin=.;

if dayssince >=0 and dayssince<=20 then daycat=1;
else if dayssince>20 and dayssince<=60 then daycat=2;
else if dayssince>60 and dayssince<=91 then daycat=3;
else if dayssince>91 and dayssince<=151 then daycat=4;
else if dayssince>151 and dayssince<=183 then daycat=5;
else if dayssince>183 and dayssince<=274 then daycat=6;
else if dayssince>274 and dayssince<=366 then daycat=7;
else if dayssince>366 and dayssince<=731 then daycat=8;
else if dayssince>731 then daycat=9;

/*center age around 50, and scale by 5 years*/
agesc=(age_consent-50) / 5;

/*reduce categories*/
if INSURANCE_TYPE_R2=0 then insured=0;
else insured=1;
if RACE_ETHNICITY_4CAT=1 then white=1;
else white=0;
if RUCA_3CAT=3 then urban=0;
```

```
else urban=1;
if DMQ040=1 or DMQ040=6 then single=0;
else single=1;
if language=1 then otherlanguage=0;
else otherlanguage=1;

if (SF12010=1 or SF12010=2) then srhealth=1;
else if SF12010=3 then srhealth=2;
else if (SF12010=4 or SF12010=5) then srhealth=3;
run;

/* Export for propensity scores*/
data propexport2;
set checkup5;
keep
SPID_num
_Imputation_
AT_LEAST_COLL NO_HIGHSCHOOL SOME_COLL
COUNTYFIPS COUNTY_FIPS CTRACT10
INQ201
INSURANCE_TYPE_R2 insured
BELOW_POVLN MED_INC_13 MED_INC10K
OCQ100
POVERTY_IR2 pN1 pN2 pN3 ppN1 ppN2
PSH050
RACE_ETHNICITY_4CAT white
RUCA_3CAT urban
SF12010
age_consent aN1 aN2 aN3 aaN1 aaN2
check
chronic
eduyrs4 eN1 eN2 eN3 educat
homeowner
language otherlanguage
male
marital
own_car;
run;

data propexport3;
set propexport2;
rename
pN1=povN1
pN2=povN2
pN3=povN3

eN1=eduN1
eN2=eduN2
eN3=eduN3

aN1=ageN1
aN2=ageN2
aN3=ageN3;
run;
```

```
/* Export for multinomial analysis in R (general checkup)*/
data genexport;
set checkup5;
keep
_imputation_
SPID_num
own_car
dayssince
daycat
agesc
age_consent
aN1
aN2
aN3
aaN1
aaN2
SF12010
srhealth
chronic
male
RACE_ETHNICITY_4CAT
INSURANCE_TYPE_R2
usualplace
usualdoc
educat
eduyrs4 eN1 eN2 eN3 eeN1 eeN2
POVERTY_IR2 pN1 pN2 pN3 ppN1 ppN2
marital
language
CTRACT10
RUCA_3CAT

/*binary variables*/
insured
white
otherlanguage
urban;
run;

data genexport2;
set genexport;
rename _imputation_=Imputation;
run;



/* export for mammo analysis*/
/* eligibility: export only:
-women
- no history of breast cancer
- age 50-74
*/
```

```
data mamexport1;
set checkup5;
mamfreq=.;
if SIQ180=0 then mamfreq=0;
else if (SIQ180 NE . and SIQ180 NE 0) then mamfreq=1/SIQ181;

/*center and scale age (per 5 years)*/
agesc=(age_consent-50)/5;

/*center eduyrs4*/
educent=eduyrs4 - 3;

rename
_imputation_=Imputation;
run;

data mamexport2;
set mamexport1;
mamcat=.;
if SIQ181=1 then mamcat=1; /* within 1 year*/
else if SIQ181=2 then mamcat=2; /*2 years*/
else if SIQ181=3 then mamcat=3;/*3 years*/
else if SIQ181>3 and SIQ181<=10 then mamcat=4; /*3-10 years*/
else if SIQ181>10 then mamcat=5;/* > 10 years*/
else if SIQ181=. and SIQ180=0 then mamcat=6; /*never use*/
run;

proc sort data=mamexport2;
by Imputation; run;

data mamexport;
set mamexport2;
keep
Imputation
SPID_num
own_car
mamcat
mamfreq
agesc
age_consent aN1 aN2 aN3 aaN1 aaN2
RACE_ETHNICITY_4CAT
white
insured
educent
educat
eduyrs4 eN1 eN2 eN3 eeN1 eeN2
POVERTY_IR2 pN1 pN2 pN3 ppN1 ppN2
language
otherlanguage
CTRACT10
RUCA_3CAT
urban;
if male=0 and age_consent>=50 and age_consent<75 and mamfreq NE . and HHQ481_14=2;
run;
```

```
/*************************************************************************
******* Logistic Regression ***********************************************
*************************************************************************/
data checkup7;
set checkup5;
atleast2y=.;

/*Dichotomize outcome for logistic regression*/
if (dayssince NE . and dayssince < 730.5) then atleast2y=0;
else if dayssince >= 730.5 then atleast2y=1;
run;

proc sort data=checkup7;
by _imputation_;
run;

/*Run final proportional odds model as logistic regression also*/

/* with/without car ownership*/
proc glimmix data=checkup7 Method=LAPLACE;
class CTRACT10 srhealth(ref=first) RUCA_3CAT;
model atleast2y =
srhealth male white insured
eduyrs4
POVERTY_IR2 ppN1 ppN2
agesc
otherlanguage RUCA_3CAT/ dist=binomial covb solution;
Random intercept/SUB=CTRACT10;
ods exclude solutionr;
by _imputation_;
ods output parameterestimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
class CTRACT10 srhealth RUCA_3CAT;
modeleffects srhealth male white insured
eduyrs4
POVERTY_IR2 ppN1 ppN2
agesc
otherlanguage RUCA_3CAT;
title 'coeffs, logistic no car';
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects srhealth male white insured
eduyrs4
POVERTY_IR2 ppN1 ppN2
agesc
otherlanguage RUCA_3CAT;
/*specific OR:*/
```

```
/* PovIR 2 vs 1*/
pov2vs1: test 1*POVERTY_IR2 + 0.43333*ppN1 + 0.01923*ppN2;
/* PovIR 3 vs 1*/
pov3vs1: test 2*POVERTY_IR2 + 2.04394*ppN1 + 0.4965*ppN2;
/* PovIR 5 vs 2*/
pov5vs2: test 3*POVERTY_IR2 + 8.85909*ppN1 + 3.73601*ppN2;
/* PovIR 5 vs 3*/
pov5vs3: test 2*POVERTY_IR2 + 7.24848*ppN1 + 3.25874*ppN2;
/* PovIR 8 vs 5*/
pov8vs5: test 3*POVERTY_IR2 + 16.69091*ppN1 + 8.24476*ppN2;
title 'pov ORs, logictic no car';
run;

/* restrict to people below 200% FPL*/
data checkup8;
set checkup7;
health2=.;
if POVERTY_IR2 <=2;
run;


/*with/without car ownership*/
proc glimmix data=checkup8 Method=LAPLACE;
class CTRACT10 srhealth(ref=first) RUCA_3CAT;
model atleast2y =
srhealth male white insured
eduyrs4
POVERTY_IR2
agesc
otherlanguage RUCA_3CAT/ dist=binomial covb solution;
Random intercept/SUB=CTRACT10;
ods exclude solutionr;
by _imputation_;
ods output parameterestimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
class CTRACT10 srhealth RUCA_3CAT;
modeleffects srhealth male white insured
eduyrs4
POVERTY_IR2
agesc
otherlanguage RUCA_3CAT;
title 'coeffs, logistic no car';
run;


/*********************************************/
/*mammography analysis as logistic regression*/
data mamlog;
set mamexport;
if mamcat>2 then mambin=0; /*mamcat=6 = never use*/
else if  mamcat=1 or mamcat=2 then mambin=1;
```

```
_imputation_=Imputation;
run;

/*with/Without car*/
proc glimmix data=mamlog Method=LAPLACE;
class CTRACT10 RUCA_3CAT;
model mambin =
POVERTY_IR2 ppN1 ppN2
age_consent
eduyrs4
white insured urban
/ dist=binomial covb solution;
Random intercept/SUB=CTRACT10;
ods exclude solutionr;
by _imputation_;
ods output parameterestimates=parms CovB=cov solutionr=rand_parms;
*where POVERTY_IR2<=2;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects POVERTY_IR2 ppN1 ppN2
age_consent
eduyrs4
white insured urban ;
/*specific OR:*/
/* PovIR 2 vs 1*/
pov2vs1: test 1*POVERTY_IR2 + 0.43333*ppN1 + 0.01923*ppN2;
/* PovIR 3 vs 1*/
pov3vs1: test 2*POVERTY_IR2 + 2.04394*ppN1 + 0.4965*ppN2;
/* PovIR 5 vs 2*/
pov5vs2: test 3*POVERTY_IR2 + 8.85909*ppN1 + 3.73601*ppN2;
/* PovIR 5 vs 3*/
pov5vs3: test 2*POVERTY_IR2 + 7.24848*ppN1 + 3.25874*ppN2;
/* PovIR 8 vs 5*/
pov8vs5: test 3*POVERTY_IR2 + 16.69091*ppN1 + 8.24476*ppN2;
title 'pov ORs, mam logictic no car, everyone';
run;


proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
class CTRACT10 RUCA_3CAT;
modeleffects POVERTY_IR2 ppN1 ppN2
age_consent
eduyrs4
white insured urban;
title 'coeffs, remaining ORS';
run;


/*Restrict to people <= 2 FPL*/
/*with/Without car*/
proc glimmix data=mamlog Method=LAPLACE;
```

```
class CTRACT10 RUCA_3CAT;
model mambin =
POVERTY_IR2
age_consent
eduyrs4
white insured urban own_car
/ dist=binomial covb solution;
Random intercept/SUB=CTRACT10;
ods exclude solutionr;
by _imputation_;
ods output parameterestimates=parms CovB=cov solutionr=rand_parms;
where POVERTY_IR2<=2;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
class CTRACT10 RUCA_3CAT;
modeleffects POVERTY_IR2
age_consent
eduyrs4
white insured urban own_car;
title 'coeffs, remaining ORS';
run;

/*without insurnace:*/
proc glimmix data=mamlog Method=LAPLACE;
class CTRACT10 RUCA_3CAT;
model mambin =
POVERTY_IR2
age_consent
eduyrs4
white urban own_car
/ dist=binomial covb solution;
Random intercept/SUB=CTRACT10;
ods exclude solutionr;
by _imputation_;
ods output parameterestimates=parms CovB=cov solutionr=rand_parms;
where POVERTY_IR2<=2;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
class CTRACT10 RUCA_3CAT;
modeleffects POVERTY_IR2
age_consent
eduyrs4
white urban own_car;
title 'coeffs, remaining ORS';
run;

/*Propensity Scores:*/
option ls=76 ps=500 mergenoby=ERROR ;

%include 'T:\Core\Core_main\Source File Development\Source File Create\Formats\SHOW_MACROS.sas';
```

```
libname pjewett 'T:\CORE\CORE_Main\Analytical\Data requests\2015\206_Jewett_Car_Ownership\data' ;

data prop_in;
set propexport3; /*generated in analysis2 program*/
run;

DATA prop_in3;
SET prop_in;
* spline terms for median income (per 10K);
mcub1 = max(0,(MED_INC10K-1.1)**3);
    mcub2 = max(0,(MED_INC10K-4.4)**3);
    mcub3 = max(0,(MED_INC10K-5.4)**3);
    mcub4 = max(0,(MED_INC10K-7)**3);
    mcub5 = max(0,(MED_INC10K-15.7)**3);
    md1 = (mcub1-mcub5)/(15.7-1.1);
    md2 = (mcub2-mcub5)/(15.7-4.4);
    md3 = (mcub3-mcub5)/(15.7-5.4);
    md4 = (mcub4-mcub5)/(15.7-7);
    miN1 = md1-md4;
    miN2 = md2-md4;
    miN3 = md3-md4;

* fewer knots to reduce df: median income (per 10K);
    mmcub1 = max(0,(MED_INC10K-1.1)**3);
    mmcub2 = max(0,(MED_INC10K-4)**3);
    mmcub3 = max(0,(MED_INC10K-5)**3);
    mmcub4 = max(0,(MED_INC10K-10)**3);

    mmd1 = (mmcub1-mmcub4)/(10-1.1);
    mmd2 = (mmcub2-mmcub4)/(10-4);
    mmd3 = (mmcub3-mmcub4)/(10-5);

    mmN1 = mmd1-mmd3;
    mmN2 = mmd2-mmd3;

* spline terms number of people in household;
    hhcub1 = max(0,(INQ201-1)**3);
    hhcub2 = max(0,(INQ201-2)**3);
    hhcub3 = max(0,(INQ201-3)**3);
    hhcub4 = max(0,(INQ201-11)**3);
    hhd1 = (hhcub1-hhcub4)/(11-1);
hhd2 = (hhcub2-hhcub4)/(11-2);
    hhd3 = (hhcub3-hhcub4)/(11-3);
    hhN1 = hhd1-hhd3;
    hhN2 = hhd2-hhd3;
run;

proc sort data=prop_in3;
by _imputation_;
run;

/*create datset with interaction terms*/
data prop_in4;
set prop_in3;
```

```
povmale1=.;
povmale2=.;
povmale3=.;
povmale4=.;

/*poverty and gender interaction*/
povmale1=POVERTY_IR2*male;
povmale2=povN1*male;
povmale3=povN2*male;
povmale4=povN3*male;

/*create race dummy variable*/
raceb=.; /*AfrAm*/
raceh=.; /*Hisp*/
raceo=.; /*Other*/

if RACE_ETHNICITY_4CAT=1 then do; /*ref group = whites*/
raceb=0;
raceh=0;
raceo=0;
end;

else if RACE_ETHNICITY_4CAT=2 then do;
raceb=1;
raceh=0;
raceo=0;
end;

else if RACE_ETHNICITY_4CAT=3 then do;
raceb=0;
raceh=1;
raceo=0;
end;

else if RACE_ETHNICITY_4CAT=4 then do;
raceb=0;
raceh=0;
raceo=1;
end;

/*race and poverty*/
povraceb1=POVERTY_IR2*raceb;
povraceb2=povN1*raceb;
povraceb3=povN2*raceb;
povraceb4=povN3*raceb;

povraceh1=POVERTY_IR2*raceh;
povraceh2=povN1*raceh;
povraceh3=povN2*raceh;
povraceh4=povN3*raceh;

povraceo1=POVERTY_IR2*raceo;
povraceo2=povN1*raceo;
povraceo3=povN2*raceo;
```

```
povraceo4=povN3*raceo;

/*ruca and poverty*/
/*create ruca3cat dummy variable*/
ruca1=.; /**/
ruca2=.;

if RUCA_3CAT=1 then do;
ruca1=1;
ruca2=0;
end;

else if RUCA_3CAT=2 then do;
ruca1=0;
ruca2=1;
end;

else if RUCA_3CAT=3 then do;
ruca1=0;
ruca2=0;
end;

povruca1_1=POVERTY_IR2*ruca1;
povruca1_2=povN1*ruca1;
povruca1_3=povN2*ruca1;
povruca1_4=povN3*ruca1;

povruca2_1=POVERTY_IR2*ruca2;
povruca2_2=povN1*ruca2;
povruca2_3=povN2*ruca2;
povruca2_4=povN3*ruca2;

/*insuranced dummy varaiables*/
insure1=0;
insure2=0;

if INSURANCE_TYPE_R2=0 then do; /*None*/
insure1=1;
insure2=0;
end;

else if INSURANCE_TYPE_R2=1 then do; /*Medicaid*/
insure1=0;
insure2=1;
end;

else if INSURANCE_TYPE_R2=2 then do; /*Medicare or private (ref)*/
insure1=0;
insure2=0;
end;

run;

proc sort data=prop_in4;
```

```
by _imputation_;
run;


/*Initial model: PIR, age, and median income nonlinear*/
proc glimmix data=prop_in4 Method=LAPLACE;
class CTRACT10 RUCA_3CAT;
model own_car =
POVERTY_IR2 povN1 povN2 povN3
eduyrs4
age_consent ageN1 ageN2 ageN3
MED_INC10K  miN1 miN2 miN3
INQ201
homeowner male  urban white insured
/ dist=binomial covb solution;
Random intercept/SUB=CTRACT10;
ods exclude solutionr;
by _imputation_;
ods output parameterestimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects POVERTY_IR2 povN1 povN2 povN3
eduyrs4
age_consent ageN1 ageN2 ageN3
MED_INC10K  miN1 miN2 miN3
INQ201
homeowner male  urban white insured;

povnl: test povN1, povN2, povN3/ mult; /*income/pov nonlinear*/
agenl: test ageN1, ageN2, ageN3 / mult; /*age nonlinear*/
mincnl: test miN1, miN2, miN3 / mult; /*median income nonlinear*/
run;

/*education and people nonlinear?*/
Proc glimmix data=prop_in4 Method=LAPLACE;
class CTRACT10 RUCA_3CAT;
model own_car =
POVERTY_IR2
eduyrs4 eduN1 eduN2 eduN3
age_consent
MED_INC10K
INQ201 hhN1 hhN2
homeowner male  urban white insured
/ dist=binomial covb solution;
Random intercept/SUB=CTRACT10;
ods exclude solutionr;
by _imputation_;
ods output parameterestimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
```

```
modeleffects POVERTY_IR2
eduyrs4 eduN1 eduN2 eduN3
age_consent
MED_INC10K
INQ201 hhN1 hhN2
homeowner male  urban white insured;

edunl: test eduN1, eduN2, eduN3/ mult; /*education nonlinear*/
hhnl: test hhN1, hhN2 / mult; /*number of people nonlinear*/
run;

/*test for interactions*/
/*poverty with race*/
proc glimmix data=prop_in4 Method=LAPLACE;
class CTRACT10 RUCA_3CAT;
model own_car =
POVERTY_IR2
eduyrs4
age_consent
MED_INC10K
INQ201
homeowner male  urban white insured
POVERTY_IR2*white
/ dist=binomial covb solution;
Random intercept/SUB=CTRACT10;
ods exclude solutionr;
by _imputation_;
ods output parameterestimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms;
class CTRACT10;
modeleffects POVERTY_IR2
eduyrs4
age_consent
MED_INC10K
INQ201
homeowner male  urban white insured
POVERTY_IR2*white;
run;

/*poverty with ruca*/
proc glimmix data=prop_in4 Method=LAPLACE;
class CTRACT10 RUCA_3CAT;
model own_car =
POVERTY_IR2
eduyrs4
age_consent
MED_INC10K
INQ201
homeowner male RUCA_3CAT white insured
povruca1_1 povruca2_1
/ dist=binomial covb solution;
Random intercept/SUB=CTRACT10;
```

```
ods exclude solutionr;
by _imputation_;
ods output parameterestimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects POVERTY_IR2
eduyrs4
age_consent
MED_INC10K
INQ201
homeowner male RUCA_3CAT white insured
povruca1_1 povruca2_1;
povruca: test povruca1_1, povruca2_1/ mult; /*poverty*ruca interaction*/
run;

/*poverty with gender*/
proc glimmix data=prop_in4 Method=LAPLACE;
class CTRACT10 RUCA_3CAT;
model own_car =
POVERTY_IR2
eduyrs4
age_consent
MED_INC10K
INQ201 urban white insured
POVERTY_IR2*male
/ dist=binomial covb solution;
Random intercept/SUB=CTRACT10;
ods exclude solutionr;
by _imputation_;
ods output parameterestimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms;
class CTRACT10 RUCA_3CAT;
modeleffects POVERTY_IR2
eduyrs4
age_consent
MED_INC10K
INQ201 urban white insured
POVERTY_IR2*male;
run;

/*take out:
1. nonlinear terms for education years, number of people in
household, and interaction terms*/
/*Final model: nonlinear effects for POVIR, age, and median income*/
/*use fewer knots for poverty, age, and median income*/

proc glimmix data=prop_in4 Method=LAPLACE;
class CTRACT10 RUCA_3CAT RACE_ETHNICITY_4CAT(ref=first) INSURANCE_TYPE_R2;
model own_car =
POVERTY_IR2 ppN1 ppN2
```

```
eduyrs4
age_consent aaN1 aaN2
MED_INC10K  mmN1 mmN2
INQ201
male urban white insured
/ dist=binomial covb solution;
Random intercept/SUB=CTRACT10;
ods exclude solutionr;
by _imputation_;
ods output parameterestimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects POVERTY_IR2 ppN1 ppN2
eduyrs4
age_consent aaN1 aaN2
MED_INC10K  mmN1 mmN2
INQ201
male urban white insured;

povnl: test ppN1, ppN2 / mult; /*income/pov nonlinear*/
agenl: test aaN1, aaN2 / mult; /*age nonlinear*/
mincnl: test mmN1, mmN2 / mult; /*median income nonlinear*/

/*Overall tests*/
povov: test POVERTY_IR2, ppN1, ppN2/ mult; /*income/pov overall*/
ageov: test age_consent, aaN1, aaN2 / mult; /*age overall*/
mincov: test MED_INC10K, mmN1, mmN2 / mult; /*median income overall*/
run;

/*ORs for non-spline terms*/
proc mianalyze parms=parms;
class CTRACT10 RUCA_3CAT NBRHD006_0_FIN RACE_ETHNICITY_4CAT INSURANCE_TYPE_R2;
modeleffects POVERTY_IR2 ppN1 ppN2
eduyrs4
age_consent aaN1 aaN2
MED_INC10K  mmN1 mmN2
INQ201
male urban white insured;
run;

/*Calculate specific odds ratios for nonlinear terms*/
proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects POVERTY_IR2 ppN1 ppN2
eduyrs4
age_consent aaN1 aaN2
MED_INC10K  mmN1 mmN2
INQ201
male urban white insured;
/* PovIR 2 vs 1*/
pov2vs1: test 1*POVERTY_IR2 + 0.43333*ppN1 + 0.01923*ppN2;
/* PovIR 3 vs 1*/
```

```
pov3vs1: test 2*POVERTY_IR2 + 2.04394*ppN1 + 0.4965*ppN2;
/* PovIR 5 vs 2*/
pov5vs2: test 3*POVERTY_IR2 + 8.85909*ppN1 + 3.73601*ppN2;
/* PovIR 5 vs 3*/
pov5vs3: test 2*POVERTY_IR2 + 7.24848*ppN1 + 3.25874*ppN2;
/* PovIR 8 vs 5*/
pov8vs5: test 3*POVERTY_IR2 + 16.69091*ppN1 + 8.24476*ppN2;

/* age 35 vs 25*/
age35vs25: test 10*age_consent + 18.18182*aaN1 + 0*aaN2;
/* age 50 vs 35*/
age50vs35: test 20*age_consent + 284.09091*aaN1 + 25*aaN2;
/* age 75 vs 50*/
age75vs50: test 25*age_consent + 1467.80303*aaN1 + 526.04167*aaN2;

/*median income:*/
/*25,000 vs 15,000*/
min25vs15: test 1*MED_INC10K + 0.30112*mmN1 + 0*mmN2;
/*50,000 vs 25,000*/
min50vs25: test 2.5*MED_INC10K + 6.35674*mmN1 + 0.16667*mmN2;
/*75,000 vs 50,000*/
min70vs50: test 2.5*MED_INC10K + 19.66433*mmN1 + 3.85417*mmN2;
run;


/*calculate and store propensity scores for each individual*/

proc glimmix data=prop_in4 Method=LAPLACE;
class CTRACT10 RUCA_3CAT RACE_ETHNICITY_4CAT(ref=first) INSURANCE_TYPE_R2;
model own_car =
POVERTY_IR2 ppN1 ppN2
eduyrs4
age_consent aaN1 aaN2
MED_INC10K  mmN1 mmN2
INQ201
male urban white insured
/ dist=binomial covb solution;
Random intercept/SUB=CTRACT10;
ods exclude solutionr;
by _imputation_;
output OUT= scores_out
PREDICTED(BLUP ILINK)= Treat_Prob_Pred
STDERR(BLUP ILINK)= Treat_Prob_SE;
run;

proc contents data=scores_out;run;

data scores_noformat;
set scores_out;
FORMAT _all_;
INFORMAT _all_;
run;

data scores_out2;
```

```
set scores_noformat;
keep
SPID_num
_Imputation_
AT_LEAST_COLL NO_HIGHSCHOOL SOME_COLL
COUNTYFIPS COUNTY_FIPS CTRACT10
INQ201
INSURANCE_TYPE_R2 insured
BELOW_POVLN MED_INC_13 MED_INC10K
OCQ100
POVERTY_IR2 pN1 pN2 pN3 ppN1 ppN2
PSH050
RACE_ETHNICITY_4CAT white
RUCA_3CAT urban
SF12010
age_consent aN1 aN2 aN3 aaN1 aaN2
check
chronic
eduyrs4 eN1 eN2 eN3 educat
homeowner
language otherlanguage
male
marital
own_car
Treat_Prob_Pred
Treat_Prob_SE;
run;

PROC EXPORT DATA=work.scores_out2
OUTFILE="T:\CORE\CORE_Main\Analytical\Data requests\2015\206_Jewett_Car_Ownership\data\Non-SAS data\sco
DBMS=csv
REPLACE;
run;

/*peopl with high incomes but no car:*/
data nocar_high_income;
set scores_out2;
if POVERTY_IR2>=2;
run;

PROC EXPORT DATA=work.nocar_high_income
OUTFILE="T:\CORE\CORE_Main\Analytical\Data requests\2015\206_Jewett_Car_Ownership\data\Non-SAS data\noca
DBMS=csv
REPLACE;
run;

/*descriptive stats: home ownership by age group*/
proc freq data=prop_in4;
tables own_car /list missing;
where age_consent<20;
run;
```

F.0.3 R Code used for analysis in chapter 3 (car ownership and preventive and mammography care)

```
# poportional odds model for general checkup:
library(ordinal)
library(mitools)

input=read.csv("genexport2.csv", header=TRUE)

#  categorcal variables
input$days_factor=as.factor(input$dayssince)
input$daycatf=as.factor(input$daycat)
input$ctract_fac=as.factor(input$CTRACT10)
input$usual_fac=as.factor(input$usualplace)
input$marit_fac=as.factor(input$marital)
input$lang_fac=as.factor(input$language)
input$edu_fac=as.factor(input$educat)
input$ruca_fac=as.factor(input$RUCA_3CAT)
input$healthfac=as.factor(input$srhealth)
# reference levels
input <- within(input, INSURANCE_TYPE_R2 <- relevel(INSURANCE_TYPE_R2, ref = 2))
input <- within(input, edu_fac <- relevel(edu_fac, ref = 4))
input <- within(input, ruca_fac <- relevel(ruca_fac, ref = 3))
input <- within(input, healthfac<- relevel(healthfac, ref = 1))
# imputed datasets individually
imp1=input[which(input$Imputation==1),]
imp2=input[which(input$Imputation==2),]
imp3=input[which(input$Imputation==3),]
imp4=input[which(input$Imputation==4),]
imp5=input[which(input$Imputation==5),]

#############################################################
#  model selection
m1 <-clmm2(days_factor ~ healthfac + male + white + insured
             + eduyrs4 + eeN1 + eeN2 + agesc
             + POVERTY_IR2 + ppN1 + ppN2
             + otherlanguage + urban,
             data = imp1, random=ctract_fac, Hess=TRUE)

m2 <-clmm2(days_factor ~ healthfac + male + white + insured
           + eduyrs4 + eeN1 + eeN2 + agesc
           + POVERTY_IR2 + ppN1 + ppN2
           + otherlanguage + urban,
           data = imp2, random=ctract_fac, Hess=TRUE)

m3 <-clmm2(days_factor ~ healthfac + male + white + insured
           + eduyrs4 + eeN1 + eeN2 + agesc
           + POVERTY_IR2 + ppN1 + ppN2
           + otherlanguage + urban,
           data = imp3, random=ctract_fac, Hess=TRUE)

m4 <-clmm2(days_factor ~ healthfac + male + white + insured
           + eduyrs4 + eeN1 + eeN2 + agesc
           + POVERTY_IR2 + ppN1 + ppN2
```

```
           + otherlanguage + urban,
           data = imp4, random=ctract_fac, Hess=TRUE)

m5 <-clmm2(days_factor ~ healthfac + male + white + insured
           + eduyrs4 + eeN1 + eeN2 + agesc
           + POVERTY_IR2 + ppN1 + ppN2
           + otherlanguage + urban,
           data = imp5, random=ctract_fac, Hess=TRUE)

models=list(m1,m2,m3,m4,m5)
betas<-MIextract(models,fun=coef)
vars<-MIextract(models, fun=vcov)
combined1=MIcombine(betas,vars)

# Get combined covariance matrix
covmatrix1=vcov(combined1)
#dim(covmatrix1)[1]
# keep only the non-intercept elements from matrix
covmatrix2=covmatrix1[-(1:54),-(1:54)]
# get rid of last entry in matrix
lastentry=dim(covmatrix2)[1]
covmatrix=covmatrix2[-lastentry,-lastentry]

# Get betas
coeffs1=coefficients(combined1)[-(1:54)]
lastcoeff=length(coeffs1)
coeffs=coeffs1[-lastcoeff]
parnum=length(coeffs)
coeffvec=as.matrix(coeffs)

# test for nonlinearity
######### ##########################
# poverty nonlinear
coeffs[11:12]
lower=11
upper=12
initialvec=rep(0,parnum)
contrastmatrix=c()
tmp=c()
for (i in lower:upper) {
  tmp=initialvec
  tmp[i]=1
  contrastmatrix=rbind(contrastmatrix,tmp)
}
# contrastmatrix * covmatrix * t(contrastmatrix) = SE^2 of estimate
Vmatrix=contrastmatrix %*% covmatrix %*% t(contrastmatrix)

# Calculate Chisquare
testcoeff=as.matrix(coeffvec[lower:upper])
Chisq=t(testcoeff) %*% solve(Vmatrix) %*% testcoeff
df=dim(testcoeff)[1]
(P=pchisq(Chisq,df,lower.tail=FALSE))

######### ##########################
```

```
# education nonlinear
coeffs[7:8]
lower=7
upper=8
initialvec=rep(0,parnum)
contrastmatrix=c()
tmp=c()
for (i in lower:upper) {
  tmp=initialvec
  tmp[i]=1
  contrastmatrix=rbind(contrastmatrix,tmp)
}

Vmatrix=contrastmatrix %*% covmatrix %*% t(contrastmatrix)
# Calculate Chisquare
testcoeff=as.matrix(coeffvec[lower:upper])
dim(testcoeff)
Chisq=t(testcoeff) %*% solve(Vmatrix) %*% testcoeff
df=dim(testcoeff)[1]
(P=pchisq(Chisq,df,lower.tail=FALSE))

######## Take nonlinear education out, put age in nonlinearly #####
#####################################################################

m6 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4
           + age_consent + aaN1 + aaN2
           + POVERTY_IR2 + ppN1 + ppN2
           + otherlanguage + urban,
           data = imp1, random=ctract_fac, Hess=TRUE)

m7 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4
           + age_consent + aaN1 + aaN2
           + POVERTY_IR2 + ppN1 + ppN2
           + otherlanguage + urban,
           data = imp2, random=ctract_fac, Hess=TRUE)

m8 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4
           + age_consent + aaN1 + aaN2
           + POVERTY_IR2 + ppN1 + ppN2
           + otherlanguage + urban,
           data = imp3, random=ctract_fac, Hess=TRUE)

m9 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4
           + age_consent + aaN1 + aaN2
           + POVERTY_IR2 + ppN1 + ppN2
           + otherlanguage + urban,
           data = imp4, random=ctract_fac, Hess=TRUE)

m10 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4
            + age_consent + aaN1 + aaN2
            + POVERTY_IR2 + ppN1 + ppN2
            + otherlanguage + urban,
           data = imp5, random=ctract_fac, Hess=TRUE)
```

```
models=list(m6,m7,m8,m9,m10)
betas<-MIextract(models,fun=coef)
vars<-MIextract(models, fun=vcov)
combined1=MIcombine(betas,vars)

# Get combined covariance matrix
covmatrix1=vcov(combined1)
covmatrix2=covmatrix1[-(1:54),-(1:54)]
# get rid of last entry in matrix
lastentry=dim(covmatrix2)[1]
covmatrix=covmatrix2[-lastentry,-lastentry]

# Get betas
coeffs1=coefficients(combined1)[-(1:54)]
lastcoeff=length(coeffs1)
coeffs=coeffs1[-lastcoeff]
parnum=length(coeffs)
coeffvec=as.matrix(coeffs)

# test for nonlinearity
######### ##########################
# age nonlinear
coeffs[8:9]
lower=8
upper=9
initialvec=rep(0,parnum)
contrastmatrix=c()
tmp=c()
for (i in lower:upper) {
  tmp=initialvec
  tmp[i]=1
  contrastmatrix=rbind(contrastmatrix,tmp)
}

Vmatrix=contrastmatrix %*% covmatrix %*% t(contrastmatrix)
# Calculate Chisquare
testcoeff=as.matrix(coeffvec[lower:upper])
Chisq=t(testcoeff) %*% solve(Vmatrix) %*% testcoeff
df=dim(testcoeff)[1]
(P=pchisq(Chisq,df,lower.tail=FALSE)) # age not nonlinearly significant; P=0.7034

#############################################################################################
# take out nonlinear effects of age
# Test for interaction: pov*sex, pov*race
m11 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
           + POVERTY_IR2 + ppN1 + ppN2
           + otherlanguage + urban
           + POVERTY_IR2*white + ppN1*white + ppN2*white,
           data = imp1, random=ctract_fac, Hess=TRUE)

m12 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
            + POVERTY_IR2 + ppN1 + ppN2
            + otherlanguage + urban
            + POVERTY_IR2*white + ppN1*white + ppN2*white,
```

```
           data = imp2, random=ctract_fac, Hess=TRUE)

m13 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
            + POVERTY_IR2 + ppN1 + ppN2
            + otherlanguage + urban
            + POVERTY_IR2*white + ppN1*white + ppN2*white,
           data = imp3, random=ctract_fac, Hess=TRUE)

m14 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4
            + age_consent
            + POVERTY_IR2 + ppN1 + ppN2
            + otherlanguage + urban
            + POVERTY_IR2*white + ppN1*white + ppN2*white,
           data = imp4, random=ctract_fac, Hess=TRUE)

m15 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
            + POVERTY_IR2 + ppN1 + ppN2
            + otherlanguage + urban
            + POVERTY_IR2*white + ppN1*white + ppN2*white,
            data = imp5, random=ctract_fac, Hess=TRUE)

models=list(m11,m12,m13,m14,m15)
betas<-MIextract(models,fun=coef)
vars<-MIextract(models, fun=vcov)
combined1=MIcombine(betas,vars)

# Get combined covariance matrix
covmatrix1=vcov(combined1)
covmatrix2=covmatrix1[-(1:54),-(1:54)]
# get rid of last entry in matrix
lastentry=dim(covmatrix2)[1]
covmatrix=covmatrix2[-lastentry,-lastentry]

# Get betas
coeffs1=coefficients(combined1)[-(1:54)]
lastcoeff=length(coeffs1)
coeffs=coeffs1[-lastcoeff]
parnum=length(coeffs)
coeffvec=as.matrix(coeffs)

# test for interaction
######### ##########################
coeffs[13:15]
lower=13
upper=15
initialvec=rep(0,parnum)
contrastmatrix=c()
tmp=c()
for (i in lower:upper) {
  tmp=initialvec
  tmp[i]=1
  contrastmatrix=rbind(contrastmatrix,tmp)
}
```

```
Vmatrix=contrastmatrix %*% covmatrix %*% t(contrastmatrix)
# Calculate Chisquare
testcoeff=as.matrix(coeffvec[lower:upper])
Chisq=t(testcoeff) %*% solve(Vmatrix) %*% testcoeff
df=dim(testcoeff)[1]
(P=pchisq(Chisq,df,lower.tail=FALSE)) # interaction poverty* race not significant; P=0.2431083

#############################################################################################
# Test for interaction: pov*sex
m16 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
            + POVERTY_IR2 + ppN1 + ppN2
            + otherlanguage + urban
            + POVERTY_IR2*male + ppN1*male + ppN2*male,
            data = imp1, random=ctract_fac, Hess=TRUE)

m17 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
            + POVERTY_IR2 + ppN1 + ppN2
            + otherlanguage + urban
            + POVERTY_IR2*male + ppN1*male + ppN2*male,
            data = imp2, random=ctract_fac, Hess=TRUE)

m18 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
            + POVERTY_IR2 + ppN1 + ppN2
            + otherlanguage + urban
            + POVERTY_IR2*male + ppN1*male + ppN2*male,
            data = imp3, random=ctract_fac, Hess=TRUE)

m19 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
            + POVERTY_IR2 + ppN1 + ppN2
            + otherlanguage + urban
            + POVERTY_IR2*male + ppN1*male + ppN2*male,
            data = imp4, random=ctract_fac, Hess=TRUE)

m20 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
            + POVERTY_IR2 + ppN1 + ppN2
            + otherlanguage + urban
            + POVERTY_IR2*male + ppN1*male + ppN2*male,
            data = imp5, random=ctract_fac, Hess=TRUE)

models=list(m16,m17,m18,m19,m20)
betas<-MIextract(models,fun=coef)
vars<-MIextract(models, fun=vcov)
combined1=MIcombine(betas,vars)

# Get combined covariance matrix
covmatrix1=vcov(combined1)
covmatrix2=covmatrix1[-(1:54),-(1:54)]
# get rid of last entry in matrix
lastentry=dim(covmatrix2)[1]
covmatrix=covmatrix2[-lastentry,-lastentry]

# Get betas
length(coefficients(combined1))
coeffs1=coefficients(combined1)[-(1:54)]
```

```
lastcoeff=length(coeffs1)
coeffs=coeffs1[-lastcoeff]
parnum=length(coeffs)
coeffvec=as.matrix(coeffs)

# test for interaction
######### ##########################
coeffs[13:15]
lower=13
upper=15
initialvec=rep(0,parnum)
contrastmatrix=c()
tmp=c()
for (i in lower:upper) {
  tmp=initialvec
  tmp[i]=1
  contrastmatrix=rbind(contrastmatrix,tmp)
}

Vmatrix=contrastmatrix %*% covmatrix %*% t(contrastmatrix)
# Calculate Chisquare
testcoeff=as.matrix(coeffvec[lower:upper])
Chisq=t(testcoeff) %*% solve(Vmatrix) %*% testcoeff
df=dim(testcoeff)[1]
(P=pchisq(Chisq,df,lower.tail=FALSE)) # interaction poverty* race not significant; P=0.3887183

#############################################################################################
# Final model without car ownership
# add car ownership for comparison results
fin1 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
             + POVERTY_IR2 + ppN1 + ppN2
             + otherlanguage + ruca_fac,
             data = imp1, random=ctract_fac, Hess=TRUE)

fin2 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
             + POVERTY_IR2 + ppN1 + ppN2
             + otherlanguage + ruca_fac,
             data = imp2, random=ctract_fac, Hess=TRUE)

fin3 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
             + POVERTY_IR2 + ppN1 + ppN2
             + otherlanguage + ruca_fac,
             data = imp3, random=ctract_fac, Hess=TRUE)

fin4 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
             + POVERTY_IR2 + ppN1 + ppN2
             + otherlanguage + ruca_fac,
             data = imp4, random=ctract_fac, Hess=TRUE)

fin5 <-clmm2(days_factor ~ healthfac + male + white + insured + eduyrs4 + age_consent
             + POVERTY_IR2 + ppN1 + ppN2
             + otherlanguage + ruca_fac,
             data = imp5, random=ctract_fac, Hess=TRUE)
```

```
models=list(fin1,fin2,fin3,fin4,fin5)
betas<-MIextract(models,fun=coef)
vars<-MIextract(models, fun=vcov)
combined1=MIcombine(betas,vars)
results=summary(combined1)
write.csv(results,"results_gencheckup_nocar.csv")

# Get combined covariance matrix
covmatrix1=vcov(combined1)
covmatrix2=covmatrix1[-(1:54),-(1:54)]
# get rid of last entry in matrix
lastentry=dim(covmatrix2)[1]
covmatrix=covmatrix2[-lastentry,-lastentry]

# Get betas
coeffs1=coefficients(combined1)[-(1:54)]
lastcoeff=length(coeffs1)
coeffs=coeffs1[-lastcoeff]
parnum=length(coeffs)
coeffvec=as.matrix(coeffs)

################ poverty ORs final model ####################
coeffs[8:10]
one=8
two=9
three=10
parnum=length(coeffvec)

# 2 vs 1 POVIR
betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=1
betavec[two]=0.43333
betavec[three]=0.01923
betavec=t(betavec)
(two_vs_1=betavec %*% coeffvec)
(two_vs_1SE=sqrt(betavec %*% covmatrix %*% t(betavec)))
# 3 vs 1 POVIR
betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=2
betavec[two]=2.04394
betavec[three]=0.49650
betavec=t(betavec)
(three_vs_1=betavec %*% coeffvec)
(three_vs_1SE=sqrt(betavec %*% covmatrix %*% t(betavec)))

# 5 vs 2 POVIR
betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=3
betavec[two]=8.85909
betavec[three]=3.73601
betavec=t(betavec)
```

```
(five_vs_2=betavec %*% coeffvec)
(five_vs_2SE=sqrt(betavec %*% covmatrix %*% t(betavec)))

# 5 vs 3 POVIR
betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=2
betavec[two]=7.24848
betavec[three]=3.25874
betavec=t(betavec)
(five_vs_3=betavec %*% coeffvec)
(five_vs_3SE=sqrt(betavec %*% covmatrix %*% t(betavec)))

# 8 vs 5 POVIR
betavec=rep(0,parnum)
betavec=as.matrix(betavec)
betavec[one]=3
betavec[two]=16.69091
betavec[three]=8.24476
betavec=t(betavec)
(eight_vs_5=betavec %*% coeffvec)
(eight_vs_5SE=sqrt(betavec %*% covmatrix %*% t(betavec)))

# export linear contrasts  for model without own_car
LnORs=c(two_vs_1,three_vs_1,five_vs_2,five_vs_3,eight_vs_5)
LnORs_SE=c(two_vs_1SE,three_vs_1SE,five_vs_2SE,five_vs_3SE,eight_vs_5SE)
LnORs_export=rbind(LnORs,LnORs_SE)

write.csv(LnORs_export,"LnORS_nocar.csv")
```

F.0.4 Code used for analysis in chapter 4 (mammography utilization by driving times and mammography facility density)

```
/*Read in data:*/
option ls=76 ps=500 mergenoby=ERROR ;
libname pjewett 'G:\Team\WHS DATA\whsstat\Project Datasets
\Mammography\Elkin\Data\WI Controls';

/*Read in all controls for analysis*/
data controls;
  set pjewett.wi_controls_for_mammography_v2;
 run ;

/* calculate ref year and only keep phases 0,3,4,5 -->10,548 participants*/
 data controls1;
  set controls;
  refyr=year(rfdate);
  if phase=0 or phase=3 or phase=4 or phase=5;
 run ;

data controls2;
set controls1;
dropflag=0;
if phase=3 and (_VERSION="4.00" OR _VERSION="4.10"
```

```
OR _VERSION="5.00"  OR _VERSION="5.10"  OR _VERSION="5.30"
OR _VERSION="5.40" OR _VERSION="6.00") then dropflag=1;
run;

data controls3;
set controls2;
if dropflag=0;
run;

 /* Restrict to ages 50-74, since mammo recommendations existed consistently for that group*/
 data controls4;
  set controls3;
 if refage>=50 and refage<75;
 run;

/* data clean-up*/
data ctrlcleaned;
set controls4;
famhist2=.;
menop2=.;
hrtever=.;
white=.;
if (refage < 10 or refage > 110 ) then refage = .;
/*assumed plausible age range for breast cancer = 10-110*/
if educ < 1 then educ = .; /*1-9*/
if income < 1 then income =.; /*1-5*/
if MENOAG<21 then MENOAG=.;
if alcohol>100 then alcohol=.;
if famhist=1 then famhist2=0;
if famhist=2 then famhist2=1;
if lact<0 then lact=.;
if ocdur<0 then ocdur=.;
if (ocever<0) then ocever=.;
if menop=1 then menop2=0;
if menop=2 then menop2=1;
if (insuranc <0 AND insuranc NE -6) then insuranc=.;
if mamn < 0 then mamn=.;
if screen=3 then screen=.;
if raceXX=1 then white=1;
if raceXX=2 then white=0;
/*Combine HRTNFC and EPEVER into one binary variable*/
if (HRTNFC=2 or HRTNFC=3) OR
(EPEVER=2 OR EPEVER=3 OR EPEVER=4 OR EPEVER=5) then hrtever=1;
if HRTNFC=1 AND EPEVER=1 then hrtever=0;
if HHAD<=0 then HHAD=.; /*>=1*/
if HHCH<0 then HHCH=.; /*>=0*/
run;

/* Exclude women with unknown mammoraphy data*/
/*Set MAMN to 0 if women stated they had no mammogram and left MAMN open*/
data controls5;
set ctrlcleaned;
IF screen=1 and MAMN=. then MAMN=0;
run;
```

```
/*If women stated they had no mammogram but
then gave a number for MAMN, set MAMN to missing (contradiction)*/
data controls6;
set controls5;
IF screen=1 and MAMN NE 0 then MAMN=.;
run;

/* Exclude women for whom mammo use is unknown*/
 data controls7;
  set controls6;
 if mamn>=0;
 run;

/*keep only relevant variables*/
data controls9;
set controls7;
drop
afb
epever
famhist
HRTNFC
lact
lactdur
mammon
mamno
menoag
menop
ocdur
ocever
parity
raceXX
screen;
run;

/*************************************************************************************************/
/**** add transit scores per participant ****/

proc import datafile="G:\Team\WHS DATA\whsstat\Project Datasets\
Mammography\Elkin\Methods and outputs\R\transit scores\transit_scores_IDNUM.csv"
     out=transitscores
     dbms=csv
     replace;
     getnames=yes;
run;

proc sort data=transitscores;
by IDNUM;
run;

/**** add travel time per participant *****/
proc import datafile="G:\Team\WHS DATA\whsstat\Project Datasets
\Mammography\Elkin\Data\Non-SAS data\idnum_drivetimes.csv"
     out=trvltime
```

```
     dbms=csv
     replace;
     getnames=yes;
run;

proc sort data=trvltime;
by IDNUM;
run;

/**** add mammo buffer  ****/
proc import datafile="G:\Team\WHS DATA\whsstat\Project Datasets\
Mammography\Elkin\Data\Non-SAS data\mammobuffer_IDNUM.csv"
     out=mambuffer
     dbms=csv
     replace;
     getnames=yes;
run;

proc sort data=mambuffer;
by IDNUM;
run;

/*add travel time and mammography buffer values*/
data controls10;
set controls9;
transit_score=.;
driveFac1=.;
driveFac2=.;
shortestdrive=.;
mambuffer_20=.;
mambuffer_10=.;
MAMFAC1=.;
MAMFAC2=.;
transit_MAM1=.;
transit_MAM2=.;
run;

proc sort data=controls10;
by IDNUM;
run;

data merge1;
MERGE controls10 (in=a) transitscores trvltime mambuffer mammo_transit;
by IDNUM;
if a=1;
run;

/**** add contextual data per 2000 census tract and per county ****/

/* IDNUMs with corresponding census 2000 tract*/
proc import datafile="G:\Team\WHS DATA\whsstat\Project Datasets\
Mammography\Elkin\Data\Non-SAS data\idnum_ctract00.csv"
     out=idnumctract00
     dbms=csv
```

```
     replace;
     getnames=yes;
run;

proc sort data=idnumctract00;
by IDNUM;
run;

/* IDNUMs with corresponding county*/
proc import datafile="G:\Team\WHS DATA\whsstat\Project Datasets\
Mammography\Elkin\Data\Non-SAS data\idnum_county.csv"
     out=idnumcounty
     dbms=csv
     replace;
     getnames=yes;
run;

proc sort data=idnumcounty;
by IDNUM;
run;

/*merge census tract and county into data*/
proc sort data=merge1;
by IDNUM;
run;

data merge2;
MERGE merge1 (in=a) idnumctract00 idnumcounty;
by IDNUM;
if a=1;
run;

/********** SES per census tract **************/
proc import datafile="G:\Team\WHS DATA\whsstat\Project Datasets\
Mammography\Elkin\Data\Non-SAS data\census00_SES.csv"
     out=census00_SES
     dbms=csv
     replace;
     getnames=yes;
run;

proc sort data=census00_SES;
by CTRACT00;
run;

/*********** RUCA code per census tract *****************/
proc import datafile="G:\Team\WHS DATA\whsstat\Project Datasets\
Mammography\Elkin\Data\Non-SAS data\RUCA_CTR00.csv"
     out=RUCA_CTR00
     dbms=csv
     replace;
     getnames=yes;
run;
```

```
proc sort data=RUCA_CTR00;
by CTRACT00;
run;

/*********** % of households w/o a vehicle per census tract ****/
proc import datafile="G:\Team\WHS DATA\whsstat\Project Datasets\
Mammography\Elkin\Data\Non-SAS data\noveh_ctract00.csv"
     out=NOVEH
     dbms=csv
     replace;
     getnames=yes;
run;

proc sort data=NOVEH;
by CTRACT00;
run;

/********** mammography capacity per county ***************/
proc import datafile="G:\Team\WHS DATA\whsstat\Project Datasets\
Mammography\Elkin\Data\Non-SAS data\mammo_cap_county.csv"
     out=mammo_cap_county
     dbms=csv
     replace;
     getnames=yes;
run;

proc sort data=mammo_cap_county;
by COUNTYFIPS REFYR;
run;

/************* uninsurance rate per county *******************/
proc import datafile="G:\Team\WHS DATA\whsstat\Project Datasets\
Mammography\Elkin\Data\Non-SAS data\SAHIE05_uninsured_per_county.csv"
     out=CNTY_UNINSURED
     dbms=csv
     replace;
     getnames=yes;
run;

proc sort data=CNTY_UNINSURED;
by COUNTYFIPS;
run;

/*merge contextual data into original individual-level data*/
proc sort data=merge2;
by CTRACT00;
run;

data merge3;
MERGE merge2 (in=a) census00_SES  RUCA_CTR00 NOVEH;
by CTRACT00;
if a=1;
run;
```

```
proc sort data=merge3;
by COUNTYFIPS REFYR;
run;

data merge4;
MERGE merge3 (in=a) mammo_cap_county;
by COUNTYFIPS REFYR;
if a=1;
run;

data merge5;
MERGE merge4 (in=a) CNTY_UNINSURED;
by COUNTYFIPS;
if a=1;
run;

/*Set mambuffer to 0 if IDNUM was actually geocoded and no mammo facility within buffer was found
(ArcGIS does not return the IDNUMS in whose buffer radius there were 0 facilities)*/
data controls11;
set merge5;
if mambuffer_20=. and CTRACT00 NE . then mambuffer_20=0;
if mambuffer_10=. and CTRACT00 NE . then mambuffer_10=0;
run;

/**** impute missing values ****/
/* imputation step 0 */
proc mi nimpute=0 data=controls11;
var phase refage refyr mamn educ white hrtever
BELOW_POVLN CT_FRAC_UNINS MED_INC_99 NO_HIGHSCHOOL GT_COLL RUCA_2CAT
MAMMO_CAP perc_noveh shortestdrive mambuffer_10 CTRACT00
famhist2 menop2 income HHCH HHAD;
run;


/* imputation step 1*/
proc mi nimpute=5 data=controls11 out=Pattern seed=100
/*round  = 1 1  1 1 1 1 1 1 1 1  1
min = 0 50 1995 2 0 0 0 0 1 0  1
max  = 5 74 2007 9 1 1 1 1 5 10 10;*/

round = 1 1 1 1 1 1 1 0.1 0.1 1 0.1
0.1 1 0.1 0.1 1 1 1 1 1 1 1 1
Min = 0 50 1995 0 2 0 0 0 7.1 8892 0.6 1
1 0 0 1 0 55001950100 0 0 1 0 1
Max = 5 74 2007 7 9 1 1 62.3 17.9 161292 62.6
85.3 2 9.9 67.2 8327 39 55141011700 1 1 5 10 10
;
* em maxiter=500;
mcmc impute = monotone;
var  phase refage refyr mamn educ white hrtever
BELOW_POVLN CT_FRAC_UNINS MED_INC_99 NO_HIGHSCHOOL GT_COLL RUCA_2CAT
MAMMO_CAP perc_noveh shortestdrive mambuffer_10 CTRACT00
famhist2 menop2 income HHCH HHAD;
run;
```

```
/* imputation step 2 */
proc sort data=Pattern;
by _imputation_ IDNUM;
run;

proc mi nimpute=1 data=Pattern seed=10 out=Imputed;
class  phase educ white hrtever RUCA_2CAT shortestdrive
mambuffer_10 CTRACT00 famhist2 menop2 income HHCH HHAD;
var  phase refage refyr mamn educ white hrtever BELOW_POVLN
CT_FRAC_UNINS MED_INC_99 NO_HIGHSCHOOL
GT_COLL RUCA_2CAT shortestdrive perc_noveh mambuffer_10
MAMMO_CAP CTRACT00
famhist2 menop2 income HHCH HHAD;
by _imputation_;

monotone logistic (EDUC=refage refyr mamn);
monotone discrim (white=refage refyr mamn);
monotone discrim (hrtever=refage refyr mamn);
monotone regression (BELOW_POVLN=refage refyr mamn EDUC white hrtever);
monotone regression (CT_FRAC_UNINS=refage refyr mamn EDUC white hrtever
BELOW_POVLN);
monotone regression (MED_INC_99=refage refyr mamn EDUC white hrtever
BELOW_POVLN CT_FRAC_UNINS);
monotone regression (NO_HIGHSCHOOL=refage refyr mamn EDUC white hrtever
BELOW_POVLN CT_FRAC_UNINS MED_INC_99);
monotone regression (GT_COLL=refage refyr mamn EDUC white hrtever BELOW_POVLN
CT_FRAC_UNINS MED_INC_99 NO_HIGHSCHOOL);
monotone logistic   (RUCA_2CAT=refage refyr mamn BELOW_POVLN CT_FRAC_UNINS
 MED_INC_99 NO_HIGHSCHOOL GT_COLL);
monotone logistic  (shortestdrive=refage refyr mamn EDUC white hrtever BELOW_POVLN
CT_FRAC_UNINS MED_INC_99 NO_HIGHSCHOOL GT_COLL);
monotone regression (perc_noveh=refage refyr mamn EDUC white hrtever BELOW_POVLN
CT_FRAC_UNINS MED_INC_99 NO_HIGHSCHOOL GT_COLL RUCA_2CAT);
monotone logistic  (mambuffer_10=refage refyr mamn EDUC white hrtever BELOW_POVLN
CT_FRAC_UNINS MED_INC_99 NO_HIGHSCHOOL GT_COLL perc_noveh);
monotone regression (MAMMO_CAP=refage refyr mamn EDUC white hrtever BELOW_POVLN
CT_FRAC_UNINS MED_INC_99 NO_HIGHSCHOOL GT_COLL RUCA_2CAT);
monotone logistic   (CTRACT00=refage refyr mamn BELOW_POVLN CT_FRAC_UNINS
 MED_INC_99 NO_HIGHSCHOOL GT_COLL RUCA_2CAT MAMMO_CAP perc_noveh);
monotone discrim (famhist2=refage refyr mamn BELOW_POVLN CT_FRAC_UNINS
MED_INC_99 NO_HIGHSCHOOL GT_COLL MAMMO_CAP perc_noveh);
monotone discrim (menop2=refage refyr mamn BELOW_POVLN CT_FRAC_UNINS
MED_INC_99 NO_HIGHSCHOOL GT_COLL MAMMO_CAP perc_noveh);
monotone logistic (income=refage refyr mamn BELOW_POVLN CT_FRAC_UNINS
MED_INC_99 NO_HIGHSCHOOL GT_COLL MAMMO_CAP perc_noveh);
monotone logistic (HHCH=refage refyr mamn BELOW_POVLN CT_FRAC_UNINS
MED_INC_99 NO_HIGHSCHOOL GT_COLL MAMMO_CAP perc_noveh);
monotone logistic (HHAD=refage refyr mamn BELOW_POVLN CT_FRAC_UNINS
MED_INC_99 NO_HIGHSCHOOL GT_COLL MAMMO_CAP perc_noveh);
run;

/*transform variables*/
```

```
data tmp1;
set Imputed;
mamfreq=.;

/*numericalmammography frequency*/
if mamn=0 then mamfreq=0;
else if mamn=1 then mamfreq=1/5;
else if mamn=2 then mamfreq=2/5;
else if mamn=3 then mamfreq=3/5;
else if mamn=4 then mamfreq=4/5;
else if mamn=5 then mamfreq=5/5;
else if mamn=6 then mamfreq=0.5; /*every other year*/
else if mamn=7 then mamfreq=1.5; /*more than 5 mammograms in past 5 years*/
run;

/*mammography use categories*/
data tmp2;
set tmp1;

PERSHH= HHCH + HHAD; /*persons in household*/

if mamfreq<0.5 then mamcat=1;
else if (mamfreq>=0.5 and mamfreq<1) then mamcat=2;
else if (mamfreq>=1 and mamfreq<2) then mamcat=3;
else if mamfreq>=2 then mamcat=4;

/*create numerical variable with years of education from educ (test code)*/
eduyrs = .;
if educ = 1 then eduyrs =0;
else if educ = 2 then eduyrs = 4;
else if educ = 3 then eduyrs = 8;
else if educ = 4 then eduyrs = 10;
else if educ = 5 then eduyrs = 12;
else if educ = 6 then eduyrs = 14;
else if educ = 7 then eduyrs = 16;
else if educ = 8 then eduyrs = 18;
else if educ = 9 then eduyrs = 20;

/*create numerical income variable  (test code)*/
incnum = .;
if income=1 then incnum = 8000;
if income=2 then incnum = 22500;
if income=3 then incnum = 40000;
if income=4 then incnum = 75000;
if income=5 then incnum = 150000;
run;

/*permanent dataset*/
data pjewett.step1_wi_controls_mammo;
set tmp3;
run;


/*Data analysis:*/
```

```
option ls=76 ps=500 mergenoby=ERROR ;
libname pjewett 'G:\Team\WHS DATA\whsstat\Project Datasets\
Mammography\Elkin\Data\WI Controls';

/*read dataset*/
data mammo;
set pjewett.step1_wi_controls_mammo;
run;

/*rescaling*/
data mammo2;
set mammo;
trvltm10min = shortestdrive/600; /*rescale by 10 minutes*/
trvl10 = shortestdrive/600; /*rescale by 10 minutes*/
incnum_10K = incnum/10000; /*per $10,000*/
inc10K = incnum/10000; /*per $10,000*/
eduyrs_4=eduyrs/4; /*per 4 years*/
edu_4=eduyrs/4; /*per 4 years*/
age_5=refage/5; /*per 5 years of age*/
medianinc_10K=MED_INC_99/10000; /*per $10,000*/
belpov_5=BELOW_POVLN/5; /*per 5%*/
perc_noveh5=perc_noveh/5; /* per 5%*/
NO_HIGHSCHOOL5=NO_HIGHSCHOOL/5; /*per 5%*/
GT_COLL5=GT_COLL/5; /*per 5%*/
run;

data mammo4;
set mammo2;
/*create binary ruca variable */
rucabin=.;
if RUCA_2CAT = 1 then rucabin=1; *urban;
else if RUCA_2CAT = 2 then rucabin=0; *rural;

/*cubic splines for travel time and mammo density*/
cub1 = max(0,(trvl10-0.1)**3);
cub2 = max(0,(trvl10-0.47)**3);
cub3 = max(0,(trvl10-0.80)**3);
cub4 = max(0,(trvl10-1.41)**3);
*cub5 = max(0,(trvl10-13.87)**3);
cub5 = max(0,(trvl10-4)**3);
/*d1 = (cub1-cub5)/(13.87-0.1);
d2 = (cub2-cub5)/(13.87-0.47);
d3 = (cub3-cub5)/(13.87-0.80);
d4 = (cub4-cub5)/(13.87-1.41);*/
d1 = (cub1-cub5)/(4-0.1);
d2 = (cub2-cub5)/(4-0.47);
d3 = (cub3-cub5)/(4-0.80);
d4 = (cub4-cub5)/(4-1.41);
N1 = d1-d4;
N2 = d2-d4;
N3 = d3-d4;

dcub1 = max(0,(mambuffer_10-0)**3);
dcub2 = max(0,(mambuffer_10-2)**3);
```

```
dcub3 = max(0,(mambuffer_10-6)**3);
dcub4 = max(0,(mambuffer_10-10)**3);
dd1 = (dcub1-dcub4)/(10-0);
dd2 = (dcub2-dcub4)/(10-2);
dd3 = (dcub3-dcub4)/(10-6);
dN1 = dd1-dd3;
dN2 = dd2-dd3;
run;

proc sort data=mammo4;
by _imputation_ idnum;
run;

/************************ Multinomial Regression ***********************************/
/*Proportional odds model with cubic splines*/
proc glimmix data=mammo4 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
ods exclude solutionr;
by _imputation_;
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh;
test1: test N1, N2, N3 / mult; /*travel time nonlinear*/
test2: test dN1, dN2 / mult;  /*mambuffer nonlinear*/
test3: test trvl10 , N1, N2, N3 / mult; /*travel time overall*/
test4: test mambuffer_10, dN1, dN2 / mult; /*mambuffer overall*/
title 'model 1: minimla model, test for nonlinear effects';
run;

/*few confounders, only linear effecs*/
proc glimmix data=mammo4 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 mambuffer_10
MAMMO_CAP eduyrs_4 incnum_10K white pershh
/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
ods exclude solutionr;
by _imputation_;
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 mambuffer_10
MAMMO_CAP eduyrs_4 incnum_10K white pershh;
title 'model 1b: few confounders;  only linear effects';
```

```
run;

/* model 2: include more confounders */
proc glimmix data=mammo4 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
ods exclude solutionr;
by _imputation_;
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh;
test1: test N1, N2, N3 / mult; /*travel time nonlinear*/
test2: test dN1, dN2 / mult;  /*mambuffer nonlinear*/
test3: test trvl10 , N1, N2, N3 / mult; /*travel time overall*/
test4: test mambuffer_10, dN1, dN2 / mult; /*mambuffer overall*/
title 'model 2: more confounders; including nonlinear effects';
run;

/* model 3: only linear effects */
proc glimmix data=mammo4 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 mambuffer_10
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
ods exclude solutionr;
by _imputation_;
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 mambuffer_10
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5;
title 'model 3: more confounders;  only linear effects';
run;


/*stratify by urban/rural*/
proc glimmix data=mammo4 Method=LAPLACE;
```

```
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
ods exclude solutionr;
by _imputation_;
where rucabin=0;
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh;
test1: test N1, N2, N3 / mult; /*travel time nonlinear*/
test2: test dN1, dN2 / mult;  /*mambuffer nonlinear*/
test3: test trvl10 , N1, N2, N3 / mult; /*travel time overall*/
test4: test mambuffer_10, dN1, dN2 / mult; /*mambuffer overall*/
title 'model 3a: more confounders; rural';
run;


/*urban*/
proc glimmix data=mammo4 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
ods exclude solutionr;
by _imputation_;
where rucabin=1;
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh;
test1: test N1, N2, N3 / mult; /*travel time nonlinear*/
test2: test dN1, dN2 / mult;  /*mambuffer nonlinear*/
test3: test trvl10 , N1, N2, N3 / mult; /*travel time overall*/
test4: test mambuffer_10, dN1, dN2 / mult; /*mambuffer overall*/
title 'model 3b: more confounders; urban';
run;

/*create interaction terms*/
data mammo5;
set mammo4;
```

```
trinter1=rucabin*trvl10;
trinter2=rucabin*N1;
trinter3=rucabin*N2;
trinter4=rucabin*N3;
maminter1=rucabin*mambuffer_10;
maminter2=rucabin*dN1;
maminter3=rucabin*dN2;

incinter1=incnum_10K*trvl10;
incinter2=incnum_10K*N1;
incinter3=incnum_10K*N2;
incinter4=incnum_10K*N3;

eduinter1=eduyrs_4*trvl10;
eduinter2=eduyrs_4*N1;
eduinter3=eduyrs_4*N2;
eduinter4=eduyrs_4*N3;

incinter5=incnum_10K*mambuffer_10;
incinter6=incnum_10K*dN1;
incinter7=incnum_10K*dN2;

eduinter5=eduyrs_4*mambuffer_10;
eduinter6=eduyrs_4*dN1;
eduinter7=eduyrs_4*dN2;

run;

proc sort data=mammo5;
by _imputation_ idnum;
run;

/*test for interaction by urban/rural*/
proc glimmix data=mammo5 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5 GT_COLL5 perc_noveh5
rucabin
trinter1 trinter2 trinter3 trinter4
maminter1 maminter2 maminter3
/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
ods exclude solutionr;
by _imputation_;
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
```

```
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5
rucabin
trinter1 trinter2 trinter3 trinter4
maminter1 maminter2 maminter3;
trvlrucainteract: test trinter1, trinter2, trinter3, trinter4 / mult; /*travel time ruca interaction*/
mamrucainteract: test maminter1, maminter2, maminter3 / mult; /*mambuffer ruca interaction*/
title 'model 4: urban/rural interaction';
run;


/*test for interaction of travel time and mambuffer by income or education*/
proc glimmix data=mammo5 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5
incinter1 incinter2 incinter3 incinter4 incinter5 incinter6 incinter7
eduinter1 eduinter2 eduinter3 eduinter4 eduinter5 eduinter6 eduinter7
/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
ods exclude solutionr;
by _imputation_;
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5 GT_COLL5 perc_noveh5
incinter1 incinter2 incinter3 incinter4 incinter5 incinter6 incinter7
eduinter1 eduinter2 eduinter3 eduinter4 eduinter5 eduinter6 eduinter7;
test1: test incinter1, incinter2, incinter3, incinter4 / mult; /*travel time income interaction*/
test2: test eduinter1, eduinter2, eduinter3, eduinter4 / mult;  /*travel time edu interaction*/
test3: test incinter5, incinter6, incinter7 / mult; /*mambuffer income interaction*/
test4: test eduinter5, eduinter6, eduinter7 / mult;  /*mAMBUFFER edu interaction*/
title 'model 5: SES interaction';
run;

/*test for interaction only by education*/
proc glimmix data=mammo5 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5
eduinter1 eduinter2 eduinter3 eduinter4 eduinter5 eduinter6 eduinter7
/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
```

```
ods exclude solutionr;
by _imputation_;
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5 GT_COLL5 perc_noveh5
eduinter1 eduinter2 eduinter3 eduinter4 eduinter5 eduinter6 eduinter7;
test2: test eduinter1, eduinter2, eduinter3, eduinter4 / mult;  /*travel time edu interaction*/
test4: test eduinter5, eduinter6, eduinter7 / mult;  /*mAMBUFFER edu interaction*/
title 'model 5: education interaction';
run;

/*Final model: more confounders, nonlinear terms, no interactions, stratif by urban/rural*/
proc glimmix data=mammo5 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5
/ dist=multinomial link=cumlogit;
Random intercept/ SUB=CTRACT00;
where rucabin=0 and _imputation_=1; /*rural*/
run;


/*Final model: more confounders, nonlinear terms, no interactions, stratif by urban/rural*/
proc glimmix data=mammo5 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5
/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
ods exclude solutionr;
by _imputation_;
where rucabin=0; /*rural*/
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

/*Calculate ORs for rural*/
proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5 GT_COLL5 perc_noveh5;
```

```
/*OR trvlt=20 vs 10 mins*/
test20vs10: test 1*trvl10 + 1.4925*N1 + 0.89314*N2 + 0.45820*N3;

/*OR  trvlt=30 vs 20 mins*/
test30vs20: test 1*trvl10 + 3.02217*N1 + 2.10030*N2 + 1.31480*N3 ;
/*OR trvlt=40 vs 20 mins*/
test40vs20: test 2*trvl10 + 6.82248*N1  + 4.81749*N2 + 3.07120*N3;
/*OR 1 vs 0 facilities*/
test1vs0: test 1*mambuffer_10 + 0.1*dN1 + 0*dN2;
/*OR 2 vs 1 facilities*/
test2vs1: test 1*mambuffer_10 + 0.7*dN1 + 0*dN2;
/*OR 3 vs 2 facilities*/
test3vs2: test 1*mambuffer_10 + 1.9*dN1 + 0.125*dN2;
/*OR 4 vs 3 facilities*/
test4vs3: test 1*mambuffer_10 + 3.7*dN1 + 0.875*dN2;
title 'model 1: final model rural, ORs';
run;

/*urban*/
proc glimmix data=mammo5 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5
/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
ods exclude solutionr;
by _imputation_;
where rucabin=1; /*urban*/
*where _imputation_=1; /*urban*/
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

/*Calculate ORs for urban*/
proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
/*modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh;*/
modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5;

/*OR trvlt=20 vs 10 mins*/
test20vs10: test 1*trvl10 + 1.4925*N1 + 0.89314*N2 + 0.45820*N3;
/*OR  trvlt=30 vs 20 mins*/
test30vs20: test 1*trvl10 + 3.02217*N1 + 2.10030*N2 + 1.31480*N3 ;
/*OR trvlt=40 vs 20 mins*/
test40vs20: test 2*trvl10 + 6.82248*N1  + 4.81749*N2 + 3.07120*N3;
/*OR 1 vs 0 facilities*/
```

```
test1vs0: test 1*mambuffer_10 + 0.1*dN1 + 0*dN2;
/*OR 2 vs 1 facilities*/
test2vs1: test 1*mambuffer_10 + 0.7*dN1 + 0*dN2;
/*OR 3 vs 2 facilities*/
test3vs2: test 1*mambuffer_10 + 1.9*dN1 + 0.125*dN2;
/*OR 4 vs 3 facilities*/
test4vs3: test 1*mambuffer_10 + 3.7*dN1 + 0.875*dN2;
title 'model 1: final model urban, ORs';
run;


/*Final model; fewer confounders*/
/*Calculate ORs for rural*/
proc glimmix data=mammo5 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
ods exclude solutionr;
by _imputation_;
where rucabin=0; /*rural*/
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh;

/*OR trvlt=20 vs 10 mins*/
test20vs10: test 1*trvl10 + 1.4925*N1 + 0.89314*N2 + 0.45820*N3;
/*OR  trvlt=30 vs 20 mins*/
test30vs20: test 1*trvl10 + 3.02217*N1 + 2.10030*N2 + 1.31480*N3 ;
/*OR trvlt=40 vs 20 mins*/
test40vs20: test 2*trvl10 + 6.82248*N1  + 4.81749*N2 + 3.07120*N3;
/*OR 1 vs 0 facilities*/
test1vs0: test 1*mambuffer_10 + 0.1*dN1 + 0*dN2;
/*OR 2 vs 1 facilities*/
test2vs1: test 1*mambuffer_10 + 0.7*dN1 + 0*dN2;
/*OR 3 vs 2 facilities*/
test3vs2: test 1*mambuffer_10 + 1.9*dN1 + 0.125*dN2;
/*OR 4 vs 3 facilities*/
test4vs3: test 1*mambuffer_10 + 3.7*dN1 + 0.875*dN2;
title 'model 1: final minimal model rural, ORs';
run;

/*urban*/
proc glimmix data=mammo5 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model mamfreq (descending) = trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
/ dist=multinomial link=cumlogit covb solution;
Random intercept/ SUB=CTRACT00;
```

```
ods exclude solutionr;
by _imputation_;
where rucabin=1; /*urban*/
*where _imputation_=1; /*urban*/
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

/*Calculate ORs for urban*/
proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh;

/*OR trvlt=20 vs 10 mins*/
test20vs10: test 1*trvl10 + 1.4925*N1 + 0.89314*N2 + 0.45820*N3;
/*OR  trvlt=30 vs 20 mins*/
test30vs20: test 1*trvl10 + 3.02217*N1 + 2.10030*N2 + 1.31480*N3 ;
/*OR trvlt=40 vs 20 mins*/
test40vs20: test 2*trvl10 + 6.82248*N1  + 4.81749*N2 + 3.07120*N3;
/*OR 1 vs 0 facilities*/
test1vs0: test 1*mambuffer_10 + 0.1*dN1 + 0*dN2;
/*OR 2 vs 1 facilities*/
test2vs1: test 1*mambuffer_10 + 0.7*dN1 + 0*dN2;
/*OR 3 vs 2 facilities*/
test3vs2: test 1*mambuffer_10 + 1.9*dN1 + 0.125*dN2;
/*OR 4 vs 3 facilities*/
test4vs3: test 1*mambuffer_10 + 3.7*dN1 + 0.875*dN2;
title 'model 1: final minimal model urban, ORs';
run;

/*Startification by SES*/
/* include above: where-conditions:
where educ<7; /*no college degree*/
where educ>=7; /*at least college degree*/

/*Calculate ORs where per capita income < 11250 (Q1)*/
data mammo6;
set mammo5;
/*per capita income:*/
incpercap=incnum/PERSHH;
run;

/* rerun above models stratifying by:*/
where incpercap<=11250; /*low incomes*/
where incpercap>11250; /*higher incomes*/

/*Run final model also as logistic regression:
Comparing non-adherent users with adherent users
Adherent users: at least 1 mammogram in 2 years
Non-adherent users: less than 1 mammogram in 2 years
*/

data mammo7;
set mammo6;
```

```
adherent=.;
if (mamfreq NE . and mamfreq < 0.5) then adherent=0;
else adherent=1;
run;

proc sort data=mammo7;
by _imputation_ idnum;
run;

/*Logistic regression for rural */
proc glimmix data=mammo7 Method=LAPLACE;
CLASS IDNUM CTRACT00;
model adherent (descending) =  trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5
/ dist=binomial link=logit covb solution;
by _imputation_;
where rucabin=0; /*rural*/
ods output ParameterEstimates=parms CovB=cov solutionr=rand_parms;
run;

/*Calculate ORs for rural*/
proc mianalyze parms=parms
covb(effectvar=rowcol)=cov;
modeleffects trvl10 N1 N2 N3 mambuffer_10 dN1 dN2
MAMMO_CAP eduyrs_4 incnum_10K white pershh
age_5 famhist2 belpov_5
medianinc_10K CT_FRAC_UNINS NO_HIGHSCHOOL5
GT_COLL5 perc_noveh5;
/*OR trvlt=20 vs 10 mins*/
test20vs10: test 1*trvl10 + 1.4925*N1 + 0.89314*N2 + 0.45820*N3;
/*OR  trvlt=30 vs 20 mins*/
test30vs20: test 1*trvl10 + 3.02217*N1 + 2.10030*N2 + 1.31480*N3 ;
/*OR trvlt=40 vs 20 mins*/
test40vs20: test 2*trvl10 + 6.82248*N1  + 4.81749*N2 + 3.07120*N3;
/*OR 1 vs 0 facilities*/
test1vs0: test 1*mambuffer_10 + 0.1*dN1 + 0*dN2;
/*OR 2 vs 1 facilities*/
test2vs1: test 1*mambuffer_10 + 0.7*dN1 + 0*dN2;
/*OR 3 vs 2 facilities*/
test3vs2: test 1*mambuffer_10 + 1.9*dN1 + 0.125*dN2;
/*OR 4 vs 3 facilities*/
test4vs3: test 1*mambuffer_10 + 3.7*dN1 + 0.875*dN2;
title 'logistic regression: rural';
run;

/*Logistic regression for urban as above*/
```

F.0.5 Code used for analysis in chapter 5 (long-term mammography utilization among DCIS survivors)

```
/*Read in data*/
option ls=76 ps=500 mergenoby=ERROR ;

/** STEP 1a) reading in recurrence dataset */

/*reading in recurrence dataset*/
libname z 'G:\Team\WHS DATA\whsstat\BCIS_Recontact\Data\RECURRENCE\2014' ;

data recurrence;
set z.recurrence2014_v1;
run;

data recurrence2;
set recurrence;

/*need to exclude the following IDNUMS from BCIS 1 and BCIS2 because ineligible:*/
ineligible=0;
if idnum in (542845, 542852, 542885, 542915, 542946, 543447, 543492,
545074, 545522, 545672, 546120, 546922, 547349, 550508, 550668, 565307,
 566416, 566589, 566629, 568571, 568693, 568956, 570763, 570795, 570828,
 570840, 570894, 570910, 571019, 571157, 571178, 572139, 572238, 572731,
 574924, 576461, 577774, 579226, 580209, 581450, 581560) then ineligible=1 ;

/*IDNUM 570701 with missing recurrence status; should be recurrence=0 (see email from John to Trish on M
if idnum=570701 then recurrence=0;
run;

data recurrence3;
set recurrence2;
/*exclude ineligible subjects and subjects with unknown recurrence status*/
if ineligible=0;
if recurrence NE .;
run;

/**** BCIS1 *****/
libname ph3 'G:\Team\WHS DATA\whsstat\Phase3\Data' ;
filename ph3conf 'G:\Team\WHS DATA\whsstat\Phase3\Confound' ;

data bcis1 ;
  set ph3.BCIS_wi;

  intyy=year(_intdate) ;
  intmm=month(_intdate) ;
  intyymm=intyy*100+intmm ;

  refyy=year(_refdate) ;
  refmm=month(_refdate) ;
  refyymm=refyy*100+refmm ;

  if status in (1,2) then do ;
  if r_histo in (8500, 85003, 85002) then histgp=1 ;  ** ductal ;
```

```
  if r_histo in (8520, 85203, 85202) then histgp=2 ;  ** lobular ;
  if histgp=. then histgp=3 ;
 end ;

  agerf=(_refdate-_dob)/365.25;
  if 0 lt agerf lt 40 then agegp=1 ;
  if 40 le agerf lt 45 then agegp=2 ;
  if 45 le agerf lt 50 then agegp=3 ;
  if 50 le agerf lt 55 then agegp=4 ;
  if 55 le agerf lt 60 then agegp=5 ;
  if 60 le agerf lt 65 then agegp=6 ;
  if 65 le agerf lt 70 then agegp=7 ;
  if agerf ge 70 then agegp=8 ;

  agegp1=agegp=1 ;  agegp2=agegp=2 ;
  agegp3=agegp=3 ;  agegp4=agegp=4 ;
  agegp5=agegp=5 ;  agegp6=agegp=6 ;
  agegp7=agegp=7 ;  agegp8=agegp=8 ;

  %include ph3conf(psycho1a.sas) ;
  %include ph3conf(nsaid2a.sas) ;
  %include ph3conf(alch2.sas) ;
  %include ph3conf(bmi2.sas) ;
  %include ph3conf(menop2X.sas) ;
  %include ph3conf(smoke.sas) ;
  %include ph3conf(famhist2.sas) ; /*adds famhist*/
  %include ph3conf(parity3.sas) ; /*adds parity and afb= age at first birth*/

/*need to exclude the following IDNUMS from BCIS 1 and BCIS2 because ineligible:*/
  ineligible=0;
if idnum in (542845, 542852, 542885, 542915, 542946, 543447, 543492, 545074, 545522, 545672,
             546120, 546922, 547349, 550508, 550668, 565307, 566416, 566589, 566629, 568571,
             568693, 568956, 570763, 570795, 570828, 570840, 570894, 570910, 571019,
             571157, 571178, 572139, 572238, 572731, 574924, 576461, 577774, 579226,
             580209, 581450, 581560) then ineligible=1 ;
  run ;


/**** BCIS2 *****/
libname b2 'G:\Team\WHS DATA\whsstat\BCIS2\Data' ;
filename b2conf 'G:\Team\WHS DATA\whsstat\BCIS2\Confound' ;
filename ph3conf 'G:\Team\WHS DATA\whsstat\Phase3\Confound' ;
filename ph4conf 'G:\Team\WHS DATA\whsstat\Phase4\Confound' ;
data bcis2 ;
  set b2.BCIS2_v1a ;

  * NOTE 1:  if analyzing case-control data, drop out cases 70+
  ** if rfage ge 70 then delete ;
  * NOTE 2:  if analyzing case-control data, consider dropping out
              cases without Driver's License ;
  ** if eligbad2=50 then delete ;

 /*include to create smoking and alcohol use variables*/
  %include ph3conf(famhist2.sas) ;
```

```
  %include b2conf(alch2_b2.sas) ;
  %include b2conf(bmi.sas) ;
  %include ph3conf(smoke.sas) ;
  %include ph3conf(parity3.sas) ; /*adds parity and afb=age at first birth*/

   intyy=year(_intdate) ;
 ** menopause confound code ;
  if pdrf1=1 and pdrf1hr=1 then pdrf=1 ;
  if pdrf1=1 and pdrf1hr ne 1 then do ; pdrf=3 ; pdnt=. ; pdunt=. ; end ;
  if pdrf1=2 then pdrf=4 ;
  if pdrf1=1 and meno=3 then pdrf=5 ;
*  age=round((_intdate-_dob)/365.25) ;
  %include ph3conf(menop2x) ;
  if pdrf1=2 and meno=2 then do ; menop=3 ; menoag=. ; end ;
  if menop=. then menop=3 ;

    format _refdate bcayr bcayr2 bcayr3 dob _dob
        otcayr1 otcayr2 otcayr3 otcayr4 mmddyy10. ;

/*need to exclude the following IDNUMS from BCIS 1 and BCIS2 because ineligible:*/
  ineligible=0;
if idnum in (542845, 542852, 542885, 542915, 542946, 543447, 543492, 545074, 545522, 545672,
             546120, 546922, 547349, 550508, 550668, 565307, 566416, 566589, 566629, 568571,
             568693, 568956, 570763, 570795, 570828, 570840, 570894, 570910, 571019,
             571157, 571178, 572139, 572238, 572731, 574924, 576461, 577774, 579226,
             580209, 581450, 581560) then ineligible=1 ;
  run ;

/*** STEP 2: Keep only relevant variables and exclude ineligible subjects */
/**** BCIS1 *****/
  data bcis1_2;
  set bcis1;
  keep
  idnum /*ID*/
  status /*vase or control at baseline*/
  _REFDATE   /*date of first diagnosis*/
   INTDATE /*baseline interview date*/
  _VERSION /*version of interview questionnaire*/
  ineligible /*ineligible subjects*/
  R_HISTO /*cancer cell 5 digit histology*/
  /*some demographics*/
  agerf /*age at diagnosis*/
  racef /*race father*/
  racem /*race mother*/
  HR /*any use of hormones (hrt)*/
  famhist /*family history --> absent (1), present (2), unknown (3)  */
  INCOME /*annual household income (ref); first questionnaire versions*/
  INCRF1 /*annual household income (ref-1); last questionnaire versions*/
  EDUC /*highest school degree completed*/
  hhch /*number of children per household*/
  hhad   /*number of adults per household*/
  mamn; /*number of mammograms in past 5 years*/

/*exclude ineligible subjects*/
```

```
  if ineligible=0;
  run;

/*phase; merge income baseline variables into one variable*/
data bcis1_3;
set bcis1_2;
incmbsl=.;

phase ="bcis1";

if (INCOME=. and INCRF1 NE .) then incmbsl=INCRF1;
else if (INCRF1=. and INCOME NE .) then incmbsl=INCOME;
run;

/**** BCIS2 *****/
  data bcis2_2;
  set bcis2;
  keep
  idnum /*ID*/
  status /*vase or control at baseline*/
  _REFDATE   /*date of first diagnosis*/
  INTDATE /*baseline interview date*/
  _VERSION /*version of interview questionnaire*/
  ineligible /*ineligible subjects*/

  R_HISTO /*cancer cell 5 digit histology*/

  /*some demographics*/
  agerf /*age at diagnosis*/
  racedes
  HR /*any use of hormones (hrt)*/
  famhist /*family history --> absent (1), present (2), unknown (3)  */

  INCRF /*annual household income (ref-1)*/
  EDUC /*highest school degree completed*/
  hhch /*number of children per household*/
  hhad /*number of adults per household*/

  mamn; /*number of mammograms in past 5 years*/

/*exclude ineligible subjects*/
if ineligible=0;
  run;


  data bcis2_3;
  set bcis2_2;
  phase ="bcis2";
  rename
  INCRF =incmbsl
  INTDATE=INTDATE_bsl;
  run;

  data bcis1_4;
```

```
  set bcis1_3;
  rename
  INTDATE=INTDATE_bsl;
  run;

data bcisall;
set bcis1_4 bcis2_3;
run;

/*** read in recontact1 and recontact2 datasets */
libname here 'G:\Team\WHS DATA\whsstat\BCIS_Recontact\Data' ;

data recon1;
  set here.bcis_recon1_raw;
  run;

  proc contents data=recon1;run;

data recon2 ;
  set here.bcis_recon2_raw;
  run;

/* Keep only relevant variables*/
data recon1_2;
set recon1;
keep
idnum
_INTDATE /*date of interview*/
bca2
bcayr  /*month/year of first diagnosis*/
mamsdx
mamsdxn /*number of mammograms since most recent diagnosis*/
mamrecm /*month od most recent mammogram*/
mamrecy; /*year of most recent mammogram*/
run;

/*if mamsdx=2, set mamsdxn to 0 (no mammograms)*/
data recon1_2b;
set recon1_2;
if mamsdx=2 then mamsdxn=0;
run;

proc sort data=recon1_2b;
by IDNUM _INTDATE;
run;

data recon1_3;
set recon1_2b;
by IDNUM _INTDATE;
if first.IDNUM;
run;

/**** recontact 2 *****/
data recon2_2;
```

```
set recon2;
keep
idnum
_INTDATE /*date of interview*/
mamsdx
mamsdxn /*number of mammograms since last interview*/
mamrecm /*month od most recent mammogram*/
mamrecy; /*year of most recent mammogram*/
run;

/*if mamsdx=2 then set mamsxn to 0 (no mammograms)*/
data recon2_2b;
set recon2_2;
if mamsdx=2 then mamsdxn=0;
run;

data recon2_3;
set recon2_2b;
rename
_INTDATE=_INTDATE2
mamsdxn=mamsdxn2
mamrecm=mamrecm2
mamrecy=mamrecy2;
run;

/*set recontact 1 and 2 flags*/
data recon1_4;
set recon1_3;
rec1=1;
run;

data recon2_4;
 set recon2_3;
 rec2=1;
 run;

/* add survey data*/
libname new 'W:\WHS QDS\BCIS3 Data\SAS export' ;
libname library 'W:\WHS QDS\BCIS3 Data\SAS export\Library' ;
filename b 'W:\WHS QDS\BCIS3 Data\SAS export' ;
options nofmterr validvarname=upcase fmtsearch=(library.bcis3x2);

data survey1 ;
  *length comm $300. ; informat comm $300. ;
  set new.bcis3_survey1_v1  ;

  %include b(bcis3x2.sas) ;
  format dob mmddyy10. ;
  est_survey_age=2010-doby ;
  walkX=index(patpyoa, "walk") ;
  if cigwork =0 then cigworkX=0 ; else if cigwork  > 0 then cigworkX=1 ;
  if cigsoc=0 then cigsocX=0 ; else if cigsoc > 0 then cigsocX=1 ;

  if wght18 < 0 then wght18missing=1 ;
```

```
    else wght18missing=0 ;
  run ;

proc sort data=survey1 ; by idnum ; run ;

libname bcis2013 'W:\WHS QDS\BCIS3 Data\SAS export 2013' ;
libname lib2013 'W:\WHS QDS\BCIS3 Data\SAS export 2013\Library' ;
filename b2013 'W:\WHS QDS\BCIS3 Data\SAS export 2013' ;
options nofmterr validvarname=upcase fmtsearch=(lib2013.formats);

data survey2013 ;
  *length comm $300. ; informat comm $300. ;
  set bcis2013.bcis3_survey2_v1  ;

  %include b2013("bcis format 2013.sas") ;
  format dob mmddyy10. ;
  run ;

proc sort data=survey2013 ; by idnum ; run ;


data survey1_2;
set survey1;
keep
IDNUM
MAMNUM /*how many mams in past 5 years:
1=0 times, 2= 1 time; 3=2 times; 4=3 times; 5=4 times; 6=5 times; 7= more than 5 times*/
/*MAMRECM*/
MAMRECMO /*month of most recent mammogram*/
/*MAMRECY*/
MAMRECYR; /*year of most recent mammogram*/
run;

data survey1_3;
set survey1_2;
rec3=1;
run;
proc sort data=survey1_3; by idnum ; run ;

data survey2013_2;
set survey2013;
keep
IDNUM
MAMMO /* how many mams in past 4 years:
1=0 times, 2= 1 time; 3=2 times; 4=3 times; 5=4 times; 6= more than 4 times*/
MAMRECM4 /*month of most recent mammogram*/
MAMRECY4 /*year of most recent mammogram*/
MRI4 /*how many MRIs in past 4 years: 1=0 times, 2= 1 time; 3=2 times; 4=3 times; 5= more than 4 times*/
RECMRIM /*month of most recent MRI*/
RECMRIY  /*year of most recent MRI*/
RECMRI; /*date of most recent MRY*/
run;

data survey2013_3;
```

```
set survey2013_2;
rec4=1;
run;
proc sort data=survey2013_3; by idnum ; run ;

/* add interview dates to survey*/
libname new 'W:\WHS QDS\BCIS3 Data\SAS export' ;

data survey1_return_date ;
  set new.bcis_v2010_survey_return_date ;
  run ;
proc sort data=survey1_return_date ; by idnum ; run ;

libname bcis2013 'W:\WHS QDS\BCIS3 Data\SAS export 2013' ;

data survey2013_return_date ;
  set bcis2013.bcis_v2013_survey_return_date ;
  run ;
proc sort data=survey2013_return_date; by idnum ; run ;

data survey1_4;
MERGE survey1_3(in=a)survey1_return_date ;
by IDNUM;
if a=1;
run;

data survey2013_4;
MERGE survey2013_3(in=a)survey2013_return_date;
by IDNUM;
if a=1;
run;

/* end add survey data*/

/*merge all follow up points*/
proc sort data=recon1_4; by idnum ; run ;
proc sort data=recon2_4; by idnum ; run ;
proc sort data=survey1_4; by idnum ; run ;
proc sort data=survey2013_4; by idnum ; run ;

data reconfinal;
MERGE recon1_4 recon2_4 survey1_4 survey2013_4;
by IDNUM;
run;

/**** STEP 4: merge recurrence and BCIS datasets */

proc sort data=recurrence3; by IDNUM; run;
proc sort data=bcisall; by IDNUM; run;
proc sort data=reconfinal; by IDNUM; run;

data recurr_merge;
MERGE bcisall(in=a) recurrence3 reconfinal;
by IDNUM;
```

```
if a=1;
run;

/*drop unnnecessary variables*/
data recurr_drop;
set recurr_merge;
drop
BC_ER_STATUSX
DUKE_ER_STATUS
INCOME
INCRF1
INELIGIBLE
LATERALITYX
PATH_CONTRADICTS_SELFREPORT
RECURRENCE2_CONFIRMED
RECURRENCE2_ER_STATUS
RECURRENCE2_EXTENT
RECURRENCE2_LATERALITY
RECURRENCE2_SOURCE
RECURRENCE_CONFIRMED
RECURRENCE_ER_STATUS
RECURRENCE_EXTENT
RECURRENCE_LATERALITY
RECURRENCE_SOURCE
RECUR_BUT_NO_SELFREPORT
REGISTRY_AFTER_LAST_CONTACT
STATUS
;
run;

/*DCIS case or not?*/
data recurr_tmp01;
set recurr_drop;
ductal=0;
if r_histo in (80502, 82012, 82102, 82302, 84012, 85002,
85012, 85032, 85042, 85072, 85222, 85232, 85402, 85432)
 then ductal=1 ;
if r_histo=. then ductal=.;
/*set missing recontact flags to 0 instead of missing*/
if rec1=. then rec1=0;
if rec2=. then rec2=0;
if rec3=. then rec3=0;
if rec4=. then rec4=0;
run;

/*only keep participamts who participate din at least one follow-up interview*/
data recurr_tmp1;
set recurr_tmp01;
if rec1=1 or rec2=1 or rec3=1 or rec4=1;
run;

/********** treatment data *****************/
libname surgZ 'G:\Team\WHS DATA\whsstat\BCIS_Recontact\Data\Tx Surgery\2014' ;
```

```
/*treatment initial diagnosis*/
data surgery ; set surgZ.bcis_surgery_2014_v1 ;
  if surg_dx=0 ; run ;
proc sort ; by idnum ; run ;

/*treatment recurrence=1*/
data surgery01 ; set surgZ.bcis_surgery_2014_v1 ;
  if surg_dx=1 ; run ;
proc sort ; by idnum ; run ;

data surgery02 ; set surgZ.bcis_surgery_2014_v1 ;
  if surg_dx=2 ; run ;
proc sort ; by idnum ; run ;

/*treatment initial diagnosis*/
data surgery2;
set surgery;
surgerydx0=.;

if (SURG_DX=0 AND SURG_TYPE="none" ) then  surgerydx0=0;
else if (SURG_DX=0 AND SURG_TYPE="biopsy" ) then  surgerydx0=1;
else if (SURG_DX=0 AND SURG_TYPE="lumpectomy" ) then  surgerydx0=2;
else if (SURG_DX=0 AND SURG_TYPE="mastectomy" and
(surg_lat="left" OR surg_lat="right")) then do;
surgerydx0=3;/*unilateral mastectomy*/
surg_latdx0=surg_lat;
end;
else if (SURG_DX=0 AND SURG_TYPE="mastectomy" and surg_lat="both")
then  surgerydx0=4; /*bilateral mastectomy*/
run;

/*IDNUM 579094 had 2 unilateral mastectomy procedures and
was therefore wrongly assigned to surgerydx0=3 before,
but really shoukld be assigned to surgerydx0=4 (bilateral mastectomy)*/
data surgery3;
set surgery2;
if IDNUM=579094 then surgerydx0=4;
run;

proc sort data=surgery3;
by IDNUM surgerydx0;
run;

/*only keep most aggressive treatment per IDNUM*/
data surgery4;
set surgery3;
by IDNUM surgerydx0;
if last.IDNUM;
run;

data surgery5;
set surgery4;
keep
IDNUM
```

```
surgerydx0
surg_latdx0;
run;
proc sort data=surgery5; by IDNUM; run;

/*treatment 1st recurrence*/
data surgery1_2;
set surgery01;
surgerydx1=.;

if (SURG_DX=1 AND SURG_TYPE="none" ) then  surgerydx1=0;
else if (SURG_DX=1 AND SURG_TYPE="biopsy" ) then  surgerydx1=1;
else if (SURG_DX=1 AND SURG_TYPE="lumpectomy" ) then  surgerydx1=2;
else if (SURG_DX=1 AND SURG_TYPE="mastectomy"
and (surg_lat="left" OR surg_lat="right")) then do;
surgerydx1=3;/*unilateral mastectomy*/
surg_latdx1=surg_lat;
end;
else if (SURG_DX=1 AND SURG_TYPE="mastectomy" and surg_lat="both")
then  surgerydx1=4; /*bilateral mastectomy*/
run;

data surgery1_3;
set surgery1_2;
keep
IDNUM
surgerydx1
surg_latdx1;
run;
proc sort data=surgery1_3; by IDNUM; run;

/*treatment 2nd recurrence*/
data surgery2_2;
set surgery02;
surgerydx2=.;

if (SURG_DX=2 AND SURG_TYPE="none" ) then  surgerydx2=0;
else if (SURG_DX=2 AND SURG_TYPE="biopsy" ) then  surgerydx2=1;
else if (SURG_DX=2 AND SURG_TYPE="lumpectomy" ) then  surgerydx2=2;
else if (SURG_DX=2 AND SURG_TYPE="mastectomy"
and (surg_lat="left" OR surg_lat="right")) then  do;
surgerydx2=3;/*unilateral mastectomy*/
surg_latdx2=surg_lat;
end;
else if (SURG_DX=2 AND SURG_TYPE="mastectomy"
and surg_lat="both") then  surgerydx2=4; /*bilateral mastectomy*/
run;

data surgery2_3;
set surgery2_2;
keep
IDNUM
surgerydx2
surg_latdx2;
```

```
run;
proc sort data=surgery2_3; by IDNUM; run;

proc sort data=recurr_tmp1; by IDNUM; run;

data recurr_tmp1_1;
MERGE recurr_tmp1(in=a)surgery5 surgery1_3 surgery2_3;
by IDNUM;
if a=1;
run;

data recurr_tmp1_01;
set recurr_tmp1_1;
/*for all people without a recurrence, set surgerydx1 and surgerydx2 to zero*/
if recurrence=0 then do;
surgerydx1=0;
surgerydx2=0;
end;
run;

/*if there were 2 unilateral mastectomies (on both sides), set to bilateral mastectomy*/
data recurr_tmp1_2;
set recurr_tmp1_01;
if ((surgerydx0=3 and surg_latdx0="left") and (surgerydx1=3 and surg_latdx1="right"))
OR ((surgerydx0=3 and surg_latdx0="right")
and (surgerydx1=3 and surg_latdx1="left")) then surgerydx1=4;

else if ((surgerydx0=3 and surg_latdx0="right") and (surgerydx2=3 and surg_latdx2="left"))
OR ((surgerydx0=3 and surg_latdx0="left") and (surgerydx2=3 and surg_latdx2="right"))
OR ((surgerydx1=3 and surg_latdx1="left") and (surgerydx2=3 and surg_latdx2="right"))
OR ((surgerydx1=3 and surg_latdx1="right")
and (surgerydx2=3 and surg_latdx2="left")) then surgerydx2=4;
run;

/***** radiation, endocrine treatment ********/
libname radZ 'G:\Team\WHS DATA\whsstat\
BCIS_Recontact\Data\Tx Treatment\Radiation\2014' ;
data radiation; set radZ.bcis_radiation_2014_v1 ;  run ;
proc sort ; by idnum ; run ;

libname txZ 'G:\Team\WHS DATA\whsstat\
BCIS_Recontact\Data\Tx Treatment\Tamoxifen Raloxifene and AIs\2014' ;
data AIs;  set txZ.bcis_Raloxifene_AIs_2014_v1 ;   run ;
proc sort ; by idnum ; run ;

data tamox; set txZ.bcis_tamoxifen_2014_v1 ;  run ;
proc sort ; by idnum ; run ;

data tamox2;
set tamox;
keep IDNUM
tamox_dx0;
run;
```

```
proc sort data=tamox2; by idnum ; run ;

/**** STEP 4: merge recurrence and other datasets */

proc sort data=recurr_tmp1_2; by IDNUM; run;

/*Merge everything with BCIS 1 and 2 data*/
data all_merged;
MERGE recurr_tmp1_2(in=a) radiation AIs tamox2;
by IDNUM;
if a=1;
run;

/************* data clean up *****************/

data recurr_tmp2;
set all_merged;

if BCA2=2 then BCA2=0; /* set 'no' to value 0 instead of 2*/
if educ<1 then educ=.; /*1-9*/
if famhist=3 then famhist=.; /* "3" = unknown*/
else if famhist=2 then famhist=0; /* set 'no' to value 0 instead of 2*/
if hhad <=0 then hhad=.; /* >=1 */
if hhch <0 then hhch=.; /* >=0 */
if HR<0 then HR=.; /*1,2*/
if HR=2 then HR=0; /* set 'no' to value 0 instead of 2*/
if INCMBSL<1 then INCMBSL=.; /* 1-5 */
if mamn<1 then mamn=.;
if MAMNUM<0 or MAMNUM>25 then MAMNUM=.; /* >=0; get rid of "D", "N", "R"*/
if ( MAMRECM<1 OR MAMRECM>12 ) then MAMRECM =.;
if ( MAMRECM2<1 OR MAMRECM2>12 ) then MAMRECM2 =.;
if ( MAMRECMO<1 OR MAMRECMO>12 ) then MAMRECMO =.;
if ( MAMRECM4<1 OR MAMRECM4>12 ) then MAMRECM4 =.;
if MAMRECY<1975 then MAMRECY=.;
if MAMRECY2<1975 then MAMRECY2=.;
if MAMRECY4>2014 then MAMRECY4=.;
if (MAMRECYR<1985 or MAMRECYR>2011) then MAMRECYR=.;
if MAMMO=7 or MAMMO=8 then MAMMO=.; /* 1-6 */
if MAMSDXN<0 then MAMSDXN=.; /* >=0 */
if MAMSDXN2<0 then MAMSDXN2=.; /* >=0 */
if RACEDES<1 then RACEDES=.; /*1-4*/
if RACEF<1 then RACEF=.; /*1-4*/
if RACEM<1 then RACEM=.; /*1-4*/
if RECMRIM>12 then RECMRIM=.;
if (RECMRIY<1997 or RECMRIY>2013) then RECMRIY=.;
if MRI4> 6 then MRI4=.;
if RECURRENCE=0 then RECURRENCE2=0;
/*if people had no first recurrence, they could not have a 2nd*/

if rad_dx0=-1 then rad_dx0=.;
if rad_dx1=-1 then rad_dx1=.;
if rad_dx2=-1 then rad_dx2=.;
if ralox_dx0=-1 then ralox_dx0=.;
if ralox_dx0=9 then ralox_dx0=2; /*recoding of late raloxifen use*/
```

```
if ralox_dx1=-1 then ralox_dx1=.;
if ralox_dx1=9 then ralox_dx1=2; /*recoding of late raloxifen use*/
if ai_dx0=-1 then ai_dx0=.;
if ai_dx0=9 then ai_dx0=2; /*recoding of late AI use*/
if ai_dx1=-1 then ai_dx1=.;
if ai_dx1=9 then ai_dx1=2; /*recoding of late AI use*/
if tamox_dx0=-1 then tamox_dx0=.;
if tamox_dx0=9 then tamox_dx0=2; /*recoding of late tamoxifen use*/

/*recurent cases: if trweatment at dx1<dx0 or dx2<dx1, then set to most aggressive treatment value*/
if (RECURRENCE=1 and SURGERYDX1 NE .
and SURGERYDX1 < SURGERYDX0) then SURGERYDX1=SURGERYDX0;
if (RECURRENCE2=1 and SURGERYDX2 NE .
and SURGERYDX2 < SURGERYDX1) then SURGERYDX2=SURGERYDX1;
run;


/*************************************************************************
Change of coding procedure as of April 2017:
     Export dataset to .csv and implememt steps below in R instead
(bootstrap sampling before imputations and subsequent data transformations)
*************************************************************************/

/*Mark the records that get imputed treatment values*/
data recurr_tmp2_11;
set recurr_tmp2_1;
imputedx0=0;
imputedx1=0;
imputedx2=0;
imputedmam1=0;
imputedmam2=0;
imputedmam3=0;
imputedmam4=0;

if mamsdxn=. then imputedmam1=1;
if mamsdxn2=. then imputedmam2=1;
if mamnum=. then imputedmam3=1;
if mammo=. then imputedmam4=1;
if SURGERYDX0=. then imputedx0=1;
if RECURRENCE=1 and SURGERYDX1=. then imputedx1=1;
if RECURRENCE2=1 and SURGERYDX2=. then imputedx2=1;
run;

/*date formatting*/
data mamlong_export2;
set recurr_tmp2_11;
format INTDATE_BSL MMDDYY10.;
format _intdate MMDDYY10.;
format _intdate2 MMDDYY10.;
format BCAYR MMDDYY10.;

/*create numerical date variables*/
/*objective diagnosis date*/
diagdt=year(_REFDATE) * 10000 +
```

```
month(_REFDATE) * 100 + day(_REFDATE);
/*subjective diagnosis date*/
subjdiagdt=year(BCAYR) * 10000 +
month(BCAYR) * 100 + day(BCAYR);
/*baseline interview date*/
intbsldt=year(INTDATE_BSL) * 10000 +
month(INTDATE_BSL) * 100 + day(INTDATE_BSL);
/*first FU date*/
int1dt=year(_INTDATE) * 10000 +
month(_INTDATE) * 100 + day(_INTDATE);
/*second FU date*/
int2dt=year(_INTDATE2) * 10000 +
month(_INTDATE2) * 100 + day(_INTDATE2);
/*third FU date*/
int3dt=year(SURVEY_RETURNEDX2010) * 10000 +
month(SURVEY_RETURNEDX2010) * 100 + day(SURVEY_RETURNEDX2010);
/*4th FU date*/
int4dt=year(SURVEY_RETURNEDX2013) * 10000 +
month(SURVEY_RETURNEDX2013) * 100 + day(SURVEY_RETURNEDX2013);
/*first recurrence date*/
rec1dt=year(RECURRENCE_DATEX) * 10000 +
month(RECURRENCE_DATEX) * 100 + day(RECURRENCE_DATEX);
/*second recurrence date*/
rec2dt=year(RECURRENCE2_DATEX) * 10000 +
month(RECURRENCE2_DATEX) * 100 + day(RECURRENCE2_DATEX);
run;

/* indicator of endocrine therapy*/
data mamlong_export3;
set mamlong_export2;
endocrinedx0=.;
endocrinedx1=.;
endocrinedx2=.;

/*initial diagnosis*/
if (ralox_dx0=1 OR ralox_dx0=2 OR ai_dx0=1 OR ai_dx0=2
OR tamox_dx0=1 OR tamox_dx0=2) then endocrinedx0=1;
else if (ralox_dx0=0 AND ai_dx0=0 AND tamox_dx0=0) then endocrinedx0=0;

/*1st recurrence*/
if (ralox_dx1=1 OR ralox_dx1=2 OR ai_dx1=1 OR ai_dx1=2) then endocrinedx1=1;
else if (ralox_dx1=0 AND ai_dx1=0) then endocrinedx1=0;

/*2nd recurrence*/
if (ralox_dx2=1 OR ralox_dx2=2 OR ai_dx2=1 OR ai_dx2=2) then endocrinedx2=1;
else if (ralox_dx2=0 AND ai_dx2=0) then endocrinedx2=0;
run;

/*lump together 'none' and 'biopsy' in surgery variables*/
data mamlong_export4;
set  mamlong_export3;
if surgerydx0=0 then surgerydx0=1;
if surgerydx1=0 then surgerydx1=1;
if surgerydx2=0 then surgerydx2=1;
```

```
run;

/*get rid of unnecessary variables*/
data mamlong_export5;
set  mamlong_export4;
drop
RACEF
RACEM
HHAD
HHCH
BCA2
RALOX_B4DX0
AI_B4DX0
ralox_dx0
ai_dx0
tamox_dx0
ralox_dx1
ai_dx1
ralox_dx2
ai_dx2
SURG_LATDX0
SURG_LATDX1
SURG_LATDX2
_VERSION
PHASE
R_HISTO
/*drop original date variables in export*/
_REFDATE
BCAYR
INTDATE_BSL
_INTDATE
_INTDATE2
SURVEY_RETURNEDX2010
SURVEY_RETURNEDX2013
RECURRENCE_DATEX
RECURRENCE2_DATEX
LAST_CONTACT
HR
EDUC
INCMBSL
;
run;

/*Data analysis in R:*/
setwd("G:/Team/WHS ADMIN/Public/Jewett/Mammo longitudinal/methods and output/R")
library(Hmisc)

orig=read.csv("mamlongexport.csv", header=TRUE)

# keep only women with at least 1 follow-up
orig2=orig[which(orig$REC1==1 | orig$REC2==1 | orig$REC3==1 | orig$REC4==1 ) ,]

# only keep DCIS cases
DCIS=orig2[which(orig2$DUCTAL==1) ,] #dim(orig):1858; dim(DCIS): 1616
```

```
# Create date variabkles from numeric date fields
DCIS$DIAGDT=as.character(DCIS$DIAGDT) # objective diagnosis date
DCIS$diagdate=as.Date(DCIS$DIAGDT,"%Y%m%d")

DCIS$SUBJDIAGDT=as.character(DCIS$SUBJDIAGDT) # subjective diagnosis date
DCIS$subjdiagdate=as.Date(DCIS$SUBJDIAGDT,"%Y%m%d")

# baseline interview date
DCIS$INTBSLDT=as.character(DCIS$INTBSLDT)
DCIS$INTBSDT=as.Date(DCIS$INTBSLDT,"%Y%m%d")

# follow-up interview 1 and 2
DCIS$int1date=as.character(DCIS$INT1DT)
DCIS$int1date=as.Date(DCIS$int1date,"%Y%m%d")

DCIS$int2date=as.character(DCIS$INT2DT)
DCIS$int2date=as.Date(DCIS$int2date,"%Y%m%d")

# Time between 2nd interview and previous interview:
# If participated in FU1, then time from Fu1 until Fu2, otherwise bsl interview until FU2
DCIS$prevint2_1=as.numeric(difftime(DCIS$int2date,DCIS$INTBSDT, units ="days"))/365.25
DCIS$prevint2_2=as.numeric(difftime(DCIS$int2date,DCIS$int1date, units ="days"))/365.25
DCIS$prevint2_1[DCIS$prevint2_2>0] <- 0
DCIS$prevint2_2[is.na(DCIS$prevint2_2)]=0
DCIS$prevint=DCIS$prevint2_2 + DCIS$prevint2_1
DCIS$prevint[is.na(DCIS$prevint)]=0

# get rid of old numeric date fields
#DCIS2<- subset(DCIS, select = -c(DIAGDT,SUBJDIAGDT,INTBSLDT) )

DCIS2=DCIS

# absolute difference objective vs. subjective date of diagnosis
DCIS2$subvsob=abs(as.numeric(difftime(DCIS2$diagdate,DCIS2$subjdiagdate, units ="days"))/365.25)
DCIS2$subvsob[is.na(DCIS2$subvsob)]=0 # set missing ones to 0, otherwise wrongly excluded below

# exclude women whose subjective date of diagnosis differs by > 2 years from objective date of diagnosis
DCIS3=DCIS2[which(DCIS2$subvsob < 2), ]

# next steps:
# 1. draw 1 bootstrap sample at a time from original data
# 2. impute missing values in each sample
# 3. data transformations and derivations as implemented in SAS so far
# 4. transform to long format
# 5. further restrictions/datsets for sensituvuity analyses: only women
# - without bilateral mastectomy at time t
# - without recurrence at time t
# - no women who misreported date of initial diagnosis by >= 2 years

#####################################################
# 1. Draw bootstrap sample from original dataset #
#####################################################
#test=DCIS[1:10,1:10]
```

```
#test2=test[sample(nrow(test), 10,replace=T), ]
#(test2 <- test2[order(test2$IDNUM),])

# bootstrap samples: create 5 (change to more eventually)
# test samples with replacement

distPE12=c()
distPE34=c()
distcoeffs1=c()
distcoeffs2=c()

bootiter=1
while (bootiter<101){

newset=DCIS3[sample(nrow(DCIS3), 1580, replace=T), ]
#newset <- newset[order(newset$IDNUM),]
newset$IDNUM2= 1:1580

##############################################
# 2. 5 imputations for each bootstrap sample #
##############################################
# Convert to factor variables:
newset$racefac=as.factor(newset$RACEDES)
newset$surg1fac=as.factor(newset$SURGERYDX1)
newset$surg0fac=as.factor(newset$SURGERYDX0)
newset$surg2fac=as.factor(newset$SURGERYDX2)

#DCIS$grp[is.na(DCIS$grp)]=4

# Do 5 imputations and run loop below for each imputation

PE12imp=c()
PE34imp=c()
coeffs1imp=c()
coeffs2imp=c()

newset$grp=newset$surg0fac
newset$grp[is.na(newset$grp)]=2
#table(newset$grp)
#table(newset$surg0fac)

impiter=1
while (impiter< 6){

imptd <- aregImpute(~ AGERF + RECURRENCE + racefac + RAD_DX0 + FAMHIST
                    + surg0fac + ENDOCRINEDX0 + MAMN + MAMNUM
                    + MAMMO + MAMSDXN + MAMSDXN2,
                    data=newset, n.impute=1, nk=0, group=newset$grp)

compimp1=newset

imp1 <- impute.transcan(imptd, imputation=1,
data=newset, list.out=TRUE,pr=FALSE, check=FALSE)
compimp1[names(imp1)] <- imp1
```

```
#####################################################################
# 3. Create full datasets with imputed values and loop over each #
#####################################################################
# Set variable to indicate imputation
#for (i in 1:nrow(imp1)) {
#  imp1$imputation=1
#} # do for other imputed sets analogously

# RBIND all 5 imputed datsets into 1 large dataset
#implarge=rbind(imp1,imp2,imp3,imp4,imp5)

#set imputed mammography variables to missing again
loopset=compimp1
loopset$MAMSDXN[loopset$IMPUTEDMAM1==1] <- NA
loopset$MAMSDXN2[loopset$IMPUTEDMAM2==1] <- NA
loopset$MAMNUM[loopset$IMPUTEDMAM3==1] <- NA
loopset$MAMMO[loopset$IMPUTEDMAM4==1] <- NA
loopset$surg0=as.numeric(loopset$surg0fac)
loopset$MRI4[is.na(loopset$MRI4)]=1
# 1 equals 0; if NA aren't replaced by "0", sum below is not calculated correctly

#loopset$RECURRENCE2[is.na(loopset$RECURRENCE2)]=0
# Loop over one imputed dataset at a time:
# for (j in 1:5) {}
# loopset=implarge[which(implarge$imputation==j) ,]

###################################
# 4. data transformations in loop #
###################################


#########################################################################
#set mammogram values in 2010 and 2013 survey to the real number instead of number+1
# and combine mammograms and MRIs in 2013 survey into 1 number
loopset$mamint3=loopset$MAMNUM-1
loopset$mamint4=loopset$MAMMO-1
loopset$mamtotint4=loopset$mamint4 + (loopset$MRI4 - 1)

for (i in 1:nrow(loopset)){ # switch from DCIS to each imputed dataset
#########################################################################
# Combine surgical treatment with radiation data*/
# treatment initial diagnosis
loopset$surgrad_dx0[i]= (loopset$surg0[i]==0 | loopset$surg0[i]==1) * 1 + # none or biopsy
                        (loopset$surg0[i]==2 & loopset$RAD_DX0[i]==0) * 2 + # lumpectomy/no radiation
                        (loopset$surg0[i]==2 & loopset$RAD_DX0[i]==1) * 3 + # lumpectomy with radiation
                        (loopset$surg0[i]==3) * 4 + # unilateral mastectomy
                        (loopset$surg0[i]==4) * 5 # bilateral mastectomy

# treatment first recurrence
loopset$surgrad_dx1[i]= (loopset$SURGERYDX1[i]==0 | loopset$SURGERYDX1[i]==1) * 1 +
# none or biopsy
                        (loopset$SURGERYDX1[i]==2 & loopset$RAD_DX1[i]==0) * 2 +
                        # lumpectomy/no radiation
```

```
                        (loopset$SURGERYDX1[i]==2 & loopset$RAD_DX1[i]==1) * 3 +
                        # lumpectomy with radiation
                        (loopset$SURGERYDX1[i]==3) * 4 + # unilateral mastectomy
                        (loopset$SURGERYDX1[i]==4) * 5 # bilateral mastectomy

# treatment second recurrence
loopset$surgrad_dx2[i]= (loopset$SURGERYDX2[i]==0 | loopset$SURGERYDX2[i]==1) * 1 +
# none or biopsy
                        + (loopset$SURGERYDX2[i]==2 & loopset$RAD_DX2[i]==0) * 2 +
                        # lumpectomy/no radiation
                        + (loopset$SURGERYDX2[i]==2 & loopset$RAD_DX2[i]==1) * 3 +
                        # lumpectomy with radiation
                        + (loopset$SURGERYDX2[i]==3) * 4 + # unilateral mastectomy
                        + (loopset$SURGERYDX2[i]==4) * 5 # bilateral mastectomy


# if recurrence=0 or recurrence2=0, set respective treatments to missing
loopset$surgrad_dx1[i]= (loopset$RECURRENCE[i] == 1) * loopset$surgrad_dx1[i]
loopset$surgrad_dx2[i]= (loopset$RECURRENCE2[i] == 1) * loopset$surgrad_dx2[i]


##############################################################################
# recurrence status at each interview
loopset$recurrint1[i]= (loopset$RECURRENCE[i] == 1 & loopset$REC1DT[i] < loopset$INT1DT[i]) * 1
loopset$recurrint2[i]= (loopset$RECURRENCE[i] == 1 & loopset$REC1DT[i] < loopset$INT2DT[i]) * 1
loopset$recurrint3[i]= (loopset$RECURRENCE[i] == 1 & loopset$REC1DT[i] < loopset$INT3DT[i]) * 1
loopset$recurrint4[i]= (loopset$RECURRENCE[i] == 1 & loopset$REC1DT[i] < loopset$INT4DT[i]) * 1

loopset$recurrint1[i][loopset$REC1[i]==0] <- NA # were not there for 1st interview
loopset$recurrint2[i][loopset$REC2[i]==0] <- NA # were not there for 1st interview
loopset$recurrint3[i][loopset$REC3[i]==0] <- NA # were not there for 1st interview
loopset$recurrint4[i][loopset$REC4[i]==0] <- NA # were not there for 1st interview

# recurrence 2 status at each interview
loopset$recurr2int1[i]= (loopset$RECURRENCE2[i] == 1 & loopset$REC2DT[i] < loopset$INT1DT[i]) * 1
loopset$recurr2int2[i]= (loopset$RECURRENCE2[i] == 1 & loopset$REC2DT[i] < loopset$INT2DT[i]) * 1
loopset$recurr2int3[i]= (loopset$RECURRENCE2[i] == 1 & loopset$REC2DT[i] < loopset$INT3DT[i]) * 1
loopset$recurr2int4[i]= (loopset$RECURRENCE2[i] == 1 & loopset$REC2DT[i] < loopset$INT4DT[i]) * 1

loopset$recurr2int1[i][loopset$REC1[i]==0] <- NA # were not there for 1st interview
loopset$recurr2int2[i][loopset$REC2[i]==0] <- NA # were not there for 2nd interview
loopset$recurr2int3[i][loopset$REC3[i]==0] <- NA # were not there for 3rd interview
loopset$recurr2int4[i][loopset$REC4[i]==0] <- NA # were not there for 4th interview


#####################################################################
# Set treatment at each interview
# 1st interview:
loopset$treat_int1[i]= (loopset$recurr2int1[i]==1) * max(loopset$surgrad_dx0[i],
loopset$surgrad_dx1[i],loopset$surgrad_dx2[i], na.rm = TRUE) +
# people with 2nd recurrence at interview 1
                       (loopset$recurr2int1[i]==0 & loopset$recurrint1[i]==1) *
                       max(loopset$surgrad_dx0[i],loopset$surgrad_dx1[i], na.rm = TRUE) +
                       # people with 1st recurrence at interview 1
```

```
                      (loopset$recurr2int1[i]==0 & loopset$recurrint1[i]==0) *
                      loopset$surgrad_dx0[i] # people without recurrence at interview 1
loopset$treat_int1[i][loopset$REC1[i]==0] <- NA # were not there for 1st interview

# 2nd interview
loopset$treat_int2[i]= (loopset$recurr2int2[i]==1) * max(loopset$surgrad_dx0[i],
loopset$surgrad_dx1[i],loopset$surgrad_dx2[i], na.rm = TRUE) +
# people with 2nd recurrence at interview 1
                      (loopset$recurr2int2[i]==0 & loopset$recurrint2[i]==1) *
                      max(loopset$surgrad_dx0[i],loopset$surgrad_dx1[i], na.rm = TRUE) +
                      # people with 1st recurrence at interview 1
                      (loopset$recurr2int2[i]==0 & loopset$recurrint2[i]==0) *
                      loopset$surgrad_dx0[i] # people without recurrence at interview 1
loopset$treat_int2[i][loopset$REC2[i]==0] <- NA # were not there for 2nd interview

# 3rd interview
loopset$treat_int3[i]= (loopset$recurr2int3[i]==1) * max(loopset$surgrad_dx0[i],
loopset$surgrad_dx1[i],loopset$surgrad_dx2[i], na.rm = TRUE) +
# people with 2nd recurrence at interview 1
                      (loopset$recurr2int3[i]==0 & loopset$recurrint3[i]==1) *
                      max(loopset$surgrad_dx0[i],loopset$surgrad_dx1[i], na.rm = TRUE) +
                      # people with 1st recurrence at interview 1
                      (loopset$recurr2int3[i]==0 & loopset$recurrint3[i]==0) *
                      loopset$surgrad_dx0[i] # people without recurrence at interview 1
loopset$treat_int3[i][loopset$REC3[i]==0] <- NA # were not there for 3rd interview

# 3rd interview
loopset$treat_int4[i]= (loopset$recurr2int4[i]==1) * max(loopset$surgrad_dx0[i],
loopset$surgrad_dx1[i],loopset$surgrad_dx2[i], na.rm = TRUE) +
# people with 2nd recurrence at interview 1
                      (loopset$recurr2int4[i]==0 & loopset$recurrint4[i]==1) *
                      max(loopset$surgrad_dx0[i],loopset$surgrad_dx1[i], na.rm = TRUE) +
                      # people with 1st recurrence at interview 1
                      (loopset$recurr2int4[i]==0 & loopset$recurrint4[i]==0) *
                      loopset$surgrad_dx0[i] # people without recurrence at interview 1
loopset$treat_int4[i][loopset$REC4[i]==0] <- NA # were not there for 4th interview

################################################################
# Binary treatment variables (=risk factors in regression analysis)
loopset$TREAT_BIOP1[i]=(loopset$treat_int1[i]==1) * 1
loopset$TREAT_BIOP2[i]=(loopset$treat_int2[i]==1) * 1
loopset$TREAT_BIOP3[i]=(loopset$treat_int3[i]==1) * 1
loopset$TREAT_BIOP4[i]=(loopset$treat_int4[i]==1) * 1

loopset$TREAT_LUMP_NORAD1[i]=(loopset$treat_int1[i]==2) * 1
loopset$TREAT_LUMP_NORAD2[i]=(loopset$treat_int2[i]==2) * 1
loopset$TREAT_LUMP_NORAD3[i]=(loopset$treat_int3[i]==2) * 1
loopset$TREAT_LUMP_NORAD4[i]=(loopset$treat_int4[i]==2) * 1

loopset$TREAT_LUMP_RAD1[i]=(loopset$treat_int1[i]==3) * 1
loopset$TREAT_LUMP_RAD2[i]=(loopset$treat_int2[i]==3) * 1
loopset$TREAT_LUMP_RAD3[i]=(loopset$treat_int3[i]==3) * 1
loopset$TREAT_LUMP_RAD4[i]=(loopset$treat_int4[i]==3) * 1
```

```
loopset$TREAT_MASTUNI1[i]=(loopset$treat_int1[i]==4) * 1
loopset$TREAT_MASTUNI2[i]=(loopset$treat_int2[i]==4) * 1
loopset$TREAT_MASTUNI3[i]=(loopset$treat_int3[i]==4) * 1
loopset$TREAT_MASTUNI4[i]=(loopset$treat_int4[i]==4) * 1

loopset$TREAT_MASTBI1[i]=(loopset$treat_int1[i]==5) * 1
loopset$TREAT_MASTBI2[i]=(loopset$treat_int2[i]==5) * 1
loopset$TREAT_MASTBI3[i]=(loopset$treat_int3[i]==5) * 1
loopset$TREAT_MASTBI4[i]=(loopset$treat_int4[i]==5) * 1

#########################################################################
#set mammogram values in 2010 and 2013 survey to the real number instead of number+1
# and combine mammograms and MRIs in 2013 survey into 1 number
#loopset$mamint3[i]=loopset$MAMNUM[i]-1
#loopset$mamint4[i]=loopset$MAMMO[i]-1
#loopset$mamtotint4[i]=loopset$mamint4[i] + (loopset$MRI4[i] - 1)

} # end loop over each imputed dataset (loopset)

############################
##### End of loop ##########
############################
# Set implausible 0 values to missing (possibly pull into loop)
loopset$surgrad_dx1[loopset$surgrad_dx1==0] <- NA
loopset$surgrad_dx2[loopset$surgrad_dx2==0] <- NA

loopset$TREAT_BIOP1[loopset$REC1==0] <- NA
loopset$TREAT_BIOP2[loopset$REC2==0] <- NA
loopset$TREAT_BIOP3[loopset$REC3==0] <- NA
loopset$TREAT_BIOP4[loopset$REC4==0] <- NA

loopset$TREAT_LUMP_NORAD1[loopset$REC1==0] <- NA
loopset$TREAT_LUMP_NORAD2[loopset$REC2==0] <- NA
loopset$TREAT_LUMP_NORAD3[loopset$REC3==0] <- NA
loopset$TREAT_LUMP_NORAD4[loopset$REC4==0] <- NA

loopset$TREAT_LUMP_RAD1[loopset$REC1==0] <- NA
loopset$TREAT_LUMP_RAD2[loopset$REC2==0] <- NA
loopset$TREAT_LUMP_RAD3[loopset$REC3==0] <- NA
loopset$TREAT_LUMP_RAD4[loopset$REC4==0] <- NA

loopset$TREAT_MASTUNI1[loopset$REC1==0] <- NA
loopset$TREAT_MASTUNI2[loopset$REC2==0] <- NA
loopset$TREAT_MASTUNI3[loopset$REC3==0] <- NA
loopset$TREAT_MASTUNI4[loopset$REC4==0] <- NA

loopset$TREAT_MASTBI1[loopset$REC1==0] <- NA
loopset$TREAT_MASTBI2[loopset$REC2==0] <- NA
loopset$TREAT_MASTBI3[loopset$REC3==0] <- NA
loopset$TREAT_MASTBI4[loopset$REC4==0] <- NA

################################
##### drop unused variables #####
################################
```

```
smaller<- subset(loopset, select = c(IDNUM,AGERF,FAMHIST,RECURRENCE,
      RECURRENCE2,MAMSDXN,MAMSDXN2,mamint3,mamint4,mamtotint4,
                surgrad_dx0,surgrad_dx1,surgrad_dx2,
                ENDOCRINEDX0,subvsob,
                recurrint1,recurrint2,recurrint3,recurrint4,
                recurr2int1,recurr2int2,recurr2int3,recurr2int4,
                treat_int1,treat_int2,treat_int3,treat_int4,
                REC1,REC2,REC3,REC4,
                DIAGDT,SUBJDIAGDT,INT1DT,INT2DT,INT3DT,INT4DT,
                TREAT_BIOP1,TREAT_BIOP2,TREAT_BIOP3,TREAT_BIOP4,
                TREAT_LUMP_NORAD1,TREAT_LUMP_NORAD2,TREAT_LUMP_NORAD3,TREAT_LUMP_NORAD4,
                TREAT_LUMP_RAD1,TREAT_LUMP_RAD2,TREAT_LUMP_RAD3,TREAT_LUMP_RAD4,
                TREAT_MASTUNI1,TREAT_MASTUNI2,TREAT_MASTUNI3,TREAT_MASTUNI4,
                TREAT_MASTBI1,TREAT_MASTBI2,TREAT_MASTBI3,TREAT_MASTBI4))

# Transform from broad to long format
longset <- reshape(smaller,
             varying = list(c("REC1","REC2","REC3","REC4"),
               c("recurrint1", "recurrint2", "recurrint3", "recurrint4"),
               c("recurr2int1","recurr2int2","recurr2int3","recurr2int4"),
               c("treat_int1","treat_int2","treat_int3","treat_int4"),
               c("MAMSDXN","MAMSDXN2","mamint3","mamtotint4"),
               c("MAMSDXN","MAMSDXN2","mamint3","mamint4"),
               c("INT1DT","INT2DT","INT3DT","INT4DT"),
               c("TREAT_BIOP1","TREAT_BIOP2","TREAT_BIOP3","TREAT_BIOP4"),
               c("TREAT_LUMP_NORAD1","TREAT_LUMP_NORAD2","TREAT_LUMP_NORAD3","TREAT_LUMP_NORAD4"),
               c("TREAT_LUMP_RAD1","TREAT_LUMP_RAD2","TREAT_LUMP_RAD3","TREAT_LUMP_RAD4"),
               c("TREAT_MASTUNI1","TREAT_MASTUNI2","TREAT_MASTUNI3","TREAT_MASTUNI4"),
               c("TREAT_MASTBI1","TREAT_MASTBI2","TREAT_MASTBI3","TREAT_MASTBI4")),
             v.names = c("PARTICIP","RECURR_T","RECURR2_T","TREAT_T","MAMMRI","MAMMOS","INTDATE",
                         "TREAT_BIOP","TREAT_LUMP_NORAD",
                         "TREAT_LUMP_RAD","TREAT_MASTUNI","TREAT_MASTBI"),
             timevar = "FU_POINT",
             times = c(1, 2, 3, 4),
             direction = "long")


longset <- longset[order(longset$IDNUM,longset$FU_POINT),]

###################################
##### Pull in person time ########
###################################
# get person time data and get rid of unused variables
ptime=read.csv("ptimedcis.csv", header=TRUE)
smallptime<- subset(ptime, select = -c(INTDATE,PARTICIP,diagdate,subjdiagdate,prevint,id,
                                       start,end,startmodulo,startfullhalf,startrest,
                                       endmodulo,endfullhalf,endrest,fullhalves))
# Merge with person time data
merged<- merge(longset,smallptime,by=c("IDNUM","FU_POINT"),sort=FALSE,all=TRUE)

# exclude records with recurrence at given interview
longset2 <- merged[which(merged$RECURR_T==0), ]

# exclude records with bilateral mammography
```

```
longset3 <- longset2[which(longset2$TREAT_MASTBI==0), ]

#Exclude records without mammography data at intreview
longset3$MAMMRI[is.na(longset3$MAMMRI)]=9999
longset4 <- longset3[which(longset3$MAMMRI<9999), ]


#################################
##### Regression Model 1 ########
#################################
# Without year 1

# in or ex-clude women who spent person time in first year after diagnosis:
data_nobil=longset4[which(longset4$yr1==0) ,]


#################################
# Regression part as while-loop #
#################################
# initialization of parameters before loop:
iter=0 # number of iterations
sum_a=sum_b=1 # threshold values exceeded or not
coeff_a_old=coeff_a_new=rep(1,17) # person time coefficients
coeff_a_diff=rep(0,17)
coeff_b_old=coeff_b_new=coeff_b_diff=rep(0,7) # risk factor coefficients, incl. intercept
coeff_a_hist=c() # store a coefficient history
coeff_b_hist=c() # store b coefficient history


while (iter < 500 & (sum_a>0 | sum_b>0)) {
#as long as changes in parameters exceed threshold and maxit has not been surpassed

  ###########################################################################
  # STEP 1: Calculate (update) riskvar*person time as covariate for first regression step #
  ###########################################################################
  # coeff_a_old, coeff_b_old
  coeff_a_old = coeff_a_new
  coeff_b_old = coeff_b_new

  # Update b1,b2,...b5
  b1=coeff_b_old[2]
  b2=coeff_b_old[3]
  b3=coeff_b_old[4]
  b4=coeff_b_old[5]
  b5=coeff_b_old[6]
  b6=coeff_b_old[7]


  for (i in 1:nrow(data_nobil))
  {
    data_nobil$riskvar[i]= exp(b1*data_nobil$AGERF[i] +
                               b2*data_nobil$FAMHIST[i] +
                               b3*data_nobil$TREAT_BIOP[i] +
                               b4*data_nobil$TREAT_LUMP_NORAD[i] +
```

```
                                b5*data_nobil$TREAT_LUMP_RAD[i] +
                                b6*data_nobil$ENDOCRINEDX0[i])

  # calculate covariates for regression using b1 and risk factor values)
  data_nobil$riskt2[i]=data_nobil$riskvar[i] * data_nobil$yr2[i]
  data_nobil$riskt3[i]=data_nobil$riskvar[i] * data_nobil$yr3[i]
  data_nobil$riskt4[i]=data_nobil$riskvar[i] * data_nobil$yr4[i]
  data_nobil$riskt5[i]=data_nobil$riskvar[i] * data_nobil$yr5[i]
  data_nobil$riskt6[i]=data_nobil$riskvar[i] * data_nobil$yr6[i]
  data_nobil$riskt7[i]=data_nobil$riskvar[i] * data_nobil$yr7[i]
  data_nobil$riskt8[i]=data_nobil$riskvar[i] * data_nobil$yr8[i]
  data_nobil$riskt9[i]=data_nobil$riskvar[i] * data_nobil$yr9[i]
  data_nobil$riskt10[i]=data_nobil$riskvar[i] * data_nobil$yr10[i]
  data_nobil$riskt11[i]=data_nobil$riskvar[i] * data_nobil$yr11[i]
  data_nobil$riskt12[i]=data_nobil$riskvar[i] * data_nobil$yr12[i]
  data_nobil$riskt13[i]=data_nobil$riskvar[i] * data_nobil$yr13[i]
  data_nobil$riskt14[i]=data_nobil$riskvar[i] * data_nobil$yr14[i]
  data_nobil$riskt15[i]=data_nobil$riskvar[i] * data_nobil$yr15[i]
  data_nobil$riskt16[i]=data_nobil$riskvar[i] * data_nobil$yr16[i]
  data_nobil$riskt17[i]=data_nobil$riskvar[i] * data_nobil$yr17[i]
  data_nobil$riskt18[i]=data_nobil$riskvar[i] * data_nobil$yr18[i]

}

 ###############################################################################
# STEP 2: run regression based on riskt-covariates: get coefficients a0,a1,a2...a7 for each ptime #
#####################+#########################################################
init1=coeff_a_old
calc_coeff1= glm(MAMMRI ~ -1 + riskt2 + riskt3 + riskt4 + riskt5 + riskt6 +
                   riskt7 + riskt8 + riskt9 + riskt10 + riskt11 + riskt12 +
                   riskt13 + riskt14 + riskt15 + riskt16 + riskt17 + riskt18,
                 data=data_nobil, family=poisson(link="identity"),start=c(init1))

#store coefficients in array -->
for (i in 1: (length(coef(calc_coeff1)))) # with intercept
{
  coeff_a_new[i]=coef(calc_coeff1)[i]
  coeff_a_diff[i]= abs(coeff_a_new[i] - coeff_a_old[i])
}

#############################################
# STEP 3: Calculate (update) offset and R1i #
#############################################
# offset = (ptime1 + a1/a0 ptime2 + a2/a0 ptime3 + a3/a0 ptime4 + a4/a0 ptime5 + a5/a0 ptime6 +)
# R1i = E[Yi] / offset since E[Yi] = R1i * offset
for (i in 1:nrow(data_nobil))
{
  data_nobil$off[i] = data_nobil$yr2[i] + ( (coeff_a_new[2]/coeff_a_new[1]) * data_nobil$yr3[i] +
          (coeff_a_new[3]/coeff_a_new[1]) * data_nobil$yr4[i] +
          (coeff_a_new[4]/coeff_a_new[1]) * data_nobil$yr5[i] +
          (coeff_a_new[5]/coeff_a_new[1]) * data_nobil$yr6[i] +
          (coeff_a_new[6]/coeff_a_new[1]) * data_nobil$yr7[i] +
          (coeff_a_new[7]/coeff_a_new[1]) * data_nobil$yr8[i] +
          (coeff_a_new[8]/coeff_a_new[1]) * data_nobil$yr9[i] +
```

```
            (coeff_a_new[9]/coeff_a_new[1]) * data_nobil$yr10[i] +
            (coeff_a_new[10]/coeff_a_new[1]) * data_nobil$yr11[i] +
            (coeff_a_new[11]/coeff_a_new[1]) * data_nobil$yr12[i] +
            (coeff_a_new[12]/coeff_a_new[1]) * data_nobil$yr13[i] +
            (coeff_a_new[13]/coeff_a_new[1]) * data_nobil$yr14[i] +
            (coeff_a_new[14]/coeff_a_new[1]) * data_nobil$yr15[i] +
            (coeff_a_new[15]/coeff_a_new[1]) * data_nobil$yr16[i] +
            (coeff_a_new[16]/coeff_a_new[1]) * data_nobil$yr17[i] +
            (coeff_a_new[17]/coeff_a_new[1]) * data_nobil$yr18[i]
    )
  }

  ########################################################################
  # STEP 4: Calculate (update) b1, b2, b3, b4 in Poisson regression #
  ########################################################################
  calc_coeff2= glm(MAMMRI ~ AGERF + FAMHIST + TREAT_BIOP + TREAT_LUMP_NORAD
                   + TREAT_LUMP_RAD + ENDOCRINEDX0 + offset(log(off)),
                   data=data_nobil, family=poisson(link="log"))
  #summary(calc_coeff2)

  #store coefficients in array -->
  # for (i in 1: (length(coef(calc_coeff2))-1)) # without intercept
  for (i in 1: (length(coef(calc_coeff2)))) # with intercept
  {
    #coeff_b_new[i]=coef(calc_coeff2)[i+1] # do not store the intercept
    coeff_b_new[i]=coef(calc_coeff2)[i] # with intercept
    coeff_b_diff[i]= abs(coeff_b_new[i] - coeff_b_old[i])
  }

  ##########################################
  # STEP 5: Calculate threshold parameters #
  # and increase iteration value           #
  ##########################################
  iter = iter + 1
  sum_a=sum(coeff_a_diff > 0.000000001)
  sum_b=sum(coeff_a_diff > 0.000000001)

  coeff_a_hist=rbind(coeff_a_hist,coeff_a_new) # track history of a-coefficients
  coeff_b_hist=rbind(coeff_b_hist,coeff_b_new) # track history of b-coefficients

}

##################
# End While-loop #
##################

# Store point estimates

PE1= coeff_b_new # Intercept, AGERF + FAMHIST + TREAT_BIOP
+ TREAT_LUMP_NORAD + TREAT_LUMP_RAD + ENDOCRINEDX0
PE2= coeff_a_new # risk parameters yr 2-18

#write.csv(PE1,"riskfactorlog_noyr1.csv")
#write.csv(PE2,"yearest_noyr1.csv")
```

```
############################
# Model 1 including year 1 #
############################
data_nobil=longset4

# initialization of parameters before loop:
iter=0 # number of iterations
sum_a=sum_b=1 # threshold values exceeded or not
coeff_a_old=coeff_a_new=rep(1,18) # person time coefficients
coeff_a_diff=rep(0,18)
coeff_b_old=coeff_b_new=coeff_b_diff=rep(0,7) # risk factor coefficients
coeff_a_hist=c() # store a coefficient history
coeff_b_hist=c() # store b coefficient history

while (iter < 500 & (sum_a>0 | sum_b>0)) {
#as long as changes in parameters exceed threshold and maxit has not been surpassed

  ##########################################################################
  # STEP 1: Calculate (update) riskvar*person time as covariate for first regression step #
  ##########################################################################
  # coeff_a_old, coeff_b_old
  coeff_a_old = coeff_a_new
  coeff_b_old = coeff_b_new

  # Update b1,b2,...b5
  b1=coeff_b_old[2]
  b2=coeff_b_old[3]
  b3=coeff_b_old[4]
  b4=coeff_b_old[5]
  b5=coeff_b_old[6]
  b6=coeff_b_old[7]

  for (i in 1:nrow(data_nobil))
  {
    data_nobil$riskvar[i]= exp(b1*data_nobil$AGERF[i] +
                               b2*data_nobil$FAMHIST[i] +
                               b3*data_nobil$TREAT_BIOP[i] +
                               b4*data_nobil$TREAT_LUMP_NORAD[i] +
                               b5*data_nobil$TREAT_LUMP_RAD[i] +
                               b6*data_nobil$ENDOCRINEDX0[i])

    # calculate covariates for regression using b1 and risk factor values)
    data_nobil$riskt1[i]=data_nobil$riskvar[i] * data_nobil$yr1[i]
    data_nobil$riskt2[i]=data_nobil$riskvar[i] * data_nobil$yr2[i]
    data_nobil$riskt3[i]=data_nobil$riskvar[i] * data_nobil$yr3[i]
    data_nobil$riskt4[i]=data_nobil$riskvar[i] * data_nobil$yr4[i]
    data_nobil$riskt5[i]=data_nobil$riskvar[i] * data_nobil$yr5[i]
    data_nobil$riskt6[i]=data_nobil$riskvar[i] * data_nobil$yr6[i]
    data_nobil$riskt7[i]=data_nobil$riskvar[i] * data_nobil$yr7[i]
    data_nobil$riskt8[i]=data_nobil$riskvar[i] * data_nobil$yr8[i]
    data_nobil$riskt9[i]=data_nobil$riskvar[i] * data_nobil$yr9[i]
    data_nobil$riskt10[i]=data_nobil$riskvar[i] * data_nobil$yr10[i]
```

```
  data_nobil$riskt11[i]=data_nobil$riskvar[i] * data_nobil$yr11[i]
  data_nobil$riskt12[i]=data_nobil$riskvar[i] * data_nobil$yr12[i]
  data_nobil$riskt13[i]=data_nobil$riskvar[i] * data_nobil$yr13[i]
  data_nobil$riskt14[i]=data_nobil$riskvar[i] * data_nobil$yr14[i]
  data_nobil$riskt15[i]=data_nobil$riskvar[i] * data_nobil$yr15[i]
  data_nobil$riskt16[i]=data_nobil$riskvar[i] * data_nobil$yr16[i]
  data_nobil$riskt17[i]=data_nobil$riskvar[i] * data_nobil$yr17[i]
  data_nobil$riskt18[i]=data_nobil$riskvar[i] * data_nobil$yr18[i]

}

#############################################################################
# STEP 2: run regression based on riskt-covariates: get coefficients a0,a1,a2...a7 for each ptime #
#####################+#######################################################
init1=coeff_a_old
calc_coeff1= glm(MAMMRI ~ -1 + riskt1 + riskt2 + riskt3 + riskt4 + riskt5 + riskt6 +
                   riskt7 + riskt8 + riskt9 + riskt10 + riskt11 + riskt12 +
                   riskt13 + riskt14 + riskt15 + riskt16 + riskt17 + riskt18,
                 data=data_nobil, family=poisson(link="identity"),start=c(init1))
# summary(calc_coeff1)

#store coefficients in array -->
for (i in 1: (length(coef(calc_coeff1)))) # with intercept
{
  coeff_a_new[i]=coef(calc_coeff1)[i]
  coeff_a_diff[i]= abs(coeff_a_new[i] - coeff_a_old[i])
}

#############################################
# STEP 3: Calculate (update) offset and R1i #
#############################################
# offset = (ptime1 + a1/a0 ptime2 + a2/a0 ptime3 + a3/a0 ptime4 + a4/a0 ptime5 + a5/a0 ptime6 +)
# R1i = E[Yi] / offset since E[Yi] = R1i * offset
for (i in 1:nrow(data_nobil))
{
  data_nobil$off[i] = data_nobil$yr1[i] + ( (coeff_a_new[2]/coeff_a_new[1]) * data_nobil$yr2[i]
                                          + (coeff_a_new[3]/coeff_a_new[1]) * data_nobil$yr3[i]
                                          + (coeff_a_new[4]/coeff_a_new[1]) * data_nobil$yr4[i]
                                          + (coeff_a_new[5]/coeff_a_new[1]) * data_nobil$yr5[i]
                                          + (coeff_a_new[6]/coeff_a_new[1]) * data_nobil$yr6[i]
                                          + (coeff_a_new[7]/coeff_a_new[1]) * data_nobil$yr7[i]
                                          + (coeff_a_new[8]/coeff_a_new[1]) * data_nobil$yr8[i]
                                          + (coeff_a_new[9]/coeff_a_new[1]) * data_nobil$yr9[i]
                                          + (coeff_a_new[10]/coeff_a_new[1]) * data_nobil$yr10[i]
                                          + (coeff_a_new[11]/coeff_a_new[1]) * data_nobil$yr11[i]
                                          + (coeff_a_new[12]/coeff_a_new[1]) * data_nobil$yr12[i]
                                          + (coeff_a_new[13]/coeff_a_new[1]) * data_nobil$yr13[i]
                                          + (coeff_a_new[14]/coeff_a_new[1]) * data_nobil$yr14[i]
                                          + (coeff_a_new[15]/coeff_a_new[1]) * data_nobil$yr15[i]
                                          + (coeff_a_new[16]/coeff_a_new[1]) * data_nobil$yr16[i]
                                          + (coeff_a_new[17]/coeff_a_new[1]) * data_nobil$yr17[i]
                                          + (coeff_a_new[18]/coeff_a_new[1]) * data_nobil$yr18[i]
  )
```

```
  }

  ########################################################################
  # STEP 4: Calculate (update) b1, b2, b3, b4 in Poisson regression #
  ########################################################################
  init2=coeff_b_old
  calc_coeff2= glm(MAMMRI ~ AGERF + FAMHIST + TREAT_BIOP + TREAT_LUMP_NORAD
                   + TREAT_LUMP_RAD + ENDOCRINEDX0 + offset(log(off)),
                   data=data_nobil, family=poisson(link="log"),start=c(init2))
  #summary(calc_coeff2)

  #store coefficients in array -->
  # for (i in 1: (length(coef(calc_coeff2))-1)) # without intercept
  for (i in 1: (length(coef(calc_coeff2)))) # with intercept
  {
    #coeff_b_new[i]=coef(calc_coeff2)[i+1] # do not store the intercept
    coeff_b_new[i]=coef(calc_coeff2)[i] # with intercept
    coeff_b_diff[i]= abs(coeff_b_new[i] - coeff_b_old[i])
  }

  ##########################################
  # STEP 5: Calculate threshold parameters #
  # and increase iteration value           #
  ##########################################
  iter = iter + 1
  sum_a=sum(coeff_a_diff > 0.000000001)
  sum_b=sum(coeff_a_diff > 0.000000001)

  coeff_a_hist=rbind(coeff_a_hist,coeff_a_new) # track history of a-coefficients
  coeff_b_hist=rbind(coeff_b_hist,coeff_b_new) # track history of b-coefficients

}


PE3= coeff_b_new
PE4= coeff_a_new

#write.csv(PE3,"riskfactorlog_withyr1.csv")
#write.csv(PE4,"yearest_withyr1.csv")


#################################
##### Regression Model 2 ########
#################################
# without year 1:
# in or ex-clude women who spent person time in first year after diagnosis:
data1_nobil=longset4[which(longset4$yr1==0) ,]

# initialization of parameters before loop:
iter=0 # number of iterations
param_old=param_new=c(1,0,0)
# initial values for C0, beta, and space-holder for deviance in position 4
param_hist=c() # store parameter history
```

```
# Update covariates to estimate C and beta for each particpant
for (i in 1:nrow(data1_nobil))
{
  data1_nobil$C0[i] = (0.25 +                #A10
                         data1_nobil$pt1.5s[i] + data1_nobil$pt1.5e[i] + #A15
                         data1_nobil$pt2s[i] + data1_nobil$pt2e[i] + #A20
                         data1_nobil$pt2.5s[i] + data1_nobil$pt2.5e[i] + #A25
                         data1_nobil$pt3s[i] + data1_nobil$pt3e[i] + #A30
                         data1_nobil$pt3.5s[i] + data1_nobil$pt3.5e[i] + #A35
                         data1_nobil$pt4s[i] + data1_nobil$pt4e[i] + #A40
                         data1_nobil$pt4.5s[i] + data1_nobil$pt4.5e[i] + #A45
                         data1_nobil$pt5s[i] + data1_nobil$pt5e[i] + #A50
                         data1_nobil$pt5.5s[i] + data1_nobil$pt5.5e[i] + #A55
                         data1_nobil$pt6s[i] + data1_nobil$pt6e[i] + #A60
                         data1_nobil$pt6.5s[i] + data1_nobil$pt6.5e[i] + #A65
                         data1_nobil$pt7s[i] + data1_nobil$pt7e[i] + #A70
                         data1_nobil$pt7.5s[i] + data1_nobil$pt7.5e[i] + #A75
                         data1_nobil$pt8s[i] + data1_nobil$pt8e[i] + #A80
                         data1_nobil$pt8.5s[i] + data1_nobil$pt8.5e[i] + #A85
                         data1_nobil$pt9s[i] + data1_nobil$pt9e[i] + #A90

                         data1_nobil$pt9.5s[i] + data1_nobil$pt9.5e[i] + #A95
                         data1_nobil$pt10s[i] + data1_nobil$pt10e[i] + #A100
                         data1_nobil$pt10.5s[i] + data1_nobil$pt10.5e[i] + #A105
                         data1_nobil$pt11s[i] + data1_nobil$pt11e[i] + #A110
                         data1_nobil$pt11.5s[i] + data1_nobil$pt11.5e[i] + #A115
                         data1_nobil$pt12s[i] + data1_nobil$pt12e[i] + #A120
                         data1_nobil$pt12.5s[i] + data1_nobil$pt12.5e[i] + #A125
                         data1_nobil$pt13s[i] + data1_nobil$pt13e[i] + #A130
                         data1_nobil$pt13.5s[i] + data1_nobil$pt13.5e[i] + #A135
                         data1_nobil$pt14s[i] + data1_nobil$pt14e[i] + #A140
                         data1_nobil$pt14.5s[i] + data1_nobil$pt14.5e[i] + #A145
                         data1_nobil$pt15s[i] + data1_nobil$pt15e[i] + #A150
                         data1_nobil$pt15.5s[i] + data1_nobil$pt15.5e[i] + #A155
                         data1_nobil$pt16s[i] + data1_nobil$pt16e[i] + #A160
                         data1_nobil$pt16.5s[i] + data1_nobil$pt16.5e[i] + #A165
                         data1_nobil$pt17s[i] + data1_nobil$pt17e[i] + #A170
                         data1_nobil$pt17.5s[i] + data1_nobil$pt17.5e[i] + #A175
                         data1_nobil$pt18s[i])  #A180


  data1_nobil$beta[i] = (0.25 +                     # A10

                           (3/2*data1_nobil$pt2s[i] + 0.5*data1_nobil$pt2s[i]**2 +
                              0.5*data1_nobil$pt2s[i]*data1_nobil$pt2e[i] +
                              (2/2 + 3/4)*data1_nobil$pt2e[i]) +                    #A20

                           (5/2*data1_nobil$pt3s[i] + 0.5*data1_nobil$pt3s[i]**2 +
                              0.5*data1_nobil$pt3s[i]*data1_nobil$pt3e[i] +
                              (3/2 + 5/4)*data1_nobil$pt3e[i]) +                    #A30

                           (7/2*data1_nobil$pt4s[i] + 0.5*data1_nobil$pt4s[i]**2 +
                              0.5*data1_nobil$pt4s[i]*data1_nobil$pt4e[i] +
                              (4/2 + 7/4)*data1_nobil$pt4e[i]) +                    #A40
```

```
(9/2*data1_nobil$pt5s[i] + 0.5*data1_nobil$pt5s[i]**2 +
   0.5*data1_nobil$pt5s[i]*data1_nobil$pt5e[i] +
   (5/2 + 9/4)*data1_nobil$pt5e[i]) +                      #A50

(11/2*data1_nobil$pt6s[i] + 0.5*data1_nobil$pt6s[i]**2 +
   0.5*data1_nobil$pt6s[i]*data1_nobil$pt6e[i] +
   (6/2 + 11/4)*data1_nobil$pt6e[i]) +                      #A60

(13/2*data1_nobil$pt7s[i] + 0.5*data1_nobil$pt7s[i]**2 +
   0.5*data1_nobil$pt7s[i]*data1_nobil$pt7e[i] +
   (7/2 + 13/4)*data1_nobil$pt7e[i]) +                      #A70

(15/2*data1_nobil$pt8s[i] + 0.5*data1_nobil$pt8s[i]**2 +
   0.5*data1_nobil$pt8s[i]*data1_nobil$pt8e[i] +
   (8/2 + 15/4)*data1_nobil$pt8e[i]) +                      #A80

(17/2*data1_nobil$pt9s[i] + 0.5*data1_nobil$pt9s[i]**2 +
   0.5*data1_nobil$pt9s[i]*data1_nobil$pt9e[i] +
   (9/2 + 17/4)*data1_nobil$pt9e[i]) +                      #A90

(19/2*data1_nobil$pt10s[i] + 0.5*data1_nobil$pt10s[i]**2 +
   0.5*data1_nobil$pt10s[i]*data1_nobil$pt10e[i] +
   (10/2 + 19/4)*data1_nobil$pt10e[i]) +                      #A100

(21/2*data1_nobil$pt11s[i] + 0.5*data1_nobil$pt11s[i]**2 +
   0.5*data1_nobil$pt11s[i]*data1_nobil$pt11e[i] +
   (11/2 + 21/4)*data1_nobil$pt11e[i]) +                      #A110

(23/2*data1_nobil$pt12s[i] + 0.5*data1_nobil$pt12s[i]**2 +
   0.5*data1_nobil$pt12s[i]*data1_nobil$pt12e[i] +
   (12/2 + 23/4)*data1_nobil$pt12e[i]) +                      #A120

(25/2*data1_nobil$pt13s[i] + 0.5*data1_nobil$pt13s[i]**2 +
   0.5*data1_nobil$pt13s[i]*data1_nobil$pt13e[i] +
   (13/2 + 25/4)*data1_nobil$pt13e[i]) +                      #A130

(27/2*data1_nobil$pt14s[i] + 0.5*data1_nobil$pt14s[i]**2 +
   0.5*data1_nobil$pt14s[i]*data1_nobil$pt14e[i] +
   (14/2 + 27/4)*data1_nobil$pt14e[i]) +                      #A140

(29/2*data1_nobil$pt15s[i] + 0.5*data1_nobil$pt15s[i]**2 +
   0.5*data1_nobil$pt15s[i]*data1_nobil$pt15e[i] +
   (15/2 + 29/4)*data1_nobil$pt15e[i]) +                      #A150

(31/2*data1_nobil$pt16s[i] + 0.5*data1_nobil$pt16s[i]**2 +
   0.5*data1_nobil$pt16s[i]*data1_nobil$pt16e[i] +
   (16/2 + 31/4)*data1_nobil$pt16e[i]) +                      #A160

(33/2*data1_nobil$pt17s[i] + 0.5*data1_nobil$pt17s[i]**2 +
   0.5*data1_nobil$pt17s[i]*data1_nobil$pt17e[i] +
   (17/2 + 33/4)*data1_nobil$pt17e[i]) +                      #A170

(35/2*data1_nobil$pt18s[i] + 0.5*data1_nobil$pt18s[i]**2) +  #A180
```

```
(1*data1_nobil$pt1.5s[i] + 0.5*data1_nobil$pt1.5s[i]**2 +
   0.5*data1_nobil$pt1.5s[i]*data1_nobil$pt1.5e[i] +
   (3/4 + 1/2)*data1_nobil$pt1.5e[i]) +                    #A15

(2*data1_nobil$pt2.5s[i] + 0.5*data1_nobil$pt2.5s[i]**2 +   #A25
   0.5*data1_nobil$pt2.5s[i]*data1_nobil$pt2.5e[i] +
   (5/4 + 2/2)*data1_nobil$pt2.5e[i]) +

(3*data1_nobil$pt3.5s[i] + 0.5*data1_nobil$pt3.5s[i]**2 +   #A35
   0.5*data1_nobil$pt3.5s[i]*data1_nobil$pt3.5e[i] +
   (7/4 + 3/2)*data1_nobil$pt3.5e[i]) +

(4*data1_nobil$pt4.5s[i] + 0.5*data1_nobil$pt4.5s[i]**2 +   #A45
   0.5*data1_nobil$pt4.5s[i]*data1_nobil$pt4.5e[i] +
   (9/4 + 4/2)*data1_nobil$pt4.5e[i]) +

(5*data1_nobil$pt5.5s[i] + 0.5*data1_nobil$pt5.5s[i]**2 +   #A55
   0.5*data1_nobil$pt5.5s[i]*data1_nobil$pt5.5e[i] +
   (11/4 + 5/2)*data1_nobil$pt5.5e[i]) +

(6*data1_nobil$pt6.5s[i] + 0.5*data1_nobil$pt6.5s[i]**2 +   #A65
   0.5*data1_nobil$pt6.5s[i]*data1_nobil$pt6.5e[i] +
   (13/4 + 6/2)*data1_nobil$pt6.5e[i]) +

(7*data1_nobil$pt7.5s[i] + 0.5*data1_nobil$pt7.5s[i]**2 +   #A75
   0.5*data1_nobil$pt7.5s[i]*data1_nobil$pt7.5e[i] +
   (15/4 + 7/2)*data1_nobil$pt7.5e[i]) +

(8*data1_nobil$pt8.5s[i] + 0.5*data1_nobil$pt8.5s[i]**2 +   #A85
   0.5*data1_nobil$pt8.5s[i]*data1_nobil$pt8.5e[i] +
   (17/4 + 8/2)*data1_nobil$pt8.5e[i]) +

(9*data1_nobil$pt9.5s[i] + 0.5*data1_nobil$pt9.5s[i]**2 +   #A95
   0.5*data1_nobil$pt9.5s[i]*data1_nobil$pt9.5e[i] +
   (19/4 + 9/2)*data1_nobil$pt9.5e[i]) +

(10*data1_nobil$pt10.5s[i] + 0.5*data1_nobil$pt10.5s[i]**2 +   #A105
   0.5*data1_nobil$pt10.5s[i]*data1_nobil$pt10.5e[i] +
   (21/4 + 10/2)*data1_nobil$pt10.5e[i]) +

(11*data1_nobil$pt11.5s[i] + 0.5*data1_nobil$pt11.5s[i]**2 +   #A115
   0.5*data1_nobil$pt11.5s[i]*data1_nobil$pt11.5e[i] +
   (23/4 + 11/2)*data1_nobil$pt11.5e[i]) +

(12*data1_nobil$pt12.5s[i] + 0.5*data1_nobil$pt12.5s[i]**2 +   #A125
   0.5*data1_nobil$pt12.5s[i]*data1_nobil$pt12.5e[i] +
   (25/4 + 12/2)*data1_nobil$pt12.5e[i]) +

(13*data1_nobil$pt13.5s[i] + 0.5*data1_nobil$pt13.5s[i]**2 +   #A135
   0.5*data1_nobil$pt13.5s[i]*data1_nobil$pt13.5e[i] +
   (27/4 + 13/2)*data1_nobil$pt13.5e[i]) +
```

```
                          (14*data1_nobil$pt14.5s[i] + 0.5*data1_nobil$pt14.5s[i]**2 +   #A145
                             0.5*data1_nobil$pt14.5s[i]*data1_nobil$pt14.5e[i] +
                             (29/4 + 14/2)*data1_nobil$pt14.5e[i]) +

                          (15*data1_nobil$pt15.5s[i] + 0.5*data1_nobil$pt15.5s[i]**2 +   #A155
                             0.5*data1_nobil$pt15.5s[i]*data1_nobil$pt15.5e[i] +
                             (31/4 + 15/2)*data1_nobil$pt15.5e[i]) +

                          (16*data1_nobil$pt16.5s[i] + 0.5*data1_nobil$pt16.5s[i]**2 +   #A165
                             0.5*data1_nobil$pt16.5s[i]*data1_nobil$pt16.5e[i] +
                             (33/4 + 16/2)*data1_nobil$pt16.5e[i]) +

                          (17*data1_nobil$pt17.5s[i] + 0.5*data1_nobil$pt17.5s[i]**2 +   #A175
                             0.5*data1_nobil$pt17.5s[i]*data1_nobil$pt17.5e[i] +
                             (35/4 + 17/2)*data1_nobil$pt17.5e[i]))


data1_nobil$off[i] =   (0.25     +            # A10

                           (-2/3 * data1_nobil$pt2s[i] +
                              (2/3 + 1/2) * data1_nobil$pt2s[i]**2 +
                              (2/3 + 1/2) * data1_nobil$pt2s[i]*data1_nobil$pt2e[i] +
                              (1/4 - 1/3) * data1_nobil$pt2e[i]) + #A20

                           (-2/5 * data1_nobil$pt3s[i] +
                              (2/5 + 1/3) * data1_nobil$pt3s[i]**2 +
                              (2/5 + 1/3) * data1_nobil$pt3s[i]*data1_nobil$pt3e[i] +
                              (1/6 - 1/5) * data1_nobil$pt3e[i]) + #A30

                           (-2/7 * data1_nobil$pt4s[i] +
                              (2/7 + 1/4) * data1_nobil$pt4s[i]**2 +
                              (2/7 + 1/4) * data1_nobil$pt4s[i]*data1_nobil$pt4e[i] +
                              (1/8 - 1/7) * data1_nobil$pt4e[i]) + #A40

                           (-2/9 * data1_nobil$pt5s[i] +
                              (2/9 + 1/5) * data1_nobil$pt5s[i]**2 +
                              (2/9 + 1/5) * data1_nobil$pt5s[i]*data1_nobil$pt5e[i] +
                              (1/10 - 1/9) * data1_nobil$pt5e[i]) + #A50

                           (-2/11 * data1_nobil$pt6s[i] +
                              (2/11 + 1/6) * data1_nobil$pt6s[i]**2 +
                              (2/11 + 1/6) * data1_nobil$pt6s[i]*data1_nobil$pt6e[i] +
                              (1/12 - 1/11) * data1_nobil$pt6e[i]) + #A60

                           (-2/13 * data1_nobil$pt7s[i] +
                              (2/13 + 1/7) * data1_nobil$pt7s[i]**2 +
                              (2/13 + 1/7) * data1_nobil$pt7s[i]*data1_nobil$pt7e[i] +
                              (1/14 - 1/13) * data1_nobil$pt7e[i]) + #A70

                           (-2/15 * data1_nobil$pt8s[i] +
                              (2/15 + 1/8) * data1_nobil$pt8s[i]**2 +
                              (2/15 + 1/8) * data1_nobil$pt8s[i]*data1_nobil$pt8e[i] +
                              (1/16 - 1/15) * data1_nobil$pt8e[i]) + #A80
```

```
(-2/17 * data1_nobil$pt9s[i] +
   (2/17 + 1/9) * data1_nobil$pt9s[i]**2 +
   (2/17 + 1/9) * data1_nobil$pt9s[i]*data1_nobil$pt9e[i] +
   (1/18 - 1/17) * data1_nobil$pt9e[i]) +  #A90

(-2/19 * data1_nobil$pt10s[i] +
   (2/19 + 1/10) * data1_nobil$pt10s[i]**2 +
   (2/19 + 1/10) * data1_nobil$pt10s[i]*data1_nobil$pt10e[i] +
   (1/20 - 1/19) * data1_nobil$pt10e[i]) +  #A100

(-2/21 * data1_nobil$pt11s[i] +
   (2/21 + 1/11) * data1_nobil$pt11s[i]**2 +
   (2/21 + 1/11) * data1_nobil$pt11s[i]*data1_nobil$pt11e[i] +
   (1/22 - 1/21) * data1_nobil$pt11e[i]) +  #A110

(-2/23 * data1_nobil$pt12s[i] +
   (2/23 + 1/12) * data1_nobil$pt12s[i]**2 +
   (2/23 + 1/12) * data1_nobil$pt12s[i]*data1_nobil$pt12e[i] +
   (1/24 - 1/23) * data1_nobil$pt12e[i]) + #A120

(-2/25 * data1_nobil$pt13s[i] +
   (2/25 + 1/13) * data1_nobil$pt13s[i]**2 +
   (2/25 + 1/13) * data1_nobil$pt13s[i]*data1_nobil$pt13e[i] +
   (1/26 - 1/25) * data1_nobil$pt13e[i]) +  #A130

(-2/27 * data1_nobil$pt14s[i] +
   (2/27 + 1/14) * data1_nobil$pt14s[i]**2 +
   (2/27 + 1/14) * data1_nobil$pt14s[i]*data1_nobil$pt14e[i] +
   (1/28 - 1/27) * data1_nobil$pt14e[i]) +  #A140

(-2/29 * data1_nobil$pt15s[i] +
   (2/29 + 1/15) * data1_nobil$pt15s[i]**2 +
   (2/29 + 1/15) * data1_nobil$pt15s[i]*data1_nobil$pt15e[i] +
   (1/30 - 1/29) * data1_nobil$pt15e[i])  +  #A150

(-2/31 * data1_nobil$pt16s[i] +
   (2/31 + 1/16) * data1_nobil$pt16s[i]**2 +
   (2/31 + 1/16) * data1_nobil$pt16s[i]*data1_nobil$pt16e[i] +
   (1/32 - 1/31) * data1_nobil$pt16e[i]) + #A160

(-2/33 * data1_nobil$pt17s[i] +
   (2/33 + 1/17) * data1_nobil$pt17s[i]**2 +
   (2/33 + 1/17) * data1_nobil$pt17s[i]*data1_nobil$pt17e[i] +
   (1/34 - 1/33) * data1_nobil$pt17e[i])  + #A170

(-2/35 * data1_nobil$pt18s[i] +
   (2/35 + 1/18) * data1_nobil$pt18s[i]**2) + #A180


(1/1*data1_nobil$pt1.5s[i] -
   (1/1 + 2/3)*data1_nobil$pt1.5s[i]**2 -
   (1/1 + 2/3)*data1_nobil$pt1.5s[i]*data1_nobil$pt1.5e[i] +
   (-1/3 + 1/2)*data1_nobil$pt1.5e[i]) + #A15
```

```
(1/2*data1_nobil$pt2.5s[i] -
   (1/2 + 2/5)*data1_nobil$pt2.5s[i]**2 -
   (1/2 + 2/5)*data1_nobil$pt2.5s[i]*data1_nobil$pt2.5e[i] +
   (-1/5 + 1/4)*data1_nobil$pt2.5e[i]) + #A25

(1/3*data1_nobil$pt3.5s[i] -
   (1/3 + 2/7)*data1_nobil$pt3.5s[i]**2 -
   (1/3 + 2/7)*data1_nobil$pt3.5s[i]*data1_nobil$pt3.5e[i] +
   (-1/7 + 1/6)*data1_nobil$pt3.5e[i]) + #A35

(1/4*data1_nobil$pt4.5s[i] -
   (1/4 + 2/9)*data1_nobil$pt4.5s[i]**2 -
   (1/4 + 2/9)*data1_nobil$pt4.5s[i]*data1_nobil$pt4.5e[i] +
   (-1/9 + 1/8)*data1_nobil$pt4.5e[i]) + #A45

(1/5*data1_nobil$pt5.5s[i] -
   (1/5 + 2/11)*data1_nobil$pt5.5s[i]**2 -
   (1/5 + 2/11)*data1_nobil$pt5.5s[i]*data1_nobil$pt5.5e[i] +
   (-1/11 + 1/10)*data1_nobil$pt5.5e[i]) + #A55

(1/6*data1_nobil$pt6.5s[i] -
   (1/6 + 2/13)*data1_nobil$pt6.5s[i]**2 -
   (1/6 + 2/13)*data1_nobil$pt6.5s[i]*data1_nobil$pt6.5e[i] +
   (-1/13 + 1/12)*data1_nobil$pt6.5e[i]) + #A65

(1/7*data1_nobil$pt7.5s[i] -
   (1/7 + 2/15)*data1_nobil$pt7.5s[i]**2 -
   (1/7 + 2/15)*data1_nobil$pt7.5s[i]*data1_nobil$pt7.5e[i] +
   (-1/15 + 1/14)*data1_nobil$pt7.5e[i]) + #A75

(1/8*data1_nobil$pt8.5s[i] -
   (1/8 + 2/17)*data1_nobil$pt8.5s[i]**2 -
   (1/8 + 2/17)*data1_nobil$pt8.5s[i]*data1_nobil$pt8.5e[i] +
   (-1/17 + 1/16)*data1_nobil$pt8.5e[i]) + #A85

(1/9*data1_nobil$pt9.5s[i] -
   (1/9 + 2/19)*data1_nobil$pt9.5s[i]**2 -
   (1/9 + 2/19)*data1_nobil$pt9.5s[i]*data1_nobil$pt9.5e[i] +
   (-1/19 + 1/18)*data1_nobil$pt9.5e[i]) + #A95

(1/10*data1_nobil$pt10.5s[i] -
   (1/10 + 2/21)*data1_nobil$pt10.5s[i]**2 -
   (1/10 + 2/21)*data1_nobil$pt10.5s[i]*data1_nobil$pt10.5e[i] +
   (-1/21 + 1/20)*data1_nobil$pt10.5e[i]) + #A105

(1/11*data1_nobil$pt11.5s[i] -
   (1/11 + 2/23)*data1_nobil$pt11.5s[i]**2 -
   (1/11 + 2/23)*data1_nobil$pt11.5s[i]*data1_nobil$pt11.5e[i] +
   (-1/23 + 1/22)*data1_nobil$pt11.5e[i]) + #A115

(1/12*data1_nobil$pt12.5s[i] -
   (1/12 + 2/25)*data1_nobil$pt12.5s[i]**2 -
   (1/12 + 2/25)*data1_nobil$pt12.5s[i]*data1_nobil$pt12.5e[i] +
```

```
                                (-1/25 + 1/24)*data1_nobil$pt12.5e[i]) + #A125

                              (1/13*data1_nobil$pt13.5s[i] -
                                 (1/13 + 2/27)*data1_nobil$pt13.5s[i]**2 -
                                 (1/13 + 2/27)*data1_nobil$pt13.5s[i]*data1_nobil$pt13.5e[i] +
                                 (-1/27 + 1/26)*data1_nobil$pt13.5e[i]) + #A135

                              (1/14*data1_nobil$pt14.5s[i] -
                                 (1/14 + 2/29)*data1_nobil$pt14.5s[i]**2 -
                                 (1/14 + 2/29)*data1_nobil$pt14.5s[i]*data1_nobil$pt14.5e[i] +
                                 (-1/29 + 1/28)*data1_nobil$pt14.5e[i]) + #A145

                              (1/15*data1_nobil$pt15.5s[i] -
                                 (1/15 + 2/31)*data1_nobil$pt15.5s[i]**2 -
                                 (1/15 + 2/31)*data1_nobil$pt15.5s[i]*data1_nobil$pt15.5e[i] +
                                 (-1/31 + 1/30)*data1_nobil$pt15.5e[i]) + #A155

                              (1/16*data1_nobil$pt16.5s[i] -
                                 (1/16 + 2/33)*data1_nobil$pt16.5s[i]**2 -
                                 (1/16 + 2/33)*data1_nobil$pt16.5s[i]*data1_nobil$pt16.5e[i] +
                                 (-1/33 + 1/32)*data1_nobil$pt16.5e[i]) + #A165

                              (1/17*data1_nobil$pt17.5s[i] -
                                 (1/17 + 2/35)*data1_nobil$pt17.5s[i]**2 -
                                 (1/17 + 2/35)*data1_nobil$pt17.5s[i]*data1_nobil$pt17.5e[i] +
                                 (-1/35 + 1/34)*data1_nobil$pt17.5e[i])) #A175

  # Add: Calculate covariates for age, treatment etc
  data1_nobil$agerisk[i]=data1_nobil$C0[i]*data1_nobil$AGERF[i]
  data1_nobil$famrisk[i]=data1_nobil$C0[i]*data1_nobil$FAMHIST[i]
  data1_nobil$bioprisk[i]=data1_nobil$C0[i]*data1_nobil$TREAT_BIOP[i]
  data1_nobil$lumpnoradrisk[i]=data1_nobil$C0[i]*data1_nobil$TREAT_LUMP_NORAD[i]
  data1_nobil$lumpradrisk[i]=data1_nobil$C0[i]*data1_nobil$TREAT_LUMP_RAD[i]
  data1_nobil$endorisk[i]=data1_nobil$C0[i]*data1_nobil$ENDOCRINEDX0[i]

}
## Poisson regression with identity link to estimate C, beta, and risk factor coefficients
coefini0=coef(glm(MAMMRI ~ -1 + C0 + offset(off), data=data1_nobil, family=poisson(link="identity")))

calc_coeff1= glm(MAMMRI ~ -1 + C0 + offset(off) + beta + agerisk + famrisk + bioprisk
                 + lumpnoradrisk + lumpradrisk + endorisk, data=data1_nobil,
                 family=poisson(link="identity"), start=c(coefini0,0,0,0,0,0,0,0))

coeffs1=coef((calc_coeff1))
#write.csv(coeffs,"coeffs_oneByX_noyr1.csv")

#################################
##### Regression Model 2 ########
#################################
# with year 1:
data1_nobil=longset4

#################################
########### Regression #########
```

```
################################
# initialization of parameters before loop:
iter=0 # number of iterations
param_old=param_new=c(1,0,0)
# initial values for C0, beta, and space-holder for deviance in position 4
#lambda=0 # initial value for lambda
#a0=0 # initial value for A0
param_hist=c() # store parameter history

# Update covariates to estimate C and beta for each particpant
for (i in 1:nrow(data1_nobil))
{
  data1_nobil$C0[i] = (0.25 +                #A10
                         data1_nobil$pt1.5s[i] + data1_nobil$pt1.5e[i] + #A15
                         data1_nobil$pt2s[i] + data1_nobil$pt2e[i] + #A20
                         data1_nobil$pt2.5s[i] + data1_nobil$pt2.5e[i] + #A25
                         data1_nobil$pt3s[i] + data1_nobil$pt3e[i] + #A30
                         data1_nobil$pt3.5s[i] + data1_nobil$pt3.5e[i] + #A35
                         data1_nobil$pt4s[i] + data1_nobil$pt4e[i] + #A40
                         data1_nobil$pt4.5s[i] + data1_nobil$pt4.5e[i] + #A45
                         data1_nobil$pt5s[i] + data1_nobil$pt5e[i] + #A50
                         data1_nobil$pt5.5s[i] + data1_nobil$pt5.5e[i] + #A55
                         data1_nobil$pt6s[i] + data1_nobil$pt6e[i] + #A60
                         data1_nobil$pt6.5s[i] + data1_nobil$pt6.5e[i] + #A65
                         data1_nobil$pt7s[i] + data1_nobil$pt7e[i] + #A70
                         data1_nobil$pt7.5s[i] + data1_nobil$pt7.5e[i] + #A75
                         data1_nobil$pt8s[i] + data1_nobil$pt8e[i] + #A80
                         data1_nobil$pt8.5s[i] + data1_nobil$pt8.5e[i] + #A85
                         data1_nobil$pt9s[i] + data1_nobil$pt9e[i] + #A90

                         data1_nobil$pt9.5s[i] + data1_nobil$pt9.5e[i] + #A95
                         data1_nobil$pt10s[i] + data1_nobil$pt10e[i] + #A100
                         data1_nobil$pt10.5s[i] + data1_nobil$pt10.5e[i] + #A105
                         data1_nobil$pt11s[i] + data1_nobil$pt11e[i] + #A110
                         data1_nobil$pt11.5s[i] + data1_nobil$pt11.5e[i] + #A115
                         data1_nobil$pt12s[i] + data1_nobil$pt12e[i] + #A120
                         data1_nobil$pt12.5s[i] + data1_nobil$pt12.5e[i] + #A125
                         data1_nobil$pt13s[i] + data1_nobil$pt13e[i] + #A130
                         data1_nobil$pt13.5s[i] + data1_nobil$pt13.5e[i] + #A135
                         data1_nobil$pt14s[i] + data1_nobil$pt14e[i] + #A140
                         data1_nobil$pt14.5s[i] + data1_nobil$pt14.5e[i] + #A145
                         data1_nobil$pt15s[i] + data1_nobil$pt15e[i] + #A150
                         data1_nobil$pt15.5s[i] + data1_nobil$pt15.5e[i] + #A155
                         data1_nobil$pt16s[i] + data1_nobil$pt16e[i] + #A160
                         data1_nobil$pt16.5s[i] + data1_nobil$pt16.5e[i] + #A165
                         data1_nobil$pt17s[i] + data1_nobil$pt17e[i] + #A170
                         data1_nobil$pt17.5s[i] + data1_nobil$pt17.5e[i] + #A175
                         data1_nobil$pt18s[i])  #A180


  data1_nobil$beta[i] = (0.25 +                     # A10

                           (3/2*data1_nobil$pt2s[i] + 0.5*data1_nobil$pt2s[i]**2 +
                              0.5*data1_nobil$pt2s[i]*data1_nobil$pt2e[i] +
```

```
   (2/2 + 3/4)*data1_nobil$pt2e[i]) +                    #A20

(5/2*data1_nobil$pt3s[i] + 0.5*data1_nobil$pt3s[i]**2 +
   0.5*data1_nobil$pt3s[i]*data1_nobil$pt3e[i] +
   (3/2 + 5/4)*data1_nobil$pt3e[i]) +                    #A30

(7/2*data1_nobil$pt4s[i] + 0.5*data1_nobil$pt4s[i]**2 +
   0.5*data1_nobil$pt4s[i]*data1_nobil$pt4e[i] +
   (4/2 + 7/4)*data1_nobil$pt4e[i]) +                    #A40

(9/2*data1_nobil$pt5s[i] + 0.5*data1_nobil$pt5s[i]**2 +
   0.5*data1_nobil$pt5s[i]*data1_nobil$pt5e[i] +
   (5/2 + 9/4)*data1_nobil$pt5e[i]) +                    #A50

(11/2*data1_nobil$pt6s[i] + 0.5*data1_nobil$pt6s[i]**2 +
   0.5*data1_nobil$pt6s[i]*data1_nobil$pt6e[i] +
   (6/2 + 11/4)*data1_nobil$pt6e[i]) +                    #A60

(13/2*data1_nobil$pt7s[i] + 0.5*data1_nobil$pt7s[i]**2 +
   0.5*data1_nobil$pt7s[i]*data1_nobil$pt7e[i] +
   (7/2 + 13/4)*data1_nobil$pt7e[i]) +                    #A70

(15/2*data1_nobil$pt8s[i] + 0.5*data1_nobil$pt8s[i]**2 +
   0.5*data1_nobil$pt8s[i]*data1_nobil$pt8e[i] +
   (8/2 + 15/4)*data1_nobil$pt8e[i]) +                    #A80

(17/2*data1_nobil$pt9s[i] + 0.5*data1_nobil$pt9s[i]**2 +
   0.5*data1_nobil$pt9s[i]*data1_nobil$pt9e[i] +
   (9/2 + 17/4)*data1_nobil$pt9e[i]) +                    #A90

(19/2*data1_nobil$pt10s[i] + 0.5*data1_nobil$pt10s[i]**2 +
   0.5*data1_nobil$pt10s[i]*data1_nobil$pt10e[i] +
   (10/2 + 19/4)*data1_nobil$pt10e[i]) +                    #A100

(21/2*data1_nobil$pt11s[i] + 0.5*data1_nobil$pt11s[i]**2 +
   0.5*data1_nobil$pt11s[i]*data1_nobil$pt11e[i] +
   (11/2 + 21/4)*data1_nobil$pt11e[i]) +                    #A110

(23/2*data1_nobil$pt12s[i] + 0.5*data1_nobil$pt12s[i]**2 +
   0.5*data1_nobil$pt12s[i]*data1_nobil$pt12e[i] +
   (12/2 + 23/4)*data1_nobil$pt12e[i]) +                    #A120

(25/2*data1_nobil$pt13s[i] + 0.5*data1_nobil$pt13s[i]**2 +
   0.5*data1_nobil$pt13s[i]*data1_nobil$pt13e[i] +
   (13/2 + 25/4)*data1_nobil$pt13e[i]) +                    #A130

(27/2*data1_nobil$pt14s[i] + 0.5*data1_nobil$pt14s[i]**2 +
   0.5*data1_nobil$pt14s[i]*data1_nobil$pt14e[i] +
   (14/2 + 27/4)*data1_nobil$pt14e[i]) +                    #A140

(29/2*data1_nobil$pt15s[i] + 0.5*data1_nobil$pt15s[i]**2 +
   0.5*data1_nobil$pt15s[i]*data1_nobil$pt15e[i] +
   (15/2 + 29/4)*data1_nobil$pt15e[i]) +                    #A150
```

```
(31/2*data1_nobil$pt16s[i] + 0.5*data1_nobil$pt16s[i]**2 +
   0.5*data1_nobil$pt16s[i]*data1_nobil$pt16e[i] +
   (16/2 + 31/4)*data1_nobil$pt16e[i]) +                    #A160

(33/2*data1_nobil$pt17s[i] + 0.5*data1_nobil$pt17s[i]**2 +
   0.5*data1_nobil$pt17s[i]*data1_nobil$pt17e[i] +
   (17/2 + 33/4)*data1_nobil$pt17e[i]) +                     #A170

(35/2*data1_nobil$pt18s[i] + 0.5*data1_nobil$pt18s[i]**2) +  #A180



(1*data1_nobil$pt1.5s[i] + 0.5*data1_nobil$pt1.5s[i]**2 +
   0.5*data1_nobil$pt1.5s[i]*data1_nobil$pt1.5e[i] +
   (3/4 + 1/2)*data1_nobil$pt1.5e[i]) +                     #A15

(2*data1_nobil$pt2.5s[i] + 0.5*data1_nobil$pt2.5s[i]**2 +   #A25
   0.5*data1_nobil$pt2.5s[i]*data1_nobil$pt2.5e[i] +
   (5/4 + 2/2)*data1_nobil$pt2.5e[i]) +

(3*data1_nobil$pt3.5s[i] + 0.5*data1_nobil$pt3.5s[i]**2 +   #A35
   0.5*data1_nobil$pt3.5s[i]*data1_nobil$pt3.5e[i] +
   (7/4 + 3/2)*data1_nobil$pt3.5e[i]) +

(4*data1_nobil$pt4.5s[i] + 0.5*data1_nobil$pt4.5s[i]**2 +   #A45
   0.5*data1_nobil$pt4.5s[i]*data1_nobil$pt4.5e[i] +
   (9/4 + 4/2)*data1_nobil$pt4.5e[i]) +

(5*data1_nobil$pt5.5s[i] + 0.5*data1_nobil$pt5.5s[i]**2 +   #A55
   0.5*data1_nobil$pt5.5s[i]*data1_nobil$pt5.5e[i] +
   (11/4 + 5/2)*data1_nobil$pt5.5e[i]) +

(6*data1_nobil$pt6.5s[i] + 0.5*data1_nobil$pt6.5s[i]**2 +   #A65
   0.5*data1_nobil$pt6.5s[i]*data1_nobil$pt6.5e[i] +
   (13/4 + 6/2)*data1_nobil$pt6.5e[i]) +

(7*data1_nobil$pt7.5s[i] + 0.5*data1_nobil$pt7.5s[i]**2 +   #A75
   0.5*data1_nobil$pt7.5s[i]*data1_nobil$pt7.5e[i] +
   (15/4 + 7/2)*data1_nobil$pt7.5e[i]) +

(8*data1_nobil$pt8.5s[i] + 0.5*data1_nobil$pt8.5s[i]**2 +   #A85
   0.5*data1_nobil$pt8.5s[i]*data1_nobil$pt8.5e[i] +
   (17/4 + 8/2)*data1_nobil$pt8.5e[i]) +

(9*data1_nobil$pt9.5s[i] + 0.5*data1_nobil$pt9.5s[i]**2 +   #A95
   0.5*data1_nobil$pt9.5s[i]*data1_nobil$pt9.5e[i] +
   (19/4 + 9/2)*data1_nobil$pt9.5e[i]) +

(10*data1_nobil$pt10.5s[i] + 0.5*data1_nobil$pt10.5s[i]**2 +   #A105
   0.5*data1_nobil$pt10.5s[i]*data1_nobil$pt10.5e[i] +
   (21/4 + 10/2)*data1_nobil$pt10.5e[i]) +

(11*data1_nobil$pt11.5s[i] + 0.5*data1_nobil$pt11.5s[i]**2 +   #A115
   0.5*data1_nobil$pt11.5s[i]*data1_nobil$pt11.5e[i] +
```

```
                          (23/4 + 11/2)*data1_nobil$pt11.5e[i]) +

                        (12*data1_nobil$pt12.5s[i] + 0.5*data1_nobil$pt12.5s[i]**2 +   #A125
                          0.5*data1_nobil$pt12.5s[i]*data1_nobil$pt12.5e[i] +
                          (25/4 + 12/2)*data1_nobil$pt12.5e[i]) +

                        (13*data1_nobil$pt13.5s[i] + 0.5*data1_nobil$pt13.5s[i]**2 +   #A135
                          0.5*data1_nobil$pt13.5s[i]*data1_nobil$pt13.5e[i] +
                          (27/4 + 13/2)*data1_nobil$pt13.5e[i]) +

                        (14*data1_nobil$pt14.5s[i] + 0.5*data1_nobil$pt14.5s[i]**2 +   #A145
                          0.5*data1_nobil$pt14.5s[i]*data1_nobil$pt14.5e[i] +
                          (29/4 + 14/2)*data1_nobil$pt14.5e[i]) +

                        (15*data1_nobil$pt15.5s[i] + 0.5*data1_nobil$pt15.5s[i]**2 +   #A155
                          0.5*data1_nobil$pt15.5s[i]*data1_nobil$pt15.5e[i] +
                          (31/4 + 15/2)*data1_nobil$pt15.5e[i]) +

                        (16*data1_nobil$pt16.5s[i] + 0.5*data1_nobil$pt16.5s[i]**2 +   #A165
                          0.5*data1_nobil$pt16.5s[i]*data1_nobil$pt16.5e[i] +
                          (33/4 + 16/2)*data1_nobil$pt16.5e[i]) +

                        (17*data1_nobil$pt17.5s[i] + 0.5*data1_nobil$pt17.5s[i]**2 +   #A175
                          0.5*data1_nobil$pt17.5s[i]*data1_nobil$pt17.5e[i] +
                          (35/4 + 17/2)*data1_nobil$pt17.5e[i]))


data1_nobil$off[i] =   (0.25     +            # A10

                         (-2/3 * data1_nobil$pt2s[i] +
                            (2/3 + 1/2) * data1_nobil$pt2s[i]**2 +
                            (2/3 + 1/2) * data1_nobil$pt2s[i]*data1_nobil$pt2e[i] +
                            (1/4 - 1/3) * data1_nobil$pt2e[i]) + #A20

                         (-2/5 * data1_nobil$pt3s[i] +
                            (2/5 + 1/3) * data1_nobil$pt3s[i]**2 +
                            (2/5 + 1/3) * data1_nobil$pt3s[i]*data1_nobil$pt3e[i] +
                            (1/6 - 1/5) * data1_nobil$pt3e[i]) + #A30

                         (-2/7 * data1_nobil$pt4s[i] +
                            (2/7 + 1/4) * data1_nobil$pt4s[i]**2 +
                            (2/7 + 1/4) * data1_nobil$pt4s[i]*data1_nobil$pt4e[i] +
                            (1/8 - 1/7) * data1_nobil$pt4e[i]) + #A40

                         (-2/9 * data1_nobil$pt5s[i] +
                            (2/9 + 1/5) * data1_nobil$pt5s[i]**2 +
                            (2/9 + 1/5) * data1_nobil$pt5s[i]*data1_nobil$pt5e[i] +
                            (1/10 - 1/9) * data1_nobil$pt5e[i]) + #A50

                         (-2/11 * data1_nobil$pt6s[i] +
                            (2/11 + 1/6) * data1_nobil$pt6s[i]**2 +
                            (2/11 + 1/6) * data1_nobil$pt6s[i]*data1_nobil$pt6e[i] +
                            (1/12 - 1/11) * data1_nobil$pt6e[i]) + #A60
```

```
(-2/13 * data1_nobil$pt7s[i] +
   (2/13 + 1/7) * data1_nobil$pt7s[i]**2 +
   (2/13 + 1/7) * data1_nobil$pt7s[i]*data1_nobil$pt7e[i] +
   (1/14 - 1/13) * data1_nobil$pt7e[i]) + #A70

(-2/15 * data1_nobil$pt8s[i] +
   (2/15 + 1/8) * data1_nobil$pt8s[i]**2 +
   (2/15 + 1/8) * data1_nobil$pt8s[i]*data1_nobil$pt8e[i] +
   (1/16 - 1/15) * data1_nobil$pt8e[i]) + #A80

(-2/17 * data1_nobil$pt9s[i] +
   (2/17 + 1/9) * data1_nobil$pt9s[i]**2 +
   (2/17 + 1/9) * data1_nobil$pt9s[i]*data1_nobil$pt9e[i] +
   (1/18 - 1/17) * data1_nobil$pt9e[i]) +  #A90

(-2/19 * data1_nobil$pt10s[i] +
   (2/19 + 1/10) * data1_nobil$pt10s[i]**2 +
   (2/19 + 1/10) * data1_nobil$pt10s[i]*data1_nobil$pt10e[i] +
   (1/20 - 1/19) * data1_nobil$pt10e[i]) +  #A100

(-2/21 * data1_nobil$pt11s[i] +
   (2/21 + 1/11) * data1_nobil$pt11s[i]**2 +
   (2/21 + 1/11) * data1_nobil$pt11s[i]*data1_nobil$pt11e[i] +
   (1/22 - 1/21) * data1_nobil$pt11e[i]) +  #A110

(-2/23 * data1_nobil$pt12s[i] +
   (2/23 + 1/12) * data1_nobil$pt12s[i]**2 +
   (2/23 + 1/12) * data1_nobil$pt12s[i]*data1_nobil$pt12e[i] +
   (1/24 - 1/23) * data1_nobil$pt12e[i]) + #A120

(-2/25 * data1_nobil$pt13s[i] +
   (2/25 + 1/13) * data1_nobil$pt13s[i]**2 +
   (2/25 + 1/13) * data1_nobil$pt13s[i]*data1_nobil$pt13e[i] +
   (1/26 - 1/25) * data1_nobil$pt13e[i]) +  #A130

(-2/27 * data1_nobil$pt14s[i] +
   (2/27 + 1/14) * data1_nobil$pt14s[i]**2 +
   (2/27 + 1/14) * data1_nobil$pt14s[i]*data1_nobil$pt14e[i] +
   (1/28 - 1/27) * data1_nobil$pt14e[i]) +  #A140

(-2/29 * data1_nobil$pt15s[i] +
   (2/29 + 1/15) * data1_nobil$pt15s[i]**2 +
   (2/29 + 1/15) * data1_nobil$pt15s[i]*data1_nobil$pt15e[i] +
   (1/30 - 1/29) * data1_nobil$pt15e[i])  +  #A150

(-2/31 * data1_nobil$pt16s[i] +
   (2/31 + 1/16) * data1_nobil$pt16s[i]**2 +
   (2/31 + 1/16) * data1_nobil$pt16s[i]*data1_nobil$pt16e[i] +
   (1/32 - 1/31) * data1_nobil$pt16e[i]) + #A160

(-2/33 * data1_nobil$pt17s[i] +
   (2/33 + 1/17) * data1_nobil$pt17s[i]**2 +
   (2/33 + 1/17) * data1_nobil$pt17s[i]*data1_nobil$pt17e[i] +
   (1/34 - 1/33) * data1_nobil$pt17e[i])  + #A170
```

```
(-2/35 * data1_nobil$pt18s[i] +
   (2/35 + 1/18) * data1_nobil$pt18s[i]**2) + #A180


(1/1*data1_nobil$pt1.5s[i] -
   (1/1 + 2/3)*data1_nobil$pt1.5s[i]**2 -
   (1/1 + 2/3)*data1_nobil$pt1.5s[i]*data1_nobil$pt1.5e[i] +
   (-1/3 + 1/2)*data1_nobil$pt1.5e[i]) + #A15

(1/2*data1_nobil$pt2.5s[i] -
   (1/2 + 2/5)*data1_nobil$pt2.5s[i]**2 -
   (1/2 + 2/5)*data1_nobil$pt2.5s[i]*data1_nobil$pt2.5e[i] +
   (-1/5 + 1/4)*data1_nobil$pt2.5e[i]) + #A25

(1/3*data1_nobil$pt3.5s[i] -
   (1/3 + 2/7)*data1_nobil$pt3.5s[i]**2 -
   (1/3 + 2/7)*data1_nobil$pt3.5s[i]*data1_nobil$pt3.5e[i] +
   (-1/7 + 1/6)*data1_nobil$pt3.5e[i]) + #A35

(1/4*data1_nobil$pt4.5s[i] -
   (1/4 + 2/9)*data1_nobil$pt4.5s[i]**2 -
   (1/4 + 2/9)*data1_nobil$pt4.5s[i]*data1_nobil$pt4.5e[i] +
   (-1/9 + 1/8)*data1_nobil$pt4.5e[i]) + #A45

(1/5*data1_nobil$pt5.5s[i] -
   (1/5 + 2/11)*data1_nobil$pt5.5s[i]**2 -
   (1/5 + 2/11)*data1_nobil$pt5.5s[i]*data1_nobil$pt5.5e[i] +
   (-1/11 + 1/10)*data1_nobil$pt5.5e[i]) + #A55

(1/6*data1_nobil$pt6.5s[i] -
   (1/6 + 2/13)*data1_nobil$pt6.5s[i]**2 -
   (1/6 + 2/13)*data1_nobil$pt6.5s[i]*data1_nobil$pt6.5e[i] +
   (-1/13 + 1/12)*data1_nobil$pt6.5e[i]) + #A65

(1/7*data1_nobil$pt7.5s[i] -
   (1/7 + 2/15)*data1_nobil$pt7.5s[i]**2 -
   (1/7 + 2/15)*data1_nobil$pt7.5s[i]*data1_nobil$pt7.5e[i] +
   (-1/15 + 1/14)*data1_nobil$pt7.5e[i]) + #A75

(1/8*data1_nobil$pt8.5s[i] -
   (1/8 + 2/17)*data1_nobil$pt8.5s[i]**2 -
   (1/8 + 2/17)*data1_nobil$pt8.5s[i]*data1_nobil$pt8.5e[i] +
   (-1/17 + 1/16)*data1_nobil$pt8.5e[i]) + #A85

(1/9*data1_nobil$pt9.5s[i] -
   (1/9 + 2/19)*data1_nobil$pt9.5s[i]**2 -
   (1/9 + 2/19)*data1_nobil$pt9.5s[i]*data1_nobil$pt9.5e[i] +
   (-1/19 + 1/18)*data1_nobil$pt9.5e[i]) + #A95

(1/10*data1_nobil$pt10.5s[i] -
   (1/10 + 2/21)*data1_nobil$pt10.5s[i]**2 -
   (1/10 + 2/21)*data1_nobil$pt10.5s[i]*data1_nobil$pt10.5e[i] +
   (-1/21 + 1/20)*data1_nobil$pt10.5e[i]) + #A105
```

```
                              (1/11*data1_nobil$pt11.5s[i] -
                                 (1/11 + 2/23)*data1_nobil$pt11.5s[i]**2 -
                                 (1/11 + 2/23)*data1_nobil$pt11.5s[i]*data1_nobil$pt11.5e[i] +
                                 (-1/23 + 1/22)*data1_nobil$pt11.5e[i]) + #A115

                              (1/12*data1_nobil$pt12.5s[i] -
                                 (1/12 + 2/25)*data1_nobil$pt12.5s[i]**2 -
                                 (1/12 + 2/25)*data1_nobil$pt12.5s[i]*data1_nobil$pt12.5e[i] +
                                 (-1/25 + 1/24)*data1_nobil$pt12.5e[i]) + #A125

                              (1/13*data1_nobil$pt13.5s[i] -
                                 (1/13 + 2/27)*data1_nobil$pt13.5s[i]**2 -
                                 (1/13 + 2/27)*data1_nobil$pt13.5s[i]*data1_nobil$pt13.5e[i] +
                                 (-1/27 + 1/26)*data1_nobil$pt13.5e[i]) + #A135

                              (1/14*data1_nobil$pt14.5s[i] -
                                 (1/14 + 2/29)*data1_nobil$pt14.5s[i]**2 -
                                 (1/14 + 2/29)*data1_nobil$pt14.5s[i]*data1_nobil$pt14.5e[i] +
                                 (-1/29 + 1/28)*data1_nobil$pt14.5e[i]) + #A145

                              (1/15*data1_nobil$pt15.5s[i] -
                                 (1/15 + 2/31)*data1_nobil$pt15.5s[i]**2 -
                                 (1/15 + 2/31)*data1_nobil$pt15.5s[i]*data1_nobil$pt15.5e[i] +
                                 (-1/31 + 1/30)*data1_nobil$pt15.5e[i]) + #A155

                              (1/16*data1_nobil$pt16.5s[i] -
                                 (1/16 + 2/33)*data1_nobil$pt16.5s[i]**2 -
                                 (1/16 + 2/33)*data1_nobil$pt16.5s[i]*data1_nobil$pt16.5e[i] +
                                 (-1/33 + 1/32)*data1_nobil$pt16.5e[i]) + #A165

                              (1/17*data1_nobil$pt17.5s[i] -
                                 (1/17 + 2/35)*data1_nobil$pt17.5s[i]**2 -
                                 (1/17 + 2/35)*data1_nobil$pt17.5s[i]*data1_nobil$pt17.5e[i] +
                                 (-1/35 + 1/34)*data1_nobil$pt17.5e[i])) #A175

  # Add: Calculate covariates for age, treatment etc
  data1_nobil$agerisk[i]=data1_nobil$C0[i]*data1_nobil$AGERF[i]
  data1_nobil$famrisk[i]=data1_nobil$C0[i]*data1_nobil$FAMHIST[i]
  data1_nobil$bioprisk[i]=data1_nobil$C0[i]*data1_nobil$TREAT_BIOP[i]
  data1_nobil$lumpnoradrisk[i]=data1_nobil$C0[i]*data1_nobil$TREAT_LUMP_NORAD[i]
  data1_nobil$lumpradrisk[i]=data1_nobil$C0[i]*data1_nobil$TREAT_LUMP_RAD[i]
  data1_nobil$endorisk[i]=data1_nobil$C0[i]*data1_nobil$ENDOCRINEDX0[i]

}

coefini=coef(glm(MAMMRI ~ -1 + C0 + offset(off), data=data1_nobil, family=poisson(link="identity")))

## Poisson regression with identity link to estimate C, beta, and risk factor coefficients
calc_coeff2= glm(MAMMRI ~ -1 + C0 + offset(off) + beta + agerisk + famrisk + bioprisk
                 + lumpnoradrisk + lumpradrisk + endorisk, data=data1_nobil,
                 family=poisson(link="identity"),start=c(coefini,0,0,0,0,0,0,0))

coeffs2=coef(calc_coeff2)
```

```
#write.csv(coeffs2,"coeffs_oneByX_withyr1.csv")

# combine separate outputs from model 1 into one ouput
PE12=c(PE1,PE2) # average model without year 1
PE34=c(PE3,PE4) # average model with year 1


#PE1imp=rbind(PE1imp,PE1)
#PE2imp=rbind(PE2imp,PE2)
#PE3imp=rbind(PE3imp,PE3)
#PE4imp=rbind(PE4imp,PE4)
PE12imp=rbind(PE12imp,PE12)
PE34imp=rbind(PE34imp,PE34)
coeffs1imp=rbind(coeffs1imp,coeffs1) # average model without year 1
coeffs2imp=rbind(coeffs2imp,coeffs2) # average model with year 1

impiter=impiter+1

} #end imputation iteration

# Calculate average across 5 imputations
avgPE12=apply(PE12imp,2, mean)
avgPE34=apply(PE34imp,2, mean)
avgcoeffs1=apply(coeffs1imp,2, mean)
avgcoeffs2=apply(coeffs2imp,2, mean)

# Collect average estimates across boot straps
distPE12=rbind(distPE12,avgPE12)
distPE34=rbind(distPE34,avgPE34)
distcoeffs1=rbind(distcoeffs1,avgcoeffs1)
distcoeffs2=rbind(distcoeffs2,avgcoeffs2)

bootiter=bootiter+1
}

write.csv(distPE12,"out12.csv")
write.csv(distPE34,"out34.csv")
write.csv(distcoeffs1,"out5.csv")
write.csv(distcoeffs2,"out6.csv")


# person time calculation:
data_nobil=read.csv("longdcis.csv", header=TRUE)
dim(data_nobil)
summary(data_nobil)


for (i in 1:nrow(data_nobil))
{


 #if (data_nobil$T_SINCE_DIAG_YRS[11] > 8 &
 data_nobil$T_SINCE_DIAG_YRS[11] <= 9){
   # calculate start and end time
```

```
   (data_nobil$start[i]=data_nobil$T_SINCE_DIAG_YRS2[i] - data_nobil$PTIMETOTAL[i])
   (data_nobil$end[i]=data_nobil$T_SINCE_DIAG_YRS2[i])
   # distribute forward until end time
   # find next 'full half number' since start time
   (data_nobil$startmodulo[i] =  (data_nobil$start[i] %/% 0.5))
   (data_nobil$startfullhalf[i] = data_nobil$startmodulo[i]*0.5 + 0.5)
   # rest time before first full half:
   (data_nobil$startrest[i] = data_nobil$startfullhalf[i] - data_nobil$start[i])
   # find last 'full half number' before end time
   (data_nobil$endmodulo[i] =  (data_nobil$end[i] %/% 0.5))
   (data_nobil$endfullhalf[i] = data_nobil$endmodulo[i]*0.5)
   # rest time after last full half
   (data_nobil$endrest[i] = data_nobil$end[i] - data_nobil$endfullhalf[i])
   # How many full halves between first and last full halves?
   (data_nobil$fullhalves[i] = (data_nobil$endfullhalf[i] - data_nobil$startfullhalf[i]) / 0.5)
  for (j in 1:data_nobil$fullhalves[i]) {
     #  assign(paste("data_nobil$pt", data_nobil$startfullhalf[11] + i, "s", sep = ""), 0.25)
     var1=paste("pt", data_nobil$startfullhalf[i] + (j*0.5), "s", sep = "")
     var2=paste("pt", data_nobil$startfullhalf[i] + (j*0.5), "e", sep = "") #starts one too early
     var3=paste("pt", data_nobil$endfullhalf[i] + 0.5, "s", sep = "")
     var4=paste("pt", data_nobil$startfullhalf[i], "e", sep = "")
     data_nobil[i,var1]=0.25
     data_nobil[i,var2]=0.25
     data_nobil[i,var3]=data_nobil$endrest[i]
     data_nobil[i,var4]=data_nobil$startrest[i]
  }

}

#summary(data_nobil)
#dim(data_nobil)
#data_nobil[1:5,]
# find the number of the column with first ptime variable

startcol=which( colnames(data_nobil)=="pt1s" )

#data_nobil[1:3,startcol:ncol(data_nobil)]

# Replace NA values with 0 in person time variables
for (i in startcol:ncol(data_nobil))
{
data_nobil[,i][is.na(data_nobil[,i])] <- 0
}

summary(data_nobil[,startcol:ncol(data_nobil)])


#### For calculation of average rates, calculate annual person times ######
for (i in 1:nrow(data_nobil)) {
  data_nobil$yr1[i] = data_nobil$pt1s[i] + data_nobil$pt1e[i] +
  data_nobil$pt0.5e[i]
  data_nobil$yr2[i] = data_nobil$pt1.5s[i] + data_nobil$pt1.5e[i] +
  data_nobil$pt2s[i] + data_nobil$pt2e[i]
  data_nobil$yr3[i] = data_nobil$pt2.5s[i] + data_nobil$pt2.5e[i] +
```

```
  data_nobil$pt3s[i] + data_nobil$pt3e[i]
  data_nobil$yr4[i] = data_nobil$pt3.5s[i] + data_nobil$pt3.5e[i] +
  data_nobil$pt4s[i] + data_nobil$pt4e[i]
  data_nobil$yr5[i] = data_nobil$pt4.5s[i] + data_nobil$pt4.5e[i] +
  data_nobil$pt5s[i] + data_nobil$pt5e[i]
  data_nobil$yr6[i] = data_nobil$pt5.5s[i] + data_nobil$pt5.5e[i] +
  data_nobil$pt6s[i] + data_nobil$pt6e[i]
  data_nobil$yr7[i] = data_nobil$pt6.5s[i] + data_nobil$pt6.5e[i] +
  data_nobil$pt7s[i] + data_nobil$pt7e[i]
  data_nobil$yr8[i] = data_nobil$pt7.5s[i] + data_nobil$pt7.5e[i] +
  data_nobil$pt8s[i] + data_nobil$pt8e[i]
  data_nobil$yr9[i] = data_nobil$pt8.5s[i] + data_nobil$pt8.5e[i] +
  data_nobil$pt9s[i] + data_nobil$pt9e[i]
  data_nobil$yr10[i] = data_nobil$pt9.5s[i] + data_nobil$pt9.5e[i] +
  data_nobil$pt10s[i] + data_nobil$pt10e[i]
  data_nobil$yr11[i] = data_nobil$pt10.5s[i] + data_nobil$pt10.5e[i] +
  data_nobil$pt11s[i] + data_nobil$pt11e[i]
  data_nobil$yr12[i] = data_nobil$pt11.5s[i] + data_nobil$pt11.5e[i] +
  data_nobil$pt12s[i] + data_nobil$pt12e[i]
  data_nobil$yr13[i] = data_nobil$pt12.5s[i] + data_nobil$pt12.5e[i] +
  data_nobil$pt13s[i] + data_nobil$pt13e[i]
  data_nobil$yr14[i] = data_nobil$pt13.5s[i] + data_nobil$pt13.5e[i] +
  data_nobil$pt14s[i] + data_nobil$pt14e[i]
  data_nobil$yr15[i] = data_nobil$pt14.5s[i] + data_nobil$pt14.5e[i] +
  data_nobil$pt15s[i] + data_nobil$pt15e[i]
  data_nobil$yr16[i] = data_nobil$pt15.5s[i] + data_nobil$pt15.5e[i] +
  data_nobil$pt16s[i] + data_nobil$pt16e[i]
  data_nobil$yr17[i] = data_nobil$pt16.5s[i] + data_nobil$pt16.5e[i] +
  data_nobil$pt17s[i] + data_nobil$pt17e[i]
  data_nobil$yr18[i] = data_nobil$pt17.5s[i] + data_nobil$pt17.5e[i] +
  data_nobil$pt18s[i]

  # first 2 years combined
  data_nobil$yr1_2[i] = data_nobil$pt1s[i] + data_nobil$pt1e[i] + data_nobil$pt0.5e[i] +
  data_nobil$pt1.5s[i] + data_nobil$pt1.5e[i] + data_nobil$pt2s[i] + data_nobil$pt2e[i]
}

write.csv(data_nobil,"ptimedcis.csv")


# calculate credible intervals bases on bootstraps
inhighyr1=read.csv("in_avgmod_highuse_yr1.csv", header=TRUE)
inlowyr1=read.csv("in_avgmod_lowuse_yr1.csv", header=TRUE)
inmedyr1=read.csv("in_avgmod_meduse_yr1.csv", header=TRUE)

# calculate highest density credible 95% interval for each variable
library(coda)

credint95_highy1=c()
for(i in 1:ncol(inhighyr1)){
  mcmcobj=as.mcmc(inhighyr1[,i])
  interval=HPDinterval(mcmcobj, 0.95)
  credint95_highy1= rbind(credint95_highy1,interval)
}
```

```
credint95_lowy1=c()
for(i in 1:ncol(inlowyr1)){
  mcmcobj=as.mcmc(inlowyr1[,i])
  interval=HPDinterval(mcmcobj, 0.95)
  credint95_lowy1= rbind(credint95_lowy1,interval)
}

credint95_medy1=c()
for(i in 1:ncol(inmedyr1)){
  mcmcobj=as.mcmc(inmedyr1[,i])
  interval=HPDinterval(mcmcobj, 0.95)
  credint95_medy1= rbind(credint95_medy1,interval)
}

colnames(credint95_highy1) <- c("highlower","highupper")
colnames(credint95_lowy1) <- c("lowlower","lowupper")
colnames(credint95_medy1) <- c("medlower","medupper")

out_CI_avgy1 =cbind(credint95_highy1,credint95_lowy1,credint95_medy1)

write.csv(out_CI_avgy1,"out_CI_avgy1.csv")

##############################
#### same without year 1 ####
##############################
inhigh=read.csv("in_avgmod_highuse_noyr1.csv", header=TRUE)
inlow=read.csv("in_avgmod_lowuse_noyr1.csv", header=TRUE)
inmed=read.csv("in_avgmod_meduse_noyr1.csv", header=TRUE)

# calculate highest density credible 95% interval for each variable
library(coda)

credint95_high=c()
for(i in 1:ncol(inhigh)){
  mcmcobj=as.mcmc(inhigh[,i])
  interval=HPDinterval(mcmcobj, 0.95)
  credint95_high= rbind(credint95_high,interval)
}

credint95_low=c()
for(i in 1:ncol(inlow)){
  mcmcobj=as.mcmc(inlow[,i])
  interval=HPDinterval(mcmcobj, 0.95)
  credint95_low= rbind(credint95_low,interval)
}

credint95_med=c()
for(i in 1:ncol(inmed)){
  mcmcobj=as.mcmc(inmed[,i])
  interval=HPDinterval(mcmcobj, 0.95)
  credint95_med= rbind(credint95_med,interval)
}
```

```
colnames(credint95_high) <- c("highlower","highupper")
colnames(credint95_low) <- c("lowlower","lowupper")
colnames(credint95_med) <- c("medlower","medupper")

out_CI_avgnoy1 =cbind(credint95_high,credint95_low,credint95_med)

write.csv(out_CI_avgnoy1,"out_CI_avgnoy1.csv")

##################################
#### same with one by X model ####
##################################
inonebyx_y1=read.csv("in_onebyx_yr1.csv", header=TRUE)
inonebyx=read.csv("in_onebyx_noyr1.csv", header=TRUE)

# calculate highest density credible 95% interval for each variable
library(coda)

credint95_onexy1=c()
for(i in 1:ncol(inonebyx_y1)){
  mcmcobj=as.mcmc(inonebyx_y1[,i])
  interval=HPDinterval(mcmcobj, 0.95)
  credint95_onexy1= rbind(credint95_onexy1,interval)
}

credint95_onex=c()

for(i in 1:ncol(inonebyx)){
  mcmcobj=as.mcmc(inonebyx[,i])
  interval=HPDinterval(mcmcobj, 0.95)
  credint95_onex= rbind(credint95_onex,interval)
}

colnames(credint95_onex) <- c("oneXlower","oneXupper")
colnames(credint95_onexy1) <- c("oneXy1lower","oneXy1upper")

out_CI_oneX1 =cbind(credint95_onex,credint95_onexy1)
write.csv(out_CI_oneX1,"out_CI_oneX1.csv")
```

References

[1] American Community Survey 2015. American factfinder. selected economic characteristics. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_15_5YR_DP03&prodType=table.

[2] National Center for Health Statistics (US et al. Health, united states, 2015: with special feature on racial and ethnic health disparities. 2016.

[3] Sandro Galea, Melissa Tracy, Katherine J Hoggatt, Charles DiMaggio, and Adam Karpati. Estimated deaths attributable to social factors in the united states. *American Journal of Public Health*, 101(8):1456–1465, 2011.

[4] Paula A Braveman, Susan A Egerter, Catherine Cubbin, and Kristen S Marchi. An approach to studying social disparities in health and health care. *American Journal of Public Health*, 94(12):2139–2148, 2004.

[5] Nancy Krieger, David R Williams, and Nancy E Moss. Measuring social class in us public health research: concepts, methodologies, and guidelines. *Annual review of public health*, 18(1):341–378, 1997.

[6] Monica E Peek and Jini H Han. Disparities in screening mammography. *Journal of general internal medicine*, 19(2):184–194, 2004.

[7] Kristen J Wells and Richard G Roetzheim. Health disparities in receipt of screening mammography in latinas: a critical review of recent literature. *Cancer Control*, 14(4):369, 2007.

[8] David Kindig and Greg Stoddart. What is population health? *American journal of public health*, 93(3):380–383, 2003.

[9] Robert G Evans and Gregory L Stoddart. Producing health, consuming health care. *Social science & medicine*, 31(12):1347–1363, 1990.

[10] David A Kindig. *Purchasing population health: paying for results*. University of Michigan Press, 1997.

[11] Paula A Braveman, Shiriki Kumanyika, Jonathan Fielding, Thomas LaVeist, Luisa N Borrell, Ron Manderscheid, and Adewale Troutman. Health disparities and health equity: the issue is justice. *American Journal of Public Health*, 101(S1):S149–S155, 2011.

[12] National Cancer Institute. Cancer disparities. https://www.cancer.gov/about-cancer/understanding/disparities.

[13] Lu Ann Aday. *Evaluating the healthcare system: effectiveness, efficiency, and equity*. Health administration press, 2004.

[14] Ana V Diez Roux. Conceptual approaches to the study of health disparities. *Annual review of public health*, 33:41–58, 2012.

[15] Bruce G Link and Jo Phelan. Social conditions as fundamental causes of disease. *Journal of health and social behavior*, pages 80–94, 1995.

[16] Diana Kuh, Yoav Ben-Shlomo, John Lynch, Johan Hallqvist, and Chris Power. Life course epidemiology. *Journal of epidemiology and community health*, 57(10):778, 2003.

[17] Archana Singh-Manoux, Jane E Ferrie, Tarani Chandola, and Michael Marmot. Socioeconomic trajectories across the life course and health outcomes in midlife: evidence for the accumulation hypothesis? *International Journal of Epidemiology*, 33(5):1072–1079, 2004.

[18] Nancy Krieger. Methods for the scientific study of discrimination and health: an ecosocial approach. *American journal of public health*, 102(5):936–944, 2012.

[19] Amy H Auchincloss and Ana V Diez Roux. A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health. *American journal of epidemiology*, 168(1):1–8, 2008.

[20] Lisa A Carey, Charles M Perou, Chad A Livasy, Lynn G Dressler, David Cowan, Kathleen Conway, Gamze Karaca, Melissa A Troester, Chiu Kit Tse, Sharon Edmiston, et al. Race, breast cancer subtypes, and survival in the carolina breast cancer study. *Jama*, 295(21):2492–2502, 2006.

[21] Ahmedin Jemal, Rebecca Siegel, Elizabeth Ward, Yongping Hao, Jiaquan Xu, Taylor Murray, and Michael J Thun. Cancer statistics, 2008. *CA: a cancer journal for clinicians*, 58(2):71–96, 2008.

[22] The Equality of Opportunity Project. How can we reduce disparities in health? http://www.equality-of-opportunity.org/health/.

[23] County Health Rankings. Health rankings in the united states. http://www.countyhealthrankings.org.

[24] County Health Rankings and Roadmaps. Mammography screening in wisconsin. http://www.countyhealthrankings.org/app/wisconsin/2013/measure/factors/50/data.

[25] County Health Rankings. Primary care physicians in wisconsin. http://www.countyhealthrankings.org/app/wisconsin/2017/measure/factors/4/map.

[26] Wisconsin Council on Children and Families 2013. Race to equity. a baseline report on the state of racial disparities in dane county. http://racetoequity.net/baseline-report-state-racial-disparities-dane-county/.

[27] Howard Frumkin. Urban sprawl and public health. *Public health reports*, 117(3):201, 2002.

[28] United Nations Department of Economic and Social Affairs. Population Division. World urbanization prospects, the 2014 revision. https://esa.un.org/unpd/wup/Country-Profiles/.

[29] Malia Jones. Assessing geographic access to primary care across the urban-rural spectrum in wisconsin: A space-time geography approach. Presented at the Monday Seminar Series, Department of Population Health Sciences, University of Wisconsin, Madison.

[30] Chris S Kochtitzky, H Frumkin, R Rodriguez, AL Dannenberg, J Rayman, K Rose, R Gillig, and T Kanter. Urban planning and public health at cdc. *MMWR supplements*, 55(2):34–38, 2006.

[31] Candace Rutt, Andrew L Dannenberg, and Christopher Kochtitzky. Using policy and built environment interventions to improve public health. *Journal of Public Health Management and Practice*, 14(3):221–223, 2008.

[32] Mary E Northridge, Elliot D Sclar, and Padmini Biswas. Sorting out the connections between the built environment and health: a conceptual framework for navigating pathways and planning healthy cities. *Journal of Urban Health*, 80(4):556–568, 2003.

[33] Richard J Jackson. The impact of the built environment on health: an emerging field, 2003.

[34] Maggie L Grabow, Scott N Spak, Tracey Holloway, Brian Stone Jr, Adam C Mednick, and Jonathan A Patz. Air quality and exercise-related health benefits from reduced car travel in the midwestern united states. *Environmental health perspectives*, 120(1):68, 2012.

[35] A Colin Bell, Keyou Ge, and Barry M Popkin. The road to obesity or the path to prevention: motorized transportation and obesity in china. *Obesity Research*, 10(4):277–283, 2002.

[36] WHO. 7 million premature deaths annually linked to air pollution. http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/.

[37] Lara J Akinbami and Kenneth C Schoendorf. Trends in childhood asthma: prevalence, health care utilization, and mortality. *Pediatrics*, 110(2):315–322, 2002.

[38] Peter L Jacobsen. Safety in numbers: more walkers and bicyclists, safer walking and bicycling. *Injury prevention*, 9(3):205–209, 2003.

[39] John Pucher and Lewis Dijkstra. Promoting safe walking and cycling to improve public health: lessons from the netherlands and germany. *American journal of public health*, 93(9):1509–1516, 2003.

[40] Intergovernmental Panel on Climate Change. Climate change 2014 synthesis report. `http://www.ipcc.ch/pdf/assessment-report/ar5/syr/SYR_AR5_FINAL_full.pdf`.

[41] United States Environmental Protection Agency. Climate change. overview of greenhouse gases and sources of emissions. `https://www.epa.gov/ghgemissions/us-greenhouse-gas-inventory-report-1990-2014`.

[42] United States Environmental Protection Agency. Heat island effect. `https://www.epa.gov/heat-islands`.

[43] Steven Raphael and Lorien Rice. Car ownership, employment, and earnings. *Journal of Urban Economics*, 52(1):109–130, 2002.

[44] Inshu Minocha, P Sriraj, Paul Metaxatos, and Piyushimita Thakuriah. Analysis of transit quality of service and employment accessibility for the greater chicago, illinois, region. *Transportation Research Record: Journal of the Transportation Research Board*, (2042):20–29, 2008.

[45] Daniel P McMillen et al. Polycentric urban structure: The case of milwaukee. *ECONOMIC PERSPECTIVES-FEDERAL RESERVE BANK OF CHICAGO*, 25(2):15–27, 2001.

[46] Jason D Boardman and Samuel H Field. Spatial mismatch and race differentials in male joblessness: Cleveland and milwaukee, 1990. *The Sociological Quarterly*, 43(2):237–255, 2002.

[47] Tracey Giang, Allison Karpyn, Hannah Burton Laurison, Amy Hillier, and R Duane Perry. Closing the grocery gap in underserved communities: the creation of the pennsylvania fresh food financing initiative. *Journal of Public Health Management and Practice*, 14(3):272–279, 2008.

[48] Linda F Alwitt and Thomas D Donley. Retail stores in poor urban neighborhoods. *Journal of consumer affairs*, 31(1):139–164, 1997.

[49] Steven Cummins. Food deserts. *The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society*, 2014.

[50] Kimberly Morland, Steve Wing, and Ana Diez Roux. The contextual effect of the local food environment on residents' diets: the atherosclerosis risk in communities study. *American journal of public health*, 92(11):1761–1768, 2002.

[51] Jim Latham and Tina Moffat. Determinants of variation in food cost and availability in two socioeconomically contrasting neighbourhoods of hamilton, ontario, canada. *Health & Place*, 13(1):273–287, 2007.

[52] Susan L Handy. Accessibility-vs. mobility-enhancing strategies for addressing automobile dependence in the us. *Institute of Transportation Studies*, 2002.

[53] Reid Ewing and Robert Cervero. Travel and the built environment: a meta-analysis. *Journal of the American planning association*, 76(3):265–294, 2010.

[54] Elaine Murakami and Jennifer Young. *Daily travel by persons with low income*. US Federal Highway Administration Washington, DC, 1997.

[55] John Pucher and John L Renne. Socioeconomics of urban travel: evidence from the 2001 nhts. *Transportation Quarterly*, 57(3):49–77, 2003.

[56] Todd Litman. Future isn't what it used to be-changing trends and their implications for transport planning. 2005.

[57] Rahaf Alsnih and David A Hensher. The mobility and accessibility expectations of seniors in an aging population. *Transportation Research Part A: Policy and Practice*, 37(10):903–916, 2003.

[58] Steven E Polzin, Xuehao Chu, and Jodi Godfrey. The impact of millennials' travel behavior on future personal vehicle travel. *Energy Strategy Reviews*, 5:59–65, 2014.

[59] Kevin Fiscella, Peter Franks, Mark P Doescher, and Barry G Saver. Disparities in health care by race, ethnicity, and language among the insured: findings from a national sample. *Medical care*, 40(1):52–59, 2002.

[60] Barry P Katz, Deborah A Freund, David A Heck, Robert S Dittus, John E Paul, James Wright, Peter Coyte, Eleanor Holleman, and Gillian Hawker. Demographic variation in the rate of knee replacement: a multi-year analysis. *Health services research*, 31(2):125, 1996.

[61] Jose A Suaya, Donald S Shepard, Sharon-Lise T Normand, Philip A Ades, Jeffrey Prottas, and William B Stason. Use of cardiac rehabilitation by medicare beneficiaries after myocardial infarction or coronary bypass surgery. *Circulation*, 116(15):1653–1662, 2007.

[62] Sheryl Dunlop, Peter C Coyte, and Warren McIsaac. Socio-economic status and the utilisation of physicians' services: results from the canadian national population health survey. *Social science & medicine*, 51(1):123–133, 2000.

[63] Thomas A Arcury, Wilbert M Gesler, John S Preisser, Jill Sherman, John Spencer, and Jamie Perin. The effects of geography and spatial behavior on health care utilization among the residents of a rural region. *Health services research*, 40(1):135–156, 2005.

[64] Thomas A Arcury, John S Preisser, Wilbert M Gesler, and James M Powers. Access to transportation and health care utilization in a rural region. *The Journal of Rural Health*, 21(1):31–38, 2005.

[65] Ronald M Andersen. Revisiting the behavioral model and access to medical care: does it matter? *Journal of health and social behavior*, pages 1–10, 1995.

[66] U.S. Preventive Services Task Force. Screening for breast cancer. `http://www.uspreventiveservicestaskforce.org/uspstf/uspsbrca.htm`.

[67] Kathryn A Phillips, Karla Kerlikowske, Laurence C Baker, Sophia W Chang, and Martin L Brown. Factors associated with women's adherence to mammography screening guidelines. *Health services research*, 33(1):29, 1998.

[68] Ami Vyas, Suresh Madhavan, Traci LeMasters, Elvonna Atkins, Sara Gainor, Stephenie Kennedy, Kimberly Kelly, Linda Vona-Davis, and Scot Remick. Factors influencing adherence to mammography screening guidelines in appalachian women participating in a mobile mammography program. *Journal of community health*, 37(3):632–646, 2012.

[69] Melanie A Price, Phyllis N Butow, Margaret Charles, Tracey Bullen, Bettina Meiser, Joanne M McKinley, Sue-Anne McLachlan, Kelly-Anne Phillips, et al. Predictors of breast cancer screening behavior in women with a strong family history of the disease. *Breast cancer research and treatment*, 124(2):509–519, 2010.

[70] Lisa C Watson-Johnson, Amy DeGroff, C Brooke Steele, Michelle Revels, Judith Lee Smith, Erin Justen, Rachel Barron-Simpson, Latasha Sanders, and Lisa C Richardson. Mammography adherence: a qualitative study. *Journal of Women's Health*, 20(12):1887–1894, 2011.

[71] Elena B Elkin, Coral L Atoria, Nicole Leoce, Peter B Bach, and Deborah Schrag. Changes in the availability of screening mammography, 2000–2010. *Cancer*, 119(21):3847–3853, 2013.

[72] Elena B Elkin, Nicole M Ishill, Jacqueline G Snow, Katherine S Panageas, Peter B Bach, Laura Liberman, Fahui Wang, and Deborah Schrag. Geographic access and the use of screening mammography. *Medical care*, 48(4):349, 2010.

[73] G Philip Barnsley, Eva Grunfeld, Douglas Coyle, and Lawrence Paszat. Surveillance mammography following the treatment of primary breast cancer with breast reconstruction: a systematic review. *Plastic and reconstructive surgery*, 120(5):1125–1132, 2007.

[74] Susan A Sabatino, Trevor D Thompson, Lisa C Richardson, and Jacqueline Miller. Health insurance and other factors associated with mammography surveillance among breast cancer survivors: results from a national survey. *Medical care*, 50(3):270–276, 2012.

[75] Yue Chen, Wendy Thompson, Robert Semenciw, and Yang Mao. Epidemiology of contralateral breast cancer. *Cancer Epidemiology and Prevention Biomarkers*, 8(10):855–861, 1999.

[76] Philippe Broet, A De la Rochefordiere, Susan M Scholl, Alain Fourquet, Veronique Mosseri, Jean-Claude Durand, Pierre Pouillart, and Bernard Asselain. Contralateral breast cancer: annual incidence and risk parameters. *Journal of Clinical Oncology*, 13(7):1578–1583, 1995.

[77] James L Khatcheressian, Patricia Hurley, Elissa Bantug, Laura J Esserman, Eva Grunfeld, Francine Halberg, Alexander Hantel, N Lynn Henry, Hyman B Muss, Thomas J Smith, et al. Breast cancer follow-up and management after primary treatment: American society of clinical oncology clinical practice guideline update. *Journal of Clinical Oncology*, 31(7):961–965, 2012.

[78] National comprehensive cancer network: Nccn guidelines for patients (stage 0 breast cancer). dcis follow-up care, 2016.

[79] Timothy L Lash, Matthew P Fox, Diana SM Buist, Feifei Wei, Terry S Field, Floyd J Frost, Ann M Geiger, Virginia P Quinn, Marianne Ulcickas Yood, and Rebecca A Silliman. Mammography surveillance and mortality in older breast cancer survivors. *Journal of Clinical Oncology*, 25(21):3001–3006, 2007.

[80] N Houssami, S Ciatto, F Martinelli, R Bonardi, and SW Duffy. Early detection of second breast cancers improves prognosis in breast cancer survivors. *Annals of oncology*, 20(9):1505–1510, 2009.

[81] Berta M Geller, Karla Kerlikowske, Patricia A Carney, Linn A Abraham, Bonnie C Yankaskas, Stephen H Taplin, Rachel Ballard-Barbash, Mark B Dignan, Robert Rosenberg, Nicole Urban, et al. Mammography surveillance following breast cancer. *Breast cancer research and treatment*, 81(2):107–115, 2003.

[82] Chyke A Doubeni, Terry S Field, Marianne Ulcickas Yood, Sharon J Rolnick, Charles P Quessenberry, Hassan Fouayzi, Jerry H Gurwitz, and Feifei Wei. Patterns and predictors of mammography utilization among breast cancer survivors. *Cancer*, 106(11):2482–2488, 2006.

[83] Marilyn M Schapira, Timothy L McAuliffe, and Ann B Nattinger. Underutilization of mammography in older breast cancer survivors. *Medical care*, 38(3):281–289, 2000.

[84] Nancy L Keating, Mary Beth Landrum, Edward Guadagnoli, Eric P Winer, and John Z Ayanian. Factors related to underuse of surveillance mammography among breast cancer survivors. *Journal of Clinical Oncology*, 24(1):85–94, 2006.

[85] Los Angeles Times. Millennials and car ownership? it's complicated. `http://www.latimes.com/business/autos/la-fi-hy-millennials-cars-20161223-story.html`.

[86] The Washington Post. The surprising reason some millennials may be buying new cars. `https://www.washingtonpost.com/news/business/wp/2017/02/28/the-surprising-reason-some-millennials-may-be-buying-new-cars/?utm_term=.b8109a67149b`.

[87] Centers for Disease Control and Prevention. National center for health statistics. national health interview survey 2015. `https://www.cdc.gov/nchs/nhis/about_nhis.htm`.

[88] Centers for Disease Control and Prevention. National center for health statistics. national health interview survey 2015. data release. `https://www.cdc.gov/nchs/nhis/nhis_2015_data_release.htm`.

[89] Frank E Harrell. R hmisc package. `https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf`.

[90] Thomas Lumley. R survey package. `https://cran.r-project.org/web/packages/survey/survey.pdf`.

[91] Thomas Lumley. R package 'mitools'. `https://cran.r-project.org/web/packages/mitools/mitools.pdf`.

[92] Edward L Glaeser, Matthew E Kahn, and Jordan Rappaport. Why do the poor live in cities? the role of public transportation. *Journal of urban Economics*, 63(1):1–24, 2008.

[93] James J Flink. *The automobile age*. MIT Press, 1990.

[94] Jeffrey R Brown, Eric A Morris, and Brian D Taylor. Planning for cars in cities: Planners, engineers, and freeways in the 20th century. *Journal of the American Planning Association*, 75(2):161–177, 2009.

[95] United Nations. Data: Germany. http://data.un.org/CountryProfile.aspx?crName=GERMANY.

[96] United Nations. Data: United states of america. http://data.un.org/CountryProfile.aspx?crName=United%20States%20of%20America.

[97] John Pucher. Urban travel behavior as the outcome of public policy: the example of modal-split in western europe and north america. *Journal of the American Planning Association*, 54(4):509–520, 1988.

[98] Oak Ridge National Laboratory. Center for Transportation Analysis. Vehicles and vehicle miles per capita, 1950–201. http://cta.ornl.gov/data/tedb34/Edition34_Chapter08.pdf.

[99] James Woodcock, Phil Edwards, Cathryn Tonne, Ben G Armstrong, Olu Ashiru, David Banister, Sean Beevers, Zaid Chalabi, Zohir Chowdhury, Aaron Cohen, et al. Public health benefits of strategies to reduce greenhouse-gas emissions: urban land transport. *The Lancet*, 374(9705):1930–1943, 2009.

[100] Lilah M Besser and Andrew L Dannenberg. Walking to public transit: steps to help meet physical activity recommendations. *American journal of preventive medicine*, 29(4):273–280, 2005.

[101] Brigid M Lynch. Sedentary behavior and cancer: a systematic review of the literature and proposed biological mechanisms. *Cancer Epidemiology and Prevention Biomarkers*, pages 1055–9965, 2010.

[102] Emma G Wilmot, Charlotte L Edwardson, Felix A Achana, Melanie J Davies, Trish Gorely, Laura J Gray, Kamlesh Khunti, Thomas Yates, and Stuart JH Biddle. Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis, 2012.

[103] Centers for Disease Control and Prevention. Motor vehicle safety. https://www.cdc.gov/motorvehiclesafety/.

[104] Todd Litman. Transportation cost and benefit analysis. *Victoria Transport Policy Institute*, 31, 2009.

[105] Dea Van Lierop Michael Grimsrud Colin Stewart Ana Tepavac Kevin Manaugh Ahmed El-Geneidy Emory Shaw, Evelyne St-Louis. Findings from the 2013 mcgill commuter survey. http://tram.mcgill.ca/Research/Publications/McGill_2013_travel_survey_full_version.pdf.

[106] Tami Gurley and Donald Bruce. The effects of car access on employment outcomes for welfare recipients. *Journal of urban Economics*, 58(2):250–272, 2005.

[107] Robert Cervero, Onésimo Sandoval, and John Landis. Transportation as a stimulus of welfare-to-work: Private versus public mobility. *Journal of Planning Education and Research*, 22(1):50–63, 2002.

[108] Tobias Kuhnimhof, Jimmy Armoogum, Ralph Buehler, Joyce Dargay, Jon Martin Denstadli, and Toshiyuki Yamamoto. Men shape a downward trend in car use among young adults—evidence from six industrialized countries. *Transport Reviews*, 32(6):761–779, 2012.

[109] US Federal Highway Administration. Distribution of licensed drivers - 1994. `https://www.fhwa.dot.gov/policyinformation/statistics/1994/dl20.pdf`.

[110] US Federal Highway Administration. Distribution of licensed drivers - 2013. `https://www.fhwa.dot.gov/policyinformation/statistics/2013/dl20.cfm`.

[111] Noreen C McDonald. Are millennials really the "go-nowhere" generation? *Journal of the American Planning Association*, 81(2):90–103, 2015.

[112] Mauro Guillén. The global economic & financial crisis: A timeline. *The Lauder Institute, University of Pennsylvania*, pages 1–91, 2009.

[113] David Haugh, Annabelle Mourougane, and Olivier Chatal. The automobile industry in and beyond the crisis. 2010.

[114] Seung-Youn Oh. Shifting gears: industrial policy and automotive industry after the 2008 financial crisis. *Business and Politics*, 16(4):641–665, 2014.

[115] Julia M Solis, Gary Marks, Melinda Garcia, and David Shelton. Acculturation, access to care, and use of preventive services by hispanics: findings from hhanes 1982-84. *American journal of public health*, 80(Suppl):11–19, 1990.

[116] Glenn Flores and Hua Lin. Trends in racial/ethnic disparities in medical and oral health, access to care, and use of services in us children: has anything changed over the years? *International Journal for Equity in Health*, 12(1):10, 2013.

[117] Alan Nelson. Unequal treatment: confronting racial and ethnic disparities in health care. *Journal of the National Medical Association*, 94(8):666, 2002.

[118] Samina T Syed, Ben S Gerber, and Lisa K Sharp. Traveling towards disease: transportation barriers to health care access. *Journal of community health*, 38(5):976–993, 2013.

[119] R Turner Goins, Kimberly A Williams, Mary W Carter, S Melinda Spencer, and Tatiana Solovieva. Perceived barriers to health care access among rural older adults: a qualitative study. *The Journal of Rural Health*, 21(3):206–213, 2005.

[120] Syed M Ahmed, Jeanne P Lemkau, Nichol Nealeigh, and Barbara Mann. Barriers to healthcare access in a non-elderly urban poor american population. *Health & social care in the community*, 9(6):445–453, 2001.

[121] Lisa I Iezzoni, Mary B Killeen, and Bonnie L O'day. Rural residents with disabilities confront substantial barriers to obtaining primary care. *Health Services Research*, 41(4p1):1258–1275, 2006.

[122] F Javier Nieto, Paul E Peppard, Corinne D Engelman, Jane A McElroy, Loren W Galvao, Elliot M Friedman, Andrew J Bersch, and Kristen C Malecki. The survey of the health of wisconsin (show), a novel infrastructure for population health research: rationale and methods. *BMC public health*, 10(1):785, 2010.

[123] Walk Score. Walkability, real estate, and public health data. `https://www.walkscore.com/professional/research.php`.

[124] SAS Support. The mi procedure. `https://support.sas.com/rnd/app/stat/procedures/mi.html`.

[125] SAS Support. The mianalyze procedure. `https://support.sas.com/rnd/app/stat/procedures/mianalyze.html`.

[126] SAS Support. The glimmix procedure. `https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#glimmix_toc.htm`.

[127] Rune Haubo Bojesen Christensen. R ordinal package. `ftp://cran.r-project.org/pub/R/web/packages/ordinal/ordinal.pdf`.

[128] David Egan-Robertson. Uw-madison applied population laboratory. wisconsin's future population projections for the state, its counties and municipalities, 2010 - 2040. `http://www.doa.state.wi.us/Documents/DIR/Demographic%20Services%20Center/Projections/FinalProjs2040_Publication.pdf`.

[129] Anjani Chandra Gladys Martinez, Kimberly Daniels. Centers for disease control an prevention. national health statistics report. `https://www.cdc.gov/nchs/data/nhsr/nhsr051.pdf`.

[130] Centers for Disease Control and Prevention. National breast and cervical cancer early detection program (nbccedp). `https://www.cdc.gov/cancer/nbccedp/index.htm`.

[131] Laura J Esserman, Ian M Thompson, Brian Reid, Peter Nelson, David F Ransohoff, H Gilbert Welch, Shelley Hwang, Donald A Berry, Kenneth W Kinzler, William C Black, et al. Addressing overdiagnosis and overtreatment in cancer: a prescription for change. *The Lancet Oncology*, 15(6):e234–e242, 2014.

[132] Sue M Moss, Howard Cuckle, Andy Evans, Louise Johns, Michael Waller, Lynda Bobrow, et al. Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years' follow-up: a randomised controlled trial. *The Lancet*, 368(9552):2053–2060, 2006.

[133] Anthony B Miller, Teresa To, Cornelia J Baines, and Claus Wall. The canadian national breast screening study-1: breast cancer mortality after 11 to 16 years of follow-up: a randomized screening trial of mammography in women age 40 to 49 years. *Annals of internal medicine*, 137(5_Part_1):305–312, 2002.

[134] Lennarth Nyström, Ingvar Andersson, Nils Bjurstam, Jan Frisell, Bo Nordenskjöld, and Lars Erik Rutqvist. Long-term effects of mammography screening: updated overview of the swedish randomised trials. *The Lancet*, 359(9310):909–919, 2002.

[135] Peter Greenwald and Edward Jay Sondik. Cancer control objectives for the nation, 1985-2000. 1986.

[136] Gerald D Dodd. American cancer society guidelines on screening for breast cancer: an overview. *CA: A Cancer Journal for Clinicians*, 42(3):177–180, 1992.

[137] Susan A Sabatino, Ralph J Coates, Robert J Uhler, Nancy Breen, Florence Tangka, and Kate M Shaw. Disparities in mammography use among us women aged 40–64 years, by race, ethnicity, income, and health insurance status, 1993 and 2005. *Medical care*, 46(7):692–700, 2008.

[138] Judith Swan, Nancy Breen, Ralph J Coates, Barbara K Rimer, and Nancy C Lee. Progress in cancer screening practices in the united states. *Cancer*, 97(6):1528–1540, 2003.

[139] Nancy Breen, Kathleen A Cronin, Helen I Meissner, Stephen H Taplin, Florence K Tangka, Jasmin A Tiro, and Timothy S McNeel. Reported drop in mammography. *Cancer*, 109(12):2405–2409, 2007.

[140] Nancy Breen, Jane F Gentleman, and Jeannine S Schiller. Update on mammography trends. *Cancer*, 117(10):2209–2218, 2011.

[141] Reid Ewing, Rolf Pendall, and Don Chen. Measuring sprawl and its transportation impacts. *Transportation Research Record: Journal of the Transportation Research Board*, (1831):175–183, 2003.

[142] Brian L Sprague, Ronald E Gangnon, John M Hampton, Kathleen M Egan, Linda J Titus, Karla Kerlikowske, Patrick L Remington, Polly A Newcomb, and Amy Trentham-Dietz. Variation in breast cancer–risk factor associations by method of detection: Results from a series of case-control studies. *American journal of epidemiology*, 181(12):956–969, 2015.

[143] Stephanie E Nelson, Michael N Gould, John M Hampton, and Amy Trentham-Dietz. A case-control study of the her2 ile655val polymorphism in relation to risk of invasive breast cancer. *Breast cancer research*, 7(3):R357, 2005.

[144] Polly A Newcomb, A Trentham-Dietz, and JM Hampton. Bisphosphonates for osteoporosis treatment are associated with reduced breast cancer risk. *British journal of cancer*, 102(5):799–802, 2010.

[145] Brian L Sprague, Amy Trentham-Dietz, Polly A Newcomb, Linda Titus-Ernstoff, John M Hampton, and Kathleen M Egan. Lifetime recreational and occupational physical activity and risk of in situ and invasive breast cancer. *Cancer Epidemiology and Prevention Biomarkers*, 16(2):236–243, 2007.

[146] Jane A McElroy, Patrick L Remington, Amy Trentham-Dietz, Stephanie A Robert, and Polly A Newcomb. Geocoding addresses from a large population-based study: lessons learned. *Epidemiology*, 14(4):399–407, 2003.

[147] Rural Health Research Center. Rural-urban commuting area codes: Using ruca data. `http://depts.washington.edu/uwruca/ruca-uses.php`.

[148] Hadley Wickham. R package 'httr'. `https://cran.r-project.org/web/packages/httr/httr.pdf`.

[149] Johannes Textor. Dagitty. `http://www.dagitty.net`.

[150] Felix Elwert. *In: Morgan SL, ed. Handbook of causal analysis for social research.* Springer, 2013.

[151] Peter J. Diggle Paulo J. Ribeiro. R geor package. `https://cran.r-project.org/web/packages/geoR/geoR.pdf`.

[152] Benedikt Graeler Edzer Pebesma. R gstat package. `https://cran.r-project.org/web/packages/gstat/gstat.pdf`.

[153] Ray Brownrigg Thomas P Minka Alex Deckmyn Richard A. Becker, Allan R. Wilks. R maps package. `https://cran.r-project.org/web/packages/maps/maps.pdf`.

[154] Barry Rowlingson Virgilio Gomez-Rubio Robert Hijmans Michael Sumner Don MacQueen Jim Lemon Josh O'Brien Edzer Pebesma, Roger Bivand. R sp package. `https://cran.r-project.org/web/packages/sp/sp.pdf`.

[155] Ege Rubak Adrian Baddeley, Rolf Turner. R spatstat package. `https://cran.r-project.org/web/packages/spatstat/spatstat.pdf`.

[156] Martin Maechler Werner Stahel. jitter() function r. `https://stat.ethz.ch/R-manual/R-devel/library/base/html/00Index.html`.

[157] Jenna A Khan-Gates, Jennifer L Ersek, Jan M Eberth, Swann A Adams, and Sandi L Pruitt. Geographic access to mammography and its relationship to breast cancer screening and stage at diagnosis: a systematic review. *Women's Health Issues*, 25(5):482–493, 2015.

[158] Stephen C Meersman, Nancy Breen, Linda W Pickle, Helen I Meissner, and Paul Simon. Access to mammography screening in a large urban population: a multi-level analysis. *Cancer Causes & Control*, 20(8):1469–1482, 2009.

[159] Kimberly K Engelman, Daniel B Hawley, Rona Gazaway, Michael C Mosier, Jasjit S Ahluwalia, and Edward F Ellerbeck. Impact of geographic barriers on the utilization of mammograms by older rural women. *Journal of the American Geriatrics Society*, 50(1):62–68, 2002.

[160] Bin Huang, Mark Dignan, Daikwon Han, and Owen Johnson. Does distance matter? distance to mammography facilities and stage at diagnosis of breast cancer in kentucky. *The Journal of Rural Health*, 25(4):366–371, 2009.

[161] MO Celaya, EM Berke, TL Onega, J Gui, BL Riddle, SS Cherala, JR Rees, et al. Breast cancer stage at diagnosis and geographic access to mammography screening (new hampshire, 1998–2004). *Rural Remote Health*, 10(2):1361, 2010.

[162] Kevin A Henry, Recinda Sherman, Steve Farber, Myles Cockburn, Daniel W Goldberg, and Antoinette M Stroup. The joint effects of census tract poverty and geographic access on late-stage breast cancer diagnosis in 10 us states. *Health & place*, 21:110–121, 2013.

[163] Tracy Onega, Andrea Cook, Beth Kirlin, Xun Shi, Jennifer Alford-Teaster, Leah Tuzzio, and Diana SM Buist. The influence of travel time on breast cancer characteristics, receipt of primary therapy, and surveillance mammography. *Breast cancer research and treatment*, 129(1):269–275, 2011.

[164] Linda S Elting, Catherine D Cooksley, B Nebiyou Bekele, Sharon H Giordano, Ya Chen Tina Shih, Kelly K Lovell, Elenir BC Avritscher, and Richard Theriault. Mammography capacity: Impact on screening rates and breast cancer stage at diagnosis. *American journal of preventive medicine*, 37(2):102–108, 2009.

[165] Selina Rahman, James H Price, Mark Dignan, Saleh Rahman, Peter S Lindquist, and Timothy R Jordan. Access to mammography facilities and detection of breast cancer by screening mammography: a gis approach. *International journal of cancer prevention*, 2(6):403, 2009.

[166] Steven S Coughlin, Steven Leadbetter, Thomas Richards, and Susan A Sabatino. Contextual analysis of breast and cervical cancer screening and factors associated with health care access among united states women, 2002. *Social science & medicine*, 66(2):260–275, 2008.

[167] Julie Marchick and Donald E Henson. Correlations between access to mammography and breast cancer stage at diagnosis. *Cancer*, 103(8):1571–1580, 2005.

[168] VL Ernster, J Barclay, K Kerlikowske, H Wilkie, and R Ballard-Barbash. Mortality among women with ductal carcinoma in situ of the breast in the population-based surveillance, epidemiology and end results program. *Archives of Internal Medicine*, 160(7):953–958, 2000.

[169] Ellen Warner. Breast-cancer screening. *New England Journal of Medicine*, 365(11):1025–1032, 2011.

[170] Timothy L Lash and Rebecca A Silliman. Medical surveillance after breast cancer diagnosis. *Medical care*, 39(9):945–955, 2001.

[171] Diana SM Buist, Jaclyn LF Bosco, Rebecca A Silliman, Heather Taffet Gold, Terry Field, Marianne Ulcickas Yood, Virginia P Quinn, Marianne Prout, Timothy L Lash, et al. Long-term surveillance mammography and mortality in older women with a history of early stage invasive breast cancer. *Breast cancer research and treatment*, 142(1):153–163, 2013.

[172] M Robyn Andersen and Nicole Urban. The use of mammography by survivors of breast cancer. *American journal of public health*, 88(11):1713–1714, 1998.

[173] Patricia Carcaise-Edinboro, Cathy J Bradley, and Bassam Dahman. Surveillance mammography for medicaid/medicare breast cancer patients. *Journal of Cancer Survivorship*, 4(1):59–66, 2010.

[174] Heidi S Wirtz, Denise M Boudreau, Julie R Gralow, William E Barlow, Shelly Gray, Erin JA Bowles, and Diana SM Buist. Factors associated with long-term adherence to annual surveillance mammography among breast cancer survivors. *Breast cancer research and treatment*, 143(3):541–550, 2014.

[175] Kate Cowles Karen Vines-Deepayan Sarkar Douglas Bates Russell Almond Arni Magnusson Martyn Plummer, Nicky Best. R coda package. `https://cran.r-project.org/web/packages/coda/coda.pdf`.

[176] Jeanne S Mandelblatt, William F Lawrence, Jennifer Cullen, Annette L Stanton, Janice L Krupnick, Lorna Kwan, and Patricia A Ganz. Patterns of care in early-stage breast cancer survivors in the first year after cessation of active treatment. *Journal of clinical oncology*, 24(1):77–84, 2006.

[177] livability.com. Top 10 healthiest cities. http://livability.com/top-10/health/top-10-healthiest-cities/2015/wi/madison.

[178] Karen E Joynt, Yael Harris, E John Orav, and Ashish K Jha. Quality of care and patient outcomes in critical access rural hospitals. *Jama*, 306(1):45–52, 2011.

[179] Nakela L Cook, LeRoi S Hicks, A James O'Malley, Thomas Keegan, Edward Guadagnoli, and Bruce E Landon. Access to specialty care and medical services in community health centers. *Health Affairs*, 26(5):1459–1468, 2007.

[180] Ashley Meilleur, SV Subramanian, Jesse J Plascak, James L Fisher, Electra D Paskett, and Elizabeth B Lamont. Rural residence and cancer outcomes in the united states: issues and challenges. *Cancer Epidemiology and Prevention Biomarkers*, 22(10):1657–1667, 2013.

[181] GIS Population Science. An introduction to the modifiable areal unit problem. http://gispopsci.org/maup/.

[182] Betty Hart and Todd R Risley. The early catastrophe: The 30 million word gap by age 3. *American educator*, 27(1):4–9, 2003.

[183] Centers for Disease Control and Prevention. National health and nutrition examination survey 2015-2016. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/questionnaires.aspx?BeginYear=2015.