Title
Three Studies of U.S. Mortality

Permalink
https://escholarship.org/uc/item/5027841r

Author
Breen, Casey

Publication Date
2023

Peer reviewed|Thesis/dissertation

Three Studies of U.S. Mortality

By

Casey F. Breen

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Demography

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Joshua R. Goldstein, Chair
Professor Dennis Feehan
Professor Ayesha Mahmud
Professor Christopher Muller

Spring 2023

Three Studies of U.S. Mortality

Abstract

Three Studies of U.S. Mortality

by

Casey F. Breen

Doctor of Philosophy in Demography

University of California, Berkeley

Professor Joshua R. Goldstein, Chair

We are far from a complete understanding of the social determinants of mortality in the United States. Despite the longstanding interest in racial and class-based inequalities in health and mortality in the United States, research is often hampered by data limitations. However, new advances in data linkage have allowed mortality researchers to construct administrative datasets with millions of mortality records and demographic covariates. The unprecedented scale and richness of these administrative datasets allow social scientists to make new discoveries into the contours of mortality disparities in the United States.

This dissertation is comprised of three studies of mortality using large-scale, linked U.S. Census and administrative death records. In my second chapter, I investigate the relationship between owning a home in early adulthood and life expectancy, demonstrating that owning a home in early adulthood has a causal effect on life expectancy. My second study assesses the predictability of individual-level longevity, demonstrating the challenges of predicting individual-level mortality. The final study of the dissertation examines the Black-White crossover, finding evidence that the crossover is not a data artifact and cannot be uncrossed using sociodemographic variables alone.

Acknowledgments

I am indebted to many people who have provided support and guidance over the course of my Ph.D. First and foremost, I am grateful to the members of my dissertation committee. Joshua Goldstein, my chair, has always been a sounding board, full of encouragement, insights, and wisdom that have shaped me as a scholar. Dennis Feehan has been instrumental in my academic development and has always challenged me intellectually. I am thankful to both of them for their generosity of time. Ayesha Mahmud has been a great support, chairing my qualifying exams and providing incisive feedback on my research ideas. I am also thankful to Chris Muller for agreeing to serve as the outside member of my committee, giving helpful and encouraging comments on my dissertation chapters.

Before graduate school, I had the privilege of learning from dedicated mentors. My undergraduate advisors at Pomona College, Erica Flapan and Pierangelo De Pace, introduced me to the joys of academic research. After graduating and beginning as a full-time researcher at IPUMS, it quickly became clear I could not have asked for a better place to start my career as a demographer. Thank you to Kristen Jeffers, Mathew Sobek, Steve Ruggles, and Rob Warren for impressing upon me the value of creating public data infrastructure for the scientific community.

I entered graduate school with the wonderful Berkeley Demography cohort of 39. I am grateful to Elizabeth Breen, Andrea Miranda-González, Mallika Snyder, Payal Hathi, Felipe Menares, Ethan Roubenoff, and Rae Willis-Conger for their camaraderie. I have learned so much from you all over the years, and I look forward to learning more from you in the future.

I have been fortunate to serve as a graduate student researcher on the CenSoc team, where I worked collaboratively to construct the datasets I used in this dissertation. I benefited from enriching exchanges with Jason Fletcher, Leslie Root, Serge Atherwood, Ugur Yildrim, Jordan Weiss, Maria Osborne, and other researchers on the project. A special thanks to Nathan Seltzer for co-authoring my second dissertation chapter.

The broader Berkeley community has been central to my development as a scholar. I would like to express my gratitude to Sophia Rabe-Hesketh and Maya Petersen for their guidance on causal inference and Bayesian hierarchical modeling while I completed my ancillary master's degree in biostatistics. I appreciate Ken Wachter and Ron Lee for taking the time to give detailed and insightful comments on my dissertation chapters. I am also thankful to Robert Chung for teaching me demographic methods and Magali Barbieri for sharing her knowledge about mortality. I would also like to thank Leora Lawton and Michelle Poulin for their leadership in fostering a vibrant academic community at the Berkeley Population Center. I am indebted to Monique Verrier and David Murphy for their administrative help and Carl Boe and Carl Mason for their help with technical computing.

As a Ph.D. intern at Microsoft Research, it was a privilege to collaborate with my mentors, Elissa Redmiles and Cormac Herley. I learned so much about the value of interdisciplinary collaboration and the field of computer science more broadly. As a Research Scientist intern at Meta (Facebook), I am grateful to Angela Lee and the broader Demography and Survey Science team for teaching me how to use Facebook and Instagram as platforms for data

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Life expectancy in the United States increased by a remarkable 30 years over the course of the 20th century. This impressive progress was driven primarily by advances in the treatment of infectious diseases and delayed mortality for those living with chronic illness (Crimmins and Zhang, 2019), but the benefits accrued unevenly. Inequality in mortality between the most advantaged and the least advantaged actually increased over time (Preston and Elo, 1995), and by the beginning of the 21st century, the gap in life expectancy between the top and bottom 1% of income earners was over 14.6 years (Chetty et al., 2016).

Despite the longstanding interest in racial and class-based inequalities in health and mortality in the United States (Schwandt et al., 2021; Elo, 2009), research is often hampered by data limitations (Card et al., 2010; Song and Coleman, 2020). Most research into the general dimensions of mortality disparities using microdata has relied on survey data, with small sample sizes making the analysis of smaller population subgroups such as the oldest-old infeasible. In the absence of comprehensive population-level registry data such as those found in the Scandinavian countries, researchers are increasingly turning to administrative datasets from agencies such as the Social Security Administration to answer some of the most pressing questions in social science research (Chetty et al., 2016; Card, Dobkin and Maestas, 2008; Card et al., 2010; Meyer and Mittag, 2019; Ruggles, 2014).

My dissertation uses large-scale, linked administrative data to study mortality in the United States. Specifically, I use newly-available CenSoc datasets, so termed because they link the full-count 1940 Census ("Cen") with Social Security Administration mortality records ("Soc"). The CenSoc datasets represent the first nationally-representative, large-scale, publicly available data resource for researchers studying mortality (Goldstein et al., 2021; Breen and Goldstein, 2022). The unprecedented scale and detail of CenSoc data allow researchers to make new discoveries into early-life conditions and later-life mortality and racial disparities.

This dissertation is primarily devoted to presenting new empirical investigations using CenSoc data rather than providing an in-depth technical overview of how we constructed these datasets. Such details are available elsewhere. For readers interested in understanding the full technical details of how these datasets were constructed, I co-authored a technical re-

port on technical aspects of the CenSoc dataset, such as match quality (Breen and Osborne, 2022). For more detail about the original mortality records, I suggest referring to Breen and Goldstein (2022), Alexander (2018), Finlay and Genadek (2021), or Hill (2001). Finally, one shared limitation of these datasets is that they only include mortality within a limited left and right ("doubly") truncated observation window. While this limitation is discussed throughout my dissertation, I encourage researchers to consult the formative work of (Alexander, 2018) and my co-authored publication (Goldstein et al., 2021) for a more comprehensive overview.

In my second chapter, "The Longevity Benefits of Homeownership," I use linked complete-count census and Social Security mortality records to produce the first U.S.-based estimates of the association between homeownership in early adulthood and longevity. I use a sibling-based identification strategy to estimate the causal effect of homeownership on longevity for cohorts born in the first two decades of the 20$^{\text{th}}$ century. My results indicate homeownership has a significant positive impact on longevity, which I estimate at approximately 0.4 years. The findings of this study highlight the need for greater equity in the opportunities, incentives, and costs of homeownership in the United States.

My third chapter (joint with Nathan Seltzer), "The Unpredictability of Mortality," investigates the predictability of later-life, individual-level longevity using sociodemographic characteristics. We test this question using a large-scale administrative dataset combining the complete count 1940 Census with Social Security death records. We fit eight machine learning algorithms using 35 sociodemographic predictors to generate individual-level predictions of age of death for birth cohorts born at the beginning of the 20$^{\text{th}}$ century. We find that none of these algorithms are able to explain more than 1.5% of the variation in age of death. Our results point towards the challenges of predicting mortality using sociodemographic characteristics and suggest that the fundamental uncertainty around individual-level mortality needs to be better acknowledged and incorporated into demographic theory.

In my fourth chapter, "Black-White Mortality Crossover: New Evidence from Linked Administrative Data," I investigate a longstanding demographic paradox: the Black-White Mortality crossover. Black Americans experience higher age-specific mortality rates than White Americans throughout most of the life course, but this puzzlingly reverses at advanced ages. The leading explanation for the Black-White mortality crossover centers around selective mortality over the life course. Black Americans who survived higher age-specific mortality risk throughout their life course are highly selected on robustness and have lower mortality than White Americans in late life. However, skeptics argue the Black-White mortality crossover is simply a data artifact from age misreporting or related data quality issues. We use large-scale linked administrative data (N = 2.3 million) to document the Black-White mortality crossover for cohorts born in the early 20$^{\text{th}}$ century. We find evidence the crossover is not a data artifact and cannot be uncrossed using sociodemographic characteristics alone.

These studies highlight the importance of applying insights from different methodological traditions to study mortality. My first chapter advances a causal argument about the relationships between early life exposure and later life longevity, using methods from the causal inference literature. My second chapter borrows from the computer science and biostatis-

tics literature, applying ensembling machine-learning methods to make predictions about longevity. My third chapter tests insights from formal demography's theoretical models of mortality selection.

To summarize, my dissertation uses the newly-available CenSoc datasets to make new discoveries about population-level mortality. These data allow me to study cohort mortality rather than the artificially constructed period mortality experiences common in much of the literature. This dissertation describes striking racial and class-based differences in life expectancy in the United States but also makes a broader point: studying group-level differences in life expectancy cannot fully explain differential mortality experiences in the United States. Group-level averages in life expectancy can obscure substantial heterogeneity within groups and age-specific trends in mortality disparities, such as mortality crossovers. Studying both individual-level and population-level differences in mortality is critical for a complete understanding of mortality in the United States.

Chapter 2

Longevity Benefits of Homeownership

Owning a home has long been touted as a key component of the idealized "American Dream." Homeownership is associated with greater wealth and better health, but the causal impact of homeownership on health remains unclear. Using linked complete-count census and Social Security mortality records, we document Black-White disparities in homeownership rates and produce the first U.S.-based estimates of the association between homeownership in early adulthood and longevity. We then use a sibling-based identification strategy to estimate the causal effect of homeownership on longevity for cohorts born in the first two decades of the 20th century. Our results indicate homeownership has a significant positive impact on longevity, which we estimate at approximately 0.3 years.

2.1   Introduction

Owning a home is considered a key component of the idealized "American Dream" (Samuel, 2012), and the home is the single largest asset class for personal wealth in the United States (Apgar and Di, 2006). Despite the large cultural and economic significance of homeownership in the U.S., evidence on the health and mortality benefits of homeownership is relatively thin. Homeownership is associated with positive health outcomes (Rolfe et al., 2020; Finnigan, 2014; Laaksonen, Tarkiainen and Martikainen, 2009); however, less is known about whether this relationship is causal: does owning a home cause people to live longer lives? Or is the observed association driven entirely by unmeasured shared confounders between homeownership and longevity, such as individual or familial income and wealth or social capital?

    Understanding the relationship between homeownership and longevity has both scientific and policy implications. There are striking historical and contemporary disparities in Black-White homeownership in the U.S., with White Americans owning homes at nearly twice the rate of Black Americans for much of the 20th century (Collins and Margo, 2011). If homeownership has a causal effect on longevity, social policies that equitably expand homeownership opportunities for racial minorities in the U.S. may help mitigate the profound

racial disparities in mortality. On the other hand, if the association between homeownership and mortality is driven entirely by shared confounders such as family wealth, such policies would have little or no effect on narrowing Black-White mortality gaps.

In this study, we use complete-count 1920 and 1940 Census records linked to Social Security mortality records to investigate the relationship between owning a home in early adulthood and later-life longevity for the birth cohorts of 1905-1915. We first quantify the unadjusted difference in life expectancy between homeowners and renters. In separate analyses by race, we find a positive relationship between homeownership in early adulthood and later-life longevity for both Black and White men ($N = 1.4$ million). We then use a sibling-based identification strategy to ascertain whether this relationship is causal, finding that homeownership has a causal effect of 4 months on longevity. Further, we find homeownership has similar longevity benefits for both Black and White men, suggesting that differential *rates* of homeownership—rather than differential *benefits* of homeownership—is the larger contributor to racial inequality in mortality.

2.2 Background

Past Studies on Homeownership and Health and Mortality

Studies of the benefits of homeownership have centered on overall household wealth (Killewald and Bryan, 2016; Turner and Luea, 2009), wealth accumulation among low- and middle-income groups (Boehm and Schlottmann, 2008), savings in retirement (Apgar and Di, 2006), cognitive benefits for children (Haurin, Parcel and Haurin, 2002), and social capital and civic engagement (Rohe and Stewart, 1996; Manturuk, Lindblad and Quercia, 2010). Recently, there has been increasing interest in the health benefits of homeownership. There are theoretical reasons to expect that owning a home should be associated with better health outcomes, and some empirical evidence indicates that this is the case. Homeownership—in line with expectations from Fundamental Cause Theory (Link and Phelan, 1995)—is associated with higher self-rated health in England (Munford, Fichera and Sutton, 2020) and the U.S., with White Americans having stronger associations than non-White Americans (Finnigan, 2014). Other research links homeownership to improved mental health (Manturuk, 2012; Elsinga, 2007; Rossi and Weber, 1996; Kearns et al., 2000). Few studies have investigated the relation between homeownership and mortality, and none in the U.S. In Finland, the magnitude of the association between homeownership and longevity is substantial: a study of Finnish homeowners found that owning a home—as opposed to renting—is associated with a 29% lower all-cause mortality hazard rate after adjusting for socioeconomic factors, level of urbanization, and household composition (Laaksonen, Tarkiainen and Martikainen, 2009).

Due to the large positive selection into homeownership, identifying any causal effect of homeownership on health is challenging. Munford, Fichera and Sutton (2020) used an innovative strategy, exploiting exogenous variation in the "Right to Buy" policy implemented in

England. The "Right to Buy" policy allows long-term renters of publicly-owned properties the right to buy their homes at a large discount. Using geographic heterogeneity in maximum discount caps as an instrument, the study found homeownership increases self-assessed health by 0.19 points on a 5-point scale. However, it is unclear whether this causal relationship between homeownership and self-rated health would also extend to the U.S. context or whether a similar causal relationship would also be found between homeownership and longevity. As summarized by Dietz and Haurin (2003): "Drawing conclusions about the causal relationship between housing tenure status and health requires additional empirical investigation using rigorous methods." To date, there is no evidence on whether or not homeownership in the U.S. has a causal effect on longevity. This study addresses that gap.

Historical Context

Homeownership in the United States has a long, fraught history and is shaped by generational wealth, class privilege, and racism. This has resulted in striking Black-White disparities in homeownership rates stemming from racist, exclusionary policies, as well as disparities in overall home quality and lending terms driven by predatory marketing and loan practices targeting Black Americans (Taylor, 2019). In order to contextualize the findings of this study within the broader literature on homeownership, it is essential to consider this historical legacy.

In the aftermath of the Civil War, Black Americans stood to inherit little land or property, as their parents were predominantly enslaved. During the Reconstruction Era of 1863–1877, land seized by the federal government during the Civil War was not set aside for the exclusive settlement of Black families as many advocated but rather was returned to its former Confederate owners by presidential decree, dispossessing thousands of temporary Black landowners (Davis, 1992). Most Black Americans had little choice but to resume working on White plantations as wage laborers or sharecroppers and had limited hope of upward economic mobility. In 1870, only 7.7% of Black male household heads owned homes, while 57% of White male household heads owned homes (Collins and Margo, 2011).[1] During this period, the modern mortgage markets were practically non-existent, and purchasing a home often required down payments of 50% or more, effectively barring most Black Americans from owning homes.

Black homeownership rates did increase modestly between 1870 to 1910 (Figure 2.1), likely driven by modest increases in educational and occupational attainment among Black Americans (Collins and Margo, 2011). Yet, beginning in 1910, homeownership among Black Americans stagnated. The first wave of the Great Migration saw a dramatic rise in the Black population in the North, with most Black Americans settling in urban areas proximal

---

[1]In this study, we focus on household heads, the primary renter or owner of a dwelling. This allows us to better isolate the relation between homeownership and longevity by excluding adults living in units owned or rented by someone else (e.g., parents). Our analysis is also limited to men due to the difficulties linking women in the 1940 Census to mortality records due to surname changes at marriage. Exploring the relationship between homeownership and longevity for women is an important avenue for future research.

to business districts with employment opportunities (Boustan, 2017). While these urban jobs paid better than agricultural employment in the South, this often did not translate into higher rates of homeownership. The nature of urban jobs being situated in densely populated areas led to significant commuting costs, prompting most workers to seek housing in urban regions with primarily rental units available. Additionally, Black Americans faced racial discrimination and bias when attempting to purchase homes or relocate, as White Americans opposed the influx of new Black residents.
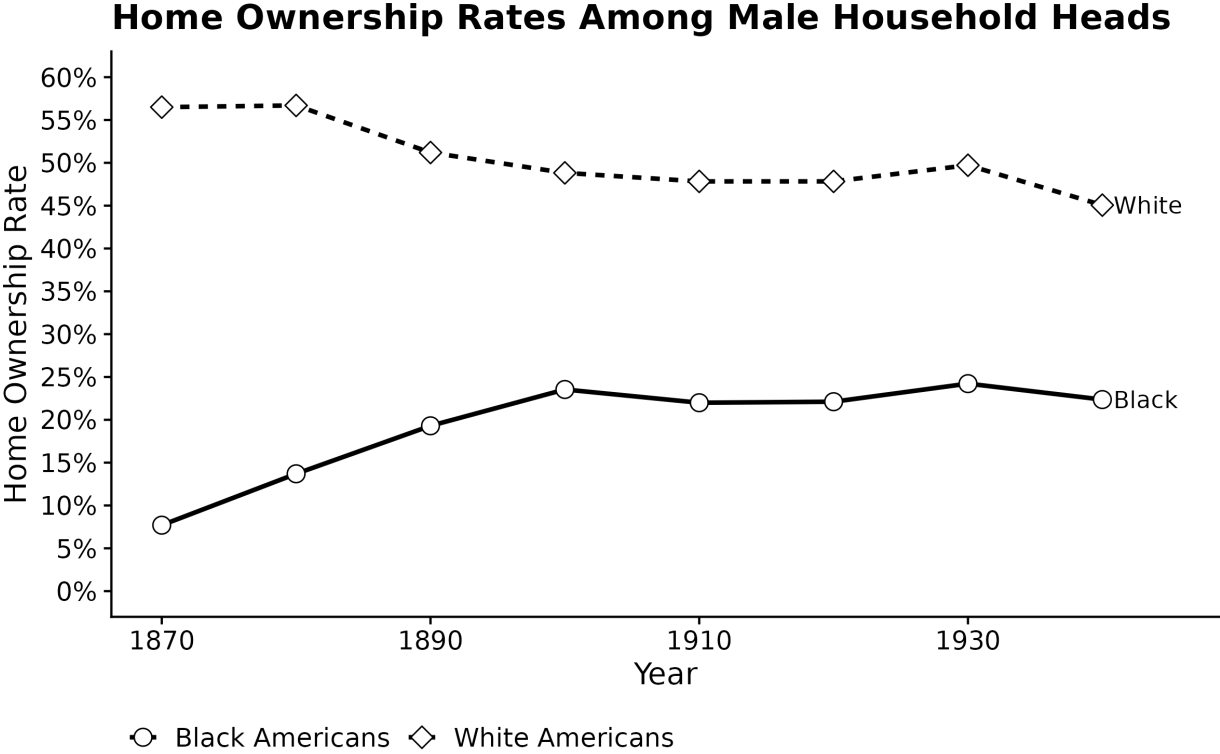
Figure 2.1: Black-White differences in homeownership among male household heads between 1870 and 1940. To estimate homeownership rates for 1900, 1910, 1920, 1930, and 1940, we use Full-Count Census records (Ruggles et al., 2020). For earlier census decades, we use estimates of the homeownership rates from Munford, Fichera and Sutton (2020).

Between 1870 and 1930, rates of homeownership for Whites declined modestly. This decrease was primarily caused by increases in real income being offset by migration away from rural areas and farms into dense urban areas that offered less opportunity for home-ownership (Collins and Margo, 2011). The gap between homeownership rates for Black and White Americans was nearly cut in half during this period. Yet the gap still remained stark, with White male household heads owning homes at rates approximately 25 percentage points higher than their Black counterparts from 1900 to 1940.
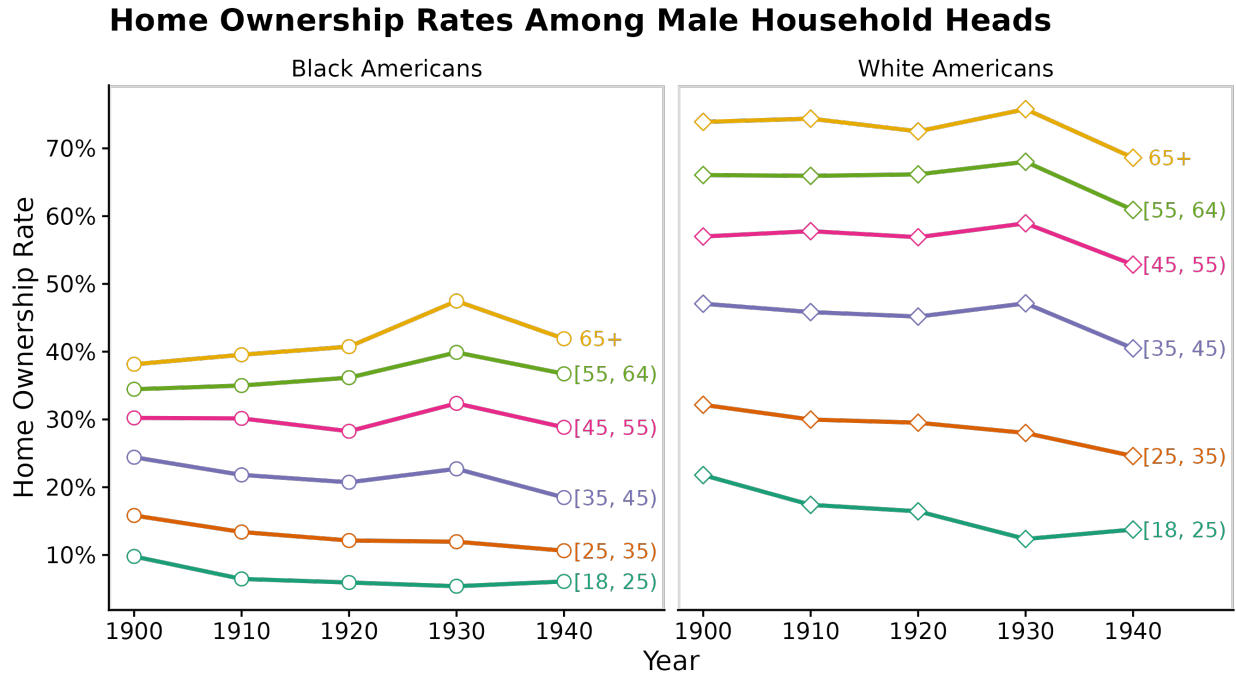
Figure 2.2: Black-White differences in homeownership rates in the early 20th century. Estimates are from the complete-count Decennial Census records (Ruggles et al., 2020)

This study focuses on homeownership measured in the 1940 Census. The major historical event leading up to the 1940 Census was the Great Depression, which saw the greatest evaporation of wealth in U.S. history. During this period, many Americans lost their home: the homeownership rates observed in the 1940 Census were lower than the homeownership rates in any other 20th century decennial census (Collins and Margo, 2011). Mortgage foreclosure rates were highest in "boom" cities that had experienced the highest rates of construction in the mid-late 1920s.

In the 1930s, racism in mortgage markets was rampant. During this period, the federal government first entered the residential mortgage market with two programs: the Home Owners Loan Corporation (HOLC) and the Federal Housing Administration (FHA). The HOLC produced a series of color-coded maps (redlining), which have been widely cited as increasing racial segregation in housing and a source of contemporary wealth inequality (Rothstein, 2017). The long-term effects of redlining are now an ongoing area of debate (Fishback et al., 2021; Aaronson et al., 2022), but there is conclusive evidence that Black Americans were discriminated against in the mortgage markets (Taylor, 2019; Michney and Winling, 2020; Chivers, 1949). Redlining did create disparately lower home values and drive economic isolation in Black neighborhoods. However, the ultimate effect of redlining on Black-White mortality disparities was modest, suggesting that redlining is only one feature of a larger public-private effort conflating the racial composition of neighborhoods with risk that affects

mortality (Graetz and Esposito, 2022).

Figure 2.2 shows the powerful age gradient in homeownership for both Black and White Americans in 1940. While fewer than 10% of Black male household heads between the ages of 18–25 were homeowners in 1940, over 40% of Black male household heads over the age of 65 owned homes. In this study, we only observe homeownership status between the ages of 24–35, and many people categorized as a renter in our study will become homeowners later in their life course. However, homeownership exit, the transition from homeowner to renter, is rare: only 6% of household-head homeowners became renters over a 40-year observation period between 1968–2009 (Sharp and Hall, 2014). In this sense, our paper builds an evidentiary base, demonstrating differences in life expectancy for those who own homes and those who rent in early adulthood. The homeowners we observe generally will continue to be homeowners throughout their life course, while renters may or may not become homeowners later in life.

Theoretical Framework

There are several compelling theoretical reasons to expect the relationship between homeownership and longevity to be causal. Figure 2.3 shows several potential mechanisms linking homeownership with increased life expectancy. Although this study advances a causal argument, due to the challenges of untangling various interconnected pathways using available data, it is not possible to pinpoint the precise contribution of alternative mechanisms. Instead, it is useful to consider the primary theoretical pathways that might, individually or in concert, account for the longevity benefits of homeownership. Given that we are observing homeownership early in the life course, we see the longevity benefits of homeownership through a cumulative advantage perspective (DiPrete and Eirich, 2006). That is, many of these potential mechanisms will compound and interact throughout the life course to create mortality disparities between homeowners and renters.

Wealth Accumulation

One causal pathway through which homeownership can affect longevity is wealth accumulation. The home is the single largest component of non-pension wealth in the United States (Apgar and Di, 2006), and homeownership is a key vessel for wealth accumulation. There are three primary reasons why homeownership is wealth-enhancing. First, homeownership may reduce housing costs by allowing homeowners to save on high rental prices and enjoy tax benefits in the form of tax deductions on mortgage interest and no capital gains tax (Killewald and Bryan, 2016). Second, homes generally appreciate in value over time slightly more rapidly than inflation, although this is location and period dependent and may be offset by maintenance costs (Harding, Rosenthal and Sirmans, 2007). Finally, monthly mortgage payments for homeowners encourage savings, which can buttress economic security in retirement (Boehm and Schlottmann, 2008).
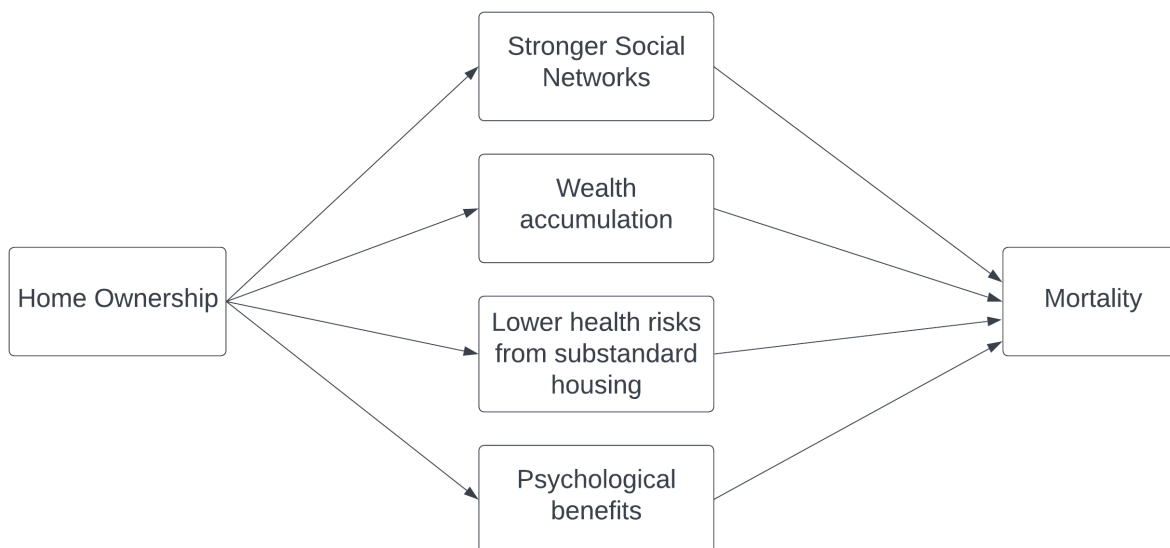
Figure 2.3: Causal pathways between homeownership and mortality.

Estimating the causal effect of homeownership on wealth is challenging because wealth prior to homeownership is a large confounder of the relationship between owning a home and wealth. Further, it is impossible to observe an individual's counterfactual use of financial resources (e.g., spending and investment) if they did not purchase a home. However, the empirical evidence to date confirms that homeownership does indeed have a causal effect on wealth. Turner and Luea (2009) find an additional year of homeownership is associated with a $15k annual increase in wealth holdings for high-income groups and a $6–10k increase for low-income groups. However, this analysis does not control for wealth prior to owning a home and likely overstates the relationship between wealth and homeownership. Di, Belsky and Liu (2007) estimate the effect of homeownership on personal wealth over a 12-year period, controlling for baseline wealth and pre-trends in wealth accumulation using data from the Panel Survey of Income Dynamics. They find that each additional year of homeownership causes a $3k–$14k increase in wealth annually. Herbert, McCue and Sanchez-Moyano (2013) also control for baseline wealth, estimating an association of homeownership on wealth of approximately $9,500, with some modest discrepancies for White Americans ($10,542) and Black Americans ($8,474). Finally, Killewald and Bryan (2016) use marginal structure models and the National Longitudinal Survey of Youth 1979 (NLSY79) between 1985 and 2008 to investigate the effect of homeownership on wealth. They find homeownership is wealth-enhancing, but to a smaller extent than other estimates: an additional year of homeownership creates a $6,800 increase in wealth. In sum, across different data sources, time periods, and estimation strategies, owning a home has repeatedly been found to be wealth-enhancing.

Social Networks

A second pathway between homeownership and lower mortality is stronger community integration and social support networks. Homeowners stay in a unit longer than their renter counterparts (Rohe and Stewart, 1996), fostering stronger feelings of community attachment, integration, and commitment to their neighborhood. The empirical evidence to date indicates that homeowners are more likely to be socially and politically involved, even after controlling for socioeconomic characteristics (Wandersman, 1981; Rohe and Stegman, 1994). Additionally, homeowners' vested interest in the conditions of their neighborhood often leads to higher rates of participation in local neighborhood organizations and interactions with other members of the neighborhood (Davis and Fine-Davis, 1981).

Such neighborhood interactions help homeowners build more social capital and become more socially integrated than their renter counterparts (Manturuk, Lindblad and Quercia, 2010). Higher social integration and social capital have a clear positive impact on health outcomes (Berkman, Leo-Summers and Horwitz, 1992; Berkman and Syme, 1979; Smith and Christakis, 2008). Social support can serve as a moderator of life stress and lead to higher compliance with medical regimes, faster recovery, and protective effects against depression and other conditions (Cobb, 1976).

Housing Conditions

A more direct pathway between homeownership and longevity is through superior housing conditions. Living in adequate housing is strongly linked to longevity. Renters are more likely than homeowners to experience infectious diseases, injuries, and chronic conditions (Krieger and Higgins, 2002). Rental units were often overcrowded, dirty, and improperly ventilated, conditions that can facilitate the spread of tuberculosis, influenza, pneumonia, and other infectious diseases. Racial differences in mortality at the beginning of the 20th century were largely attributed to disparities in respiratory diseases (Feigenbaum et al., 2022), highlighting the impact of racial segregation in housing on population health. Quotas that reduced overcrowding in rental units had a considerable impact on infectious disease mortality (Ager et al., 2021), demonstrating the connection between adequate housing and longevity.

Compared to renters, homeowners have more autonomy to make improvements and modifications to increase the living conditions of their homes. Homeowners also have more financial incentives to make these modifications, as such modifications can increase their home's value. Controlling for the characteristics of the occupant and unit, homeowners invest more in home maintenance and repairs than renters (Galster, 1983). Owning a home is associated with a 13–23% increase in the quality of the home environment, attributable to home investments lowering levels of lead-based paint, unhygienic living conditions, structural hazards, and other factors (Haurin, Parcel and Haurin, 2002).

Psychological Benefits

Finally, homeownership has the capacity to promote well-being through higher levels of self-efficacy, self-control, and stability. Homeowners also report greater feelings of control and self-determinism over their lives because their external environment is predictable and dependable (Manturuk, 2012). Additionally, homeowners are at lower risk of the instability caused by eviction; frequent relocations are associated with higher risks of anxiety and other mental health issues (Acharya, Bhatta and Dhakal, 2022; Manturuk, 2012). Homeowners also report higher levels of satisfaction in their living conditions than renters, which factors into overall levels of life satisfaction.

Owning a home is also a key way of communicating social status. The higher social status of homeowners may increase their self-esteem and overall life satisfaction (Tremblay and Dillman, 1983; Rakoff, 1977). Additionally, being part of the dominant "successful" group can avoid associated stressful downward comparisons with others (Elsinga, 2007). Together, these psychological well-being benefits promote physical and mental health and, ultimately, longevity.

2.3 Data

This study uses digitized complete-count census records, mortality records, and record linkage techniques to construct a longitudinal panel of male siblings. Specifically, we use the CenSoc-DMF file (Goldstein et al., 2021), which links the IPUMS complete-count 1940 Census (Ruggles et al., 2020) with mortality records from the Social Security Death Master File (DMF). The DMF is a collection of over 85 million death records reported to the Social Security Administration, capturing 95%+ of deaths occurring after the age of 65 between 1975–2005 (Alexander, 2018; Hill, 2001). We limit our analysis to men born in 1905–1915 who were between the ages of 24–35 when they were enumerated in the 1940 Census. We focus on these cohorts as we can both observe these cohorts as household heads in the 1940 Census and capture much of their mortality in our mortality observation window. We use this sample to study the association between homeownership and longevity ($N = 1{,}361{,}883$).

To identify brothers, we link men in the 1940 Census—which contains no information allowing for the systematic identification of adult brothers—back to the 1920 Census. In the 1920 Census, we use information from household rosters to identify biological brothers aged 4–15 living together in the same household in 1920. We use this sample to identify the causal effect of homeownership on longevity ($N = 84{,}945$). This process is illustrated in the Lexis diagram shown in Figure 2.5.

The 1940 Census

The 1940 Census was conducted in April 1940, at the tail end of one of the most eventful decades in U.S. history. The 1930s began with the Great Depression, the longest and deepest depression in the 20th century. Between 1929 and 1933, employment decreased 17.4% (Margo,
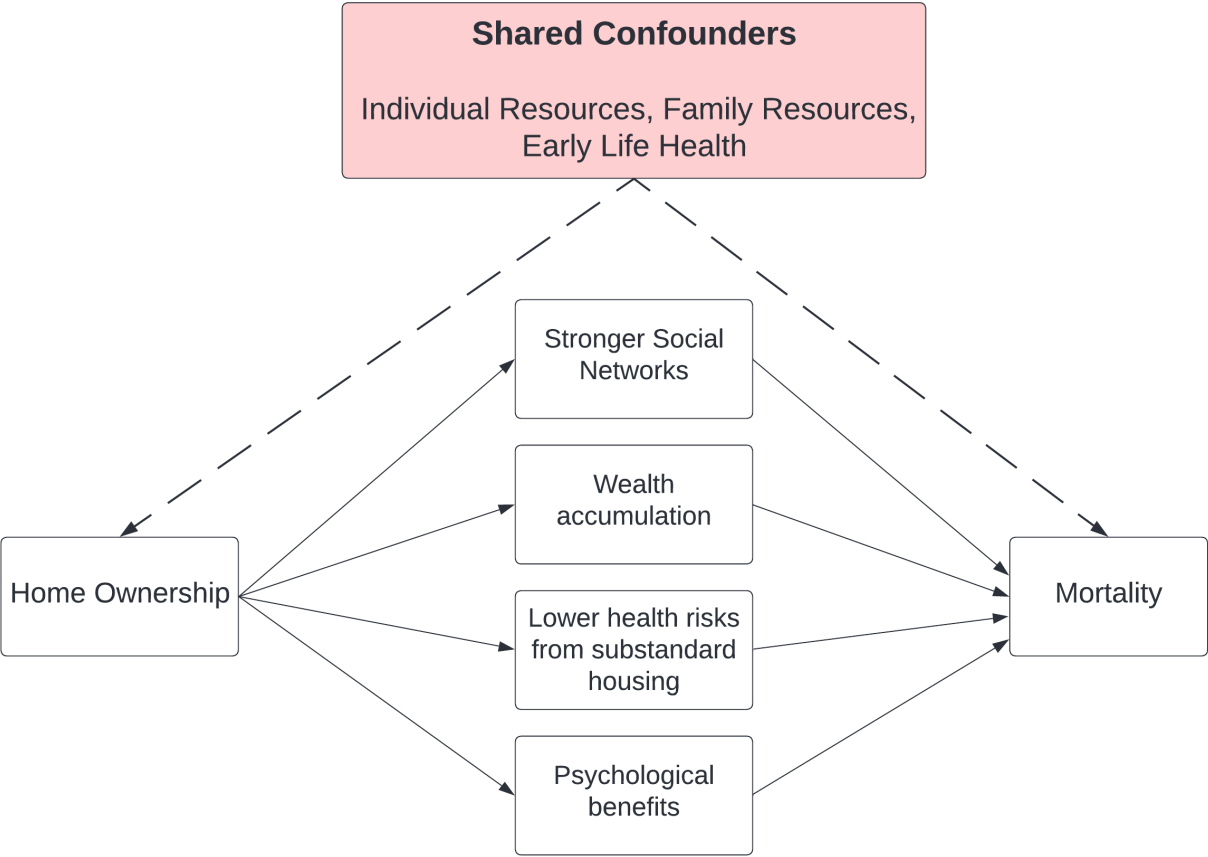
Figure 2.4: The solid lines show the causal pathways between homeownership and mortality. The dashed lines show the potential confounding pathways between homeownership and longevity.

1993) and manufacturing contracted over 30% (Lee and Mezzanotti, 2017). To combat the Great Depression, President Franklin Roosevelt introduced a series of unprecedented expansions of government-allocated aid to provide immediate relief and promote economic recovery in his First and Second New Deal. The questions asked in the 1940 Census reflected this time of heightened social awareness: for the first time, a decennial census included questions on wage and salary income, educational attainment, and employment as part of an emergency relief program.

The 1940 Census also included several questions on homeownership. First, 1940 Census enumerators asked whether the home or dwelling unit was owned or rented, regardless of whether it was still being paid for by a mortgage (Ruggles et al., 2020).[2] Second, enumerators collected information on the value of the home for homeowners and the amount of monthly rent paid for renters. Unless the home was recently purchased, enumerators were instructed to estimate the current market value of the home. Specifically, enumerators were instructed
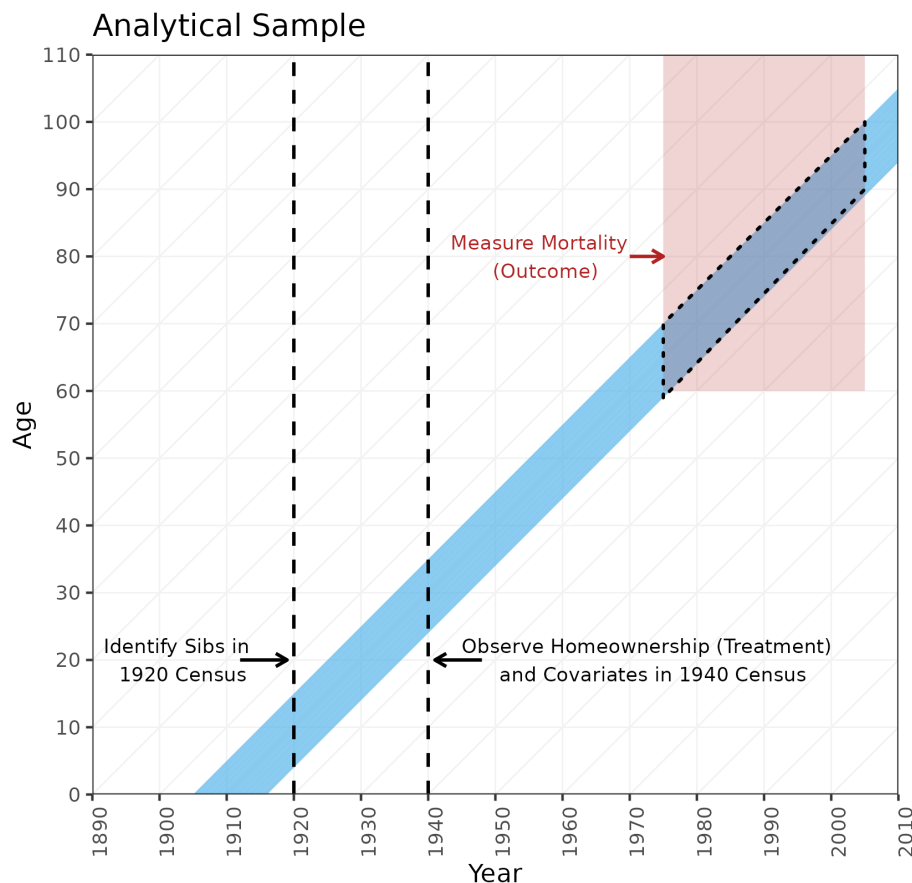
Figure 2.5: A Lexis diagram illustrating our analytic sample. The focal birth cohorts of this study (1905–1915) are highlighted in blue. We first observe these men in childhood between the ages of 4–15 living together in the same household in the 1920 Census, where we use information from the household roster to establish sets of brothers. Next, we observe these individuals in the 1940 Census, where we observe homeownership status and other characteristics such as educational attainment, wage and salary income, occupation, and marital status. Finally, we observe these men dying after the age of 60 in the DMF mortality coverage window of 1975–2005.

to "represent the amount for which the home, including (except on a farm) such land as belongs to it, would sell under ordinary conditions not at forced sale" (Ruggles et al., 2020).

Finally, the 1940 Census included a question about whether the household was on a farm.

Sibling Sample

The 1940 Census does not include information allowing for the systematic identification of adult siblings living in different households. To overcome this, we identify brothers in the 1920 Census living in the same household using household rosters. We focus on the birth cohorts of 1905–1915, who are between the ages of 4–15 in the 1920 Census. We link the brothers identified in the 1920 Census to the 1940 Census using linkages provided by the Census Linking Project (Abramitzky et al., 2020). The linkages were constructed using the ABE algorithm (Abramitzky, Boustan and Eriksson, 2012, 2014; Abramitzky et al., 2021), which links on first name, last name, place of birth, and year of birth. This algorithm first standardizes first names to account for common misspellings or nicknames (e.g., Robbie → Robert). It then establishes matches based on an exact match on first name, last name, and place of birth while allowing for some flexibility (± 2 years) on birth year. This choice to use a relatively conservative "exact match" results in a smaller analytic sample but minimizes the number of false matches, which pose the greatest threat to statistical inference (Ruggles, Fitch and Roberts, 2018).[3] For our study, false matches would mean that we are not doing actual within-sibling comparisons, threatening our sibling-based identification strategy.

To link our set of siblings to their mortality records, we use the publicly-available CenSoc-DMF (V2.1) dataset (Goldstein et al., 2021). This file links men in the 1940 Census to Social Security mortality records in the Death Master File (DMF) again using the ABE record linkage algorithm (Abramitzky et al., 2020). The DMF includes nearly complete death coverage from 1975–2005 over the age of 65 (Alexander, 2018; Hill, 2001), and approximately 30% of these deaths are successfully matched back onto a 1940 Census record (Breen and Osborne, 2022). For brothers to be included in our sibling sample, two or more brothers living in the same household in 1920 must have been successfully linked to both the 1940 Census and the DMF mortality records.[4]

In Table 2.1, we compare the composition of our matched samples to the composition of all men aged 24–35 in the 1940 Census. Overall, our matched samples are reasonably representative of the general population, although men with higher socioeconomic status are slightly overrepresented. Further, consistent with other linkage efforts, Black Americans are underrepresented. This underrepresentation can be attributed to lower rates of linkage due

---

[2]A household was defined in the 1940 Census as "a family or any other group of persons living together, with common housekeeping arrangements, in the same living quarters." One member of each household—almost exclusively a man—was designated as the "household head." The household head reported on the other members of the household to the enumerator. A home or dwelling unit was marked as "owned" if the household head or another member of the family owned the home.

[3]As a robustness check, we use the more strict ABE-Conservative algorithm, which requires names to be unique within a ± 2 year window. Our results are robust across linkage algorithms (see Appendix A.1 for details).

to higher rates of age misreporting and levels of name homogeneity (Goeken et al., 2011).

To summarize, we create a longitudinal panel of brothers by first extracting all records for children ages 4–15 in the 1920 Census ($N = 16{,}321{,}702$). We then link the 1920 Census to the 1940 Census using ABE linkages ($N = 4{,}446{,}863$), corresponding to a linkage rate of 27% (Abramitzky et al., 2020). Next, we link these individuals to their death records in the DMF ($N = 1{,}443{,}728$). Finally, we identify sets of brothers in these data ($N = 84{,}945$) and restrict to household heads. Of these sibling pairs, 34.5% were discordant on homeownership status.

---

[4]If we do not successfully link an individual from the 1940 Census onto the mortality records, this may be due to (1) the individual died outside of our mortality observation window or (2) we were not able to establish a link using our record linkage strategy. Therefore, conventional methods for working with censored data are not appropriate in this setting, and our analysis is restricted to sibships where two or more siblings were successfully linked to both the 1940 Census and the DMF mortality records.

Educational Attainment						
<High School	4,951,782	67.3	608,639	64.7	56,591	66.6
High School or some college	1,783,203	24.3	247,103	26.3	22,176	26.1
Bachelors Degree	339,072	4.6	48,024	5.1	3,504	4.1
Advanced Degree	162,122	2.2	24559	2.6	1,673	2.0
Not Available	117,086	1.6	12,091	1.3	1,001	1.2
Race						
Black	656,027	8.9	34,159	3.6	1,020	1.2
Other	27,778	0.4	3,296	0.4	97	0.1
White	6,669,460	90.7	902,961	96.0	83,828	98.7
Marital Status						
Married	7,013,184	95.4	905,924	96.3	82,444	97.1
Not married	340,081	4.6	34,492	3.7	2,501	2.9
Home Ownership						
Homeowner in 1940	1,780,906	24.2	249,379	26.5	24,032	28.3
Not Homeowner in 1940	5,572,359	75.8	691,037	73.5	60,913	71.7
Socioeconomic Status						
Sei 1-9	1,293,523	17.6	138,209	14.7	12,966	15.3
Sei 10-14	1,170,543	15.9	149,673	15.9	16,924	19.9
Sei 15-25	1,862,967	25.3	246,484	26.2	22,004	25.9
Sei 26+	2,776,321	37.8	380,226	40.4	30,969	36.5
Not Available	249,911	3.4	25,824	2.7	2,082	2.5
Rural						
Rural	3,183,160	43.3	397,739	42.3	43,320	51.0
Urban	4,170,105	56.7	542,677	57.7	41,625	49.0
Region String						
East North Central Div.	1,485,519	20.2	235,080	25.0	24,224	28.5
East South Central Div.	629,263	8.6	51,407	5.5	3,670	4.3
Middle Atlantic Division	1,420,842	19.3	193,013	20.5	13,734	16.2
Mountain Division	257,901	3.5	33,581	3.6	3,773	4.4
New England Division	396,893	5.4	54,448	5.8	4,093	4.8
Pacific Division	618,800	8.4	88,569	9.4	7,871	9.3
South Atlantic Division	1,004,976	13.7	82,833	8.8	6,028	7.1
West North Central Div.	724,762	9.9	118,185	12.6	14,957	17.6
West South Central Div.	814,309	11.1	83,300	8.9	6,595	7.8
Total	7,353,265	100	940,416	100	84,945	100

Table 2.1: Representativeness of our analytic samples. The first two columns report the sociodemographic composition of all men between the ages of 24-35 in the 1940 Census. The third and fourth columns report the sociodemographic composition of men aged 24-35 in the 1940 Census who were successfully matched to a mortality record in the Death Master file. Finally, the fifth and sixth columns report the sociodemographic composition of our subsample of brothers.

2.4 Methods

Unadjusted Difference in Life Expectancy Between Homeowners and Renters

We fit separate OLS regression models for Black and White men with our full linked sample (N = 1.4 million) to estimate the unadjusted difference in life expectancy between homeowners and renters. Because the CenSoc-DMF only contains deaths for the left and right ("doubly") truncated window of 1975–2005, the magnitude of the reported differences across groups will be smaller than if we had the complete window of deaths (Goldstein et al., 2023). For each birth cohort, we observe a different window of ages of death. To account for this, we include birth-year fixed effects to control for the different distribution of birth years across population subgroups. We fit models of the form:

$$D_i = \beta_0 + \lambda_{byear} + \delta_{homeown} + \epsilon \tag{2.1}$$

$$D_i = \beta_0 + \lambda_{byear} + \delta_{homeown} + \boldsymbol{\beta} X_{controls} + \epsilon \tag{2.2}$$

where D_i is age of death, β_0 is the general intercept, λ_{byear} is a fixed effect for a given year of birth, $\delta_{homeown}$ is a dummy variable for whether an individual is a renter or a homeowner, $\boldsymbol{\beta}$ is a set of regression coefficients, and $X_{controls}$ is a vector of adjustment variables. The model described by Equation (2.1) only includes fixed effects for year of birth. Our second model described by Equation (2.2) additionally adjusts for educational attainment in years, race, wage and salary income, occupation, urbanicity, state of residence in 1940, and marital status. We fit these models separately for Black ($N = 65,053$) and White ($N = 1,296,830$) Americans.

Identification Strategy

Theoretical Estimand

The primary inferential goal of this study is to estimate the causal effect of homeownership on longevity. To estimate this quantity, we first define our theoretical estimand, which is the precise quantity we are interested in estimating (Lundberg, Johnson and Stewart, 2020). Theoretical estimands are composed of two key building blocks: *unit specific quantity* and *target population*. The *unit specific quantity* is a quantity defined for each unit of the population. The *target population* is the set of units over which the *unit specific quantity* is aggregated.

In our analysis, the *unit specific quantity* is the counterfactual difference in life expectancy if a male household head owned a home in early adulthood versus if they rented a home in early adulthood. The *target population* is all men born in the U.S. between 1905 and 1915. Using the potential outcome framework (Imbens and Rubin, 2015), we can define our causal estimand (Ψ) as an average treatment effect (ATE):

$$\Psi_{\text{ATE}} = \underbrace{\underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\substack{\text{Mean over every } i \text{ among} \\ \text{male household heads}}}}_{\text{Target Population}} \left(\underbrace{D_i(homeowner)}_{\substack{\text{Life expectancy if} \\ \text{homeowner in early adulthood}}} - \underbrace{D_i(renter)}_{\substack{\text{Life expectancy if} \\ \text{renter in early adulthood}}} \right) \quad (2.3)$$

$$\underbrace{}_{\text{Unit-Specific Quantity}}$$

Empirical Estimand

It is impossible to estimate our theoretical estimand Ψ using only our observational data, as we can only observe one potential outcome. To convert the theoretical estimand into an empirical estimand,[5] we must define a quantity that can be estimated from our data alone. This requires us to formulate an empirical estimand that can be measured with our available data (Lundberg, Johnson and Stewart, 2020).

To estimate our empirical estimand from our observed sample of siblings, we fit an OLS regression with sibling fixed effects:

$$D_i = \beta_0 + \lambda_{byear} + \delta_{homeown} + \boldsymbol{\beta} X_{controls} + \Omega_{SiblingFE} + \epsilon \quad (2.4)$$

This class of models has been used to estimate the effect of education on longevity (Halpern-Manners et al., 2020), neighborhood effects on social mobility (Chetty and Hendren, 2018), and the effect of social programs on educational policies (Currie and Thomas, 1995). The idea behind the sibling fixed-effect models is to control for unobserved heterogeneity within families (Conley, Pfeiffer and Velez, 2007):

$$D_{ij} - D_{is} = \beta \left(X_{ij} - X_{is} \right) + (\alpha_i - \alpha_i) + (\mu_{ij} - \mu_{is}) \quad (2.5)$$

where D_{ij} is the age of death of individual i in household j, D_{is} is the average age of death of siblings in household j, X_{ij} is a vector of control variables for individual i in household j and X_{is} is a vector of control variables for siblings in household j. For both the individual and sibling vectors of control variables, we include birth year fixed effects and birth order fixed effects. In this equation, the error term is now broken down into two different components: α_i, the sibling fixed effect, and μ_{ij}, the individual-level error term for individual j in family i. By differencing across siblings in each family, we effectively eliminate the unobserved sibling effects. In this setting, the inclusion of sibling fixed effects is critical because they partially eliminate unobserved contextual factors such as family wealth and genetic factors from our model (Halpern-Manners et al., 2020).

Like all attempts at establishing causality with observational data, there are several limitations to this approach. The sibling fixed effect design does not capture variation between

---

[5]The key difference between the theoretical estimand and the empirical estimand is that the theoretical estimand is the theoretical quantity we are interested in learning about. Stating the theoretical estimand is helpful for precisely stating the research goal. The empirical estimand states the empirical quantity that our data and statistical methods allow us to estimate.

families, only variation within families. If there is unobserved individual-level heterogeneity that is correlated with homeownership, our estimates of the causal effect of homeownership on longevity may be biased. Moreover, although siblings provide a broader representation of general environments and experiences compared to twins, they also share only half as many genes, which introduces the potential for genetic confounding. In other words, there could be genes associated with both homeownership and longevity that are not accounted for in our analysis using sibling fixed effects. Simply put, if the reason that one sibling owns a home and the other(s) does not is related to longevity, this presents a threat to our causal interpretation.

2.5 Results

We first analyze the association between homeownership and longevity using our full pooled sample (N = 1.4 million). We fit models described by Equation (2.1) and Equation (2.2) on Black and White Americans separately. Figure 2.6 shows a clear mortality advantage for both Black and White homeowners. The unadjusted difference in life expectancy for Black Americans (0.50 years) and White Americans (0.53 years) is highly comparable. These regression estimates correspond to conservative estimates of the difference in life expectancy conditional on living to age 65 ($e65$).[6] After controlling for educational attainment in years, race, wage and salary income, occupation, state of residence in 1940, urbanicity, and marital status, the difference in life expectancy is 0.36 years for Black Americans and 0.42 years for White Americans.

Next, we turn to our subsample of siblings identified from our linkage to the 1920 Census. Here, we fit a model described by Equation (2.4) on a pooled sample; we are unable to fit separate models for Black and White Americans separately due to sample size limitations. Figure 2.7 again shows a strong positive association between homeownership and longevity for our baseline model without controls. Owning a home in 1940 is associated with a mortality advantage of 0.64 years.[7] After adjusting for covariates (educational attainment in years, race, occupation, urbanicity, state of residence in 1940, and marital status), the association is slightly attenuated: the mortality advantage is 0.53 years. Our final model, which includes sibling fixed effects to account for shared conditions in childhood such as family wealth, genetic endowment, and other hard-to-measure confounders, still shows a statistically significant mortality advantage of 0.31 years. The full regression tables for Figure 2.6

---

[6]Throughout the paper, we report our estimates from regression models. However, our regression models are fit on doubly truncated data, which downwardly biases our estimated regression coefficients (Goldstein et al., 2023). For interpretability, we convert these regression estimates into difference in life expectancy at age 65, $e(65)$ using a conventional parametric Gompertz approach assuming proportional hazards. This gives us an inflation factor of approximately 17% in converting our regression coefficients into estimates of differences in $e(65)$. For more technical details, see Appendix A.1.
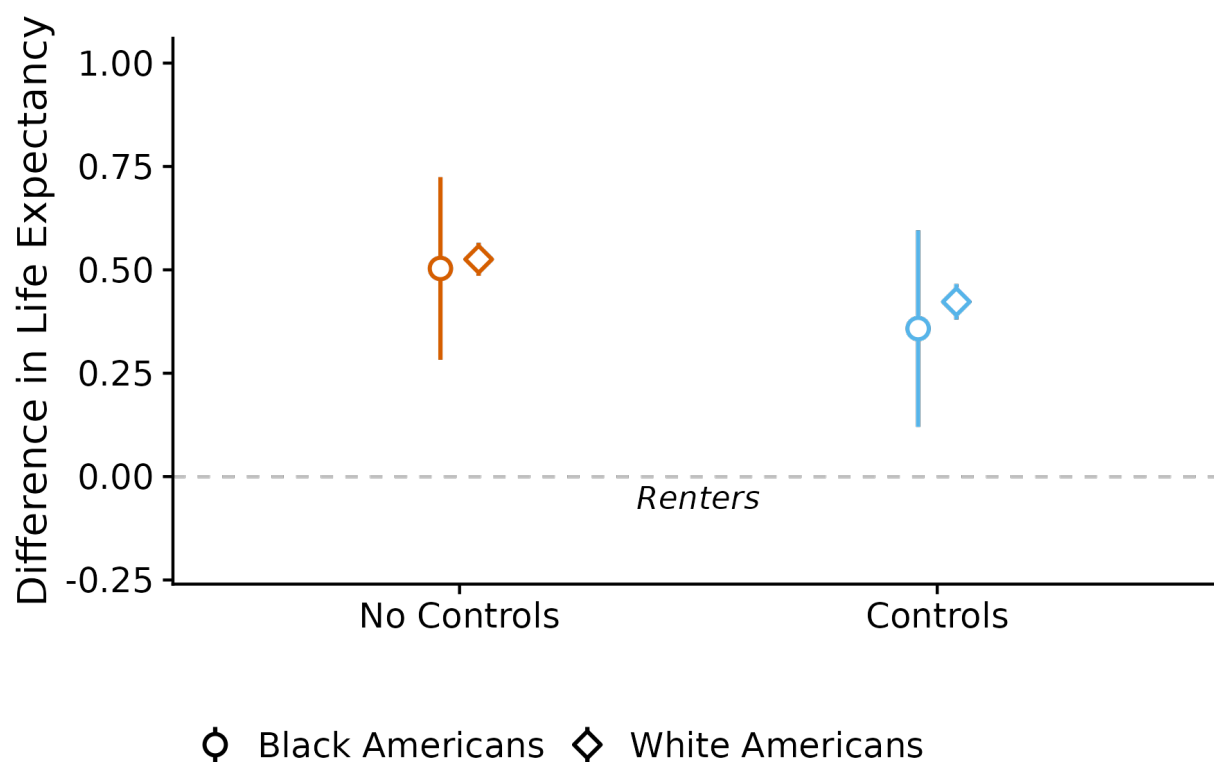
Figure 2.6: The unadjusted difference in life expectancy between homeowners and renters for Black and White Americans in the CenSoc-DMF. Uncertainty bars show 95% confidence intervals; for White Americans, the uncertainty bars are smaller than graph point size.

and Figure 2.7 are presented in Appendix A.1.

Sample Restriction

To address potential concerns of residual confounding within siblings, we refit our sibling fixed effect model on a series of different subsamples. The goal of this sensitivity analysis is to reduce potential residual confounding within siblings. For example, by restricting to siblings who are still living in the same county or have comparable educational attainment, we may account for unobserved heterogeneity, such as early life health, between siblings. For this analysis, we create five different subsamples: (1) brothers who are both married; (2)

---

[7]The estimated unadjusted difference in life expectancy between homeowners and renters in our subsample of siblings (0.52, 95% CI = [0.48, 0.56]) is lower than in our full sample (0.64, 95% CI = [0.53, 0.75]), but the difference in the associations is not statistically significant. This suggests our sibling subsample is not specially selected in a way that would bias the relationship between homeownership and longevity.
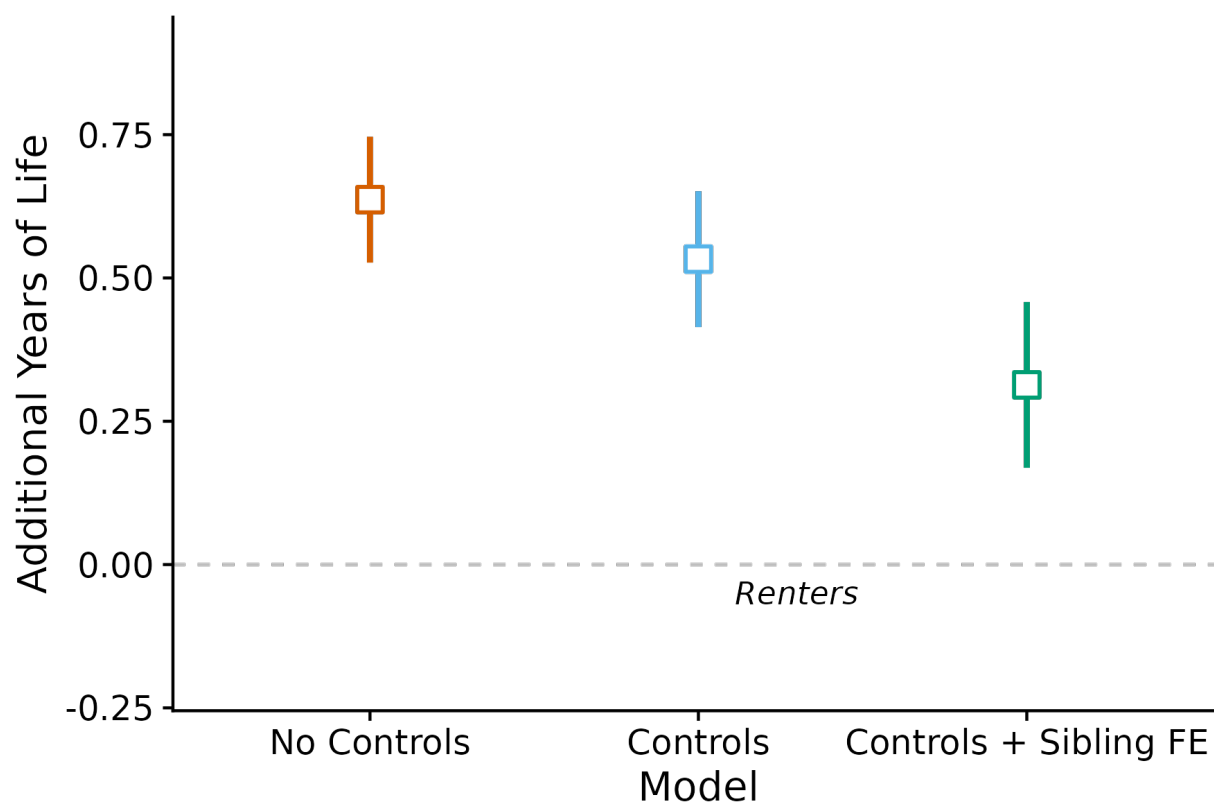
Figure 2.7: The relationship between owning a home and longevity from three regression models for our sibling sample. The "no controls" model adjusts only for birth cohort (left). The "controls" model adjusts for educational attainment, race, occupation, birth cohort, marital status, state of residence in 1940, and urban-rural status (middle). The "controls + Sibling FE" additionally includes sibling and birth order fixed effects (right). Uncertainty bars show 95% confidence interval.

brothers with ± 2 years of educational attainment; (3) brothers living in the same county in 1940; (4) brothers born within five years of each other; and (5) a sample with all of the aforementioned restrictions. Figure 2.8 shows our key result that owning a home has a causal effect on longevity is robust across all subsamples. We interpret this as evidence that our findings are unlikely to be driven by residual within–sibling confounding.

Does the Effect of Homeownership on Longevity Vary by Home Price?

In Figure 2.9, we examine whether more valuable homes have a stronger relationship with longevity compared to less valuable homes. We assign each home a quartile score based
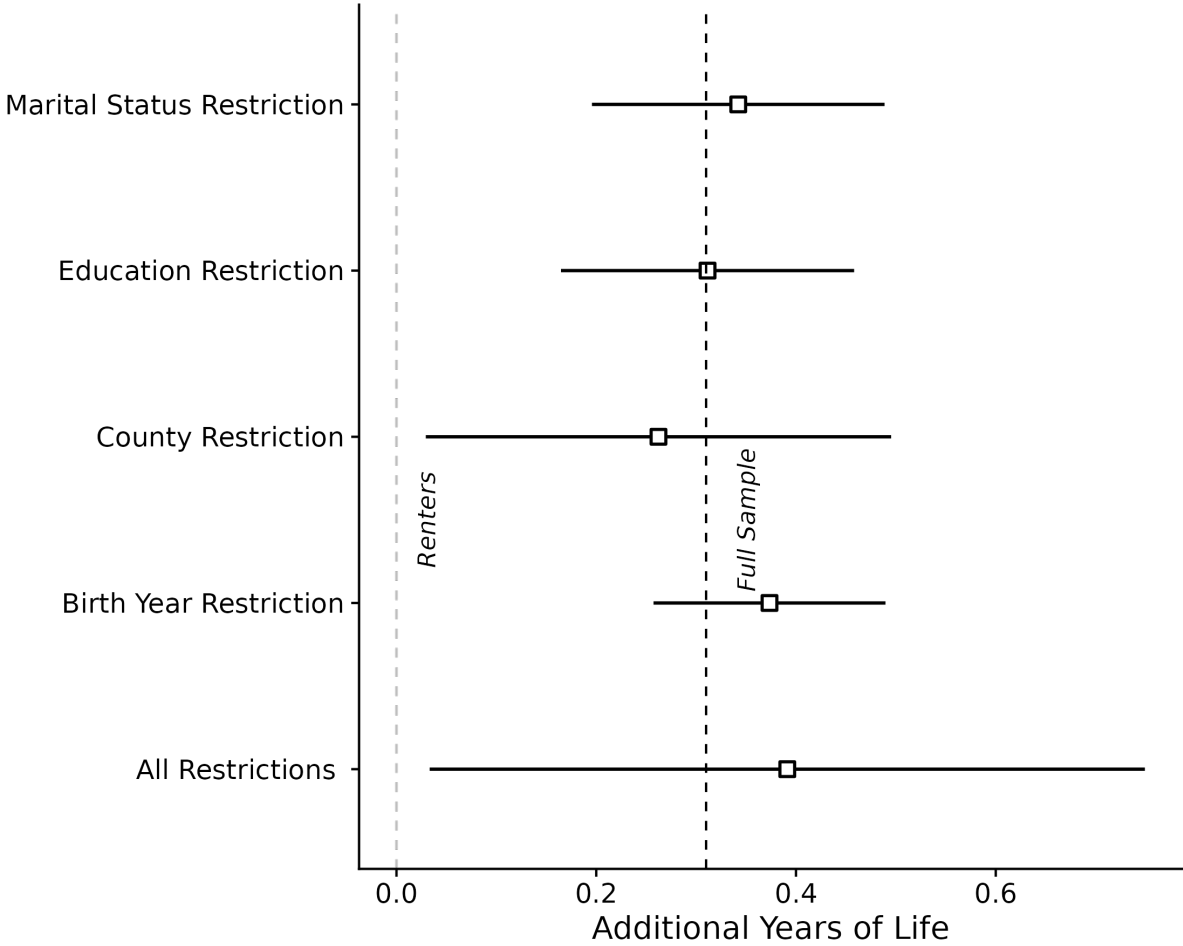
Figure 2.8: Estimates of the effect of homeownership on longevity from refitting our sibling fixed effect model on a subsample to account for potential within-sibling confounding. Our estimates do not meaningfully change across subsamples. The vertical black dashed line shows the estimate from the full sibling sample of 0.31 additional years of life.

on its value relative to other homes in the same state, as assessed in the 1940 Census. We then incorporate these quartile scores as four dummy predictors in our models. In our model without controls, we observe a linear increase in the unadjusted difference in life expectancy between homeowners and renters as home value rises. This association becomes less prominent in our model with controls and sibling fixed effects, indicating that higher-valued homes yield only a modest increase in life expectancy compared to lower-valued homes. However, the large uncertainty bands preclude us from making any definitive claims about effect heterogeneity with respect to home value.
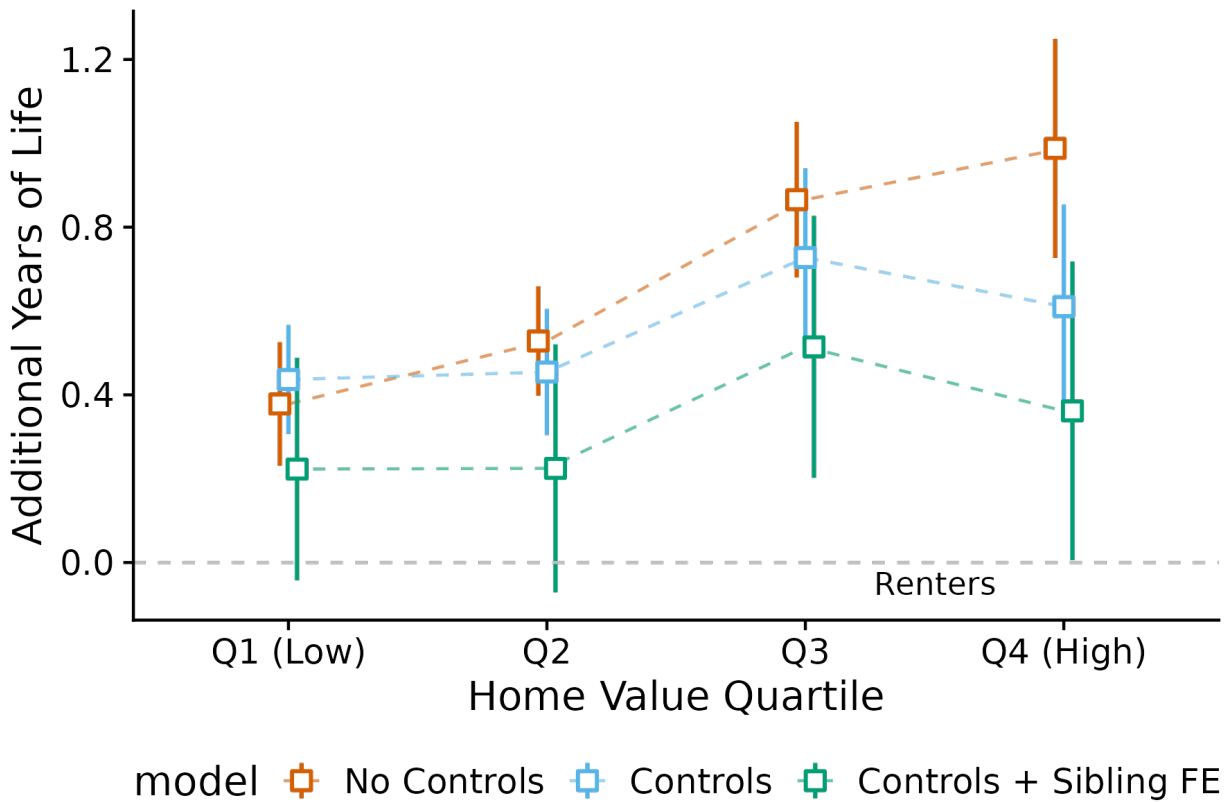
Figure 2.9: The relationship between owning a home and longevity for different levels of home value in the sibling sample. Each home was categorized according to its value relative to other homes in the same state. Uncertainty bars show 95% confidence intervals.

2.6 Discussion

We find a meaningful, statistically significant difference in life expectancy between homeowners and renters. Those who own a home in early adulthood live approximately six months longer at age 65 than those who rent, and controlling for a large set of observable demographic and socioeconomic characteristics only slightly attenuates this estimate. These results align with our theoretical expectations from Fundamental Cause Theory: homeowners are wealthier than renters, and wealth is almost universally associated with higher life expectancy.

In our second analysis, we use a sibling-based identification strategy, which partially controls for shared family environment and genetic endowment, to estimate the causal effect of owning a home on longevity. We find that owning a home in early adulthood has an effect of 0.31 years on life expectancy, equivalent to approximately one-third the Black-White mortality gap observed in our sample. We interpret this as evidence that homeownership

causes people to, on average, live meaningfully longer lives. This aligns closely with our theoretical framework that homeownership in early adulthood in the U.S. context will have compounded financial, social, and psychological advantages across the life course, culminating in a longer life. To give this estimate a causal interpretation requires making several assumptions, including that there is no residual confounding within sibling pairs (Boardman and Fletcher, 2015). However, the robustness of our results and our theoretical framework support our conclusion that homeownership has a causal effect on longevity.

Despite huge disparities in homeownership rates between Black and White Americans, we find no significant Black-White differences in the relationship between homeownership and longevity.[8] This is a surprising finding, as the median value of homes in our sample for White Americans ($2,000) is four times that of Black Americans ($500), and other studies have found stratified racial and ethnic differences in both health (Finnigan, 2014) and wealth-enhancing effects of homeownership (Killewald and Bryan, 2016). Further, many observable social determinants of mortality, such as educational attainment (Card and Krueger, 1992), have a weaker correlation with longevity for Black Americans than for White Americans.

The puzzlingly similar results for Black and White Americans may be explained by the differential timing of homeownership. Black Americans who were able to purchase a home before 1940 may have been spared some of the exploitative real estate practices that arose in the following decades, such as racist exclusion from the housing market and predatory inclusion characterized by overpriced, older housing in Black neighborhoods with unfavorable or exploitative loan conditions (Taylor, 2019). These policies and practices meant that many Black Americans did not benefit from homeownership to the same extent as White Americans due to lack of investment in Black neighborhoods (Sugrue, 2014). However, those who were able to purchase a home before 1940 may have been able to escape predatory inclusion practices that became more widespread in the following decades (Taylor, 2019).

We observe a relatively weak, positive correlation between the value of a home and its impact on longevity. Coupled with our analysis indicating similar mortality advantages for both Black and White homeowners, despite White Americans having substantially more valuable homes, this suggests that homeownership has an effect on longevity that goes beyond wealth. In other words, being a homeowner is, on average, beneficial for longevity, regardless of the value of the home.

There are several important limitations and caveats to this analysis that warrant discussion. Several methodological issues have been identified in the context of within-siblings designs (McGue, Osler and Christensen, 2010). Most importantly, while the results of this study suggest a causal effect of homeownership on longevity, residual confounding within sibling pairs—unobserved differences between siblings not accounted for in our model—is a valid concern. For instance, the sibling more interested in long-term financial security might buy a home earlier in their life course, and interest in long-term financial security

---

[8]Our sibling sample is not large enough to estimate a causal effect separately for Black and White Americans. However, it is plausible that the causal effects for Black and White Americans would be comparable, given the similar unadjusted differences in life expectancy between homeowners and renters.

may also be a correlate of healthy lifestyle choices. While we cannot *confirm* that there is no residual confounding within siblings, our robustness checks are reassuring. Specifically, we re-estimated our sibling fixed effect models on different subsamples (e.g., a subsample where both siblings lived in the same county in 1940) to account for potential unobserved heterogeneity between siblings. We find no evidence our findings are driven by residual within-sibling confounding.[9]

It is also important to note our study focuses only on male household heads, excluding women, lodgers, and men living with their parents. In particular, investigating the impact of homeownership on longevity for women is a crucial direction for future research. Our linked samples also overrepresent White men with higher socioeconomic status and under-represent Black men and men from the South. While our regression models can adjust for socioeconomic status, the small resulting linked sample of Black American siblings makes fitting race-stratified sibling fixed effects models infeasible.

In this study, we only consider the relationship between homeownership and longevity in the U.S. context. This relationship may differ in other countries with different levels of public housing, homeownership rates, and social conditions surrounding both homeownership and health. For example, Vienna has a high level of public housing (Kadi, 2015), and Singapore's homeownership rates are over 90% due to heavily subsidized housing (Phang and Helble, 2016). Given the significant differences in rates and societal norms surrounding homeownership, our results can only tell us definitively about the U.S. context.

There are many promising avenues for future research. First, we are only able to observe homeownership in early adulthood, and future work could consider longitudinally tracking entrance and exit from homeownership over the life course. Second, while our theoretical framework suggests various potential causal pathways between homeownership and longevity, examining the contribution of each plausible causal mechanism is a promising area of research. Finally, future work could consider studying different periods, cohorts, or population subgroups to investigate the extent to which the results of our study generalize to other contexts. These are important components of future research, which will deepen our understanding of the relationship between homeownership and longevity.

Despite its limitations, this study offers compelling evidence that owning a home in early adulthood has a causal effect on later-life longevity. This finding aligns with our theoretical framework that owning a home promotes longevity and health more broadly through a combination of stronger social networks, wealth accumulation, lower health risks due to substandard housing, and improved psychological well-being. Our results speak to the need for greater equity in the opportunities, incentives, and costs of homeownership. Policies

---

[9]Using a similar identification strategy and research design, Halpern-Manners et al. (2020) found evidence of a causal effect of education on longevity. To bolster their causal claim, they used simulation methods to demonstrate that the amount of residual confounding required to invalidate their causal effect would be extreme and unlikely. Further, they found no statistically significant difference between parallel estimates of the education-mortality relationship using fixed effects models on samples of twins and non-twin siblings. This suggests that there is little remaining residual genetic confounding in siblings, at least in the case of education and mortality.

successful in responsibly and equitably expanding homeownership opportunities for Black Americans could bring substantial longevity benefits.

Chapter 3

The Unpredictability of Individual-level Longevity

How accurately can age of death be predicted using basic sociodemographic characteristics? We test this question using a large-scale administrative dataset combining the complete count 1940 Census with Social Security death records. We fit eight machine learning algorithms using 35 sociodemographic predictors to generate individual-level predictions of age of death for birth cohorts born at the beginning of the 20th century. We find that none of these algorithms are able to explain more than 1.5% of the variation in age of death. Our results point towards the unpredictability of mortality and underscore the challenges of using algorithms to predict major life outcomes.

3.1 Introduction

Social scientists studying mortality in the United States have made remarkable progress in describing racial and class-based inequalities mortality (Schwandt et al., 2021; Elo, 2009), developing theories about the social origins of mortality disparities (Link and Phelan, 1995; Dannefer, 2003; van Raalte, 2021), and investigating the causal determinants of longevity (Chetty et al., 2016; Chetty and Hendren, 2018). Yet a fundamental question remains unanswered: how predictable is individual-level longevity? Characterizing the predictability of longevity is crucial for understanding the extent to which mortality theories can tell us about individual-level longevity. Further, understanding the predictability of major outcomes like longevity can aid policymakers in making informed choices about using predictive algorithms for decision-making (Salganik et al., 2020).

To assess the predictability of individual longevity using demographic characteristics observed early in life, we use data from the CenSoc-DMF. The CenSoc-DMF is a large-scale, nationally representative dataset that links the complete count U.S. 1940 Census with the Social Security Death Master File (DMF) (Goldstein et al., 2021). Our study draws upon a large sample of 130,000 individuals and employs eight statistical and machine learning

algorithms incorporating 35 sociodemographic predictors to predict age of death.

We find that none of the algorithms we tested can make accurate predictions about age of death, confirming the unpredictability of individual-level longevity. Our results highlight the unpredictability of individual-level longevity, the limits of prediction, and the modest gains of complex machine learning algorithms over simple linear models. Mortality researchers need to reconcile the lack of predictability with existing mortality theories. Additionally, our findings raise concerns about the appropriateness of individual-level prediction for policy decision-making.

3.2   Background

Prediction in the Sciences

In the physical sciences, prediction is an indispensable tool for evaluating scientific evidence. A theory can be evaluated based on its ability to make falsifiable predictions about future observations (Breiman, 2001). The predictive accuracy of a theory can be used to benchmark its explanatory power. Scientific theories contradicted by new evidence—from experimental or observational studies—can be refined or replaced by new theories. Predictive modeling has been applied extensively in the field of computer science, where standard processes and straightforward metrics for success have allowed researchers to make major advances in areas such as speech recognition, machine vision, and translation (Hofman, Sharma and Watts, 2017).

In contrast, social scientists have generally relied on interpreting coefficients in idealized models (Molina and Garip, 2019). This approach – termed generative modeling[1] – involves assuming the observed data is generated from an underlying data-generating process described by a parametric model. The researcher chooses a parametric model (e.g., linear regression), fits this model to the observed data, and checks whether the model coefficients are statistically significant and in the direction predicted by theory. The core assumption behind generative modeling is that the researcher can accurately choose a parametric model that describes the complex social process that generated the observed data. This practice of "searching for significance" is beginning to draw criticism in light of the ongoing crisis of replication (Breznau et al., 2022; Open Science Collaboration, 2015).

---

[1]The two mainstream cultures of statistical analysis have been described as "generative modeling" and "predictive modeling" (Breiman, 2001; Donoho, 2017). In generative modeling, researchers assume a parametric model that could have generated the observed data and estimate the different parameters of the model using their observed data. This leads to simple and interpretable models, but such models have not had much out-of-sample explanatory power. Predictive modeling, on the other hand, prioritizes prediction. The data-generating process is assumed to be an unknown "black box" and evaluates models based on their predictive accuracy on new data. Predictive modeling favors complex machine learning models that perform well when predicting on out of sample data. However, such models generally give little insight into the mechanisms linking predictors with outcomes (Molina and Garip, 2019).

Social scientists have historically avoided prediction exercises due to data and computational limitations. Data are often sparse, social systems are complex, and models attempting to capture the complexity of the social world can be computationally intensive to train. Moreover, social scientists fear that complex "black box" models are challenging to interpret and are not useful for studying the social world (Hofman, Sharma and Watts, 2017). However, in recent years, researchers have increasingly acknowledged the complementary nature of predictability and interpretation, and advances in computing, data collection, digitization, and record linkage have ushered in a new data ecosystem for computational social scientists (Kashyap, 2021). These new data allow social scientists to pursue prediction rather than only interpreting whether a coefficient in a model is statistically significant and is in the direction predicted by theory.[2] Yet the extent to which major social science outcomes are predictable is an open area of research.

Group vs. Individual Mortality Differences

When studying differences in mortality, researchers generally partition the population into G mutually exclusive groups and estimate and compare group-level averages $\mu_1, \mu_2, \ldots \mu_G$. The key advantage of investigating group differences in mortality is that it is straightforward to make comparisons across groups. However, studying group differences often hides substantial heterogeneity within groups. On the other hand, studying individual-level mortality provides insight into individual-level heterogeneity but cannot provide insight into group differences in mortality.

Predictability of life outcomes

Our results echo a growing body of research that uses predictive modeling to confirm the unpredictability of life outcomes. Salganik et al. (2020) used a mass scientific collaboration to assess the predictability of major life outcomes. Over 160 research teams used statistical and machine learning algorithms to predict six outcomes: material hardship, GPA, grit, eviction, job training, and layoff. Each research team used the same training and hold-out data from the Fragile Families and Child Wellbeing study, which includes over 1,000 variables on early life (Families, 2021). However, none of the research teams were able to produce accurate predictions for any of the outcomes, suggesting that life outcomes are fundamentally unpredictable. This unpredictability could not be attributed to specific methodological choices by any one team, highlighting the complexity and unpredictability of life outcomes and the limitations of relying solely on predictive models for decision-making in areas such as social policy.

---

[2]Prediction in the social sciences is becoming increasingly common, but there are still several outstanding methodological issues, such as standardized metrics for evaluating the performance of predictive algorithms and characterizing the limits to prediction, that need to be addressed before prediction reaches its full potential (Hofman, Sharma and Watts, 2017).

Similarly, Arpino, Le Moglie and Mencarini (2022) used machine learning algorithms to predict union dissolution in Germany. The study highlights the benefits of using machine learning methods such as random survival forests for prediction and identifies key predictors of union dissolution. Yet this study also ultimately finds union dissolution is highly random and unpredictable.

Machine learning in public policy

The results of these studies are concerning as predictive algorithms are widely applied in public policy. Decision-making in criminal justice, computer security, higher education, targeted advertisement, and bank and mortgage qualification increasingly relies on algorithmic predictions (Peet et al., 2022). Within the criminal justice system, predictive models are widely applied to forecast potential offenders, victims of crime, offenders most likely to miss a court date, and which criminals will be most likely to re-offend in the future. For instance, over 1 million offenders in the U.S. have been assessed with the commercial risk software COMPAS. One part of this assessment is a recidivism risk score, which reports a person's risk of committing another crime in the next two years (Dressel and Farid, 2018). In child protective services, predictive risk models have been used to identify cases at higher risk of adverse outcomes (Chouldechova et al., 2018). Further, the US Department of Veteran Affairs is currently using a predictive model to identify veterans at higher risk of suicide to better target care using electronic medical records (Kessler et al., 2017).

The explosion of machine learning in both academic and industry settings has been accompanied by a growing skepticism (Narayanan, 2021). In theory, flexible machine learning algorithms are appealing due to their ability to discover interactions and higher-level non-linear effects (Rose, 2013). However, existing evidence suggests that more complex machine learning models may not necessarily perform better than simpler models. For instance, Dressel and Farid (2018) found that a simple linear model with only two predictors had comparable performance to a complex AI model with over 137 predictors. Similarly, Salganik et al. (2020) observed that their best-performing models only slightly outperformed their simple benchmark model. These studies suggest that machine learning may only produce modest gains over traditional parametric models for predicting social outcomes.

Moreover, despite the widespread application of predictive models in public policy for decision-making, the predictive accuracy of these algorithms is unclear. Are these algorithms making accurate predictions, and if not, what are the implications of inaccurate predictions being used for decision-making? A stronger characterization of the individual-level social outcomes that can and cannot be accurately predicted is essential for a complete scientific understanding of the implications of using the predictions from such algorithms for decision-making.

Health and epidemiological studies of mortality risk

A handful of applied epidemiology studies have predicted individual-level mortality risk, but the aim of such studies is very different from ours. In epidemiology and medicine, generating accurate individual-level predictions of mortality risk is an important area of research, because mortality risk scores are valuable for adjusting for risk between treatment groups in both clinical and observational studies (Austin and van Walraven, 2011). To maximize their predictive accuracy, these studies generally rely on health or behavioral characteristics such as self-rated health. These direct health measures are generally more predictive of mortality than the sociodemographic characteristics considered in this study.

Even with these behavioral and health-related predictors, past studies have had limited success in predicting mortality. Rose (2013) used data from the Study of Physical Performance and Age-Related Changes in Sonoma (SPPARCS) dataset and Superlearner machine learning algorithms to predict 5-year mortality. The SPPARCS dataset includes 2,092 individuals aged 54 or older in Sonoma, California. The study collected information on gender, age at baseline, self-rated health at baseline, smoking behavior, pre-existing health conditions, self-rated health, and a leisure time physical activity score. The Superlearner algorithm was employed to forecast death in a 5-year period from 1993-2005, and the study's outcomes revealed that the algorithm could explain 20.1% of the variation in 5-year mortality risk.

Puterman et al. (2020) used data from the nationally-representative Health and Retirement Study (HRS) to predict mortality from 57 economic, health, behavioral, social, and psychological factors. This study used both conventional methods (i.e., multivariate Cox regressions) and machine learning (i.e., lasso, random forest analysis) to identify the leading predictors of mortality. Together these 57 predictors explained a total of 11.4% of the variance in mortality; however, baseline demographic characteristics (gender, race/ethnicity, and place of birth) only explained 1.9% of the variation in mortality.

Our research objectives diverge significantly from these prior studies. These investigations use health-related behavioral predictors, such as alcohol abuse, self-rated health, and current smoker status, to predict mortality outcomes. In contrast, our study only uses basic sociodemographic characteristics observed early in the life course to predict longevity later in life. Our approach aims to outline the boundaries of prediction and to provide more realistic expectations of what major life outcomes can and cannot be predicted.

Mortality theory

Social scientists studying mortality in the United States have a long-standing interest in describing aggregate disparities in mortality. There are striking class-based (Elo, 2009; Chetty et al., 2016) and racial (Feigenbaum, Muller and Wrigley-Field, 2019; Wrigley-Field, 2020*b*) disparities in mortality. While overall longevity has increased over the course of the 20$^{\text{th}}$ century, inequality in mortality has increased over time (Preston and Elo, 1995). Researchers have also documented demographic paradoxes, such as the Black-White mortality crossover (Wrigley-Field, 2020*a*) and the Hispanic mortality paradox (Elo et al., 2004).

To explain these striking population-level disparities in mortality, social scientists have developed mortality theories. Life-course theories investigate how later-life mortality is impacted by events in early life or other critical periods through mechanisms such as *scarring*, *selection*, and *immunity* (van Raalte, 2021). Theories on the influence of macroeconomic conditions and social inequalities on mortality, such as fundamental cause theory, were introduced to explain the social determinants of mortality disparities (Link and Phelan, 1995). Further, theoretical mortality models have been introduced and tested to describe death rates at the most advanced ages (Vaupel, Manton and Stallard, 1979; Feehan, 2018; Gompertz, 1825; Barbi et al., 2018). Yet it remains unclear to what extent mortality theories can help us predict individual-level morality.

3.3 Data

We use the recently released CenSoc-DMF dataset (Goldstein et al., 2021), which links the complete-count 1940 Census (Ruggles et al., 2020) with mortality records from the Social Security Death Master File (DMF). The DMF captures nearly all deaths in the U.S. between the years of 1975–2005 (Alexander, 2018). For each death record, the DMF contains full first and last name, exact date of birth (d/m/y), and exact date of death (d/m/y) but contains no information on gender, place of birth, or other sociodemographic characteristics. The 1940 Census includes 36 population questions and 31 household questions. Notably, the 1940 Census was the first decennial census to include questions on wage and salary income, educational attainment, and employment status.

As there is no shared unique identifier in both the 1940 Census and DMF mortality records (e.g., Social Security Number), we link on fields unlikely to change over the life course. Specifically, we use the ABE exact match algorithm (Abramitzky, Boustan and Eriksson, 2012, 2014, 2019), which links on first name, last name, and age of birth. This algorithm first standardizes first names to account for common misspellings or nicknames (e.g., Robbie $\rightarrow$ Robert). The algorithm then establishes matches based on an exact match on first and last name while allowing for some flexibility (± 2 years) on birth year. To reduce the number of false matches, we further restrict to a set of "conservative" matches, which requires first and last names to be unique within ± 2 years around year of birth (a 5-year band). We are able to match approximately 20% of records between the 1940 Census and Death Master File, a match rate comparable to other historical linking efforts (Abramitzky and Boustan, 2017).

The resulting CenSoc-DMF dataset is a large-scale, nationally representative dataset with a range of covariates observed in 1940 and exact age of death. For parsimony, we focus on the birth cohort of 1910 (N = 133,000),[3] who were observed in the 1940 Census at age 29 or 30 and died between the ages of 64–95. The CenSoc-DMF file does not include women due to challenges linking unmarried women whose surnames may have changed after the

1940 Census.
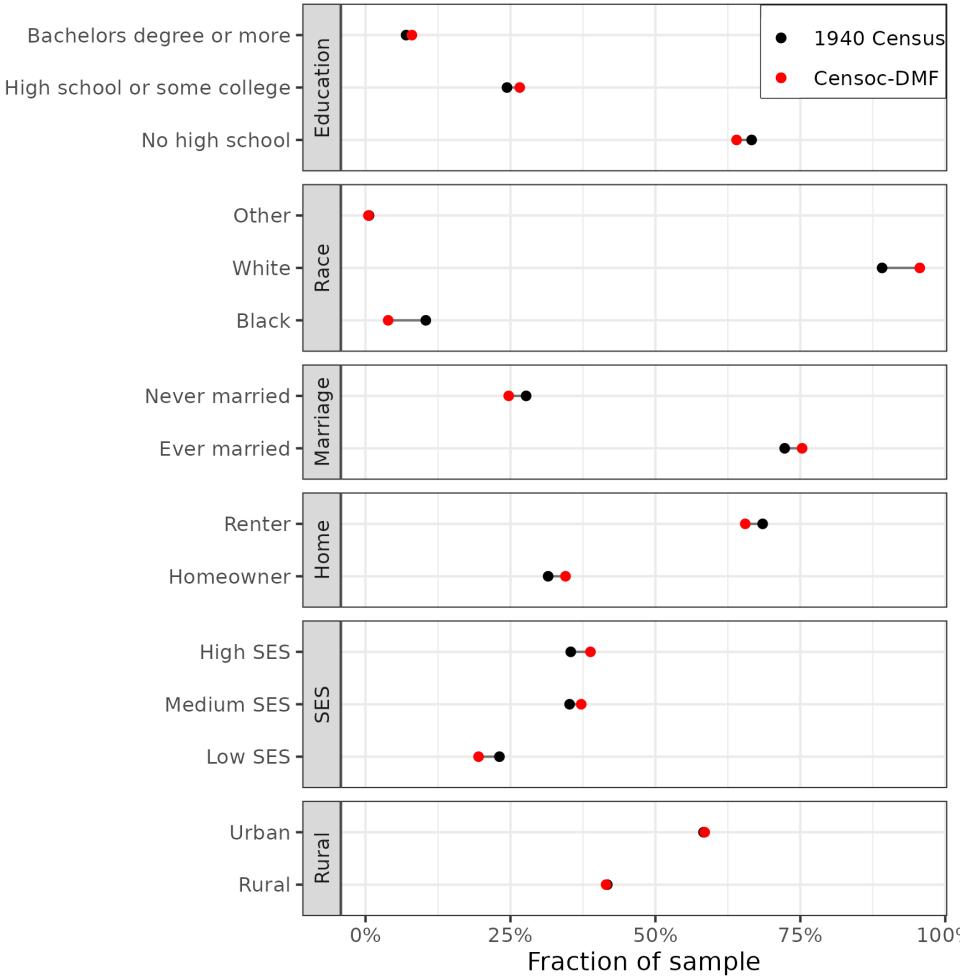
Figure 3.1: Each facet shows the composition of the CenSoc-DMF (red) and the complete count 1940 Census (black) for a given covariate. The Censoc-DMF is broadly representative of the 1940 Census, except with slightly higher socioeconomic status and a slightly higher proportion of whites.

To assess the representativeness of our matched CenSoc-DMF sample, Figure 3.1 compares the composition of our matched CenSoc-DMF sample and the full population of men in the 1940 Census age 29 or 30. Broadly, our CenSoc-DMF sample is representative of the general population, albeit comprised of a higher proportion of white, higher socioeconomic status individuals.

---

[3]For our main analysis, we only present results for the cohort of 1910. However, in supplementary analyses, we find highly comparable results for both earlier and later birth cohorts.

From the CenSoc-DMF, we select 36 key predictors from the 1940 Census and age of death in years, our outcome of interest. We restrict our analysis to individuals with no missing values for any predictors.[4] For modeling purposes, we normalize all variables to ensure they share a common scale without distorting relative differences between variables. For dichotomous variables, we do not apply any normalization or standardization techniques.
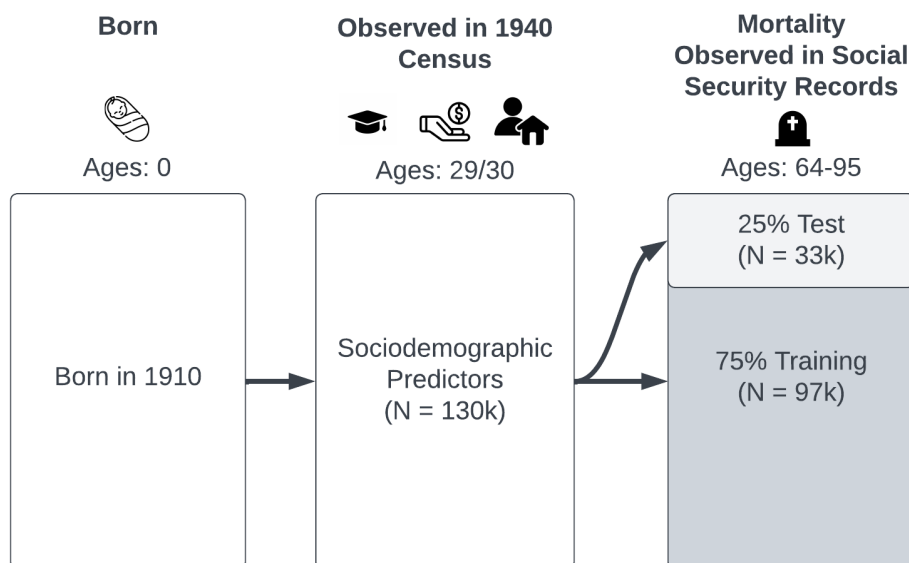
Figure 3.2: Overview of our analytical sample. We focus on the birth cohort of 1910. We observe their early life characteristics in the 1940 census at age 29/30, and we observe their mortality between the ages of 64–95.

3.4 Methods

Predictive Modeling

We use both machine learning approaches and conventional linear regression to evaluate how accurately we can predict longevity. Flexible machine learning algorithms are appealing in this setting because of their ability to detect interactions and higher order effects and better handle multiple, highly-correlated predictors (Rose, 2013; Puterman et al., 2020). We test a variety of conventional and machine learning algorithms, including random forests,

---

[4]As a sensitivity analysis, we reran our full analysis imputing missing data instead of dropping missing data. Our results were nearly identical.

linear regression, gradient boosting machines, lasso regression, extreme gradient boosting machines, and support vector machines.

For most prediction tasks, it is impossible to know in advance which algorithm will have the best performance. To overcome this, we use Superlearning. Superlearning—also known as weighted ensembling or stacking—is a statistical technique for prediction that combines many machine learning algorithms into a single algorithm (Van der Laan, Polley and Hubbard, 2007). This single algorithm is expected to perform at least as well as the best of the individual algorithms. This avoids the issue of picking the "right" algorithm by systematically picking the best algorithm using a pre-defined set of decision rules. The motivation behind Superlearning is that a weighted combination of different algorithms may outperform any single algorithm by smoothing out the limitations of any specific models. The Superlearner algorithm finds the best weighted combination of algorithms using a k-fold cross-validation procedure to minimize cross-validated risk (Van der Laan, Polley and Hubbard, 2007).

We fit the Superlearner algorithm on our training data using the following implementation:

1. Split the data into $k = 10$ different folds (partitions) for k-fold cross-validation.

2. Using the first fold as the hold-out data, fit each algorithm on the 9 other folds. Make predictions on the hold-out fold using each algorithm. Repeat this process for each fold, using a different fold as the hold-out data.

3. Choose a loss function (e.g., mean squared error) and compute a metalearner. The metalearner is a principled approach to combining multiple machine learning algorithms by fitting a regression of outcome variables on the predicted variables to minimize the cross-validated risk of the set of potential weighted combinations.

4. Fit each of the algorithms on the full data.

In addition to producing a weighted algorithm that theoretically performs as well or better than the top algorithm, the Superlearner has the added advantage of training many machine learning algorithms at the same time. This allows researchers to assess the performance of many different machine learning algorithms with an a priori specified evaluation criterion, minimum cross-validated mean squared error.

Table 3.1 shows the full set of machine learning algorithms used in our Superlearner. The cross-validated risk is lowest for the generalized boosted regression and highest for the random forest and extreme gradient boosting. The Superlearner algorithm most heavily weighted the GBM and XGBOOST models.

We randomly split our dataset into a training partition (75%) and a holdout partition (25%). Our training dataset includes our 36 predictors and outcome of interest, age of death. We use this training dataset to train our predictive models. To evaluate the performance of our models, we use the models trained on our training dataset to predict longevity for each

Algorithm	Description	Cross-validated Risk	Coefficient
gbm	Generalized boosted machines	63.61	0.80
lm	Linear model	63.73	0.00
xgboost	Extreme gradient boosting	63.81	0.12
ranger	Random forest regression	64.98	0.02
svm	Support vector machine	65.50	0.06
lasso	Lasso regression	63.86	0.00
mars	Multivariate adaptive regression splines	63.85	0.00
mean	Arithmetic mean	64.49	0.00
Superlearner	Ensemble Superlearner	63.59	—

Table 3.1: Full set of algorithms included in the Superlearner. The cross-validated risk refers to the risk calculated by the Superlearner, the cross-validated mean squared error. The coefficient gives the total weight (contribution) of each algorithm towards the full Superlearner ensemble model.

person in the holdout dataset, withholding information about the true age of death. We then compare our predicted age of death with the true, withheld age of death.

Benchmarch OLS Regression Model

As a benchmark model, we also fit an OLS regression model using three predictors, educational attainment (in years), income, and race:

$$\text{Death Age}_i = \text{Education}_i + \text{Income}_i + \text{Race}_i + \epsilon_i. \tag{3.1}$$

We chose these predictors based on substantive knowledge about key social determinants of mortality. This allows us to assess the performance of our model against a simple benchmark model with only three predictors.

3.5 Results

To contextualize our findings, we first present results on aggregate mortality disparities for our focal cohort of 1910 in the CenSoc-DMF. We observe substantial aggregate-level disparities that align closely with our theoretical expectations. We then present our results on the predictability of individual-level longevity using models trained on the exact same data. Our major finding is that none of our statistical or machine learning algorithms were able to accurately predict individual-level longevity.

Aggregate Mortality Disparities

In Figure 3.3a, we see a clear educational gradient, with higher educational attainment being associated with higher life expectancy. Further, we see the "education staircase" – large jumps at years 12 and 16, which correspond to terminal years of education (Goldstein and Edwards, 2020). In Figure 3.3b, we see a clear income gradient, with higher wage and salary income being associated with a higher life expectancy. These clear and demographically meaningful aggregate disparities between socioeconomic groups align closely with our theoretical expectations.

To illustrate variation in life expectancy within groups, Figure 3.3c shows the distribution of age of death across three levels of educational attainment: "No High School," "High School," and "College." We focus here on educational attainment as it has a well-established relationship with mortality (Halpern-Manners et al., 2020; Lleras-Muney, Price and Yue, 2020) and is a common proxy for social class (Muller and Roehrkasse, 2022; Pettit and Western, 2004). Three insights emerge from our education analysis. First, there are substantial education-based disparities in mortality at the aggregate level (Figure 3.3a). Second, within education categories, there is substantial heterogeneity in age of death. Finally, for any given individual, knowing their level of educational attainment will be a weak predictor of their individual-level longevity. Aggregate mortality patterns hide the substantial individual-level mortality within groups.

Individual-level predictions

We split our sample into a training dataset (75% of observations) and a holdout dataset (25% of observations). We fit models predicting age of death on the training dataset and use these models to generate predictions on the holdout dataset. We quantify each model's performance using the coefficient of determination (R^2), the proportion of variation in the outcome variable explained by the predictor variables in the holdout dataset (Chicco, Warrens and Jurman, 2021). In other words, the R^2 value describes the total variation in age of death (in years) an algorithm was able to explain.

As shown in Figure 3.4a, none of the machine learning algorithms was able to explain more than 1.5% of the total variation in age of death. The Superlearner algorithm and the gradient boosted machines had the best out-of-sample performance, explaining 1.26% and 1.25% of the total variation in age of death. The worst-performing algorithm was random forest, which only explains 0.4% of the total variation in age of death. Our linear model explains 1.10% of the variation; our benchmark linear model described in Equation (3.1) explains 1.02% of the variation.

Figure 3.4b plots the correlation between the predicted age of death of our best-performing Superlearner algorithm and the actual values of age of death. There is no discernible relationship between our predicted age of death and the true observed age of death. Further, most of the predictions are narrowly concentrated between the ages of 76–82, suggesting that
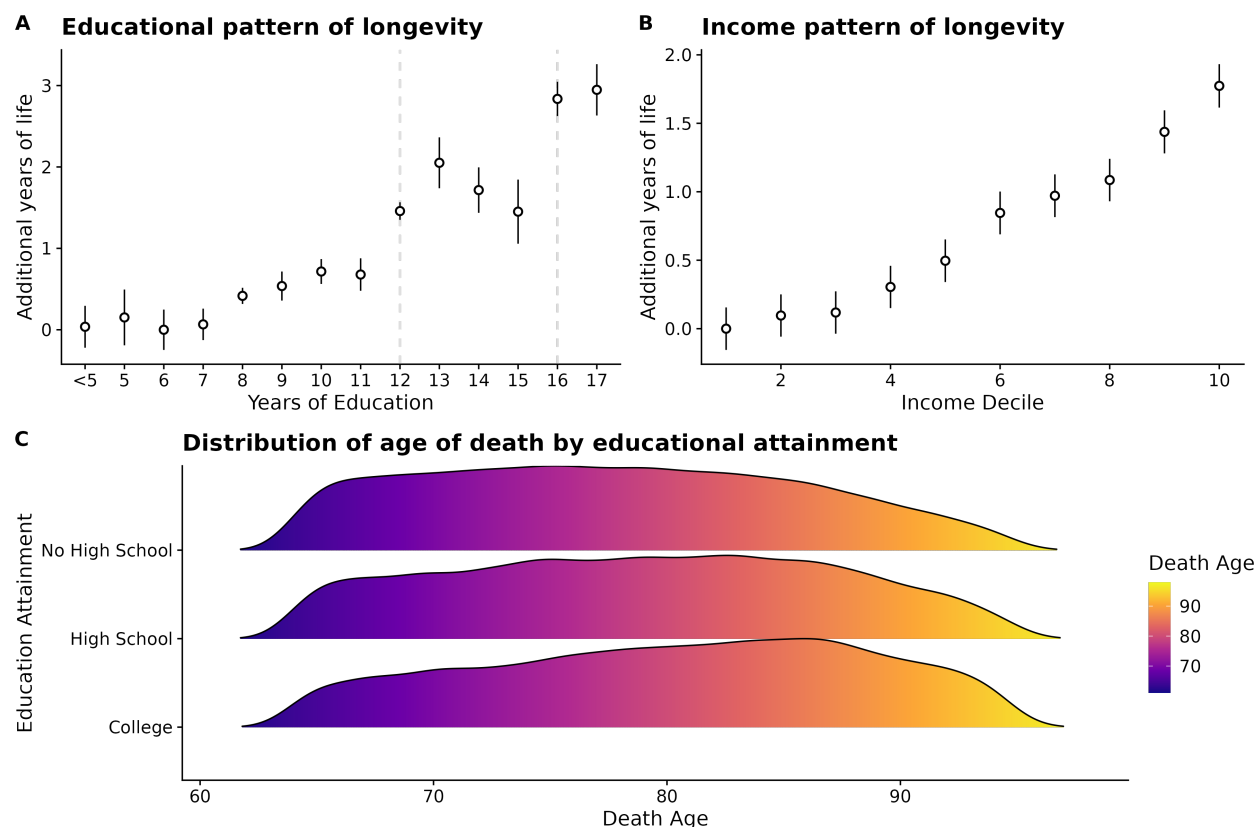
Figure 3.3: Panel A shows the educational gradient of longevity. Panel B shows the educational pattern of income. Panel C shows the distribution of death age by educational attainment. Error bars show 95% uncertainty intervals.

our best model was not able to accurately predict whether someone would die at a relatively earlier or later age.

There is a marginal improvement in predictive accuracy for Superlearner over the other algorithms. Compared to our top-performing individual algorithm, gradient boosted machines, Superlearner only increased predictive accuracy by 0.8%. Compared to our simple benchmark with three predictors, the Superlearner increased predictive accuracy by 22%.

Figure 3.4c plots variable importance for the Superlearner algorithm. Variable importance is a measure of how much an algorithm relies on a given variable for making its predictions. Here, variable importance is calculated as the reduction in predictive accuracy (as measured by mean squared error) attributable to each variable if that variable was removed from our model (Friedman, 2001). Education in years and occupational prestige[5] were the two most important predictors in the Superlearner model, likely reflecting class-based
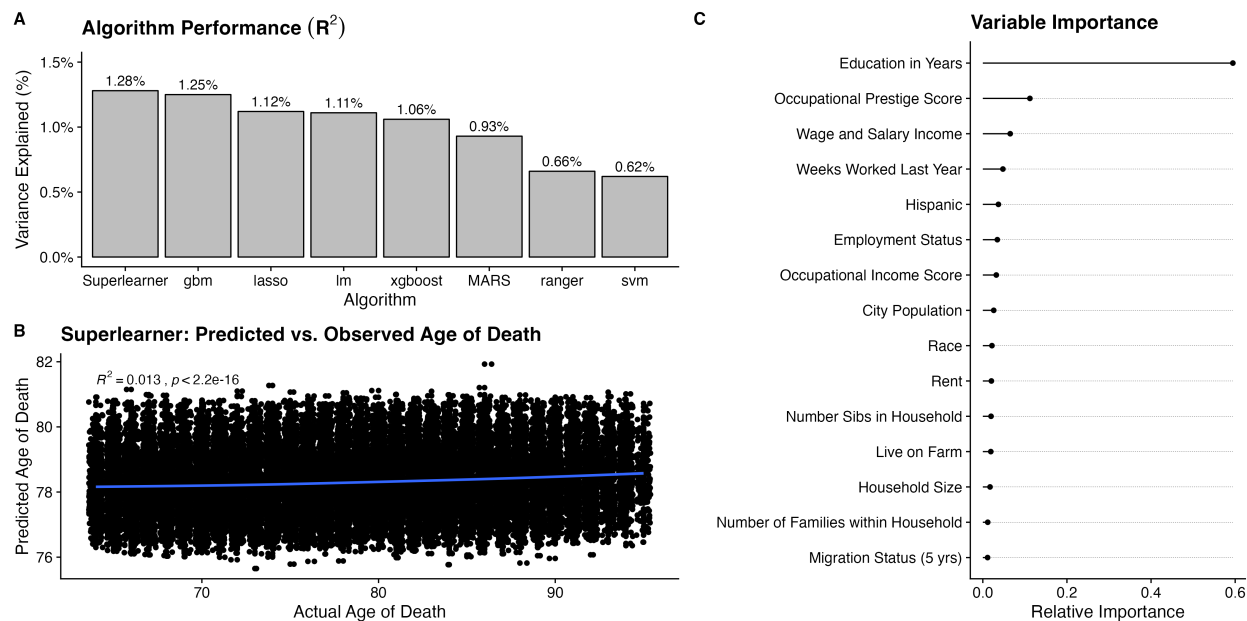
Figure 3.4: Panel A shows the R^2 performance of each machine learning algorithm. Panel B shows the predicted vs. observed predictions for Superlearner, our best-performing algorithm. Panel C shows the relative variable importance for the top 15 variables in our Superlearner model.

differences in life expectancy (Elo, 2009; Ostan et al., 2016).

3.6   Discussion

Our findings add to a growing body of evidence that individual-level social outcomes are challenging to predict (Salganik et al., 2020; Arpino, Le Moglie and Mencarini, 2022; Dressel and Farid, 2018). We use large-scale administrative data and a range of machine learning algorithms to predict a single outcome: age of death. Our data allows us to study cohort lifespans, rather than artificially constructed measures of period life expectancy much of the literature is confined to. We find that all our algorithms performed poorly at predicting individual-level age of death. Our best-performing algorithm only explained 1.26% of the overall variation in age of death, despite being trained on a large dataset with 35 different individual-level predictors.

The ensemble Superlearner algorithm outperforms any individual algorithm. While the performance over other top-performing algorithms is marginal, the Superlearner far outper-

---

[5]This measure of occupational prestige is based on the Siegel occupational prestige score, which calculates an occupational prestige score for each occupation based on their perceived status according to a survey of the general population.

formed our worst-performing algorithms, extreme gradient boosted machines, and random forest. However, none of the machine learning algorithms greatly outperformed conventional linear regression. The Superlearner only had a modest 22% overall increase in R^2 over our three-predictor benchmark model. The relatively modest increase between our three-predictor benchmark model and our top-performing algorithm aligns closely with findings from past studies (Dressel and Farid, 2018; Salganik et al., 2020). Additionally, fitting 8 of the most popular and widely applied machine learning algorithms, and having none perform well, gives reassurance that our low predictability is not simply driven by choice of algorithm.

The relatively modest increase of our complex, machine-learning algorithms over a three-predictor linear model has several implications. First, complex machine learning algorithms can give a modest boost in predictive accuracy. This boost in predictive accuracy helps explain the widespread application of machine learning in fields such as quantitative finance or social media feeds, where small increases in predictive power have large financial implications. Second, when an outcome is unpredictable using a simple linear model, it will also be unpredictable with a more complex machine learning algorithm. Finally, both social scientists and policymakers must weigh the relatively modest gains of complex machine learning algorithms with the difficulty of training and interpreting such models.

For policymakers seeking to use predictive algorithms in policy decision-making, these results suggest caution. The poor performance of predictive models for estimating a well-studied outcome like mortality raises concerns about the accuracy of predicting other types of social outcomes. Policymakers should require clear evidence of an algorithm's accuracy and effectiveness before using it to inform policy decisions and should carefully consider the implications of inaccurate predictions in their specific policy setting.

These results also have implications for social scientists studying mortality. Why do we puzzlingly observe such stark aggregate mortality disparities but cannot predict individual-level longevity? Mortality theories about group differences in mortality have low explanatory power for individual-level mortality, suggesting mortality researchers may need to develop new theories that better explain why individual-level mortality is so hard to predict. The fundamental uncertainty around individual-level mortality needs to be better acknowledged and incorporated into demographic theory.

This study's findings can provide practical guidance for individuals seeking to make better decisions and plans for their future. The substantial variation in life expectancy within sociodemographic groups and the randomness of individual-level longevity suggest individuals should largely ignore their own life expectancy. For example, individuals in groups with lower predicted average life expectancy should not plan on a shorter retirement: individual-level longevity is highly random, and the likelihood that they far outlive their expected life expectancy is high. On the other hand, life insurance companies should be less concerned that individual-level prediction is so challenging. While individual-level prediction may be highly random, the insurance companies' primary goal is to accurately predict life expectancy on average.

There are several limitations to this study that warrant discussion. Our sample is limited to men born in 1910 and dying between 1975–2005. We only observe deaths between the

ages of 64–95, and it is possible deaths earlier in the life course would be more predictive than those observed in our sample. However, our sample captures approximately 65% of the deaths occurring after age 30 (Breen and Osborne, 2022). The main prediction exercise presented in this paper is restricted to men, as surname changes for some women during marriage preclude accurate record linkage. In Appendix B.1, we replicate our analysis in the CenSoc-Numident, which includes all genders but with a shorter observed window of deaths (1988–2005). Qualitatively, our results are nearly identical, with our mortality predictions being slightly less accurate in the CenSoc-Numident.

We tested a wide range of different statistical and machine learning algorithms, but we cannot definitively rule out that a different algorithm would be more predictive than those we tested. Yet given the modest gains of machine learning models over simple linear models in other settings (Salganik et al., 2020), it is unlikely that any new algorithm would meaningfully improve our ability to predict longevity. It is also possible that there are other sociodemographic covariates that would help improve the predictive power of our model. Further, it is possible that alternative feature engineering approaches may help improve predictive accuracy. Finally, we note that predicting other measures of mortality—such as whether an individual died within a 5-year window—would lead to higher predictive accuracy. However, changing our outcome to make prediction easier does not change the fact that our predictions of individual longevity are inaccurate.

Despite these limitations, our results speak to the unpredictability of one of life's most major life outcomes: age of death. While demographers have long observed striking mortality disparities across population subgroups—ethnoracial disparities, place-based disparities, and class-based disparities, individual-level longevity is stochastic and challenging to predict. Sociodemographic characteristics alone cannot be used to predict individual-level longevity.

Chapter 4

Black-White Mortality Crossover: New Evidence from Administrative Data

The Black-White mortality crossover is a well-studied demographic paradox. Black Americans experience higher age-specific mortality rates than White Americans throughout most of the life course, but this puzzlingly reverses at advanced ages. The leading explanation for the Black-White mortality crossover centers around selective mortality over the life course. Black Americans who survived higher age-specific mortality throughout their life course are a select population composed of individuals who are exceptionally robust. However, skeptics argue the Black-White mortality crossover is simply a data artifact from age misreporting or related data quality issues. We use large-scale linked administrative data (N = 2.3 million) to document the Black-White mortality crossover for cohorts born in the early 20$^{\text{th}}$ century. We find evidence the crossover is not a data artifact and cannot be uncrossed using sociodemographic characteristics alone.

4.1   Introduction

The Black-White mortality crossover is a long-standing demographic paradox. The crossover occurs when non-Hispanic Black Americans experience higher age-specific mortality rates than Non-Hispanic White Americans until very late in life. At advanced ages, the age-specific mortality rates first converge and then cross over, with Black mortality being lower than White mortality. The crossover has been repeatedly documented in the United States (Sautter et al., 2012; Dupre, Franzese and Parrado, 2006; Masters, 2012; Lynch, Brown and Harmsen, 2003; Hummer, 1996). However, there is little consensus on the explanation for the Black-White mortality crossover: critics have questioned these findings, suggesting that the apparent crossover is simply an artifact of sparse or poor-quality mortality data at the most advanced ages (Preston and Elo, 2006; Lynch, Brown and Harmsen, 2003; Preston

et al., 1996; Preston, Elo and Preston, 1999). Others have theorized that the crossover is the product of selective mortality over the life course (Vaupel, Manton and Stallard, 1979; Vaupel and Yashin, 1985; Wrigley-Field, 2014, 2020*a*).

Understanding the Black-White mortality crossover is important for several reasons. First, the mortality crossover has implications for our understanding of inequality at the most advanced ages. Is there really a narrowing of mortality conditions for Black and White Americans among the oldest old? Or is the crossover just a data artifact or a ruse of heterogeneity in susceptibility to mortality? Second, the Black-White mortality crossover is a useful empirical example for developing theoretical frameworks of mortality selection. Finally, insights gained from studying the Black-White mortality crossover can be applied to related research areas, such as mortality compression and deceleration of mortality rates at advanced ages (Lynch, Brown and Harmsen, 2003).

In this study, we use linked administrative mortality data from the CenSoc-DMF (N = 2.3 million) to investigate the Black-White mortality crossover. The unprecedented size of the CenSoc-DMF dataset, along with its rich array of covariates, allows us to empirically assess two of the main explanations for the Black-White crossover. We find a mortality crossover for the male birth cohorts of 1890–1905 at age 85 and a crossover for the male birth cohorts of 1906–1915 at age 90. Our analysis is restricted to men, as surname changes for some women during marriage make linking women between the 1940 Census and the DMF mortality records infeasible. The quality of our mortality data, paired with a sensitivity analysis, allows us to rule out that our observed crossovers are simply an artifact of age misreports or exaggerations. We then stratify for observed heterogeneity to test whether the crossover can be uncrossed using sociodemographic characteristics, finding that the crossover persists across all subgroups. We conclude that unobserved heterogeneity may still be responsible, or there are indeed as-yet unknown protective factors that influence race differentials at older ages in ways that are different than at younger ages.

4.2 Background

Past Studies on the Black-White Mortality Crossover

Since its original discovery by Sibley (1930), the Black-White mortality crossover has been repeatedly documented in the United States (Manton, Poss and Wing, 1979; Berkman, Singer and Manton, 1989; Lynch, Brown and Harmsen, 2003; Dupre, Franzese and Parrado, 2006; Sautter et al., 2012; Kestenbaum, 1992; Masters, 2012). The Black-White mortality crossover has also served as a motivating example for a growing body of methodological work on theoretical models of mortality selection (Vaupel and Yashin, 1985; Vaupel, Manton and Stallard, 1979; Wrigley-Field, 2014, 2020*a*). More recently, a handful of empirical studies have investigated the contribution of covariates such as socioeconomic status or religious attendance to the Black-White mortality crossover (Dupre, Franzese and Parrado, 2006; Sautter et al., 2012; Yao and Robert, 2011; Berkman, Singer and Manton, 1989).

Table 4.1 presents several of the major empirical studies documenting the Black-White crossover. Across studies, the "age of crossover"—the age at which Black age-specific mortality rates first become lower than White age-specific mortality rates—occurs between the ages of 74 and 90, generally centered around 85. However, the age at crossover has been trending upwards over the course of the 20th century (Masters, 2012). In the 1960s, the crossover was observed at age 75 for men and age 77 for women (Kestenbaum, 1992). In the 1970s, the age at crossover was observed at ages of 78 for men and 80 for women (Masters, 2012). More recently, the crossover has been observed at ages 88 for men and 87 for women in U.S. lifetables from 2003 (Arias, 2006). This upward trend in the timing of the age of crossover suggests that differential cohort experiences are an important consideration for any study of the Black-White crossover.

Data Source	Age of Crossover	Covariates	Age Verification	Citation
Tennessee Vital Statistics	74			Sibley (1930)
Evans County Study	85 (f); 80 (m)			Wing et al. (1985)
Medicare Enrollment	88 (f); 86 (m)			Kestenbaum (1992)
U.S. Death Certificates	90 (f); 85 (m)		✓	Preston et al. (1996)
Medicare Enrollment	85-86			Parnell and Owens (1999)
Survey on Asset and Health Dynamics Among the Oldest Old	81			Johnson (2000)
Berkeley Mortality Database	79-87		✓	Lynch, Brown and Harmsen (2003)
Medicare Enrollment	80-85			Arias (2006)
Established Populations for Epidemiologic Studies of the Elderly	83 (f); 79 (m)	Religious Attendance		Dupre, Franzese and Parrado (2006)
Americans' Changing Lives study	80	Education, Income, Neighborhood Disadvantage Index		Yao and Robert (2011)
National Health Interview Survey-Linked Mortality Files	85			Masters (2012)
Established Populations for Epidemiologic Studies of the Elderly	83 (f); 79 (m)			Sautter et al. (2012)
NCHS Multiple Cause-of-Death public-use files	87	Education, Income		Fenelon (2013)
National Longitudinal Mortality Study	85			Şahin and Heiland (2017)

Table 4.1: Past studies of the Black-White mortality crossover.

Explanations for the Black-White Crossover

There are three prominent explanations for the Black-White mortality crossover. The evidence to date is not yet seen as conclusive, and population scholars are increasingly seeking explanations for the Black-White mortality crossover. These competing explanations are outlined below.

Data Artifact

One explanation for the Black-White crossover is that there is no crossover at all. Rather, differential age misreporting or exaggeration, uncounted or unmatched deaths, and other inaccuracies can lead to a spurious crossover. According to this perspective, once these data errors are accounted for, the crossover disappears or is delayed until even more advanced ages (Preston and Elo, 2006; Lynch, Brown and Harmsen, 2003; Preston et al., 1996; Preston, Elo and Preston, 1999).

This perspective was most clearly advanced by Preston et al. (1996), who linked death certificates to both decennial census records (1900, 1910, and 1920) and the Social Security Death Master File (DMF). This linkage exercise demonstrated that misreporting was common; over 50% of Black women decedents had disagreement between the ages of death on their death certificate and their Social Security record. Upon correcting for misreporting in these death rates for Black Americans, the crossover disappeared. As further evidence of age misreporting,Preston and Elo (2006) in a follow-up study demonstrated that the age-specific mortality rates for Black Americans above 85 were lower than the age-specific mortality rates in the lowest-mortality countries.

Age-As-A-Leveler

The "naive" theoretical explanation for the Black-White mortality crossover is that for the oldest-old, mortality conditions converge for Black and White Americans. According to this *age-as-a-leveler* hypothesis, older adults are increasingly separated from the unequal social institutions that contribute to racial health disparities, such as the education system, the labor market, and the criminal justice system. The departure from these stressors of daily living may cause mortality rates to converge in later life (Kim and Miech, 2009). Further, increased availability of a social safety net in later life, including Medicare and Social Security, and stronger kin and support networks, could cause age-specific mortality rates to converge in the oldest ages.

In this sense, old age acts as a "leveler" and causes a convergence in age-specific mortality rates; real racial disadvantage attenuates at the most advanced age. However, it is unclear why such attenuation of disadvantage at the most advanced ages would cause a crossover rather than simply a convergence. Further, this hypothesis is at odds with a large body of research documenting racial inequality in the U.S. (Bryan L. Sykes and Michelle Maroto, 2016; Alexander, 2010; Riddle and Sinclair, 2019; Perry and Morris, 2014).

Heterogeneity in Frailty

The most famous explanation for the mortality crossover comes from theoretical models of mortality selection. Mortality selection models begin with the premise that people vary systematically in mortality risk. In this frailty modeling tradition, as a cohort ages, it becomes increasingly composed of robust individuals. This mortality selection can occur unequally across population subgroups and has been hypothesized to explain mortality crossovers, mortality deceleration, and mortality compression (Lynch, Brown and Harmsen, 2003; Wrigley-Field, 2014).

In the case of the Black-White crossover, Black Americans who faced higher mortality risks in early and midlife will be composed of a greater proportion of robust individuals in later life, resulting in their age-specific mortality rates becoming lower than those of White Americans, who faced lower mortality risks earlier in their life course. In other words, the Black Americans who survive to the most advanced ages are more highly selected for robustness than their White counterparts and will have lower mortality at advanced ages (Wrigley-Field, 2020*a*; Vaupel and Yashin, 1985).

Skeptics of this heterogeneity-in-frailty explanation point out that poor health conditions in early life can "scar" survivors, leading to higher mortality in later life (Preston and Elo, 2006). The limited number of empirical investigations have suggested that the dominant direction of mortality conditions at different points in the life course is positive: higher mortality risk in early life is associated with higher mortality later in the life course (Finch and Crimmins, 2004; Janssen et al., 2004; Preston, 1970).

This study has two specific aims. First, we establish the mortality crossover as real, not a data artifact. Second, we provide empirical evidence that the crossover cannot be uncrossed using sociodemographic characteristics alone. The remainder of the paper proceeds as follows. In the Section 4.3, we describe the complete count census data and mortality records used in our analysis. We then describe our methods for mortality estimation in the absence of denominators in Section 4.4. In Section 4.5 and Section 4.6, we present and interpret our findings and discuss their implications for our understanding of mortality selection.

4.3 Data

This study uses complete count 1940 Census data, mortality records from the Social Security Death Master File (DMF), and record linkage techniques to construct a large-scale dataset with rich covariates and mortality outcomes. This dataset, termed the CenSoc-DMF (Goldstein et al., 2021), links the complete count 1940 Census (Ruggles et al., 2020) to the DMF. The DMF is a collection of over 83 million death records reported to the Social Security Administration, with nearly complete mortality coverage between 1975–2005 (Alexander, 2018; Hill, 2001). However, the DMF does not contain any socioeconomic or demographic variables. To obtain individual-level covariates, we link the DMF mortality records to 1940

Census records. The resulting matched file includes only men, as surname changes due to marriage for some women make the systematic linkage of women infeasible.

We link individual records in the complete count 1940 Census to the DMF using first name, last name, and year of birth using the ABE exact match record linkage algorithm (Abramitzky, Boustan and Eriksson, 2012, 2014; Abramitzky and Boustan, 2017; Abramitzky et al., 2021). To reduce false matches, we restrict to matches where names are unique within and across datasets for a $\pm$ 2-year window. This approach prioritizes minimizing the number of false matches over maximizing the overall match rate; this minimizes the amount of systematic bias introduced by false matches (Ruggles, Fitch and Roberts, 2018).

Representativeness of Matches

Our mortality-adjusted match rate is approximately 20% (Breen and Osborne, 2022). To demonstrate that our matched sample is representative of the general population within racial groups, we compare the composition of our matched sample to the general population in the 1940 Census. Figure 4.1 shows our matched sample is broadly representative of the general population within racial groups, except with slightly higher socioeconomic status.

Reliability of DMF

A key consideration for our study is the reliability of the DMF mortality records. The DMF is extracted from the Social Security Numident and contains over 75 million death records. The death coverage between 1975–2005 is nearly complete, containing approximately 95% coverage for deaths occurring after the age of 65 (Hill, 2001; Alexander, 2018). Death coverage rates drop after 2005, and the DMF has substantial coverage gaps beginning in 2011 (Maynard, 2019). Our analysis is restricted to deaths occurring in our mortality observation window of 1975–2005.

The DMF does not explicitly include information on age of death. Rather, the DMF contains information on date of birth and date of death from which age of death can be imputed (Preston et al., 1996). Therefore, to assess the reliability of the imputed age of death, we need to investigate the quality of the reported date of birth and date of death.

Dates of death are directly reported to the Social Security Administration from a funeral director or a family member. These reports are generally made directly following a death, minimizing the likelihood of misreporting. The date of death in the DMF almost always exactly matches the date of death in the corresponding death certificates (Hill, Preston and Rosenwaike, 2000).

Information on date of birth is submitted personally by the decedent in conjunction with a benefit claim. The Social Security Administration closely tracks age to determine eligibility for benefits. Age verification is a required condition for entitlement to benefits, and stringent tests were put in place in 1965. The focal cohorts of this study would have become eligible for Social Security benefits after these age verification procedures were put in place.
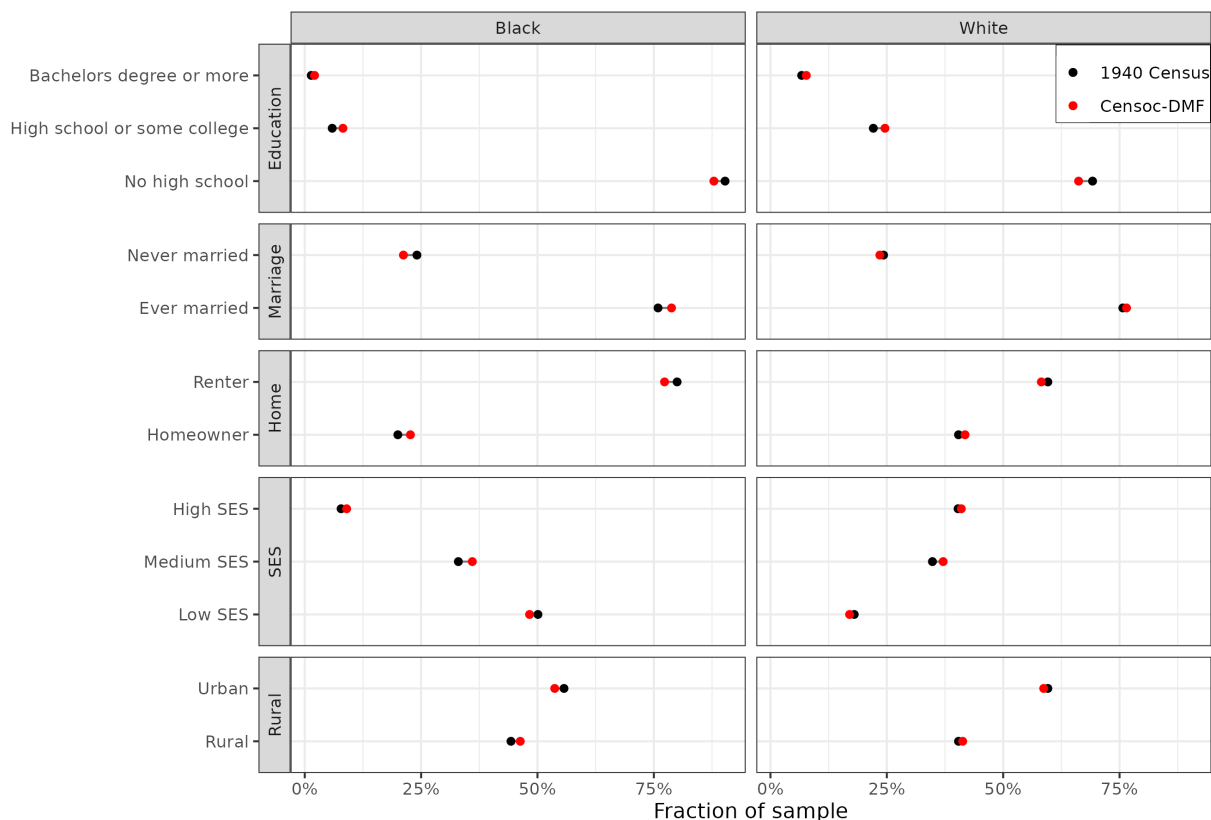
Figure 4.1: Each facet shows the composition of the CenSoc-DMF (Red) and the complete count 1940 Census (black) for a given covariate for Black and White matches. The matched sample has slightly higher socioeconomic status than the general population.

To empirically assess the reliability of the date of birth information in the DMF, we look at heaping on year of birth. Heaping, a common indicator of data quality, is the systematic misstatement of ages or dates to round or terminal ages (e.g., end in "0" or "5.") We find minimal date heaping on year of birth, as shown in Figure 4.2. However, there is slightly higher heaping for Black Americans than White Americans. To investigate whether this age heaping has any effect on our observed crossover, we conduct a sensitivity analysis by dropping years of birth that end in terminal ages and re-estimating the observed crossover.

The nature of our sample provides additional reassurance that the reported age of birth is accurate. For an individual to be successfully matched and included in our sample, their reported age in the 1940 Census must correspond to ±2 years of their year of birth reported in the DMF. Therefore, mortality records where the year of birth is misreported by over two years will be excluded from our sample. This is similar to the validation approach taken by Hill, Preston and Rosenwaike (2000); Preston et al. (1996) and gives an additional level of reassurance that the reported birth year is accurate.
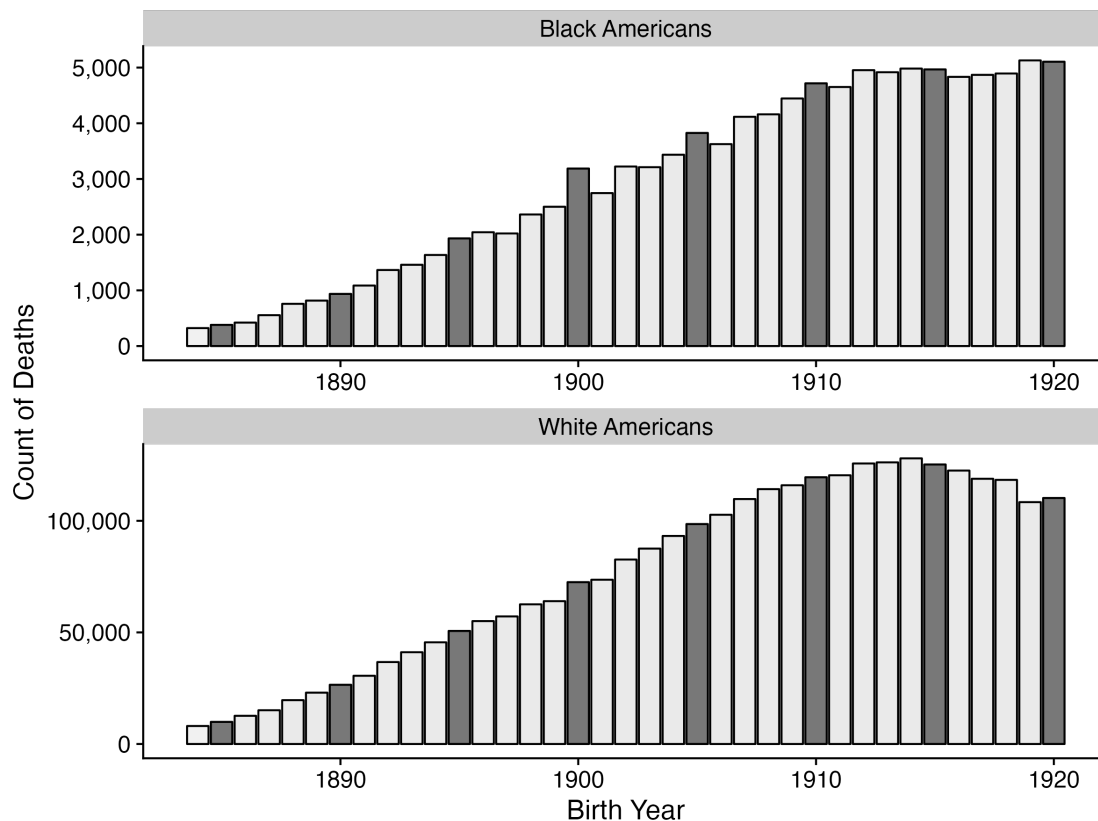
Figure 4.2: Highlighted years (dark grey) show very slight amounts of age-heaping on the terminal digits "0" or "5," suggesting the DMF has minimal misreporting of year of birth.

4.4 Methods

Estimating Mortality Rates

The CenSoc-DMF dataset only includes deaths for the left and right ("doubly") truncated window of 1975 to 2005. Further, the CenSoc-DMF does not include any measure of survivorship, as we have no way of determining whether an individual observed in the 1940 Census died outside our observation window or was not successfully matched to their death record. The absence of any measure of a denominator precludes conventional occurrence-exposure methods for estimating mortality rates (Alexander, 2018).

To overcome this, we use two different methods to estimate mortality rates in the absence of denominators. First, for the earlier cohorts of 1890–1905, we use the reverse survival method to estimate mortality rates. This approach assumes that all persons in the cohort died by the end of our mortality observation window in 2005. Specifically, we estimate the total number of survivors at a given age by summing up all the deaths occurring above

that age and then estimating the age-specific mortality rates using the age-specific ratios of
deaths to survivors. This method is only appropriate for the cohorts born before 1905, for
which only a few survivors to age 100 will die after 2005.

Second, for the later-born cohorts of 1906-1915, those that we cannot assume are extinct
by 2005, we assume the distribution of deaths within a cohort follows a Gompertz distri-
bution and use maximum likelihood estimation methods to estimate the parameters of this
distribution (Goldstein et al., 2023; Gompertz, 1825). Specifically, the hazard of dying at
age x is:

$$h_0(x) = ae^{bx} \tag{4.1}$$

where a is a background level of mortality at age x, b is the rate of mortality increase with
age, and $h(x)$ describes the hazard schedule. This approach allows us to estimate age-specific
mortality rates for both ages where we did and did not observe deaths.

Stratifying on Observed Dimensions of Heterogeneity

The classical mortality selection model used to explain the crossover is unidimensional. That
is, all heterogeneity in susceptibility to mortality is captured in a single parameter ("frailty").
A growing body of empirical research on the Black-White mortality crossover has used
individual-level covariates to study the observed dimensions of heterogeneity that constitute
frailty. Borrowing logic from unidimensional mortality selection model, these studies inves-
tigated how controlling for some piece of frailty changes the age of crossover (Sautter et al.,
2012; Dupre, Franzese and Parrado, 2006). Yet theoretical advances have demonstrated
that the unidimensional mortality selection model is not an appropriate starting point for
empirical work. When there is both observed and unobserved heterogeneity, stratifying on
observed heterogeneity can cause the age at crossover to either move up or down (Wrigley-
Field, 2020a). In other words, the age at crossover will always change when some factor
related to both race and mortality is controlled for.

One important exception occurs when an observed dimension of heterogeneity constitutes
a large portion of the overall heterogeneity. In this setting, if the crossover is caused by
heterogeneity in frailty, stratifying on a covariate that represents over 50% of total frailty
will uncross the crossover (Wrigley-Field, 2020a). For empirical researchers, this implies that
combining many covariates into a single risk measure is a promising strategy for examining
the role of observed heterogeneity in explaining the crossover.

To investigate the role that observed heterogeneity plays on the mortality crossover,
we use socioeconomic covariates available in the 1940 Census. First, we investigate the
crossover in six distinct subgroups: individuals with high education (more than eight years),
individuals with low education (less than eight years), individuals with high income (above
the median income), individuals with low income (below the median income), homeowners,
and renters. On each subgroup, we estimate age-specific mortality rates using the reverse
survival method. We then combine these covariates into a single risk score and investigate

the crossover in subgroups defined by risk. Together, these analyses allow us to investigate whether the crossover still persists when we stratify on major pieces of frailty.

4.5 Results

We first analyze the Black-White mortality crossover using both the reverse survival method and our parametric Gompertz approach. Next, we present results on observed mortality selection. Finally, we examine whether our observed heterogeneity can help explain the Black-White mortality crossover.

Black-White Mortality Crossover

We first examine the Black-White crossover for the pooled birth cohorts of 1890-1905. Figure 4.3a shows a clear mortality crossover at age 86, consistent with past findings. For this analysis, we estimated age-specific mortality rates using the reverse survival method. Because our mortality data showed very slight heaping on year of birth, as a sensitivity analysis, we recalculated our age-specific mortality rates excluding birth years with potential age heaping: 1890, 1895, 1900, and 1905. Figure 4.3b shows that the crossover persists, suggesting that low-quality mortality data is not responsible for the crossover.

For the cohorts of 1905-1915, which were not extinct by the end of our mortality observation window in 2005, we fit a parametric Gompertz model to calculate age-specific mortality rates (Goldstein et al., 2023). We perform maximum likelihood estimation for the Black and White groups separately. Figure 4.4 shows a mortality crossover at age 90, slightly higher than our observed age at crossover for the cohorts of 1890–1905. A higher age at crossover for later birth cohorts is consistent with past studies (Masters, 2012).
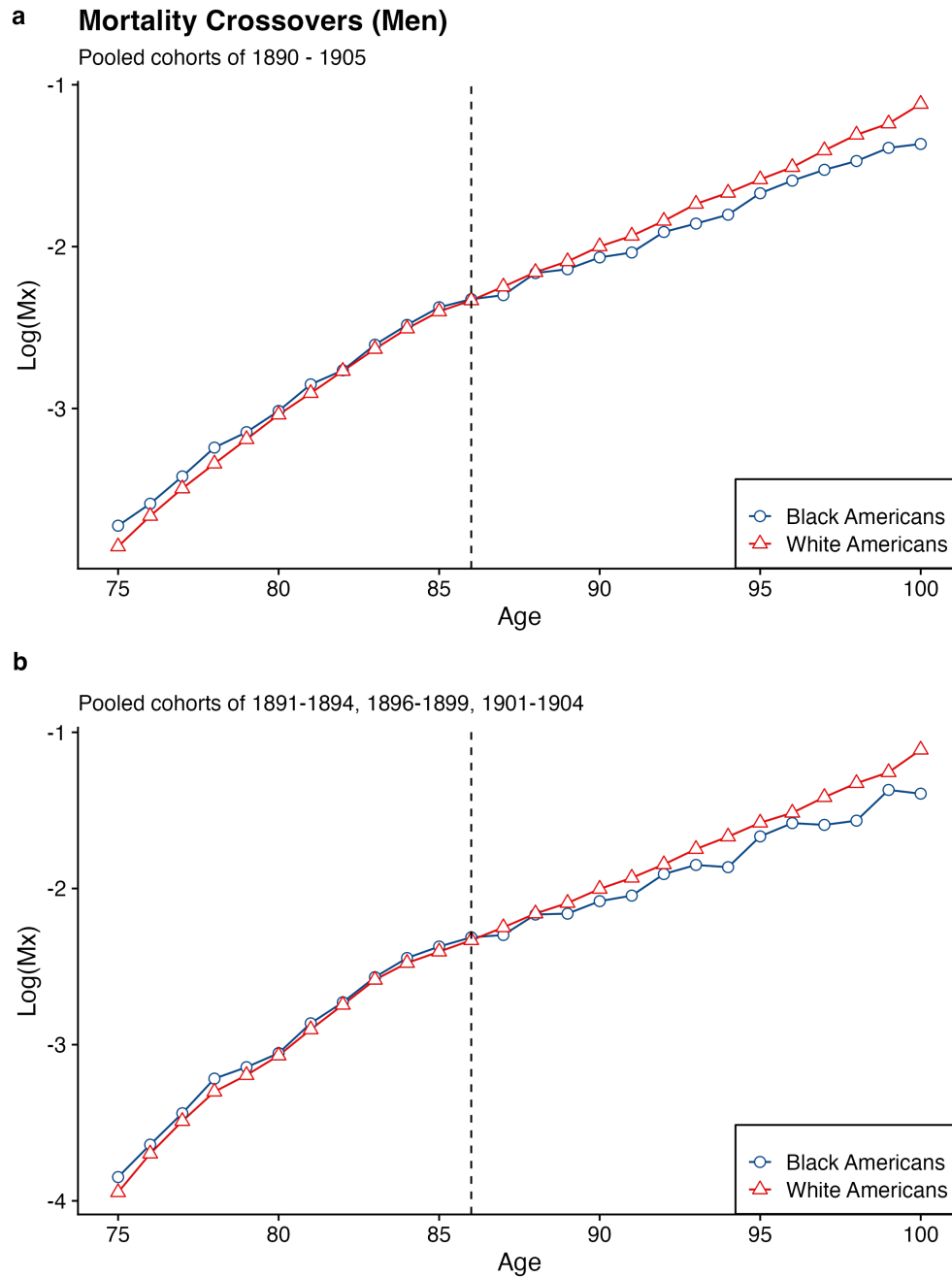
Figure 4.3: Panel (a) shows the Black-White mortality crossover for the cohorts of 1890-1905. Panel (b) shows the mortality crossover dropping the cohorts of 1890, 1895, 1900, and 1905, where we observed slight but detectable age heaping. The mortality rates were estimated using the reverse survival method.
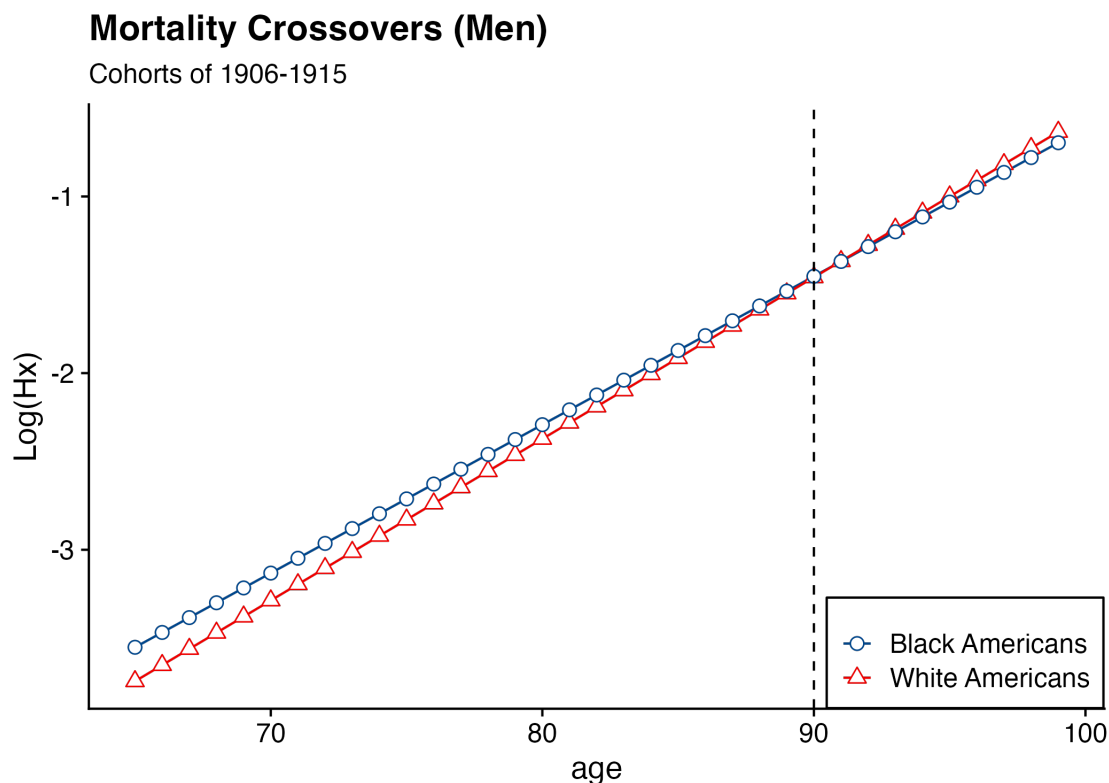
Figure 4.4: Black White mortality crossovers for cohorts of 1905-1915.

Observed Mortality Selection

To investigate mortality selection, we track how the characteristics of survivors change as a cohort ages, and members of the cohort die off. We focus on how the composition of the cohorts of 1909–1911 changes with respect to employment, educational attainment, socioeconomic status score, wage and salary income, homeownership status, and residing in the South. We interpret an increase in a dimension of socioeconomic status as a cohort ages to be evidence of selective mortality: more frail individuals are dying off at earlier ages.

As shown in Figure 4.5, we do observe selective mortality, which is more pronounced for White Americans than Black Americans. For instance, members of the cohort of 1909–1911 who survived to age 65 have approximately 10 years of education, while members of the cohort who survived to age 90 have approximately 10.6 years of education. The difference is more slight for the cohort of Black Americans: survivors at age 65 had 6.6 years of education, and survivors at age 90 had 6.7 years of education. Across all of the covariates tested, we find that the surviving members of a cohort become increasingly advantaged as the cohort ages.
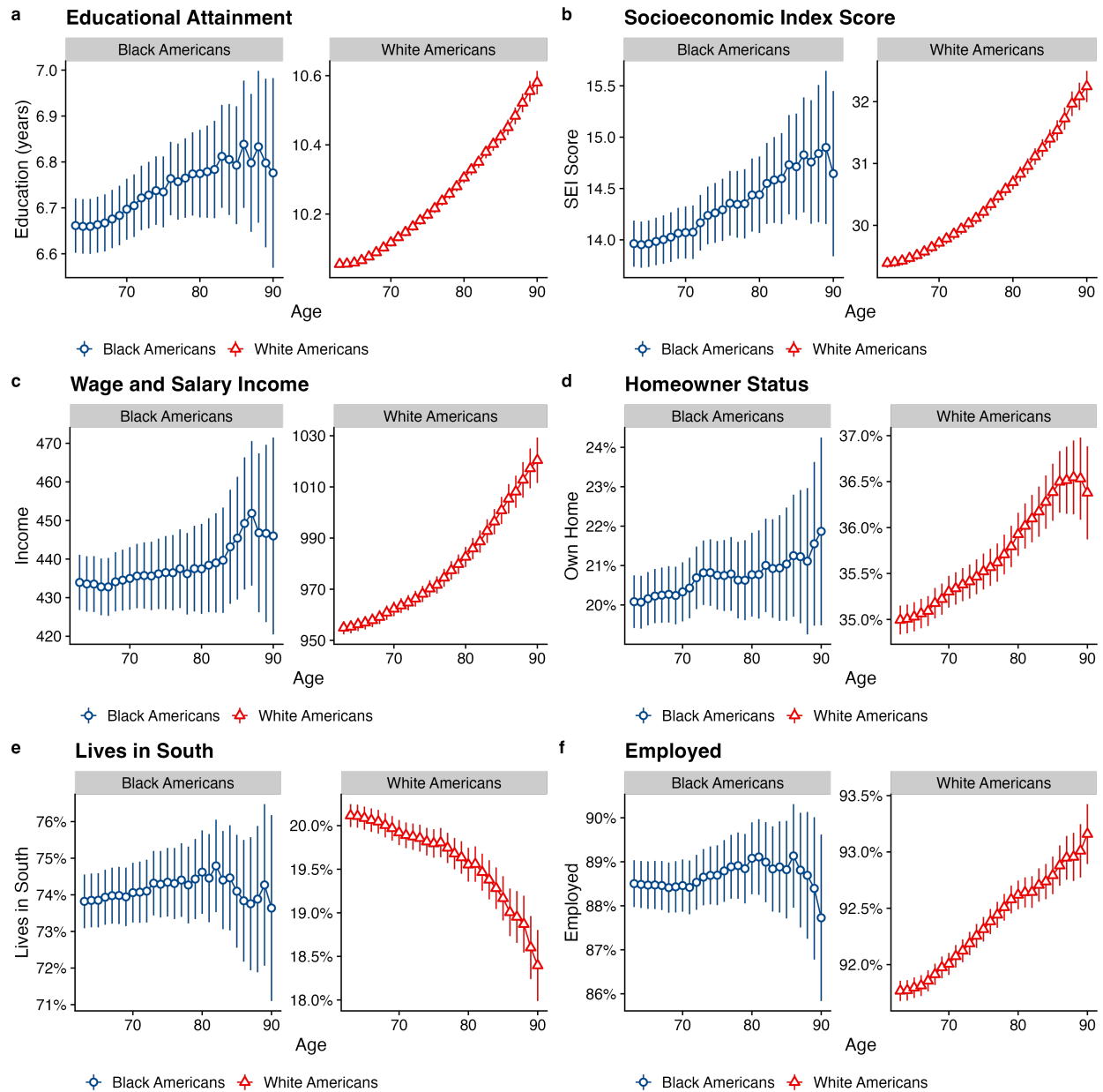
Figure 4.5: Changing composition of the survivors. We see only modest evidence of selection. Error bands show 95% uncertainty intervals.
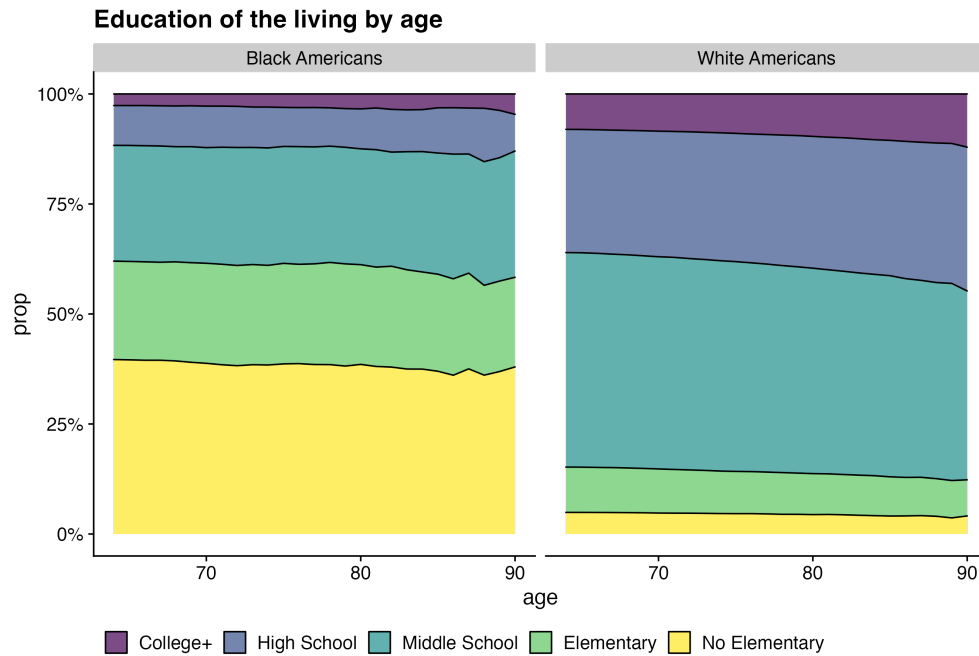
Figure 4.6: Changing educational composition of the survivors.

The Surprising Non-Effect of Observed Heterogeneity on the Mortality Crossover

Next, we investigate the effect of observed heterogeneity on the morality crossover. We split our 1890-1905 birth cohort sample into different population subgroups defined by education, homeownership, and wage and salary income. Figure 4.7 shows the result of this analysis: the crossover persists across all subgroups.
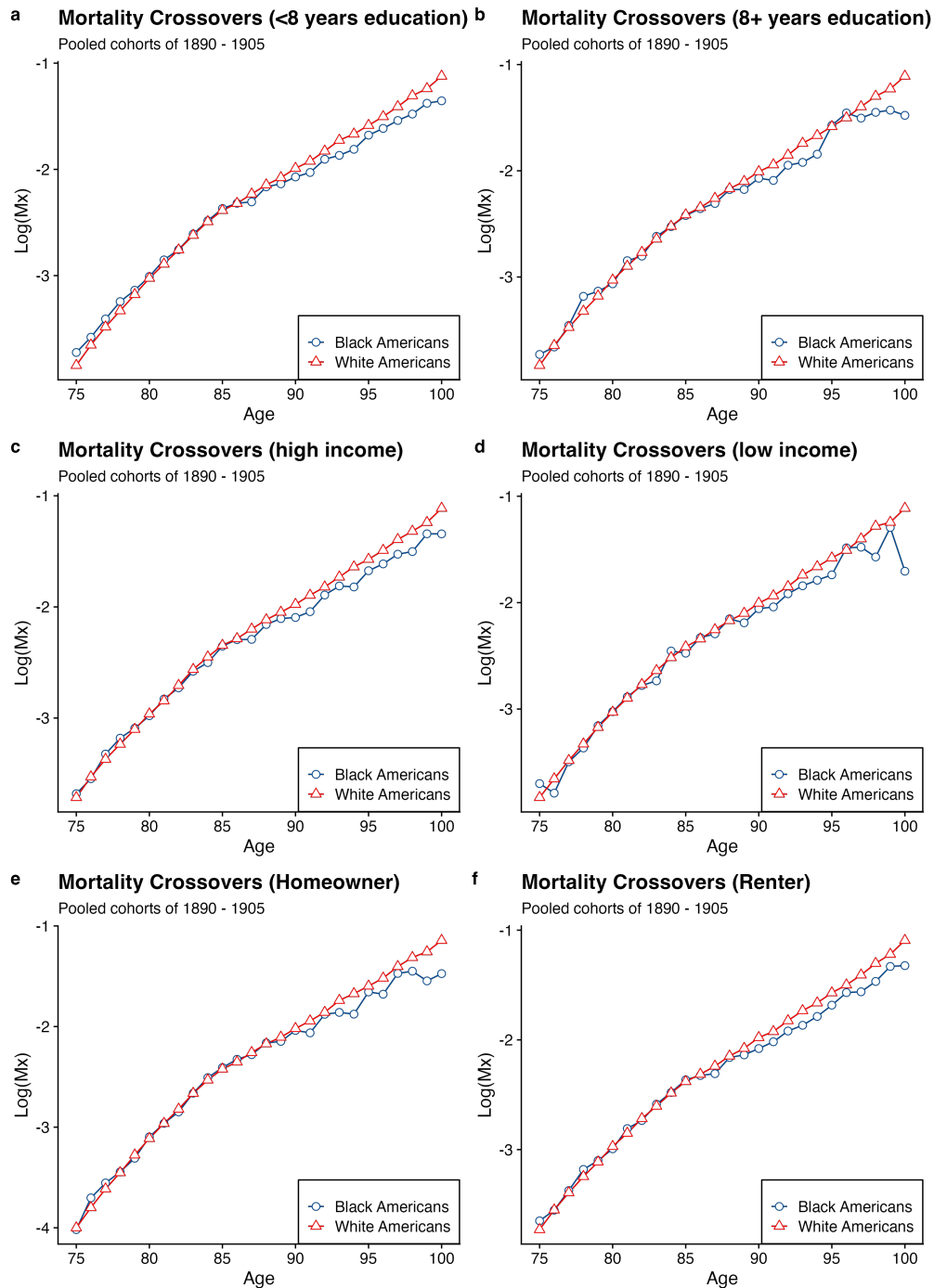
Figure 4.7: Black White mortality crossover for different subgroups defined by socioeconomic status.

Next, we follow the advice outlined in Wrigley-Field (2020*a*) and investigate the mortality

crossover stratified by risk scores. To construct the risk score, we aggregate together the
following covariates into a single score: education, wage and salary income, socioeconomic
index, marital status, employment, living in the South, and owning a home. The motivation
for constructing this risk score is to capture as much of the heterogeneity in frailty as possible
in one score. To estimate the risk score, we fit linear regressions of the form:

$$\text{death\_age}_i = educ_i + income_i + homeowner_i + marital\_status_i + southern_i + \epsilon_i. \quad (4.2)$$

We fit these regression separately for Black and White subgroups.[1] We fit separate models
for Black and White subgroups as the relationship between covariates such as education and
longevity vary across racial groups (Card and Krueger, 1992). We use these models to assign
each person in the cohorts of 1890-1905 a "low," "medium," or "high" risk group. On each
subgroup defined by aggregate risk score, we estimate age-specific mortality rates using the
reverse survival method and examine the crossover.

Figure 4.8a presents the age-specific mortality rates for Black and White men within
risk group. We see the crossover persists in all different risk groups. Figure 4.8b plots the
difference in log hazards, again finding a clear crossover for all three groups.

---

[1]We first fit this model on our analytic sample, those born between 1890-1905. To avoid potentially
over-fitting, we also fit the model on the out-of-sample cohort of 1906. Results from both models provided
highly comparable predictions of risk score.

a **Mortality Crossovers by Risk Score**
Pooled cohorts of 1890 - 1905

Black Americans   White Americans

b **Difference in Log Hazards (White - Black)**
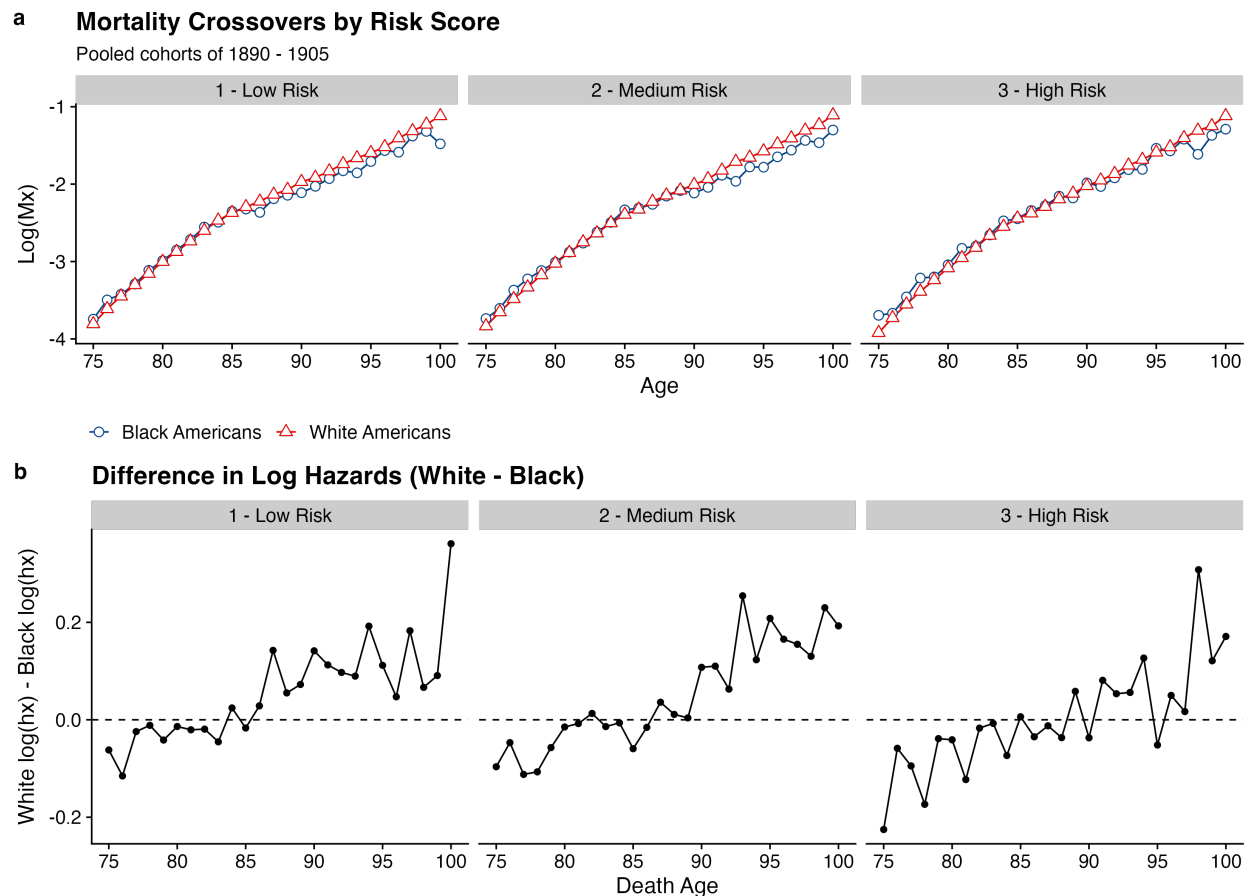
Figure 4.8: Black White mortality crossover by risk score.

4.6 Discussion

This study uses mortality records from the Death Master File (DMF) linked to the 1940 Census to investigate the Black-White mortality crossover. We find a clear mortality crossover at age 85 for men in the birth cohorts of 1890–1905 using reverse survival methods to estimate age-specific mortality rates. Using a Gompertz parametric maximum likelihood approach, we find a mortality crossover at age 90 for the birth cohorts of 1906–1915. Given the reliability of the DMF mortality data, we interpret this as evidence that the Black-White mortality crossover is not simply an artifact of sparse data or age misreporting: the crossover persists even when we restrict the sample to the highest-quality mortality data.

Using individual-level characteristics from the 1940 Census, we investigate observable mortality selection. We find clear evidence of selective mortality: as a cohort ages, the survivors have increasingly higher educational attainment, rates of homeownership, rates of employment (in 1940), and wage and salary income. However, the observable selection is

relatively modest and is more pronounced for White Americans than Black Americans. The lack of observable mortality selection for Black Americans is perhaps attributable to the weaker correlation between covariates, such as education or income, and mortality risk for Black Americans (Card and Krueger, 1992).

Our investigation of the Black-White mortality crossover for subgroups defined by socioeconomic characteristics indicated a clear crossover in every subgroup. Additionally, the crossover persisted when we stratified on a risk score that aggregated many mortality covariates. This suggests that stratifying on observed socioeconomic dimensions of heterogeneity does not explain the crossover. There are two potential explanations for this finding. First, it is possible that sociodemographic characteristics alone simply do not capture enough of the heterogeneity in frailty to really uncross the crossover. Second, it is possible that the crossover is not driven by heterogeneity in frailty at all; rather, there is actually some true narrowing of inequality at the most advanced ages.

Taken together, our results suggest that the mortality crossover is real and not an artifact of measurement or data errors. Our data allow us to study the mortality experience of real cohorts, not the synthetic period measures commonly used to study the crossover. However, our study cannot make definitive claims about the theoretical explanations for the crossover. While our study found that stratifying on observed dimensions of frailty, such as educational attainment or homeownership, does not explain the crossover, it is possible that we are simply not capturing enough of the heterogeneity in frailty to uncross the crossover.

There are several limitations and avenues for future research. First, we only observe mortality in the window of 1975–2005, so our analyses are restricted to birth cohorts that would be experiencing a crossover in our mortality observation window. Second, it is possible that the sociodemographic characteristics we observe only constitute a very small piece of frailty and therefore have limited utility for explaining the crossover. Future research could test whether covariates that capture more of the heterogeneity in frailty, such as biomarkers, anthropometric measures (weight, height), and direct measurement of subjective and objective health, have a greater impact on the crossover. However, the present study benefits from an exceptionally large sample size, making it potentially challenging for other studies to achieve comparable levels of precision. Third, while we find little evidence of age misstatement or exaggeration, it is possible there remain undetected age misreports in the DMF. Finally, this analysis is limited in scope to men. Broadening this study to include women is necessary to make complete claims about health and longevity disparities in the most advanced ages.

Chapter 5

Conclusion

In this dissertation, I draw on the newly-available CenSoc datasets to study mortality in the United States. The unprecedented scale and size of these datasets allow me to study the effect of homeownership on longevity, the Black-White mortality crossover, and the predictability of individual-level longevity. All my studies investigate cohort lifespans rather than artificially constructed period life expectancy common in the literature. Together, my chapters highlight the importance of considering heterogeneity and age-specific trends in studying mortality outcomes.

The second chapter of my dissertation investigates the relationship between homeownership and life expectancy. We first investigate the unadjusted difference in life expectancy, finding homeowners live approximately 0.6 years longer than renters. We then use a sibling-based identification strategy to estimate the first U.S.-based estimates of the effect of homeownership on longevity, finding owning a home in early life has a positive effect of an additional 0.38 years on life expectancy at age 65. Substantively, we provide a theoretical framework outlining the causal pathways between homeownership and longevity and find empirical evidence to indicate that owning a home does indeed have a causal effect on homeownership. Methodologically, this study is the first to use linked administrative data in the United States to study homeownership and longevity. However, residual confounding within sibling pairs may threaten the causal interpretation of our results. Future work could consider using a different identification strategy, such as exploiting different state-level policies as an instrument, to independently corroborate the causal estimates presented in this study.

The third chapter investigates the predictability of individual-level longevity using basic sociodemographic characteristics. Substantively, this chapter demonstrates that sociodemographic characteristics measured in early adulthood cannot accurately predict individual-level longevity. This highlights a tension between the striking group-level sociodemographic disparities in mortality and the unpredictability of individual-level longevity. The results suggest fundamental uncertainty around individual-level mortality needs to be better acknowledged and incorporated into demographic theory. Methodologically, this study highlights the limitations of using machine learning algorithms in Social Science: their performance over benchmark linear regression models is modest. Developing a more formal framework

for assessing the performance of machine learning methods is critical, particularly given the increasing application of prediction for policy decision-making.

The fourth chapter revisits a perennial topic for mortality researchers: the Black-White mortality crossover. While most investigations of the Black-White mortality crossover use synthetic period data, our dataset allows us to track real cohorts. We find evidence of a clear crossover at age 85 for the birth cohorts of 1890-1905 and rule out that the crossover is the product of age misreporting or exaggeration. Using characteristics measured in the 1940 Census, I find a crossover for several different subgroups defined by different socioeconomic characteristics. The major contribution of this study is that the crossover is a real phenomenon, not simply a data artifact of age misreports. Further, this crossover cannot be uncrossed by stratifying on sociodemographic covariates alone. However, I am not able to definitively make a claim about what is driving the crossover.

There are several shared limitations across all three studies. My investigations are limited in scope to the United States and cannot speak to important cross-national differences in mortality. I only examine all-cause mortality and not other important measures of health, such as morbidity. Further, I am not able to study specific causes of death. My analyses are restricted primarily to male deaths occurring after the age of 65 between the periods of 1975 and 2005, and I am not able to study mortality in early or mid-life. This period restriction prevents me from studying some of the most pressing topics in mortality today, such as the deaths of despair.

There are several promising avenues for future research. Identifying a factor such as homeownership that has a substantively meaningful effect on longevity is important for furthering our understanding of the social determinants of longevity. Yet identifying such causal factors is only partially satisfying. In the case of homeownership, being able to identify the specific causal mechanisms through which something like homeownership affects longevity is critical for a complete understanding of the relationships between homeownership and longevity. Future studies quantifying the relative contribution of mechanisms such as wealth accumulation to the crossover is an important direction for future research.

Social Scientists are increasingly recognizing the importance of characterizing the predictability of demographic outcomes. While my dissertation illustrates the challenges of predicting mortality at the individual level using sociodemographic characteristics, the theoretical implications could be further developed. Specifically, the fundamental uncertainty around individual-level mortality needs to be better acknowledged and incorporated into demographic theory.

There is still more empirical work to be done investigating the Black-White mortality crossover. While my study found a clear crossover in different sets of birth cohorts using a reliable, large-scale dataset, the covariates available in the 1940 Census likely did not capture the major pieces of heterogeneity that constitute frailty. Future work could investigate the other larger dimensions of frailty, such as biomarkers, anthropometric measures (weight, height), and direct measurements of subjective and objective health. Together, these are all important components of future research, which will advance our understanding of mortality in the United States.

Bibliography

Aaronson, Daniel, Daniel Hartley, Bhashkar Mazumder and Martha Stinson. 2022. "The Long-Run Effects of the 1930s Redlining Maps on Children." *Journal of Economic Literature* .

Abramitzky, Ran and Leah Boustan. 2017. "Immigration in American Economic History." *Journal of Economic Literature* 55(4):1311–1345.

Abramitzky, Ran, Leah Boustan and Katherine Eriksson. 2019. "To the New World and Back Again: Return Migrants in the Age of Mass Migration." *ILR Review* 72(2):300–322.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum and Santiago Pérez. 2021. "Automated Linking of Historical Data." *Journal of Economic Literature* 59(3):865–918.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, Santiago Pérez and Myera Rashid. 2020. "Census Linking Project: Version 1.0.".

Abramitzky, Ran, Leah Platt Boustan and Katherine Eriksson. 2012. "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review* 102(5):1832–1856.

Abramitzky, Ran, Leah Platt Boustan and Katherine Eriksson. 2014. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." *Journal of Political Economy* 122(3):467–506.

Acharya, Binod, Dependra Bhatta and Chandra Dhakal. 2022. "The Risk of Eviction and the Mental Health Outcomes among the US Adults." *Preventive Medicine Reports* 29:101981.

Ager, Philipp, James J Feigenbaum, Casper Worm Hansen and Hui Ren Tan. 2021. "How the Other Half Died: Immigration and Mortality in US Cities.".

Alexander, Michelle. 2010. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness.* New Press.

Alexander, Monica. 2018. Deaths without Denominators: Using a Matched Dataset to Study Mortality Patterns in the United States. Preprint SocArXiv.

Apgar, William C. and Zhu Xiao Di. 2006. Housing Wealth and Retirement Savings. In *The Oxford Handbook of Pensions and Retirement Income*, ed. Gordon L. Clark, Alicia H. Munnell and J. Michael Orszag. Oxford University Press p. 0.

Arias, Elizabeth. 2006. "National Vital Statistics Reports, Vol. 54, No. 14 (4/19/2006).".

Arpino, Bruno, Marco Le Moglie and Letizia Mencarini. 2022. "What Tears Couples Apart: A Machine Learning Analysis of Union Dissolution in Germany." *Demography* 59(1):161–186.

Austin, Peter C. and Carl van Walraven. 2011. "The Mortality Risk Score and the ADG Score: Two Points-Based Scoring Systems for the Johns Hopkins Aggregated Diagnosis Groups (ADGs) to Predict Mortality in a General Adult Population Cohort in Ontario, Canada." *Medical care* 49(10):940–947.

Barbi, Elisabetta, Francesco Lagona, Marco Marsili, James W. Vaupel and Kenneth W. Wachter. 2018. "The Plateau of Human Mortality: Demography of Longevity Pioneers." *Science* 360(6396):1459–1461.

Berkman, L. F., L. Leo-Summers and R. I. Horwitz. 1992. "Emotional Support and Survival after Myocardial Infarction. A Prospective, Population-Based Study of the Elderly." *Annals of Internal Medicine* 117(12):1003–1009.

Berkman, Lisa, Burton Singer and Kenneth Manton. 1989. "Black/White Differences in Health Status and Mortality Among the Elderly." *Demography* 26(4):661–678.

Berkman, Lisa F. and S. Leonard Syme. 1979. "Social Networks, Host Resistance, and Mortality: A Nine-Year Follow-up Study of Alameda County Residents." *American Journal of Epidemiology* 109(2):186–204.

Boardman, Jason D. and Jason M. Fletcher. 2015. "To Cause or Not to Cause? That Is the Question, but Identical Twins Might Not Have All of the Answers." *Social science & medicine (1982)* 127:198–200.

Boehm, Thomas P. and Alan Schlottmann. 2008. "Wealth Accumulation and Homeownership: Evidence for Low-Income Households." *Cityscape* 10(2):225–256.

Boustan, Leah Platt. 2017. *Competition in the Promised Land: Black Migrants in Northern Cities and Labor Markets.* Princeton University Press.

Breen, Casey and Joshua R. Goldstein. 2022. "Berkeley Unified Numident Mortality Database: Public Administrative Records for Individual-Level Mortality Research." *Demographic Research* 47:111–142.

Breen, Casey and Maria Osborne. 2022. An Assessment of CenSoc Match Quality. Preprint SocArXiv.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16(3):199–231.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Hung H. V. Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik K. Andersen, Daniel Auer, Flavio Azevedo, Oke Bahnsen, Dave Balzer, Gerrit Bauer, Paul C. Bauer, Markus Baumann, Sharon Baute, Verena Benoit, Julian Bernauer, Carl Berning, Anna Berthold, Felix S. Bethke, Thomas Biegert, Katharina Blinzler, Johannes N. Blumenberg, Licia Bobzien, Andrea Bohman, Thijs Bol, Amie Bostic, Zuzanna Brzozowska, Katharina Burgdorf, Kaspar Burger, Kathrin B. Busch, Juan Carlos-Castillo, Nathan Chan, Pablo Christmann, Roxanne Connelly, Christian S. Czymara, Elena Damian, Alejandro Ecker, Achim Edelmann, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andrea Forster, Chris Gaasendam, Konstantin Gavras, Vernon Gayle, Theresa Gessler, Timo Gnambs, Amélie Godefroidt, Max Grömping, Martin Groß, Stefan Gruber, Tobias Gummer, Andreas Hadjar, Jan Paul Heisig, Sebastian Hellmeier, Stefanie Heyne, Magdalena Hirsch, Mikael Hjerm, Oshrat Hochman, Andreas Hövermann, Sophia Hunger, Christian Hunkler, Nora Huth, Zsófia S. Ignácz, Laura Jacobs, Jannes Jacobsen, Bastian Jaeger, Sebastian Jungkunz, Nils Jungmann, Mathias Kauff, Manuel Kleinert, Julia Klinger, Jan-Philipp Kolb, Marta Kołczyńska, John Kuk, Katharina Kunißen, Dafina Kurti Sinatra, Alexander Langenkamp, Philipp M. Lersch, Lea-Maria Löbel, Philipp Lutscher, Matthias Mader, Joan E. Madia, Natalia Malancu, Luis Maldonado, Helge Marahrens, Nicole Martin, Paul Martinez, Jochen Mayerl, Oscar J. Mayorga, Patricia McManus, Kyle McWagner, Cecil Meeusen, Daniel Meierrieks, Jonathan Mellon, Friedolin Merhout, Samuel Merk, Daniel Meyer, Leticia Micheli, Jonathan Mijs, Cristóbal Moya, Marcel Neunhoeffer, Daniel Nüst, Olav Nygård, Fabian Ochsenfeld, Gunnar Otte, Anna O. Pechenkina, Christopher Prosser, Louis Raes, Kevin Ralston, Miguel R. Ramos, Arne Roets, Jonathan Rogers, Guido Ropers, Robin Samuel, Gregor Sand, Ariela Schachter, Merlin Schaeffer, David Schieferdecker, Elmar Schlueter, Regine Schmidt, Katja M. Schmidt, Alexander Schmidt-Catran, Claudia Schmiedeberg, Jürgen Schneider, Martijn Schoonvelde, Julia Schulte-Cloos, Sandy Schumann, Reinhard Schunck, Jürgen Schupp, Julian Seuring, Henning Silber, Willem Sleegers, Nico Sonntag, Alexander Staudt, Nadia Steiber, Nils Steiner, Sebastian Sternberg, Dieter Stiers, Dragana Stojmenovska, Nora Storz, Erich Striessnig, Anne-Kathrin Stroppe, Janna Teltemann, Andrey Tibajev, Brian Tung, Giacomo Vagni, Jasper Van Assche, Meta van der Linden, Jolanda van der Noll, Arno Van Hootegem, Stefan Vogtenhuber, Bogdan Voicu, Fieke Wagemans, Nadja Wehl, Hannah Werner, Brenton M. Wiernik, Fabian Winter, Christof Wolf, Yuki Yamada, Nan Zhang, Conrad Ziller, Stefan Zins and Tomasz Żółtak. 2022. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty." *Proceedings of the National Academy of Sciences* 119(44):e2203150119.

Bryan L. Sykes and Michelle Maroto. 2016. "A Wealth of Inequalities: Mass Incarceration,

Employment, and Racial Disparities in U.S. Household Wealth, 1996 to 2011." *RSF: The Russell Sage Foundation Journal of the Social Sciences* 2(6):129.

Card, David and Alan B. Krueger. 1992. "School Quality and Black-White Relative Earnings: A Direct Assessment." *The Quarterly Journal of Economics* 107(1):151–200.

Card, David, Carlos Dobkin and Nicole Maestas. 2008. "The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare." *American Economic Review* 98(5):2242–2258.

Card, David E., Raj Chetty, Martin S. Feldstein and Emmanuel Saez. 2010. "Expanding Access to Administrative Data for Research in the United States." *American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas.* .

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron and David Cutler. 2016. "The Association Between Income and Life Expectancy in the United States, 2001-2014." *JAMA* 315(16):1750.

Chetty, Raj and Nathaniel Hendren. 2018. "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects*." *The Quarterly Journal of Economics* 133(3):1107–1162.

Chicco, Davide, Matthijs J. Warrens and Giuseppe Jurman. 2021. "The Coefficient of Determination R-squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation." *PeerJ Computer Science* 7:e623.

Chivers, Walter R. 1949. "The Negro Ghetto. By Robert C. Weaver. New York: Harcourt, Brace and Company, 1948. 404 Pp. $3.75." *Social Forces* 27(3):328.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko and Rhema Vaithianathan. 2018. A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency.* PMLR pp. 134–148.

Cobb, Sidney. 1976. "Social Support as a Moderator of Life Stress:." *Psychosomatic Medicine* 38(5):300–314.

Collins, William J. and Robert A. Margo. 2011. "Race and Home Ownership from the End of the Civil War to the Present." *American Economic Review* 101(3):355–359.

Conley, Dalton, Kathryn M. Pfeiffer and Melissa Velez. 2007. "Explaining Sibling Differences in Achievement and Behavioral Outcomes: The Importance of within- and between-Family Factors." *Social Science Research* 36(3):1087–1104.

Crimmins, Eileen M. and Yuan S. Zhang. 2019. "Aging Populations, Mortality, and Life Expectancy." *Annual Review of Sociology* 45(1):69–89.

Currie, Janet and Duncan Thomas. 1995. "Does Head Start Make a Difference?" *The American Economic Review* 85(3):341–364.

Dannefer, Dale. 2003. "Cumulative Advantage/Disadvantage and the Life Course: Cross-Fertilizing Age and Social Science Theory." *The Journals of Gerontology: Series B* 58(6):S327–S337.

Davis, Earle. and Margret Fine-Davis. 1981. "Predictors of Satisfaction with Housing and Neighbourhood: A Nationwide Study in the Republic of Ireland." *Social Indicators Research* 9(4):477–494.

Davis, Ronald L. F. 1992. "Review of Freedom: A Documentary History of Emancipation, 1861-1867. Selected from the Holdings of the National Archives of the United States. Series I. Volume III: The Wartime Genesis of Free Labor: The Lower South." *The Journal of Southern History* 58(4):725–727.

Di, Zhu Xiao, Eric Belsky and Xiaodong Liu. 2007. "Do Homeowners Achieve More Household Wealth in the Long Run?" *Journal of Housing Economics* 16(3):274–290.

Dietz, Robert D. and Donald R. Haurin. 2003. "The Social and Private Micro-Level Consequences of Homeownership." *Journal of Urban Economics* 54(3):401–450.

DiPrete, Thomas A. and Gregory M. Eirich. 2006. "Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments." *Annual Review of Sociology* 32(1):271–297.

Donoho, David. 2017. "50 Years of Data Science." *Journal of Computational and Graphical Statistics* 26(4):745–766.

Dressel, Julia and Hany Farid. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4(1):eaao5580.

Dupre, Matthew E., Alexis T. Franzese and Emilio A. Parrado. 2006. "Religious Attendance and Mortality: Implications for the Black-White Mortality Crossover." *Demography* 43(1):141–164.

Elo, Irma T. 2009. "Social Class Differentials in Health and Mortality: Patterns and Explanations in Comparative Perspective." *Annual Review of Sociology* 35:553–572.

Elo, Irma T., Cassio M. Turra, Bert Kestenbaum and B. Reneé Ferguson. 2004. "Mortality Among Elderly Hispanics in the United States: Past Evidence and New Results." *Demography* 41(1):109–128.

Elsinga, Marja, ed. 2007. *Home Ownership beyond Asset and Security: Perceptions of Housing Related Security and Insecurity in Eight European Countries.* Number 32 *in* "Housing and Urban Policy Studies" Amsterdam, the Netherlands: IOS Press.

Families, Fragile. 2021. "Fragile Families & Child Wellbeing Study, Fragile Families Publication Search.".

Feehan, Dennis M. 2018. "Separating the Signal From the Noise: Evidence for Deceleration in Old-Age Death Rates." *Demography* 55(6):2025–2044.

Feigenbaum, James J., Christopher Muller and Elizabeth Wrigley-Field. 2019. "Regional and Racial Inequality in Infectious Disease Mortality in U.S. Cities, 1900–1948." *Demography* 56(4):1371–1388.

Feigenbaum, James J., Lauren Hoehn-Velasco, Christopher Muller and Elizabeth Wrigley-Field. 2022. "1918 Every Year: Racial Inequality in Infectious Mortality, 1906–1942." *AEA Papers and Proceedings* 112:199–204.

Fenelon, Andrew. 2013. "An Examination of Black/White Differences in the Rate of Age-Related Mortality Increase." *Demographic Research* 29:441–472.

Finch, Caleb E. and Eileen M. Crimmins. 2004. "Inflammatory Exposure and Historical Changes in Human Life-Spans." *Science* 305(5691):1736–1739.

Finlay, Keith and Katie R. Genadek. 2021. "Measuring All-Cause Mortality With the Census Numident File." *American Journal of Public Health* 111(Suppl 2):S141–S148.

Finnigan, Ryan. 2014. "Racial and Ethnic Stratification in the Relationship Between Homeownership and Self-Rated Health." *Social science & medicine (1982)* 115:72–81.

Fishback, Price V., Jonathan Rose, Kenneth A. Snowden and Thomas Storrs. 2021. "New Evidence on Redlining by Federal Housing Programs in the 1930s.".

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29(5):1189–1232.

Galster, George C. 1983. "Empirical Evidence on Cross-Tenure Differences in Home Maintenance and Conditions." *Land Economics* 59(1):107.

Goeken, Ron, Lap Huynh, Thomas Lenius and Rebecca Vick. 2011. "New Methods of Census Record Linking." *Historical methods* 44(1):7–14.

Goldstein, J. R., M. Alexander, C. Breen, A. Miranda González, F. Menares, M. Osborne, M. Snyder and U. Yildirim. 2021. "Censoc Project." *CenSoc Mortality File: Version 2.0. Berkeley: University of California* .

Goldstein, Joshua R, Maria Osborne, Casey F Breen and Serge Atherwood. 2023. "Mortality Modeling of Partially Observed Cohorts Using Administrative Death Records." *Soc Arxiv.* p. 33.

Goldstein, Joshua R. and Ryan Edwards. 2020. "The Surprising Staircase: A New, High Resolution Look at the Effect of Education on Longevity in the United States.".

Gompertz, Benjamin. 1825. "On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies." *Philosophical Transactions of the Royal Society of London* .

Graetz, Nick and Michael Esposito. 2022. "Historical Redlining and Contemporary Racial Disparities in Neighborhood Life Expectancy." *Social Forces* p. soac114.

Halpern-Manners, Andrew, Jonas Helgertz, John Robert Warren and Evan Roberts. 2020. "The Effects of Education on Mortality: Evidence From Linked U.S. Census and Administrative Mortality Data." *Demography* 57(4):1513–1541.

Harding, John P., Stuart S. Rosenthal and C. F. Sirmans. 2007. "Depreciation of Housing Capital, Maintenance, and House Price Inflation: Estimates from a Repeat Sales Model." *Journal of Urban Economics* 61(2):193–217.

Haurin, Donald, Toby L. Parcel and R. Jean Haurin. 2002. "Does Homeownership Affect Child Outcomes?" *Real Estate Economics* 30(4):635–666.

Herbert, Christopher E, Daniel T McCue and Rocio Sanchez-Moyano. 2013. "Is Homeownership Still an Effective Means of Building Wealth for Low-income and Minority Households? (Was It Ever?)." p. 56.

Hill, Mark E. 2001. "The Social Security Administration's Death Master File: The Completeness of Death Reporting at Older Ages." *Social Security Bulletin* 64(1).

Hill, Mark E., Samuel H. Preston and Ira Rosenwaike. 2000. "Age Reporting among White Americans Aged 85+: Results of a Record Linkage Study." *Demography* 37(2):175–186.

Hofman, Jake M., Amit Sharma and Duncan J. Watts. 2017. "Prediction and Explanation in Social Systems." *Science* 355(6324):486–488.

Hummer, Robert A. 1996. "Black-White Differences in Health and Mortality:." *The Sociological Quarterly* 37(1):105–125.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge: Cambridge University Press.

Janssen, I., L. Krabbendam, M. Bak, M. Hanssen, W. Vollebergh, R. de Graaf and J. van Os. 2004. "Childhood Abuse as a Risk Factor for Psychotic Experiences." *Acta Psychiatrica Scandinavica* 109(1):38–45.

Johnson, N. E. 2000. "The Racial Crossover in Comorbidity, Disability, and Mortality." *Demography* 37(3):267–283.

Kadi, Justin. 2015. "Recommodifying Housing in Formerly "Red" Vienna?" *Housing, Theory and Society* 32(3):247–265.

Kashyap, Ridhi. 2021. "Has Demography Witnessed a Data Revolution? Promises and Pitfalls of a Changing Data Ecosystem." *Population Studies* 75(sup1):47–75.

Kearns, Ade, Rosemary Hiscock, Anne Ellaway and Sally MaCintyre. 2000. "'Beyond Four Walls'. The Psycho-social Benefits of Home: Evidence from West Central Scotland." *Housing Studies* 15(3):387–410.

Kessler, Ronald C., Irving Hwang, Claire A. Hoffmire, John F. McCarthy, Maria V. Petukhova, Anthony J. Rosellini, Nancy A. Sampson, Alexandra L. Schneider, Paul A. Bradley, Ira R. Katz, Caitlin Thompson and Robert M. Bossarte. 2017. "Developing a Practical Suicide Risk Prediction Model for Targeting High-risk Patients in the Veterans Health Administration." *International Journal of Methods in Psychiatric Research* 26(3):e1575.

Kestenbaum, B. 1992. "A Description of the Extreme Aged Population Based on Improved Medicare Enrollment Data." *Demography* 29(4):565–580.

Killewald, Alexandra and Brielle Bryan. 2016. "Does Your Home Make You Wealthy?" *RSF: The Russell Sage Foundation Journal of the Social Sciences* 2(6):110.

Kim, JinYoung and Richard Miech. 2009. "The Black-White Difference in Age Trajectories of Functional Health over the Life Course." *Social science & medicine (1982)* 68(4):717.

Krieger, James and Donna L. Higgins. 2002. "Housing and Health: Time Again for Public Health Action." *American Journal of Public Health* 92(5):758–768.

Laaksonen, Mikko, Lasse Tarkiainen and Pekka Martikainen. 2009. "Housing Wealth and Mortality: A Register Linkage Study of the Finnish Population." *Social Science & Medicine (1982)* 69(5):754–760.

Lee, James and Filippo Mezzanotti. 2017. "Bank Distress and Manufacturing: Evidence from the Great Depression." p. 47.

Link, B. G. and J. Phelan. 1995. "Social Conditions as Fundamental Causes of Disease." *Journal of Health and Social Behavior* Spec No:80–94.

Lleras-Muney, Adriana, Joseph Price and Dahai Yue. 2020. "The Association Between Educational Attainment and Longevity Using Individual Level Data from the 1940 Census." *NBER Working Paper Serie.* .

Lundberg, Ian, Rebecca Johnson and Brandon Stewart. 2020. What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. Preprint SocArXiv.

Lynch, Scott M., J. Scott Brown and Katherine G. Harmsen. 2003. "Black-White Differences in Mortality Compression and Deceleration and the Mortality Crossover Reconsidered." *Research on Aging* 25(5):456–483.

Manton, K. G., S. S. Poss and S. Wing. 1979. "The Black/White Mortality Crossover: Investigation from the Perspective of the Components of Aging." *The Gerontologist* 19(3):291–300.

Manturuk, Kim, Mark Lindblad and Roberto Quercia. 2010. "Friends and Neighbors: Home-ownership and Social Capital among Low- to Moderate-Income Families." *Journal of Urban Affairs* 32(4):471–488.

Manturuk, Kim R. 2012. "Urban Homeownership and Mental Health: Mediating Effect of Perceived Sense of Control." *City & Community* 11(4):409–430.

Margo, Robert A. 1993. "Employment and Unemployment in the 1930s." *Journal of Economic Perspectives* 7(2):41–59.

Masters, Ryan K. 2012. "Uncrossing the U.S. Black-White Mortality Crossover: The Role of Cohort Forces in Life Course Mortality Risk." *Demography* 49(3):773–796.

Maynard, Charles. 2019. "The Incompleteness of the Social Security Death Master File." *JAMA Cardiology* 4(8):831.

McGue, Matt, Merete Osler and Kaare Christensen. 2010. "Causal Inference and Observational Research: The Utility of Twins." *Perspectives on psychological science : a journal of the Association for Psychological Science* 5(5):546–556.

Meyer, Bruce D. and Nikolas Mittag. 2019. "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net." *American Economic Journal: Applied Economics* 11(2):176–204.

Michney, Todd M. and LaDale Winling. 2020. "New Perspectives on New Deal Housing Policy: Explicating and Mapping HOLC Loans to African Americans." *Journal of Urban History* 46(1):150–180.

Molina, Mario and Filiz Garip. 2019. "Machine Learning for Sociology." *Annual Review of Sociology* 45(1):27–45.

Muller, Christopher and Alexander F. Roehrkasse. 2022. "Racial and Class Inequality in US Incarceration in the Early Twenty-First Century.".

Munford, Luke A., Eleonora Fichera and Matt Sutton. 2020. "Is Owning Your Home Good for Your Health? Evidence from Exogenous Variations in Subsidies in England." *Economics & Human Biology* 39:100903.

Narayanan, Arvind. 2021. "How to Recognize AI Snake Oil.".

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science (New York, N.Y.)* 349(6251):aac4716.

Ostan, Rita, Daniela Monti, Paola Gueresi, Mauro Bussolotto, Claudio Franceschi and Giovannella Baggio. 2016. "Gender, Aging and Longevity in Humans: An Update of an Intriguing/Neglected Scenario Paving the Way to a Gender-Specific Medicine." *Clinical Science* 130(19):1711–1725.

Parnell, Allan M. and Cynthia R. Owens. 1999. "Evaluation of U.S. Mortality Patterns at Old Ages Using the Medicare Enrollment Data Base." *Demographic Research* 1:2.

Peet, Evan D., Brian G. Vegetabile, Matthew Cefalu, Joseph D. Pane and Cheryl L. Damberg. 2022. Machine Learning in Public Policy: The Perils and the Promise of Interpretability. Technical report RAND Corporation.

Perry, Brea L. and Edward W. Morris. 2014. "Suspending Progress: Collateral Consequences of Exclusionary Punishment in Public Schools." *American Sociological Review* 79(6):1067–1087.

Pettit, Becky and Bruce Western. 2004. "Mass Imprisonment and the Life Course: Race and Class Inequality in U.S. Incarceration." *American Sociological Review* 69(2):151–169.

Phang, Sock Yong and Matthias Carl Helble. 2016. "Housing Policies in Singapore." *SSRN Electronic Journal* .

Preston, Samuel H. 1970. "An International Comparison of Excessive Adult Mortality." *Population Studies* 24(1):5–20.

Preston, Samuel H. and Irma T. Elo. 1995. "Are Educational Differentials in Adult Mortality Increasing in the United States?" *Journal of Aging and Health* 7(4):476–496.

Preston, Samuel H. and Irma T. Elo. 2006. "Black Mortality at Very Old Ages in Official US Life Tables: A Skeptical Appraisal." *Population and Development Review* 32(3):557–566.

Preston, Samuel H., Irma T. Elo, Ira Rosenwaike and Mark Hill. 1996. "African-American Mortality at Older Ages: Results of a Matching Study." *Demography* 33(2):193.

Preston, Samuel H., Irma T. Elo and Samuel H. Preston. 1999. "Effects of Age Misreporting on Mortality Estimates at Older Ages." *Population Studies* 53(2):165–177.

Puterman, Eli, Jordan Weiss, Benjamin A. Hives, Alison Gemmill, Deborah Karasek, Wendy Berry Mendes and David H. Rehkopf. 2020. "Predicting Mortality from 57 Economic, Behavioral, Social, and Psychological Factors." *Proceedings of the National Academy of Sciences* 117(28):16273–16282.

Rakoff, Robert M. 1977. "Ideology in Everyday Life:- The Meaning of the House." *Politics & Society* 7(1):85–104.

Riddle, Travis and Stacey Sinclair. 2019. "Racial Disparities in School-Based Disciplinary Actions Are Associated with County-Level Rates of Racial Bias." *Proceedings of the National Academy of Sciences* 116(17):8255–8260.

Rohe, William M. and Leslie S. Stewart. 1996. "Homeownership and Neighborhood Stability." *Housing Policy Debate* 7(1):37–81.

Rohe, William M. and Michael A. Stegman. 1994. "The Impact of Home Ownership on the Social and Political Involvement of Low-Income People." *Urban Affairs Quarterly* 30(1):152–172.

Rolfe, Steve, Lisa Garnham, Jon Godwin, Isobel Anderson, Pete Seaman and Cam Donaldson. 2020. "Housing as a Social Determinant of Health and Wellbeing: Developing an Empirically-Informed Realist Theoretical Framework." *BMC Public Health* 20(1):1138.

Rose, Sherri. 2013. "Mortality Risk Score Prediction in an Elderly Population Using Machine Learning." *American Journal of Epidemiology* 177(5):443–452.

Rossi, Peter H. and Eleanor Weber. 1996. "The Social Benefits of Homeownership: Empirical Evidence from National Surveys." *Housing Policy Debate* 7(1):1–35.

Rothstein, Richard. 2017. *The Color of Law: A Forgotten History of How Our Government Segregated America.* Liveright Publishing.

Ruggles, Steven. 2014. "Big Microdata for Population Research." *Demography* 51(1):287–297.

Ruggles, Steven, Catherine A. Fitch and Evan Roberts. 2018. "Historical Census Record Linkage." *Annual Review of Sociology* 44(1):19–37.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Mathew Sobek. 2020. "IPUMS USA: Version 10.0 [Dataset]." *Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V10.0.* .

Şahin, Duygu Başaran and Frank W. Heiland. 2017. Black-White Mortality Differentials at Old-Age: New Evidence from the National Longitudinal Mortality Study*. In *Applied Demography and Public Health in the 21st Century*, ed. M. Nazrul Hoque, Beverly Pecotte and Mary A. McGehee. Applied Demography Series Cham: Springer International Publishing pp. 141–162.

Salganik, Matthew J., Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts and Sara McLanahan. 2020. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *Proceedings of the National Academy of Sciences* 117(15):8398–8403.

Samuel, Lawrence R. 2012. *The American Dream: A Cultural History.* Syracuse University Press.

Sautter, Jessica M., Patricia A. Thomas, Matthew E. Dupre and Linda K. George. 2012. "Socioeconomic Status and the Black–White Mortality Crossover." *American Journal of Public Health* 102(8):1566–1571.

Schwandt, Hannes, Janet Currie, Marlies Bär, James Banks, Paola Bertoli, Aline Bütikofer, Sarah Cattan, Beatrice Zong-Ying Chao, Claudia Costa, Libertad González, Veronica Grembi, Kristiina Huttunen, René Karadakic, Lucy Kraftman, Sonya Krutikova, Stefano Lombardi, Peter Redler, Carlos Riumallo-Herl, Ana Rodríguez-González, Kjell G. Salvanes, Paula Santana, Josselin Thuilliez, Eddy van Doorslaer, Tom Van Ourti, Joachim K. Winter, Bram Wouterse and Amelie Wuppermann. 2021. "Inequality in Mortality between Black and White Americans by Age, Place, and Cause and in Comparison to Europe, 1990 to 2018." *Proceedings of the National Academy of Sciences* 118(40):e2104684118.

Sharp, Gregory and Matthew Hall. 2014. "Emerging Forms of Racial Inequality in Homeownership Exit, 1968–2009." *Social Problems* 61(3):427–447.

Sibley, Elbridge. 1930. *Differential Mortality in Tennessee, 1917-1928*. Fisk University Press.

Smith, Kirsten P. and Nicholas A. Christakis. 2008. "Social Networks and Health." *Annual Review of Sociology* 34(1):405–429.

Song, Xi and Thomas S Coleman. 2020. "Using Administrative Big Data to Solve Problems in Social Science and Policy Research." *University of Pennsylvania Population Center Working Paper (PSC/PARC), 2020-58.* .

Sugrue, Thomas J. 2014. *The Origins of the Urban Crisis: Race and Inequality in Postwar Detroit - Updated Edition*. Princeton University Press.

Taylor, Keeanga-Yamahtta. 2019. *Race for Profit: How Banks and the Real Estate Industry Undermined Black Homeownership*. UNC Press Books.

Tremblay, Kenneth R. and Don A. Dillman. 1983. *Beyond the American Housing Dream: Accommodation to the 1980s*. University Press of America.

Turner, Tracy and Heather Luea. 2009. "Homeownership, Wealth Accumulation and Income Status." *Journal of Housing Economics* 18(2):104–114.

Van der Laan, Mark J., Eric C. Polley and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6(1).

van Raalte, Alyson A. 2021. "What Have We Learned about Mortality Patterns over the Past 25 Years?" *Population Studies* 75(sup1):105–132.

Vaupel, J. W. and A. I. Yashin. 1985. "Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics." *The American Statistician* 39(3):176–185.

Vaupel, James W., Kenneth G. Manton and Eric Stallard. 1979. "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality." *Demography* 16(3):439–454.

Wandersman, Abraham. 1981. "A Framework of Participation in Community Organizations." *The Journal of Applied Behavioral Science* 17(1):27–58.

Wing, Steve, Kenneth G. Manton, Eric Stallard, Curtis G. Hames and H. A. Tryoler. 1985. "The Black/White Mortality Crossover: Investigation in a Community-Based Study1." *Journal of Gerontology* 40(1):78–84.

Wrigley-Field, Elizabeth. 2014. "Mortality Deceleration and Mortality Selection: Three Unexpected Implications of a Simple Model." *Demography* 51(1):51–71.

Wrigley-Field, Elizabeth. 2020*a*. "Multidimensional Mortality Selection: Why Individual Dimensions of Frailty Don't Act Like Frailty." *Demography* 57(2):747–777.

Wrigley-Field, Elizabeth. 2020*b*. "US Racial Inequality May Be as Deadly as COVID-19." *Proceedings of the National Academy of Sciences* 117(36):21854–21856.

Yao, Li and Stephanie A. Robert. 2011. "Examining the Racial Crossover in Mortality between African American and White Older Adults: A Multilevel Survival Analysis of Race, Individual Socioeconomic Status, and Neighborhood Socioeconomic Context." *Journal of Aging Research* 2011:1–8.

Appendix A

Longevity benefits of homeownership

A.1 Supplemental Analyses

Record Linkage: ABE Conservative

Our main analysis is conducted using a dataset constructed by linking together Census records with the conventional ABE record linkage algorithm. As a robustness check, we refit our main models using the conservative variant of the ABE-exact record linkage algorithm, which restricts to first and last names unique within a ±2 year window (Abramitzky et al., 2020). Approximately 50% of our matches were deemed conservative for both the 1920–1940 linkage and the 1940 to DMF linkage ($N = 39,629$). Our results are qualitatively similar, but the uncertainty errors are larger due to the smaller sample sizes.

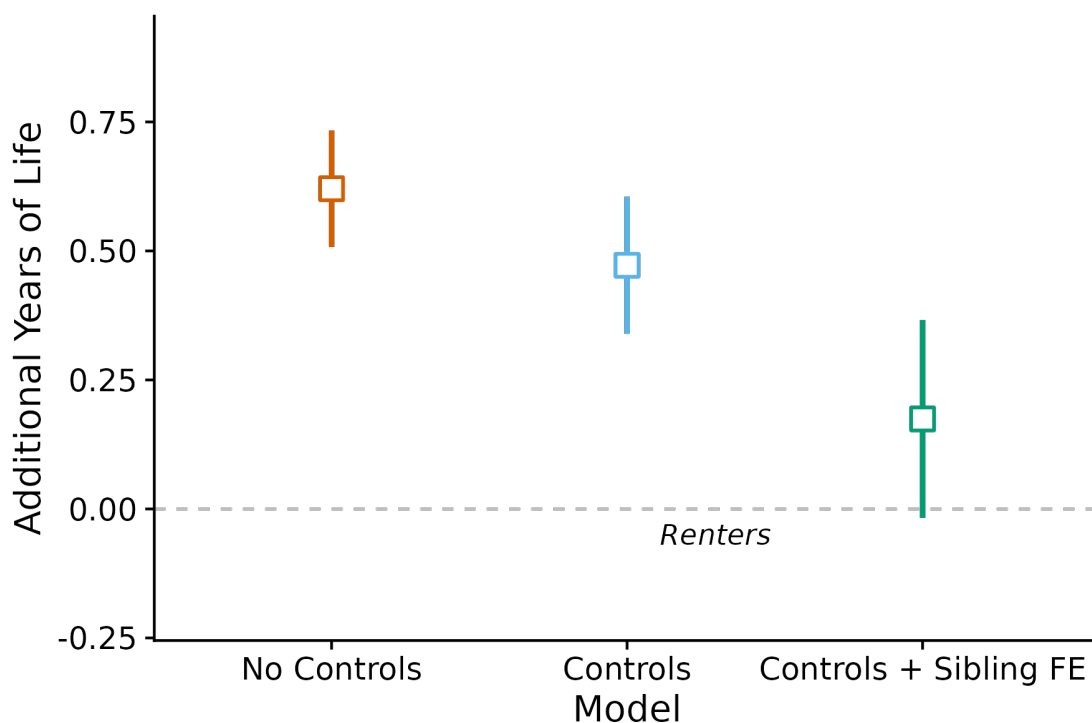Figure A.1: We refit our main models using a subsample of matches established with the ABE-Exact conservative algorithm for both the linkage between censuses and between censuses and mortality records.

Addressing Double-Truncation

In our main analysis, we do not explicitly account for the doubly-truncated nature of our sample. Double truncation occurs because we only observe deaths between the ages of

1975 and 2005, with no measure of survivorship. For example, for the birth cohort of 1910, we only see deaths between the ages of 65 and 95 but no deaths after age 95. To address this, we refit our model using a Gompertz proportional hazard model and maximum likelihood techniques (Goldstein et al., 2023). After fitting a model that explicitly accounts for truncation to our sibling sample, we estimate an $e(65)$ approximately 19% larger than our reported regression coefficient. While we cannot re-fit our full model with sibling fixed effects to due computational limitations (Goldstein et al., 2023), this suggests that we should inflate our estimate of the effect of homeownership on longevity to 0.37.

Regression Tables

As a companion to the coefficient plots presented in the main body of the paper, in this section, we present the full regression coefficient tables for Figure 2.6, Figure 2.7, and Figure 2.9.

Dependent Variable:	Death Age			
Model:	White	White (Controls)	Black	Black (Controls)
Homeowner	0.5255**	0.4228**	0.5033**	0.3580*
	(0.0204)	(0.0220)	(0.1127)	(0.1213)
Education (Years)		0.1460**		0.0541**
		(0.0099)		(0.0087)
Urban		-0.2109**		-0.1003
		(0.0179)		(0.0868)
Fixed-effects				
Birth Year	Yes	Yes	Yes	Yes
Occupation		Yes		Yes
Marital Status		Yes		Yes
State (1940)		Yes		Yes
Observations	1,296,830	1,296,830	65,053	65,053
R^2	0.03857	0.04780	0.04829	0.05542
Within R^2	0.00083	0.00270	0.00043	0.00062

Clustered (byear) standard-errors in parentheses
Signif. Codes: **: 0.01, *: 0.05, †: 0.1

Table A.1: Full regression coefficients for models reported in Figure 2.6.

Dependent Variable:			Death Age	
Model:	No Controls	Controls	Family FE	Family FE + Conrols
Homeowner	0.6366**	0.5328**	0.3132**	0.3135**
	(0.0561)	(0.0605)	(0.0734)	(0.0738)
Education (years)		0.1699**		0.1389**
		(0.0134)		(0.0187)
Race: Other		-2.654**		0.7235
		(0.9509)		(1.364)
Race: White		-0.3457		-1.580
		(0.2813)		(0.9868)
Urban		-0.3713**		-0.0964
		(0.0715)		(0.1080)
Fixed-effects				
Birth Year	Yes	Yes	Yes	Yes
Occupation		Yes		Yes
Marital Status		Yes		Yes
State (1940)		Yes		Yes
Family			Yes	Yes
Birth Order			Yes	Yes
Observations	84,945	84,945	84,945	84,945
R^2	0.03617	0.04956	0.54446	0.54876
Within R^2	0.00129	0.00380	0.00028	0.00135

Clustered (byear) standard-errors in parentheses
*Signif. Codes: **: 0.01, *: 0.05, †: 0.1*

Table A.2: Full regression coefficients for models reported in Figure 2.7.

| Dependent Variable: | | Death Age | |
Model:	No Controls	Controls	Controls + Sibling FE
Home Value Quartile 1	0.3782**	0.4365**	0.2229
	(0.0752)	(0.0664)	(0.1355)
Home Value Quartile 2	0.5284**	0.4544**	0.2245
	(0.0666)	(0.0767)	(0.1510)
Home Value Quartile 3	0.8655**	0.7275**	0.5145**
	(0.0947)	(0.1083)	(0.1595)
Home Value Quartile 4	0.9880**	0.6105**	0.3617*
	(0.1333)	(0.1240)	(0.1818)
Education (Years)		0.1666**	0.1352**
		(0.0134)	(0.0202)
Race: Other		-2.463*	0.8707
		(0.9610)	(1.455)
Race: White		-0.3728	-1.621
		(0.2753)	(0.9901)
Urban		-0.3819**	-0.0894
		(0.0782)	(0.1205)
Fixed-effects			
Birth Year	Yes	Yes	Yes
Occupation		Yes	Yes
Marital Status		Yes	Yes
State (1940)		Yes	Yes
Family			Yes
Birth Order			Yes
Observations	83,990	83,990	83,990
R^2	0.03648	0.04978	0.54927
Within R^2	0.00155	0.00382	0.00138

Clustered (byear) standard-errors in parentheses
*Signif. Codes: **: 0.01, *: 0.05, †: 0.1*

Table A.3: Full regression coefficients for models reported in Figure 2.9.

Appendix B

Unpredictability of individual-level longevity

B.1   CenSoc Numident

For our main analysis, we use the CenSoc-DMF file, which includes deaths from the wider window of 1975-2005. In this supplementary analysis, we use data from the Censoc-Numident, which links the 1940 Census onto mortality records from the Social Security Numident (Goldstein et al., 2021). The CenSoc-Numident only includes deaths for a shorter window, between 1988–2005. However, the CenSoc-Numident file includes women, allowing us to replicate our analysis on a sample more representative of the general population.

Figure B.1: Panel A shows the R^2 performance of each machine learning algorithm. Panel B shows the predicted vs. observed predictions for Superlearner, our best-performing algorithm. Panel C shows the relative variable importance for the top 15 variables in our Superlearner model.

To assess whether our results also extend to women, we replicate our prediction exercise on the CenSoc-Numident. Figure B.1 shows that our results using the CenSoc-Numident dataset are highly comparable to our CenSoc-DMF results: all models have very low predictive accuracy. The best model is only able to explain less than 1% of the variation in age of death. The Superlearner, again, is the top performing algorithm, but its performance is only modestly better than a simple linear model.

B.2   Prediction Results

Figure B.2 shows the full set of predictions (predicted vs. observed) for all 8 machine learning algorithms used in our analysis.

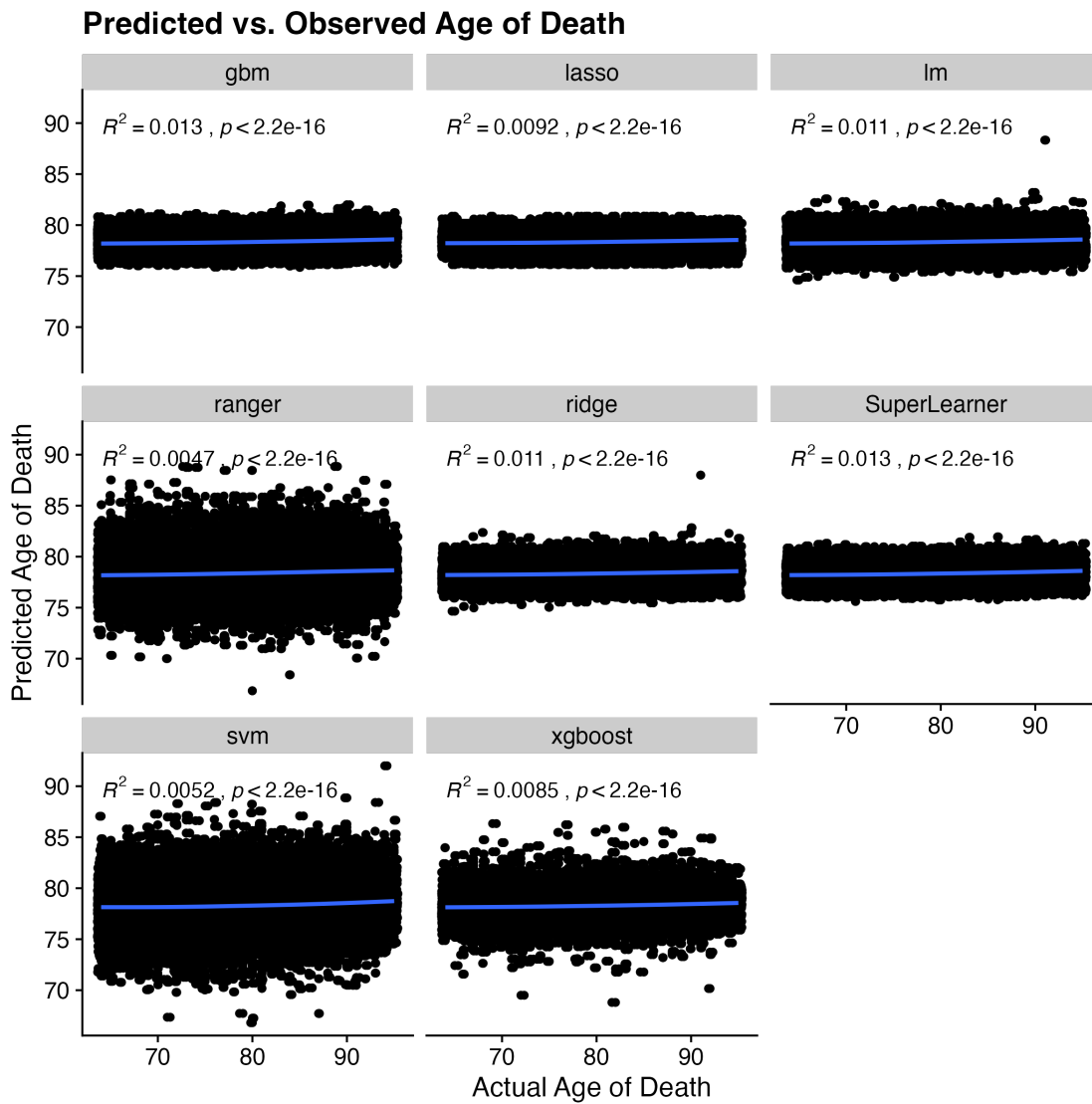**Predicted vs. Observed Age of Death**

Figure B.2: The predicted vs. observed age of death for each machine learning algorithm.