UNIVERSITY OF CALIFORNIA

Los Angeles

Educational Attainment and Hospital Admissions:
New Evidence from the Health and Retirement Study

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Health Policy and Management

by

Dahai Yue

2020

ABSTRACT OF THE DISSERTATION

Educational Attainment and Hospital Admissions:

New Evidence from the Health and Retirement Study

by

Dahai Yue

Doctor of Philosophy in Health Policy and Management

University of California, Los Angeles, 2020

Professor Ninez A. Ponce, Co-Chair

Professor Adriana Lleras-Muney, Co-Chair

Research Objective: Education is one of the most significant correlates of health. However, the extent to which this relationship is causal is yet to be established. Additionally, there is a dearth of studies investigating the effect of education on health care utilization. This dissertation's overall objective was to examine the relationship between educational attainment and hospitalizations using a large longitudinal database and more efficient estimation methods. The three specific aims were: 1) to investigate determinants of attrition due to death and non-response in the Health and Retirement Study (first study); 2) to examine the association between education and hospitalizations based on a pre-set conceptual model and assess the impact of attrition on the estimation of the education-hospitalization relationship (second study); and 3) to determine the causal effect of education on hospitalizations (third study).

Methods: The primary data source was the Health and Retirement Study (HRS) with restricted files, including state-identifiers from 1992 to 2016. This database was further merged with data consisting of 1919-1973 state-level compulsory schooling laws and the quality of schooling measures to study the causal effects of education on hospitalizations. I used a multinomial logistic regression model to investigate the determinants of attrition status in 2016 as well as the between-wave attrition. I then constructed weights to account for attrition bias in the relationship between education and hospitalizations using the inverse probability weighting approach. To determine the causal effects of education on hospitalizations, I used compulsory schooling laws as instruments for years of completed education. A Post-Double-Selection method based on the Least Absolute Shrinkage and Selection Operator (LASSO) regressions was used to select optimal instruments and a parsimonious set of controls, which yields more efficient but still consistent instrumental variable (IV) estimators.

Population Studied: The study population included eligible respondents and their spouses in the HRS survey from 1992 to 2016. The first study excluded the Later Baby Boomer cohort that entered the HRS in 2016. The second study focused on those born in the United States. The third study further restricted the study population to white respondents who had high school or lower educational attainment and were born in the 48 contiguous states and the District of Columbia (excluding Hawaii and Alaska) between 1905 and 1959.

Results: Respondents who were female, white, Hispanic, married, who had more living children, who had more years of education, and who were healthier, and financially better off during childhood were more likely to remain in the survey and respond in every follow-up wave. These variables had different impacts on attrition due to death and attrition due to non-response. On average, compared to individuals with less than a high school education, individuals with a high school education or some college had a 3.37 percentage point (pp) (95% CI, -3.93 pp to -2.80 pp) lower likelihood of being hospitalized, and individuals with a college degree or above had an 8.39 pp (95% CI, -9.10 pp to -7.67 pp) lower likelihood of hospitalization over the past two years, controlling for demographics, childhood socioeconomic conditions, childhood health status, state-of-birth fixed effects, year-of-birth fixed effects, state-specific linear time trends, and accounting for attrition bias. After age 78, the probability of hospitalization for those with a high school education was not significantly different from that of those with less than a high school education; the estimate was -0.96 pp and not statistically significant. The preferred IV estimator (LASSO-IV estimator) implies that a one year increase in schooling lowered the probability of two-year hospitalization by 6.5 pp (95% CI: - 9.1 pp to -3.9 pp), which is much larger than that of the OLS estimator (-1.1 pp, 95% CI: -1.4 pp to -0.7 pp) without correcting for the endogeneity of education.

Conclusions: Individuals with more years of schooling had a lower probability of two-year hospitalizations compared to their counterparts with fewer years of education. These effects would be underestimated if attrition bias was not accounted for. Moreover, age modifies the relationship. After age 78, the effect of a high school or some college education became indistinguishable from zero, but the effect of higher education remained statistically significant. Importantly, when accounting for the endogeneity of education, I found a relatively large and significant effect of education on hospitalizations.

Implications for Research and Policy: My main finding that educational attainment has a large effect on hospitalizations contributes to the growing literature on the social determinants of health. Results from this study should inform policymakers and suggest that providing more health care resources to the low-education group might be an effective means for reducing health disparities. It also provides rigorous evidence for health care payment reforms that consider incorporating education into the risk-adjustment models. In a broader context, it suggests that investing in the educational system could be a more cost-effective way to reduce intensive health care use and health care costs. Furthermore, the analytic framework constructed in this dissertation to account for attrition bias and produce efficient estimators by selecting optimal instruments and controls with LASSO regression models should guide further research for evaluating the effects of education in other similar studies, and, more generally, longitudinal studies involving many instruments and/or many controls.

The dissertation of Dahai Yue is approved.

Susan Louise Ettner

Jack Needleman

Ninez A. Ponce, Committee Co-Chair

Adriana Lleras-Muney, Committee Co-Chair

University of California, Los Angeles

2020

Table of Contents

Table of Figures

Table of Tables

ACKNOWLEDGEMENTS

First and foremost, I want to thank my dissertation committee co-chair Dr. Ninez Ponce. Dr. Ponce has provided me with the wonderful mentorship starting from the first day I came to UCLA. I appreciate all the talks and conversions that guide my career development in academia. Dr. Ponce always paints the big picture of health policy issues for me and encourages me to follow the most rigorous and state of the art knowledge. Specifically, I am grateful to Dr. Ponce's support for my research on social determinants of health and health equity, for my training in the UCLA Department of Economics, and my teaching in the HPM 237C class. I also thank Dr. Ponce's financial and emotional support during my stay in Los Angeles.

I also would like to thank Dr. Adriana Lleras-Muney for willing to serve on my dissertation committee as the co-chair. I deeply appreciate the opportunities to take the Econ 262P class and to work on Dr. Lleras-Muney's projects, which helped me narrow down my dissertation topics and finalize the analytic strategy. I am also grateful to other members of my committee, Dr. Susan Ettner, and Dr. Jack Needleman. They always prioritized time for me to review and discuss my dissertation. They challenged me to think deeply about the research questions, methods, and results, which pushed me to become a better researcher. Their advice and support for my career development in academia have been invaluable!

Many thanks to all my professors in the Department of Health Policy and Management (HPM) and the Department of Economics (ECON), Dr. Thomas Rice, Dr. Emmeline Chuang, Dr. Arturo Bustamante, Dr. Corrina Moucheraud, Dr. John Riley, and Dr. Rosa Matzkin, to name just a few, who have contributed to my professional development over the course of my time at UCLA. Their advice on my career in academia has been incredibly helpful! I also would like to thank the administrative staff in the HPM department, particularly Allison Kamerman and Anna Lim, for their nice support.

I also want to express appreciation to my mentors and colleagues at the UCLA Center for Health Policy Research. I enjoyed my time working with the California Health Interview Survey team. I also deeply appreciated the research opportunities I received from the Health Economics and Evaluation Research (HEER) program led by Dr. Nadereh Pourat. Special thanks to Dr. Xiao Chen for her dedicated mentorship on statistical programming and modeling! Thanks a lot for the support from the IT team. Many thanks to Dr. Ying-Ying Meng for involving me in the Health Effects Institute project. I will miss my time at the center.

I also would like to thank staff members of the US Health and Retirement Study at the University of Michigan for their support of my use of restricted data.

Thank you to my Los Angeles friends, who have made my life easier and happier here. I am grateful to my classmates in both HPM and ECON departments, my teammates on the soccer field, and all my friends here.

Finally, many thanks to my wife, Yuhui Zhu, who has inspired, encouraged, and supported me along the journey. Huge thanks to all my family members that care about me and support me, especially during the hard times.

VITA

EDUCATION

M.S. in Health Economics, Peking University	2015
M.D. in Preventive Medicine, Shandong University	2012

GRANTS

Principal Investigator. “Housing price dynamics and health outcomes in the United States.” Gilbert Foundation Research Grant. UCLA Rosalinde and Arthur Gilbert Program in Real Estate, Finance, and Urban Economics. $4,000. 2018-2019

Principal Investigator. “Racial and Ethnic Disparities in Access to Primary Care after Medicaid Expansions.” UCLA Graduate Summer Research Mentorship Program. $6,000. 2017

SELECTED PUBLICATIONS

Dahai Yue, Yuhui Zhu, Petra W. Rasmussen, James Godwin, and Ninez Ponce. (2020). Coverage, affordability, and care for low-income people with diabetes: four years after the Affordable Care Act’s Medicaid Expansion. Online First. *Journal of General Internal Medicine*.

Dahai Yue, Nadereh Pourat, Xiao Chen, Connie Lu, Weihao Zhou, Marlon Daniel, Hank Hoang, Alek Sripipatana, and Ninez Ponce. (2019). Enabling services improve access to care, preventive care use and satisfaction among health center patients. *Health Affairs*, 38(9):1468-1474.

Siyuan Liang, James Macinko, **Dahai Yue**, Qingyue Meng. (2019). The impact of health human resources on Under-5 mortality in rural China. *Human Resources for Health*, 17(1):21.

Dahai Yue, Petra W. Rasmussen, and Ninez A. Ponce. (2018). Racial/ethnic differential effects of Medicaid expansion on health care access. *Health Services Research*, 53(5):3640-3656. * *Top 20 downloaded article in Health Services Research, 2017-2018*

Dahai Yue, Shiman Ruan, Jin Xu, Weiming Zhu, Luyu Zhang, Gang Cheng, and Qingyue Meng. (2017). Impact of the China healthy cities initiative on urban environment. *Journal of Urban Health,* 94(2):149-157.

Yuhui Zhu, **Dahai Yue**, Beibei Yuan, Lianhua Zhu, and Ming Lu. (2017). Reproductive factors are associated with esophageal cancer risk: results from a meta-analysis of observational studies. *European Journal of Cancer Prevention,* 26(1):1-9.

SELECTED PRESENTATIONS

Oral Presentation

Dahai Yue. “Enabling Services Could Improve Access to Care, Preventive Services, and Satisfaction among Health Center Patients.” APHA's 2019 Annual Meeting and Expo, Philadelphia, United States. November 02 – November 06, 2019.

Dahai Yue. “Racial/Ethnic Differential Effects of Medicaid Expansion on Health Care Access in the United States.” Oral Presentation. The Fifth Global Symposium of Health Systems Research, Liverpool, United Kingdom. October 8 – October 12, 2018.

Dahai Yue, and Petra W. Rasmussen. “Which low-income group did the ACA leave behind.” UCLA Center for Health Policy Research. May 23, 2018.

Poster Presentation

Dahai Yue, Nadereh Pourat, Xiao Chen, Connie Lu, Weihao Zhou, Marlon Daniel, Hank Hoang, Alek Sripipatana, Ninez A. Ponce. Enabling Services Could Improve Access to Care, Preventive Services, and Satisfaction among Health Center Patients. AcademyHealth, ARM, Washington D.C., USA. June 02 – 04, 2019.

Dahai Yue, and Ninez Ponce. "Racial and Ethnic Disparities in Access to Primary Care after Medicaid Expansion." Poster presentation. AcademyHealth ARM, New Orleans, USA. June 24-27,2017.

SELECTED AWARDS, PRIZES, AND FELLOWSHIPS

Delta Omega Honor Society in Public Health	2020
Dean's Outstanding Student Award. UCLA Fielding School of Public Health	2020
PhD Dissertation Datasets and Software funding. UCLA Department of Health Policy and Management	2019
Travel grant for attending the Fifth Global Symposium of Health Systems Research, Liverpool, United Kingdom, by Health System Global	2018
Finalist at the Student Writing Competition for the Annual Breslow Distinguished Lecture at UCLA, Los Angeles, CA, USA	2017

TEACHING EXPERIENCE

UCLA Department of Health Policy and Management

HPM237C Issues in Health Services Methodologies

Teaching Assistant for Prof. Ninez Ponce	2017
Teaching Assistant for Prof. W. Scott Comulada	2018, 2019
Teaching Associate for Prof. W. Scott Comulada	2020

RESEARCH EXPERIENCE

Graduate Student Researcher 2016-2020
UCLA Center for Health Policy Research

Graduate Student Researcher
Division of Pediatrics, UCLA David Geffen School of Medicine 2015-2016

Chapter 1. Introduction

1.1 Background

The uneven distribution of health across the socioeconomic spectrum is one of the most recognized and well-established facts in social science. The striking difference or gradient in health by socioeconomic status does not just represent differences between individuals at the top and the bottom of the spectrum but is continuous across even small changes in social and economic advantages (Ettner 1996, Grossman 2006, Montez, Hummer, and Hayward 2012). A great deal of literature has shown that several social risk factors such as low income, low educational level, minority race/ethnicity, low language proficiency, and limited social capital are associated with both adverse health outcomes and inappropriate health care use (National Academies of Sciences & Medicine, 2017).

Of the various socioeconomic status measures, the gradient in health by years of completed schooling is particularly robust. An extensive literature review (Grossman, 2000, 2006) concludes that education has been demonstrated to be the most important correlate of good health, regardless of health measures (e.g., self-reported health, mortality, morbidity, or physiological indicators). The critical role of education in shaping good health is perhaps because it forms the future skills, income, occupation, and living environment (Ross & Mirowsky, 2010; Ross & Wu, 1995). Moreover, educational attainment is an appealing social factor to investigate, especially in comparison to other social factors, such as income and occupation. First, education measured by the highest degree or years of schooling is commonly available in most surveys. Since it is also easier for individuals to respond, self-reported educational attainment suffers from fewer measurement errors; there is evidence that measurement errors in education only reduce the return to schooling by 10% (Angrist and Krueger 1999). Second, education is typically determined in early adulthood and becomes time-constant during the later stage of the life cycle. As such, it makes causal inferences via quasi-experiment study designs possible by linking an individual's schooling experience to educational policies that were in place during the time they were in school.

There is a substantive number of studies from multiple disciplines looking at the relationship between education and health. In these studies, health is typically measured as mortality, self-reported health status, and health behaviors such as smoking, drinking, and Body Mass Index (BMI). Those studies consistently found that individuals with higher levels of education are less likely to die within five years, more likely to report good health status, and more likely to have healthier behaviors (Cutler and Lleras-Muney 2006, Cutler and Lleras-Muney 2010, Low et al. 2005, Grossman 2006). Recent attention has been paid to estimating the causal effects of education on various health outcomes at a population level using quasi-experimental and econometric methods that exploit geographic variations in policies that lead to differences in years of schooling. The most popular policies examined in the recent decades are compulsory schooling laws (CSLs), legislation that has been passed to establish a minimum number of years of schooling among school-aged children in different countries at different times (Hamad et al. 2018). However, estimates of the education-health relationship from these studies are imprecise and inconclusive, particularly among US studies in which compulsory schooling laws only had a minor impact on years of completed schooling (Mazumder 2008, Fletcher 2015, Lleras-Muney 2002).

In particular, the effects of educational attainment on curative care utilization, such as hospitalizations, have, surprisingly, been largely absent from the literature. Ceteris paribus, individuals having higher educational attainment are generally healthier and need fewer medical services, but they also have more resources like generous health insurance that make them use more. This dissertation focuses on hospital admissions because hospitalizations have low demand elasticities (Manning et al. 1987), and as such, are more likely to reflect changes in health status and corresponding health care use compared to outpatient care. Many empirical studies in health services research include educational attainment as a control variable in the function of hospitalizations, but there are few studies explicitly investigating the relationship. Although there is evidence, albeit with mixed results, from Denmark and Sweden (Arendt 2008, Meghir, Palme, and Simeonova 2013), these two countries have very different health care systems from that of the United States. To date, there is only one US study that included hospitalization as an outcome. The author finds a negative but insignificant effect of education on hospitalizations using compulsory schooling laws as an instrument (Mazumder 2008). However, this study is likely to be limited by a small sample size from the Survey of Income and Program Participation (SIPP). Actually, US studies on the education-health relationship using compulsory schooling laws as instruments are not conclusive and have drawn lots of criticism that these laws are weak instruments, especially when the state-specific trends are added to the model (Fletcher 2015, Black, Hsu, and Taylor 2015).

This dissertation extends existing literature and seeks to explicitly examine the relationship between educational attainment and hospital admissions using the Health and Retirement Study (HRS). HRS is a longitudinal panel study that surveys a representative sample of individuals aged 51 and above and their spouses in the United States. HRS's collection of rich information on an individual's childhood allows us to control for childhood health and socioeconomic status. Moreover, its longitudinal feature allows a much larger number of analytical observations and a wider range of inference (e.g., causal inference based on weak instruments) than is possible with cross-sectional data. Yet, this advantage comes with the challenge of attrition that some respondents dropped out of the survey over time, which would pose a potentially damaging threat to the validity of estimates obtained from panel data.

The challenge arising from attrition motivates the first research question, which explores factors, especially educational attainment, that influence respondents' drop out of the survey. Then, I will examine how educational attainment affects hospitalizations based on a pre-defined conceptual model and adjusting for attrition bias. Finally, the third study aims to estimate the causal effect of education by leveraging compulsory schooling laws as instrumental variables. The identification strategy is more efficient and overcomes the weak instrument problem discussed in previous studies. A detailed description of these research questions is provided below. Hypotheses and rationales associated with each question are described in the Conceptual Model chapter.

1.2 Research Questions

Question 1: Attrition analysis of HRS

What factors affect respondents' drop out of the HRS survey over time?

Question 2: Association between educational attainment and hospital admissions

Is there a negative effect of education on hospitalizations after controlling for childhood socioeconomic conditions and child health?

Question 3: Causal effect of secondary schooling on hospital admissions

Whether secondary schooling is causally linked to hospital admissions?

1.3 Contributions to Literature

First of all, there is a lack of theories and conceptual frameworks that explicitly address the relationship between educational attainment and health care utilization, though there are extensive theories about education and health. A well-established conceptual framework is critical for empirical studies examining the relationship between education and health care utilization. It points out potential confounders that correlate with both education and health care utilization and thus help researchers improve their empirical approaches and avoid spurious relationships. It also lays out potential pathways through which education affects health care utilization. After compiling theories across disciplines, the framework would present a clear picture of how education affects health care use, which should enable policymakers and researchers to propose more effective policy tools to address social determinants of health.

Another contribution of this dissertation to the literature is the attrition analysis of HRS. Although HRS collects rich information on respondents' health care use experience, there is a limited number of studies examining these health care utilization outcomes. The complex study design and attrition bias are very likely to be one of the reasons. Thus, a systematic investigation of the potential selective attrition and related factors would help researchers adjust the attrition bias more appropriately. Moreover, the framework used in this dissertation to adjust for attrition bias and within-individual correlation could create a foundation for further such analyses.

This dissertation is also contributing to the growing literature on employing compulsory schooling laws to uncover the causal effects of education on health. Prior studies using different aspects of compulsory schooling laws have provided uncertain conclusions about the causal effects of education because of imprecise estimates. Most of these studies were subject to weak instrument problems, especially when state-specific linear time trends were added to the model. To overcome these limitations, we need more efficient methods or a larger sample. This dissertation applies the LASSO regression in selecting the optimal instruments and controls within the causal inference framework, which substantially boosts the efficiency of IV estimators while maintaining the consistency. The analytic framework should facilitate the causal inference in the relationship between education and other monetary and nonmonetary outcomes.

This dissertation augments the literature of social determinants of health by adding new evidence on the relationship between education and hospital admissions using US data. Understanding this relationship has become more important, given that the rising importance of health care costs in national budgets and the fact that the health of a population is generally one of the top priorities for policymakers. Thus, results from this dissertation should provide more evidence on whether population-level policies addressing social determinants of health could help reduce the probability of hospitalizations, and thereby, reduce soaring health care costs. This work would have significant implications for state budgets as well since most of those with low socioeconomic status are covered by Medicaid, which is the largest and fastest-growing line item in the budgets of most states.

Lastly, given the recent increased attention paid to reducing health inequity, it is crucial to identify whether those with lower socioeconomic status are more likely to be hospitalized and consume more expensive health care. Results from this dissertation should inform policymakers to consider education policy is health policy. In this way, societies could determine whether public policies intended to increase educational resources and opportunities may help ameliorate health disparities attached to socioeconomic status, or whether alternative strategies are more appropriate, such as investing in health care systems. If the relationship between education and hospitalization is causal and strong, then investment in social services such as education may be a more cost-effective means to improve population health, compared to increased expenditure in the health system.

Chapter 2. Literature Review

2.1 Economic Theories of Education and Health

The relationship between education and health has been explored extensively in the last decades. It started from a discussion about human capital (Becker 1967, Ben-Porath 1967) to distinguishing health capital from human capital (Grossman 1972b), which is referred to as the demand for health model or the Grossman model. Over time, the Grossman model has been extended to several economic models by relaxing its assumptions. For example, the households (Bolin, Jacobson, and Lindgren 2001, 2002c, Jacobson 2000) and employers (Bolin, Jacobson, and Lindgren 2002b) as the producers of health investment were built into models to relax the assumption of the single individual in the Grossman model. Other developments of the Grossman model include decreasing returns to scale and the demand for longevity (Ehrlich and Chuma 1990), depreciation of health as endogenous (Liljas 1998, Muurinen 1982), and uncertainty in health investment (Laporte and Ferguson 2007, Liljas 1998). Moreover, a more recent economic model has considered education as endogenous (Galama, Lleras-Muney, and van Kippersluis 2018). Given this dissertation is focusing on how education influences hospital admissions, I emphasize those theories that are most relevant here.

Grossman proposed the earliest theory explaining the concept of health capital and demand for health (Grossman 1972b), which laid the foundation for health economics. The Grossman model considers health as a durable but nonsalable capital stock that produces an output of healthy time. Individuals inherit an initial stock of health that depreciates with age—after some stage in the life cycle—but can be enhanced by investment. Individuals maximize inter-temporal utility as a function of their stock of health and preferences for consumption of other commodities. Gross investments in health capital are produced by a household production function whose direct inputs include medical care and time inputs. Education is considered an important environmental variable that improves the efficiency of the household production function. This efficiency may come through more effective use of medical care or other activities (e.g., exercise, diet, smoking, substance use) that influence the stock of health.

The Grossman model assumes education is exogenously given and has at least two important implications for health outcomes and medical care use. First, those with more years of education would demand a larger optimal stock of health. The economic explanation for this is that education improves the production efficiency of health capital and shifts the downward sloping demand curve of health capital to the right. Second, those with more years of schooling would demand less medical care; because education raises the marginal product of the direct inputs (medical care and input time) and then reduces the quantity required to produce a certain amount of gross investments—the allocation efficiency.

Recently, the Grossman model has been advanced to study the relationship between human capital, schooling, mortality, and health behaviors (Galama, Lleras-Muney, and van Kippersluis 2018). In the updated model (hereafter referred to as "GLK model"), human capital includes both health and skills (cognitive and non-cognitive skills). Health could be produced by inputs such as medical care, whereas skills can be gained by schooling. Similar to the Grossman model, the GLK model also assumes individuals make optimal decisions concerning skills and health inputs. Decision rules derived from these maximization equations will be guiding consumers' choices in consumption and eventually determine health and mortality. In the GLK model, the depreciation rate of health stock is a function of age, the level of health, consumption, and endowment. Healthy and unhealthy behaviors are included in the model as a part of consumption. GLK model advanced the Grossman model in several aspects. It treats health, skills, health behaviors, schooling, and longevity as endogenous. It distinguishes skills development from time spent in school and includes education laws into the model. Importantly, the GLK model allows us to theoretically predict the potential causal effects of education on health outcomes by leveraging variations generated by educational and child labor laws. There are two worthy noting issues. First, the minimum school-leaving age would have positive effects on schooling through the increased relative marginal value of skill, a higher stock of skill, better health, a longer life, and ambiguous wealth effects and effects through labor laws. However, these laws only influence these marginal individuals or compliers, who are forced to stay at school by the laws beyond its initial optimal drop-out age. Second, the effect of compulsory schooling laws on consumption is ambiguous; compulsory schooling laws would exhibit positive wealth effects among compilers, which enables more unhealthy consumption (e.g., smoking, drinking) but also leads to a higher marginal value of health relative to wealth that decreases unhealthy consumption.

Besides, there are several other economic explanations and empirical tests that unveil the relationship between education and health. One of the most cited ones is that time preference might serve as common causes for education and health (Fuchs 1980). Future-oriented individuals would invest in both schooling and health improvement activities. Time preference might also be endogenous; namely, schooling may cause the rate of time discount for the present to fall (Becker and Mulligan 1997).

2.2 Quantity of Schooling and Health Outcomes

Since Kitagawa and Hauser (1973) first documented the significant differences in health by socioeconomic status in the United States, a large number of studies have been done to demonstrate the “gradient” in health (Deaton and Paxson 2001a). Among these socioeconomic variables, education has received a great deal of attention from researchers across multiple disciplines. The literature providing evidence for a connection between education and health is large, consistent, and robust. Extensive reviews of the literature conclude that education is the most important correlate of good health, regardless of measures for health—mortality, morbidity rates, self-reported health, or physiological indicators of health, and whether units of observation are individuals or groups (Grossman 2000, 2006, Grossman and Kaestner 1997). They also indicate that there is a significant portion of the gross schooling effect that cannot be traced to the relationship between education and income or occupation, since there remains a statistically significant effect of schooling after controlling for income and occupation.

In recent decades, an extremely promising line of research treats schooling as endogenous and estimates the causal effects of schooling on health by using *quasi-experimental designs* (e.g., instrumenting the endogeneity by compulsory schooling laws, child labor laws, and Vietnam war draft). However, this body of literature yields no consensus on the causal relationship between education and health. Some studies reveal significant effects of years of schooling on health while others report little or no significant effect (Galama, Lleras-Muney, and van Kippersluis 2018, Mazumder 2012, Hamad et al. 2018). A recent systematic review and meta-analysis of the effects of education on health using compulsory schooling laws, based on 89 manuscripts, demonstrated small but statistically significant beneficial effects of education on mortality (effect size: -0.01), smoking (effect size: -0.01), and obesity (-0.20) (Hamad et al. 2018). For example, on average, mortality decreases by 1 percentage point per extra additional year of schooling, all else being equal. Since education reforms were typically intended to increase the schooling of those at the lower end of the education distribution, thus, these estimates from instrumental variables only reflect the local treatment effects for the marginal subjects (compliers) and by no means represent the average treatment effects.

A different approach to identify the causal effects that is closer to the average treatment effects is the use of *twin- or sibling- fixed-effect designs*. Results from this approach are not consistent as well. Fujiwara and Kawachi (2009) use US twin data from the National Survey of Mildlife Development and find no evidence of causal effects of education (years of schooling) on a range of health outcomes (e.g., perceived physical health, perceived mental health, and smoking). Similarly, Behrman et al. (2011) find no causal effect of schooling on hospitalization and mortality by exploiting within-twin-pair variation in schooling and health outcomes among Danish Twins. Similar negative and insignificant results are reported in a study using a sample of 741 identical female twin-pairs in the UK (Amin, Behrman, and Spector 2013). Another study based on Australia identical twins report a negative effect of schooling on overweight but only for men. Moreover, a series of papers by Lundborg and colleagues provide further evidence on the favorable effects of education on health. For example, one study documents the schooling's effects on reducing mortality (Lundborg, Lyttkens, and Nystedt 2012) and another study shows that mother's education improves their son's health (Lundborg, Nordin, and Rooth 2018), both based on a Swedish twin registry that includes 9,000 identical twin pairs. Additionally, in contrast to Fujiwara and Kawachi (2009), Lundborg reports significant effects, in the expected directions, of high school completion on self-rated health, chronic conditions, and exercise behaviors among identical twin pairs in the National Survey of Midlife Development, yet no effects on smoking and BMI (Lundborg 2013).

In this subsection, I do not attempt to conduct a complete review of the previous literature, but to highlight research that laid the foundations for current studies and emphasize the evidence from the United States. Also, I only focus on formal schooling, not including pre-school education and other forms of education. In the following paragraphs, I report the results by types of health outcomes. For each health outcome, I review studies documenting an association relationship first and then turn to studies that pursue causal effects, if available.

Education and Mortality.

Early studies have consistently underscored the importance of education as a determinant of mortality (Deaton and Paxson 2001b, Rosen and Taubman 1982). For example, Deaton and Paxson (2001b) find that education is negatively related to mortality for white males based on data linking 1973 Current Population Survey to Social Security and Internal Revenue Services that traces persons through 1977. Rosen and Taubman (1982) find similar results for both men and women using two datasets: the 1996 Current Population Survey merged with 1975-1995 all-cause mortality for the United States and the National Longitudinal Mortality Study. However, there are no systematic gender differences in the relationship between education and mortality based on an analysis using 1986-2000 National Health Interview Surveys linked to the National Death Index through 2002 (Zajacova and Hummer 2009). Furthermore, Montez, Hummer, and

Hayward (2012) systematically investigate the functional form between years of schooling and mortality. This study reveals a linear decline in mortality risk between 0 and 11 years of education followed by a step-change reduction upon attaining a high school diploma, at which point mortality declines linearly but with a much steeper slope. More importantly, in the United States, low education accounts for more estimated deaths than other adverse social factors such as racial segregation, low social support, poverty, and income inequality (Galea et al. 2011).

Although the studies mentioned above about the relationship between education and mortality seem persuasive, their study designs are typically unable to identify causal effects without imposing very strong assumptions. In recent years, researchers have introduced a more compelling research design to disentangle the causal effect of education from other variables that jointly determine education and health, such as childhood environment and time preference. One of the most popular designs is to exploit the plausible exogenous temporal and geographic variations in the quality of schooling and compulsory schooling laws (CSLs)—legislation that has been passed in different states at different times to establish a minimum number of years of educational attainment among school-aged children. Others exploit only differences in years of schooling between siblings or twins to purge out many potential confounding factors that vary within families. Some other studies used the military draft as an exogenous source for educational attainment. I will review these studies in the following paragraphs.

Lleras-Muney (2005) leverages compulsory schooling laws and child labor laws in place from 1915 to 1939 that governed the ages at which children were required to attend school. The identification strategy is based on the variations in these laws for individuals born in different states and at different time points. The instrument is quite plausible and very unlikely to be correlated with unmeasured determinants of education and health, especially since the model controls for state fixed effects, cohort fixed effects, and region-specific trends. The main instrumental variables (IV) results indicate that the effect of education on mortality is about 6%; one additional year of schooling is suggested to reduce 10-year mortality by 6.1 percentage points. However, the instrument might be weak (the F statistics from the joint test of instruments is 4.69), which leads to smaller and statistically insignificant effects after controlling for state-specific linear trends (Mazumder 2008). Actually, there are few variations left in the instruments after adding region/state-specific linear trends to the regression models, as argued by a study combining US Census data with the complete Vital Statistics records for a more precise measure of mortality (Black, Hsu, and Taylor 2015). Also, the CSLs in the US had small effects on the average education of the population; one more year of compulsory schooling resulted in, on average, 0.05 years of additional schooling(Lleras-Muney 2002, 2005).

Does a larger sample size help overcome the weak instrument problem? A study explores this possibility using a large and novel survey from the NIH/AARP Diet and Health Study on several hundred thousand respondents (Fletcher 2015). This study’s results indicate similar estimates to those reported by Lleras-Muney (2005) in both OLS and IV analyses. However, the effect of education on the likelihood of death over a 10-year period, which is 6.9%, is not statistically significant. Therefore, the author concludes the results appear underpowered and suggests the use of this methodology may require a larger and potentially unavailable dataset. One caution about this study is that the birth cohorts in this NIH/AARP survey range between 1925 and 1945 and were thus affected by compulsory schooling laws prevailing between 1933 and 1953 when they were eight years old. However, the effectiveness of these laws as instrumental variables has reduced dramatically after 1940 (Goldin and Katz 2003, Lleras-Muney 2002).

The method of using compulsory schooling laws as instruments for educational attainment has been extended to other countries, especially European countries, with mixed and often null findings. The UK compulsory schooling reforms in 1947(1972) resulted in 0.45(0.35) more years of schooling, which are larger than the effects in the US and other countries. A well-done study leverages these reforms by using a regression discontinuity design (RDD) that compares individuals born right before and right after the cut-off birth dates specified by the law and finds no decrease in mortality (Clark and Royer 2013). However, a recent study using UK Biobank data examines the British 1972 CSL reform and finds that education led to statistically significant declines in mortality (Davies et al. 2016). It is worth noting that the UK Biobank data is made up of young people who volunteered to participate, and the estimates from RDD are also sensitive to the selection of bandwidth and functional form for trends. Mixed results were found in Sweden. Two studies find no significant effect of education on mortality following the Swedish 1949-1962 expansion of compulsory schooling (Lager and Torssander 2012, Meghir, Palme, and Simeonova 2012). However, another study investigated the earlier reforms in 1936 and found larger effects of education on mortality, which is statistically significant at 10 percent (Fischer, Karlsson, and Nilsson 2013).

Similarly, a study exploit a 1928 law in the Netherlands that increased the years of compulsory schooling from 6 years to 7 and find a 2.5 percentage-point reduction in mortality among men, associated with one more year of schooling (Van Kippersluis, O’Donnell, and Van Doorslaer

2011). Similar studies have also been conducted in other European countries, albeit not focusing on mortality, but there is no compelling evidence on the effects of education on health, including those from France (Albouy and Lequien 2009), Germany (Braakmann 2011, Pischke and Von Wachter 2008), and Denmark (Arendt 2005, Arendt 2008). Mazumder (2012) provides a comprehensive review of these studies (Mazumder 2012).

Given the compulsory schooling laws were more likely to affect individuals with lower educational levels, IV estimates from these laws might miss the effects of higher levels of education on mortality. A recent study exploits exogenous variation in years of completed college induced by the risk of being called to serve in the Vietnam war to examine the impact of college on adult mortality (Buckles et al. 2016). The authors find that college education reduces the mortality rate in middle-age by 2.6 percentage points, which is statistically significant.

In summary, there is no consistent evidence on the relationship between education and mortality, although some studies show a beneficial effect of education on reduced mortality. Nonetheless, there are some inherent limitations of these studies that need to be acknowledged. First, for some populations under study, mortality is sufficiently rare that it is difficult to tease out a small impact of education on mortality even with large samples and accurate data. Second, mortality is often measured with imprecision, especially for those purely relying on census data. More importantly, the estimates from those using IV and RDD should be interpreted as the average treatment effects for compliers of the legislation studied. If that group of individuals shares specific characteristics, which is very likely, then it is hard to generalize the results to the overall population. In other words, these effects are local average treatment effects rather than average treatment effects.

Education and Self-Reported Health Status

Many prior studies exploring the education-health relationship use self-rated health as an outcome. Although, as a subjective measure, self-rated health status is strongly predictive of mortality and objective health outcomes (Idler and Kasl 1995), it also has several disadvantages (e.g., limited categories of health, heterogeneous perception of health) that may lead to bias (Strauss and Thomas 1998).

Positive and significant schooling effects on self-reported health status have been consistently reported in the literature. These effects were found for both males and females from the 1987 National Medical Expenditure Survey after controlling for the past stock of health (Gilleskie and Harrison 1998). High levels (beyond high school) of schooling has been shown to have positive and significant effects on self-rated health of middle-aged white males, adjusting for parents' education, health status in the early twenties, and other variables (Grossman 1976). These results of schooling effects on the self-rated health are reinforced by estimates accounting for lagged health and reverse causality—running from health at early stages in the life cycle to years of schooling—based on panel data in other countries (Doorslaer 1987, Wagstaff 1993, Bolin, Jacobson, and Lindgren 2002a, Case, Fertig, and Paxson 2005). Moreover, these results also apply to older males as shown in studies using multiple waves of Retirement History Survey (Sickles and Taubman 1986, Taubman and Rosen 1980), and a study that controls for baseline health, current income, and wealth from Health and Retirement Study (Hurd and Kapteyn 2003). Additionally, many IV studies using compulsory schooling laws as instruments also document consistent results. These studies use several datasets, including 1992 US Health and Retirement (Adams 2002), Survey of Income and Program Participation survey (Mazumder 2008), and the National Institutions of Health (NIH) / American Association of Retired Persons (AARP) Diet and Health Study (Fletcher 2015). However, other studies using similar methods fail to identify significant effects of education on self-reported health. For example, evidence from Romania's schooling expansion found no effect of schooling on self-reported health based on its 2011 Census data (Malamud, Mitrut, and Pop-Eleches 2018).

Education and Health Conditions

Previous literature has consistently documented that individuals with fewer years of education are more likely to have chronic health conditions. Vaugh et al. (2014) show that high school dropouts were more likely to report a serious health condition (e.g., asthma, diabetes, heart disease, high blood pressure) based on a national representative survey—National Survey on Drug Use and Health. Higher educational attainment is also associated with higher levels of "good" (high-density lipoprotein) cholesterol that would decrease the risk of cardiovascular diseases. Similar findings were found in European countries where low education is associated with higher incident events of chronic diseases (Avendano, Jürges, and Mackenbach 2009).

Furthermore, there is also more convincing evidence using causal inference models. Two studies in the United States demonstrated the causal effects of education on self-reported health conditions with mixed results (Fletcher 2015, Mazumder 2008). IV results from the Survey of Income and Program Participation data showed that educational attainment has no significant effects on most of the health conditions with some exceptions (Mazumder 2008). It found that respondents with higher levels of education were less likely to have physical health problems (trouble lifting, walking, climbing stairs, getting around outside the house, getting around inside the house or getting into or out of bed), back or spine problems, stiffness or deformity of a limb, diabetes, and senility/dementia/Alzheimer's disease. The limitation of this study is that health conditions were only asked for those with working disabilities; there could be considerable selection bias. Besides, health outcomes are self-reported, which suffers from selective recall bias as well. Another study with a similar study design uses a large sample NIH/AARP Diet and Health Study and finds that education was related to self-reported health, cardiovascular outcomes, and weight outcomes (Fletcher 2015). The author did not identify significant findings for other health outcomes and argued that these might be underpowered.

In addition, two US studies using compulsory schooling laws as instruments find that education attainment reduced the probability of dementia (Nguyen et al. 2016) and increased old age memory (Glymour et al. 2008). The analytic approach employed is a separate-sample instrumental variable (SSIV) method that uses a different sample for each stage (Angrist and Krueger 1992). The authors used the 1980 US Census 5% sample in the first stage of two-stage least square regression, and the Health and Retirement Study for the second stage.

Education and Health Behaviors

Lengthy literature suggests education is related to health/risk behaviors. A comprehensive study documents a significant education gradient in a wide range of health behaviors, including smoking, diet/exercise, alcohol, illegal drugs, automobile safety, household safety, and preventive care use (Cutler and Lleras-Muney 2010). This study pooled multiple years of National Health Interview Survey as the primary datasets and explored potential mechanisms between education and health behaviors.

Studies based on quasi-experiment study designs, twin studies, and randomized controlled trials have provided more robust evidence. A review (Galama, Lleras-Muney, and van Kippersluis 2018) of these studies concludes that “there is no convincing evidence of an effect of education on obesity (current obese, self-reported), and the effects on smoking (current smoking, self-reported) are only apparent when schooling reforms affect individuals’ track or their peer group, but not when they simply increase the duration of schooling.” Specifically, there are no sizable effects on smoking prevalence by only increasing years of education in England (Clark and Royer 2013, Davies et al. 2016) and Germany (Reinhold and Jürges 2010, Kemptner, Jürges, and Reinhold 2011). However, negative and significant effects were found when students were exposed to completely different schooling; completing higher school (Kenkel, Lillard, and Mathios 2006), college vs high school (Grimard and Parent 2007, De Walque 2007, Heckman, Stixrud, and Urzua 2006), or academic track vs regular track (Jürges, Reinhold, and Salm 2011).

2.3 Quality of Schooling and Health Outcomes

As reviewed in the previous section, the health effects of the quantity of schooling, measured by years of education or the highest degree, have been extensively studied in the previous literature. However, the quantity of schooling alone cannot capture all dimensions of education, especially the quality of education received. Several studies suggest that school quality, measured in different ways, improves salaries, educational attainment, and other outcomes (Card and Krueger 1992, Chetty, Friedman, and Rockoff 2014). In recent years, the relationship between school quality and health has received rising attention.

After the landmark 1954 Supreme Court *Brown C. Board of Education*, school desegregation plans have implemented gradually during the 1960s, 70s, and 80s. These plans led to a substantial increase in school resources and quality, especially in the South of the United States. Several studies have exploited these plausible exogenous variations in school quality to estimate its effect on health, among other outcomes. For example, Johnson (2011) exploits the timing and scope of these implementation plans of school desegregation using an event-study design. His study shows that those plans resulted in increased per-pupil schooling and decreased school segregation and narrowed the black-white gaps in these measures. He finds that, for blacks, school desegregation resulted in significant improvements in adult health, among other economic outcomes. The average effect of a 5-year exposure to court-ordered school desegregation results in an 11 percentage-point increase in reporting excellent/very good health, and desegregation had no health effects on whites.

Another interesting study looked at the improvements in quality of school attended by blacks in 18 segregated southern states that remain racially segregated until the 1960s (Frisvold and Golberstein 2011). During this period, the school desegregations plans should have larger effects on the quality of these schools. The authors measured school quality using three variables compiled by Card and Krueger (1992)—the pupil-teacher ratio, length of the school year, and average teachers' wages. Their findings suggest that school quality amplify the beneficial effects of education on self-reported health, smoking, obesity, and mortality. In other words, their results imply that an additional year of schooling from high-quality schools leads to greater health improvements than one more year of schooling from low-quality schools.

Dudovitz et al. (2016) used a different set of school-level quality measures that predict high school graduation and college attendance. These measures include school average daily attendance, school promotion rate (the percentage of students in each grade who were promoted to the next grade or graduated from high school), parental involvement (percentage of children with family members in a parent-teacher or other parent organization at school), and teacher experience (the percentage of full-time classroom teachers that had worked at the school for 5+ years). The authors analyzed data from 7,037 adolescents from the National Longitudinal Study of Adolescent to Adult Health and found that school quality significantly predicted all health outcomes. Students attending a school with lower average attendance were more likely to report lower self-rated health and have depression symptoms. However, attending schools with higher promotion rates also predicts lower self-rated health. Although the study controlled for baseline health, socio-demographics, and individual academic achievement, the cross-sectional feature of the study design makes it hard to interpret the results as causal effects.

In addition, school quality also matters for health-related behaviors. Dudovitz et al. (2018) exploited a natural experiment of 1270 students who applied to high-performing public charter schools via admission lotteries in low-income minority communities in Los Angeles. The authors use the lottery winning status as instruments and find that lottery “winners” reported less marijuana misuse, fewer substance-using peers, more time studying, less truancy, great teacher support, more orderly schools, and less school mobility, compare to lottery “losers”.

2.4 Education and Health Care Utilization

Many studies have examined the association between education and health service utilization measures. However, the vast majority of previous studies consider education as one of the covariates, rather than build their analytic models explicitly focusing on educational attainment. As such, these estimates are very likely to be biased, and therefore it is not surprising to see inconclusive results.

Since health literacy is one of the critical pathways running from education to health, I briefly summarize the literature on the effects of health literacy on health care utilization. There is a great deal of literature focusing on the relationship between health literacy and health care use. One systematic literature review shows that low health literacy was consistently associated with more hospitalizations, greater use of emergency care, lower receipt of mammography screening and influenza vaccine, and poorer ability to demonstrate taking medications appropriately (Berkman et al. 2011). A study using data from 2006-2008 Medical Expenditure Panel Survey reveals that low health literacy is associated with greater health care utilization, higher health expenditure, and more spending on prescriptions (Rasu et al. 2015). However, another study used a 1963 nationally representative United States survey found a positive but insignificant association between schooling and personal medical expenditure on doctors, dentists, hospital care, prescribed and non-prescribed drugs, nonmedical practitioners, and medical appliances (Grossman 1972a). Since the author lacked information on health insurance and only had very limited medical care measures, results from this study might be suspicious. Another study in the Netherlands with a much better measure of medical care and controlling for variations in the price of a physician visit found a negative and significant effect of schooling on the number of physician visits in the past eight months (Wagstaff 1986). A negative and significant effect of schooling on visits to general practitioners was reported in German as well (Erbsland, Ried, and Ulrich 1995). Evidence from these cross-sectional survey studies only indicates an association, not causation.

Some recent studies that explicitly examine the causal effects of education on health include measures of health care utilization. One study based on Survey of Income and Program Participation data in the US used compulsory schooling laws as instruments shows a significant negative relationship between education and *hospitalizations* from the Ordinary Least Square (OLS) regression model (Mazumder 2008). However, the study does not detect significant causal effects of education on hospitalizations, the number of times hospitalized, and the number of nights in a hospital. The author explains that these outcomes tend to occur later in respondents' lives, and any apparent remaining health effects from years of schooling become harder to detect. Moreover, these self-reported outcomes are subjective and might suffer from measurement errors.

In addition, there are two studies from Denmark focusing on *hospitalizations*. Arendt (2008) leverages the change in the urban-rural differences in education due to the 1958 school reform in Denmark. The author obtained hospitalization variables from the Danish National Register of Patients and other socioeconomic variables from Statistics Denmark. In contrast to prior studies, this study used a larger dataset and a more efficient estimation strategy (a Probit model with continuous endogenous regressors). Results from the bivariate model show that having a degree higher than primary schooling significantly reduces the likelihood of being hospitalized by 1.9

percentage points for women. However, for men, the effect on overall hospitalizations is not significant, whereas it reduces the probability of hospitalizations due to five life-style diagnoses by 0.7 percentage points. Besides, Behrman, et al. (2011) examines the education-hospitalization relationship using data from the Danish Twin Registry, which is linked to population-based registries that comprise 2,500 identical twin pairs. The richness of the data permits the authors to estimate within Monozygotic (MZ) effects of schooling on hospitalizations, which further controls for the individual specific endowment of the twins. They found a strong and significant negative relationship between schooling and hospitalizations (-9.8 hospital days), but generally no causal effect (1.0 hospital days). The results are robust regardless of functional forms of schooling and how hospitalization is measured (e.g., number of hospital days per year, or number of hospital days per year up to 2 years before death or end of the observation period). The authors conclude that schooling seems to be a primary proxy for parental family background or individual-specific endowments.

In terms of *preventive care use*, there is a consistent conclusion in the literature that people with higher educational levels are more likely to use more preventive services. For instance, Kenkel (1994) finds that schooling is a significant predictor for breast cancer test and pap test, and individuals with more educational attainment are more likely to use preventive care. Another study also demonstrates that higher education is associated with a higher probability of preventive care use, including pap tests, blood pressure screening, mammograms, and cholesterol screening (Sambamoorthi and McAlpine 2003). Additionally, Cutler and Lleras-Muney (2010) show that more educational attainment is associated with women getting mammograms and pap smears more regularly, and is associated with men and women are getting colorectal screening and flu shots . However, these studies only focus on establishing correlations and left several important confounding factors uncontrolled, such as the family background. To address this issue, Fletcher and Frisvold (2009) estimate within-sibling fixed effects of education on four preventive care use measures: physical examinations, dental examinations, flu shots, and cholesterol tests. The author shows that an additional year of schooling increases the likelihood of receiving a physical exam by 1.1 percentage points, a dental exam by 1.3 percentage points, a flu shot by 1.7 percentage points, a cholesterol test by 1.1 percentage points. Inspecting the mechanisms suggests occupation and access to care as potential channels.

Several studies document a positive relationship between lower education and a higher likelihood of readmission. For example, one study follows a cohort of 1,351 patients from the Medicare Current Beneficiary Survey and Medicare claims from 2001 to 2002 (Arbaje et al. 2008). The authors show that having limited education (less than high school) is correlated with higher odds of 60-day readmission (OR=1.42, 95% CI = 1.01 – 2.02). Another study includes 577 patients and shows that less than high school education is related to higher odds of 30-day readmission (OR = 2.0, 95% CI=1.1 – 3.4) (Jasti et al. 2008). However, education is not a significant predictor of 30-day readmission in a study based on 951 low-income community-dwelling older adults (Iloabuchi et al. 2014). All these studies include education as a control variable instead of the causal variable of interest. There are good reasons to believe these estimates only indicate associations since there are many confounding variables that affect both education and readmission (e.g., time preference, and family background) left uncontrolled in the model.

2.5 Mechanisms of the Relationship Between Education and Health

The mechanisms through which education affects health and health care have been extensively explored in the literature. However, explicit mechanisms remain unclear. In a broad sense, the relationship between education and health can be explained in one of three ways that are not mutually exclusive.

First, more years of schooling causes better health. Two hypotheses have been proposed for this pathway: education enhances household productive and allocative efficiency in health production. The productive efficiency model suggests that those with more educational attainment can produce more health output from a given set of medical care and other inputs (Grossman 1972a). Lots of empirical studies reviewed in the previous sections support this hypothesis.

The allocative efficiency model indicates that those with more years of schooling are assumed to pick a better input mix to produce health than those with fewer years of schooling (Rosenzweig and Schultz 1982). For example, those with more education are more likely to use new drugs, and education only matters for those who repeatedly purchase medications given a condition (Lleras-Muney and Lichtenberg 2005). Another explanation for allocative efficiency is that schooling improves individuals' knowledge of the relationship between lifestyle and health outcomes. Allocative efficiency only accounts for parts of the education-health relationship, as a study reveals that most of the schooling's effects on health remain after controlling for a direct measure of health knowledge (Kenkel 1991).

The second explanation is that there is a causal relationship that runs from early life health to education. For instance, an individual's childhood health status affects health investments in later life (Ehrlich and Chuma 1990). Also, children who are sick or malnourished are more likely to have more missed days of school, have lower academic performance, and complete fewer years of schooling (Case, Fertig, and Paxson 2005). Recent evidence with twin fixed-effect designs suggests a negative effect of low birth weight on education (Behrman and Rosenzweig 2004, Black, Devereux, and Salvanes 2007).

Third, the correlation is caused by one or more unobserved "third variables" that affect education and health in the same direction, such as parental characteristics and genetics. In a seminal paper, Fuchs identifies time preference as a potential key third variable (Fuchs 1980). Individuals who have a high degree of time preference for future benefits are more likely to spend more time at school and make larger investments in health. If one fails to control for time preference, the effects of schooling on health outcomes are biased. Since factors such as time preferences and genetic traits are often unobserved, this creates a standard omitted variable problem.

In this section, I briefly go over some essential pathways examined in recent studies.

Lifestyle. Education is related to health behaviors, and health behaviors account for nearly half of all deaths in the United States (Mokdad et al. 2004). Lifestyle (e.g., diet, exercise, sleep, smoking, and other health behaviors) could explain part of the relationship between education and health outcomes. An empirical study (Leigh 1983) confirms that most of the effect of schooling on self-rated health could be explained by cigarette, smoking, exercise, and the choice of less hazardous occupations by those with more schooling.

Socioeconomic positions. Compared to those with fewer years of education, those with more education are more likely to have higher income, higher-rank occupation, and other resources. These resources allow individuals to purchase more consumption goods (e.g., health insurance, healthy food, gym membership) to improve health status. Based on the Grossman model, increased wage and wealth also incentivize consumers to demand more health capital. The mediation effects of socioeconomic positions between education and health behaviors have been thoroughly explored (Cutler and Lleras-Muney 2010). The authors reported that income, health insurance, and family background accounted for 30% of educational gradients in healthy behaviors. Many other empirical studies also support the relationship between socioeconomic positions and health (Ettner 1996, Winkleby et al. 1992, Sapolsky 2005). Since there are diverse pathways from different dimensions of socioeconomic status (e.g., education, financial resources, rank, and race/ethnicity) to health, it might be inappropriate to consider socioeconomic status as a unified concept (Cutler, Lleras-Muney, and Vogl 2008).

Knowledge and cognitive skills. One goal of education is to improve individual knowledge and their ability to learn. One study shows that knowledge and cognitive skills account for a 30% education gradient in health behavior (Cutler and Lleras-Muney 2010). Another study also finds that health knowledge could explain part of the relationship between schooling and health behaviors but not all effects (Kenkel 1991). Individuals with more years of education are also more responsive to new information and technology. For instance, empirical studies show that one decade after the diffusion of harmful effects of smoking, cigarette smoking initiation and participation rates fell more rapidly, and quit rates rose more rapidly among the people with higher educational levels between the middle 1960s and the 1970s (De Walque 2004, Sander 1995a, b). Individuals with more years of education also have a higher survival rate in diseases with more health-related technological progress (Glied and Lleras-Muney 2008) and a higher likelihood to use new drugs if they repeatedly buy drugs for a given condition (Lleras-Muney and Lichtenberg 2005). Besides, one study about diabetes treatment found that the Wechsler Adult Intelligence Score (WAIS), a measure of higher reasoning, fully captures the education effects on a good treatment regime (Goldman and Smith 2002).

The sense of control. Education could improve the sense of control that is positively related to self-efficacy and future orientation (Hammond 2002). Effects of schooling on physical functions are found to be significantly reduced when a measure of sense of control is included in a 1995 nationally representative US sample of adults (Ross and Mirowsky 1999). Patients with more educational attainment may control their chronic conditions better and more closely adhere to treatment regimens, for instance, for human immunodeficiency virus (HIV) infection and diabetes (Goldman and Smith 2002).

Time preference. Time preference was first proposed as an unobserved variable that leads to the correlation between education and health (Fuchs 1980). Given the fact that education may change individuals’ time preference (or tastes), time preference could also serve as a mediator in the relationship between education and health. However, the importance of time preference as an explanation is not conclusive among empirical studies. Studies based on Panel Study of Income

Dynamics show that, after adding proxies for time preference, the schooling effects slightly reduced but remain significant (Leigh 1985). In contrast, schooling effects on mortality become insignificant after including proxies for time preference in another study using the 1992 Health and Retirement survey baseline data (Ippolito 2002).

In addition, there is another study dedicated to examining the role of time preference in the relationship between education and health (Van Der Pol 2011). The author uses the Dutch DNB Household Survey, which includes a question for time preference. Respondents were asked how much they would be willing to give up today in order to get a certain amount of money next year. The author reports that the effects of schooling on self-rated health fall between 7 and 14 percent once time preference is held constant. However, the inclusion of time preference had no impact on the effects of education on smoking, BMI, and obesity.

Others. Besides those pathways mentioned above, Culter and Lleras-Muney tested other possible economic theories empirically for health outcomes (Cutler and Lleras-Muney 2006) and behaviors (Cutler and Lleras-Muney 2006, Cutler and Lleras-Muney 2010). They explored the possible mechanisms through which education affects health outcomes and behaviors, including family background, access to health care, labor market, and social networks. They conclude that differences in information and cognition and financial resources explained most of the education gradients in health behaviors.

In summary, previous studies have examined the relationship between educational attainment on a wide range of health outcomes and health behaviors. In general, there is a well-established education-health relationship. However, studies on the relationship between health care utilization are very limited. More importantly, the causal relationship between education and health care is unclear, which should be of great implications on health policy and services research. Although changes in compulsory schooling laws in the US provide a natural experiment for conducting such studies, prior analyses are plagued by small sample size and/or inefficient estimation methods. This dissertation will fill these gaps and provide more rigorous evidence on the relationship between educational attainment and health care utilization.

Chapter 3. Conceptual Model

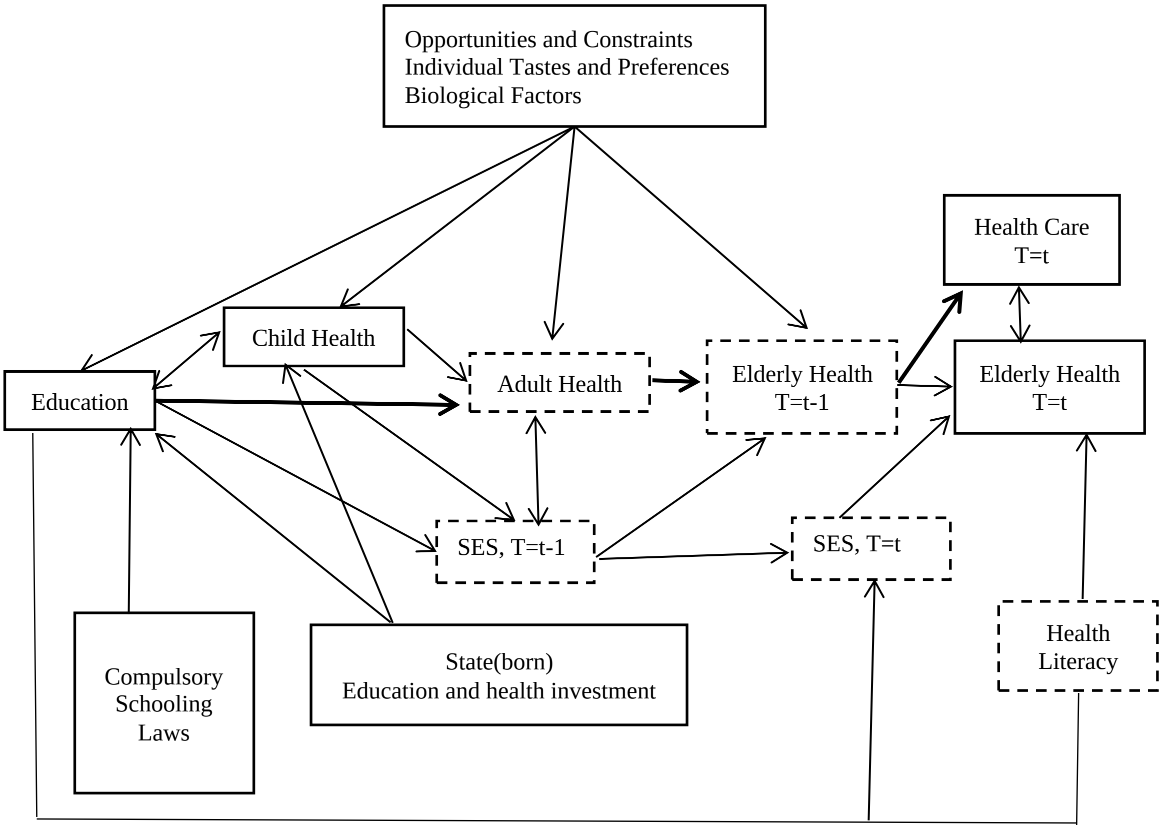
In this chapter, I shall describe the conceptual framework for this dissertation. Then, I will discuss research hypotheses and rationales.

The conceptual model is used to understand the relationship between educational attainment and health care utilization. The purpose of this conceptual model is to inform econometric model specifications that could uncover the causal relationship between education and medical care use. It lays out variables that drive both education and health outcomes over time (confounders), potential pathways through which education affects health (mediators), and possible exogenous sources that induce more years of schooling (instruments). The conceptual model also attempts to incorporate time components to reflect the dynamic relationship between socioeconomic factors, health, and health care utilization at different time points.

Based on the conceptual model, I then propose the hypotheses that will be examined in this dissertation, along with the rationales of hypothesized directions based on the conceptual framework and relevant theories.

3.1 Description of the Conceptual Model

The text below describes the conceptual model in the following order. It first describes the outcome concept for elderly health and health care at the top right of **Figure 3-1-1**. Then, starting from the left of the figure, it will explain the concept of education, which is the primary variable of interest. Continuing to the right-hand side of the figure will illustrate confounding variables at both the individual level (the top of the figure) and the state level (the bottom of the figure). It will then briefly walk through the primary pathways from education to health care in the central line of **Figure 3-1-1**. Lastly, it will introduce the concept for compulsory schooling laws at the bottom left corner of the figure and explain why it could be used as an instrumental variable for educational attainment.



Notes: The dashed boxes indicate concepts not investigated in this study.
Figure 3-1-1. Conceptual model of educational attainment and health care utilization

3.1.1 Outcome: Elderly Health and Health Care

Health Care.

Health care is depicted at the top right of the conceptual model. The concept could capture both preventive care and curative care individuals received from health care providers. However, this dissertation only focuses on curative medical services such as hospitalizations. The arrow runs from elderly health in the previous period to health care at the current period, indicating that the demand for health care is a function of the prior stock of health. For instance, individuals with poor health status have a higher need for medical care and will use more medical services if all else is equal. The double arrow between health care at time t and elderly health at time t reflects the interplay of health care and health status, especially for chronic diseases such as hypertension and diabetes. It is hard to specify the direction of this double arrow. Typically, the use of health care should improve health status and then, in turn, make individuals consume less health care. However, on the other hand, some treatment procedures like chemotherapy may significantly lower cancer patients' white blood cell count, increase the risk of infections, and ead to worse health status and more medical care use.

Elderly Health.

The concept of health status reflects the overall wellbeing, including physical, mental, and social wellbeing. Health status is highly related to health care utilization. In this conceptual model, elderly health status is a function of the previous stock of health, current socioeconomic positions, health literacy, and other factors. It is a key determinant of health care utilization—individuals consume medical services because they either need these services out of health concerns or they want to improve health status.

3.1.2 Primary explanatory variable: Education

Education.

The concept of education in this model only reflects formal years of schooling respondents received from the school. As such, it does not capture informal schooling outside of school or education at home. Since education is the primary explanatory variable in this dissertation, the conceptual model is built up to unveil its effects on health outcomes. Given educational attainment is determined in early life, individuals' childhood environment (e.g., early-life socioeconomic status and living environment), child health, and individuals' tastes and preferences, play a critical role in their educational levels. These impacts are reflected by the arrows pointed to education. Besides, these individual-level factors also have significant implications for health status. Thus, if these factors were left uncontrolled in the model, the relationship between education and health would probably be spurious.

Moreover, education shapes individuals' later life socioeconomic positions (annotated as SES in **Figure 3-1-1**) and health status (Adult Health and Elderly Health in **Figure 3-1-1**), which serves as pathways that education has impacts on health and health care utilization in later life. For example, people with higher educational attainment are more likely to obtain a more prestigious job with fewer hazardous risks for health, to have higher income, to acquire up-to-date information in health, to have more generous health insurance and better access to care, and thereby have a better health status, compared to their counterparts with less educational attainment. Nonetheless, this dissertation estimates a partial reduced-form model, which means it does not explicitly investigate these pathways.

Individual-level educational attainment is also affected by the supply of educational resources and educational policies, depicted as “State(born) Education and health investment” in the model. Previous studies have demonstrated that investments in schools such as raising teachers’ salaries or school desegregation in the 20th century have improved the average educational level (Card, Domnisoru, and Taylor 2018, Frisvold and Golberstein 2011). Studies have also shown that educational policies and laws (e.g., compulsory schooling laws, and child labor laws) have resulted in modest increases in the population’s average educational attainment (Lleras-Muney 2002).

3.1.3 Confounders

Opportunities and Constraints

Race and Ethnicity. Race/ethnicity has been historically influencing individuals’ education opportunities, especially for the blacks in the early 1900s. Schools in the South remained racially segregated until the mid-1960s, and blacks have much lower educational levels than whites (Ashenfelter, Collins, and Yoon 2006, Collins and Margo 2006). Even after the school desegregation required by the *Brown v. Board of Education*, racial/ethnic disparities in educational levels still exist with blacks having lower educational attainment than whites (Brittain and Kozlak 2007). Besides, race and ethnicity are also considered as key determinants of health due to their enduring and strong association with social and economic opportunities (Acevedo-Garcia et al. 2008, Beal 2004, Mehta, Lee, and Ylitalo 2013). Black children had the most reported health conditions, and Asian children had the lowest. More importantly, the racial/ethnic disparities in child health have barely changed over time (Mehta, Lee, and Ylitalo 2013).

Childhood Environment. Childhood environment is a broad concept that captures the socioeconomic positions (e.g., household income) and living environment (e.g., neighborhood, parental characteristics, etc.) during individuals' childhood when they were in school.

Childhood environment, typically proxied by parental educational levels in many empirical studies, plays a critical role in determining individual education levels. For instance, a one-year increase in the years of schooling of either parent reduces the likelihood that a child repeats a grade by 2-4 percentage points (Oreopoulos, Page, and Stevens 2006). Importantly, prior studies have identified child health as a potential mechanism via which intergenerational transmission of economic status takes place (Case, Lubotsky, and Paxson 2002, Case, Fertig, and Paxson 2005, Currie 2009). Those studies suggest that children from poorer households experienced poorer childhood health, lower investments in human capital (lower educational attainment), and poorer health status in adulthood.

Grow up in a family with distinct values, cultures, and background has tremendous influences on individuals concerning both academic performance and health outcomes. As an example, in the United States, race and ethnicity have been playing a significant role in shaping these outcomes. Racial/ethnic disparities in education and health status have existed for years. Black kids have historically lower educational attainment than Whites, even though the school desegregation in the 20th century resulted in significant improvements in adult attainments for blacks (Johnson 2011). In terms of health outcomes, racial/ethnic minorities have lower health status and more barriers in navigating the health care system (Fiscella et al. 2000).

As such, in **Figure 3-1-1**, arrows run from opportunities and constraints to education, child health, adult health, and elderly health.

Individual Tastes and Preferences

The concept of individuals' tastes and preferences describe an individual's predisposition to gain educational attainment and seek care for improving health status.

A commonly cited example in the economic literature is the time preference. Future-oriented individuals would be more likely to invest in schooling, such as staying in school longer (instead of jumping to the labor market too early) and perform better than their short-oriented counterparts. Meanwhile, they are also likely to do more health-enhancing improvement activities and less risky health behaviors (e.g., smoking, substance use, dangerous driving). Time preference could be shaped by certain cultures within a group or household. It is hard to measure and typically unavailable in the current dataset.

Biological Factors

Other than individual tastes and preferences, there are also other biological and genetic factors affecting both educational attainment and health, such as gender and intelligence. For example, individuals' personality traits could drive both educational attainment and health status in the same direction. For instance, psychometric intelligence is highly correlated with educational achievement, with a correlation coefficient equals to 0.81 (Deary et al. 2007). There is an extensive literature in epidemiology that shows that childhood intelligence (measured by an IQ-type test) predicts pronounced differences in adult morbidity and mortality, even controlling for socioeconomic variables (Gottfredson and Deary 2004). Also, gender is an important determinant of individual preferences. It has been found that boys are more likely than girls to delay entry into kindergarten, repeat a grade during their time in elementary school, and perform significantly worse in school. A study shows substantial sex differences in educational attainment with women outpacing men in every income group over the past seventy years (Bailey and Dynarski 2011). Boys and girls have different biological determinants of health. Studies have revealed that gender has significant effects on both the determinants and consequences of health and illnesses (Vlassoff 2007).

Childhood Health

The concept of childhood health, refers to childhood health status, typically enters the conceptual model as an argument for an inverse causality relationship between education and health.

On the one hand, childhood health status has a pronounced impact on educational attainment. For example, controlling for parental income, education, and social class, children who have experienced poorer health in childhood have significantly lower educational attainment (Case, Fertig, and Paxson 2005). On the other hand, besides learning knowledge from school, attending school itself also has an impact on children's health status. However, the direction of this effect is uncertain. It is probably positive. There is evidence that the receipt of free and reduced lunches through the National School Lunch Program improves the health outcomes of children (Gundersen, Kreider, and Pepper 2012). However, it is also likely to be negative. One typical example is school bullying. A large number of studies have documented that bullying experience (bullying and being bullies) at school was related to risky health behaviors (e.g., weapon carrying, physical fighting) and poorer psychosocial adjustment (Nansel et al. 2003, Nansel et al. 2001).

In the conceptual model, the double arrow between education and child health reflects the aforementioned reverse causality concern. Two arrows are running from childhood health to adult health and adult SES, which indicate the negative effects of poor childhood health on adult health status and socioeconomic position, respectively. Besides, those arrows to childhood health represent that child health is also a function of childhood environment, individual tastes and preferences, and state health investments.

State of Birth Education and Health Investment

The concept of the state of birth education and health investment captures the supply side of state investment that improves both schooling and childhood health outcomes. Education could be served as a pathway via which these investments in education affect health. In this conceptual model, this concept depicts those that directly affect both education and childhood health. In other words, they capture the endogenous feature of the state-level investments that drive both schooling and childhood health outcomes.

Here, I use the example of student breakfast/lunch programs to illustrate this concept. Those programs are important for kids from low-income families to get sufficient nutrition for their health. However, they could also attract kids to come to school, miss a few days of schooling, and/or stay in school longer. As those programs drive education and health in the same direction, they would create a spurious relationship between education and health if left uncontrolled in the model.

For other aspects of investments, such as the number of schools and school infrastructure salary, it is hard to tell whether they fit in this concept or not and depends on the investment decisions. If those investments were driven by some other exogenous policies/events, then they do. If those investments were driven by, for example, lower education levels of local residents, they could not fit in this concept.

3.1.4 Mediators

Adult Health.

The concept of adult health represents the overall well-being of individuals in adulthood, including physical, mental, and social well-being. It serves as a critical pathway running from education to health and health care utilization later in life. Health is accumulated over the life course, and adult health plays an integral part in this process. Many chronic diseases occur during adulthood, primarily due to unhealthy lifestyle or gene-environment interaction effects. Advances in medicine have made it possible to delay the development of a disease, but it is difficult to cure these diseases such as diabetes. Thus, it is very likely individuals will carry those diseases with some complications into later life, which increases their demand for health care utilization.

In the conceptual model, I use a dashed box for the concept of adult health to indicate that it is not the primary focus of this study. Although it is an important mediator, I do not attempt to estimate the extent to which the effect of education on later-life health care utilization is mediated by adult health. As reflected by multiple arrows pointing to the concept of adult health, it is a function of several variables that include childhood health, socioeconomic positions, and individual preferences. Notably, education could have a direct effect on adult health through channels like health literacy as well.

SES: Socioeconomic Status

The concept of SES here captures individuals' modifiable socioeconomic status variables and socioeconomic positions that have likely been shaped by education. As education allows individuals to learn new skills and knowledge, it serves as one of the most influential factors. It affects an individual's labor market outcomes. People with higher educational levels are more likely to have a high-salary job, have more occupational choices. Individuals with higher wealth and social standing are less likely to suffer from health conditions that result in hospitalization. However, given the same health conditions or the need for hospital admissions, those with more years of education have a higher chance of being hospitalized because of their social advantages such as higher income, more information, more generous health insurance, or better connections. Nonetheless, income or occupational choice explain only a part of the education effect (Cutler and Lleras-Muney 2006).

Health Literacy

Health literacy in the conceptual model reflects the direct effect of education on health. The US Institute of Medicine report (Institute of Medicine 2004) defines health literacy as "the degree to which individuals have the capacity to obtain, process and understand basic health information and services needed to make appropriate health decisions." Those with limited health literacy might lack sufficient health information and unable to navigate the complex health system effectively. Evidence on the linkage between limited health literacy and poor health are accumulating, though the causal relationship is unknown. A systematic review shows that poor health literacy is associated with more hospitalizations, more utilization of emergency care, and less use of health and screening programs (Berkman et al. 2011).

This definition of health literacy implies health literacy is knowledge based, and thus may be developed through educational intervention (Nutbeam 2008). In the United States, most elementary, middle, and high schools require health education as part of the curriculum, mostly based on the National Health Education Standards (Institute of Medicine 2004). As such, more years of formal schooling should be linked to a higher level of health literacy. There is plenty of evidence that higher educational attainment is related a higher health literacy, and health literacy mediates the relationship between educational attainment and health (Friis et al. 2016, Jansen et al. 2018).

The concept of health literacy, as denoted by the dashed box, is not included in this study, since the total effect of education on health is the main interest.

3.1.5 Compulsory Schooling Laws

The concept of compulsory schooling laws in the conceptual model represents those plausible exogenous laws that have impacts on individuals' health status but no direct effects on their health status. The purpose of this exogenous concept is to identify the causal effect of educational attainment on health care utilization.

3.2 Study Hypotheses and Rationales

Questions 1 What factors affect respondents' drop out of the HRS survey over time?

Hypothesis: Individuals with lower socioeconomic status and worse health status are more likely to drop out of the survey.

Rationales: Since the main reason for respondents to drop out of the follow-up surveys is death, factors associated with health status should be considered as correlates for attrition. There is a substantial number of evidence that those who have worse self-reported health and chronic diseases are more likely to die early. Similarly, socioeconomic status, such as income, is also significantly related to lifespan. Individuals with less income and wealth have a higher mortality rate compared to their wealthier counterparts. As such, respondents who have worse health status or lower socioeconomic status are more likely to attrite due to death.

Also, if we consider participation in the HRS survey is a signal of caring about personal health, then we can reach the same conclusion that those with higher socioeconomic status are more likely to remain in the sample. However, it is hard to tell whether individuals with worse health status are more or less likely to drop out of the survey due to non-death reasons. On the one hand, those with more diseases might be more likely to drop out because it is difficult for somebody in poor health to respond to surveys or they care less about their health. On the other hand, worsening health status might lead respondents to be more likely to remain in the survey for reasons such as seeking health-related information or getting connected to society.

Question 2. Is there a negative effect of education on hospitalizations after controlling for childhood socioeconomic conditions and child health?

Hypothesis: After controlling for childhood socioeconomic status and child health, higher educational attainment is associated with a *lower* probability of hospitalization in later life.

Rationales: As suggested by the conceptual model, there is an association between education and health care utilization, measured by hospitalizations in this study. There are at least two reasons why more years of education is associated with a lower probability of hospitalization. First, there might be a spurious relationship between education and hospitalization due to a third variable. There are other factors driving both education and health, even controlling for childhood socioeconomic status and child health. Those variables, such as time preference and intelligence, could create a false positive relationship between education and hospitalization. Second, there are both direct and indirect effects of education on reducing the probability of hospitalizations. For example, education enables individuals to eat healthy food, provides individuals with health-enhancing information, and allows individuals to work in a healthy environment. All of these effects lead individuals to have fewer health conditions and then less likely to be hospitalized. Since this study focuses on the total effect of education on hospitalizations, it only controls for socioeconomic status and health status realized in childhood.

Counterhypothesis: After controlling for childhood socioeconomic status and child health, higher educational attainment is associated with a *higher* probability of hospitalization in later life.

Rationales: There are countervailing effects of schooling on hospitalizations. People with higher educational attainment are more likely to have higher income and more generous health insurance. It gives them a higher probability of being hospitalized, given the same health conditions relative to individuals with lower levels of education. Given this study focuses on the total effect of education and only controls for childhood socioeconomic status and child health, if these effects dominate, more years of education could increase the probability of hospitalization in the later stage of the life cycle.

Question 3. Whether secondary schooling is causally linked to hospital admissions?

Hypothesis: The total effect of education on hospitalization is unclear.

Rationales: In the conceptual framework, multiple pathways are running from education to health and health care. In other words, education has a causal effect on hospitalization through different channels. However, due to the countervailing effects of these channels, the total effect of schooling on hospitalization is not determined. As discussed in the previous research question, on the one hand, education decreases the likelihood of hospitalization by reducing health conditions. On the other hand, education increases the likelihood of hospitalization due to better access to health care, higher income, or better connections, given a similar health status. This study's focus on the total effect of education on hospitalizations does not allow the reduced-form model control for any of these pathways. As such, which opposing effect dominates is unclear and depends on health policies in place and individual endowments.

Chapter 4. Research Methodology

In this chapter, I shall first describe the general data sources used for this dissertation and measures constructed for analyses. Then I will turn to the study methods employed to examine the three research topics: the attrition analysis, the longitudinal analysis of the association between education and hospitalizations, and the causal effects of education on hospitalizations based on an instrumental variables approach. For each of them, I shall give greater details about the study motivation, study population, study design, econometric model, and sensitivity analyses.

4.1 Data Sources

4.1.1 The Health and Retirement Study

This study uses data from the Health and Retirement Study (HRS) for all birth cohorts and the survey years from 1992 to 2016. Each research question includes a different sample, which shall be described in more detail in the results section.

HRS is a national and longitudinal panel study that surveys a representative sample of individuals aged 51 and above and their spouses. HRS collects data to paint an emerging portrait of an aging America's physical and mental health, insurance coverage, financial status, family support systems, labor market status, and retirement planning. The sample for the HRS has been built up over time. Starting from 1992 with biennial interviews through 2018, HRS obtains detailed information in several domains: demographics, income, assets, health, cognition, family structure, health care utilization and costs, housing, job status and history, expectations, and insurance in its core files.

As of 2018, the HRS sample is comprised of seven subsamples or cohorts:

- The initial HRS. The initial HRS cohort includes those born in 1931-1941. It was designed to follow up age-eligible individuals and their spouses as they were transitioning from active work into retirement. This cohort was first interviewed in 1992 and subsequently every two years.

- Assets and Health Dynamics Among the Oldest Old (AHEAD). AHEAD study was joined in 1993 as a companion study to examine the dynamic interactions between health, family, and economic variables in the post-retirement period at the end of life. It consists of persons born before 1924 who aged 70 and over in 1993. This cohort was first interviewed in 1993 and subsequently in 1995, 1998, and subsequently every two years.

In 1998, several major changes were made to make sure the evolution of the HRS and AHEAD studies into a single ongoing survey, which is continually representative of the complete US population over the age of 50. The original HRS and AHEAD studies were merged in 1998; respondents from each forming a cohort in a combined interview. Meanwhile, two new cohorts were added to make the sample representative of those aged 51 or above in 1998.

- Children of Depression (CODA) cohort, born 1924-1930. This cohort was first interviewed in 1998 and subsequently every two years.
- War Baby (WB) cohort, born 1942-1947. This cohort was first interviewed in 1998 and subsequently every two years.

Also, new cohorts were added every six years, which resulted in the following cohorts:

- Early Baby Boomer (EBB) cohort, born 1948-1953. This cohort was first interviewed in 2004 and subsequently every two years.
- Mid Baby Boomer (MBB) cohort, born 1954-1959. This cohort was first interviewed in 2010 and subsequently every two years.
- Late Baby Boomer (LBB) cohort, born 1960-1964. This cohort was first interviewed in 2016 and subsequently every two years.

The HRS sample is based on a multi-stage area probability design involving geographical stratification and clustering. A screening interview was conducted with each sampled housing unit to determine eligibility. A primary respondent was then randomly selected from all age-eligible households' members. The selected person's spouse or partner of any age was also included in the sample. HRS has always oversampled African American and Hispanic households at about twice the rate of Whites. In 2010, HRS undertook an expansion of the minority sample from the Baby Boomer cohorts, which is referred to as the minority oversample. Compared to other surveys, HRS maintained high response rates that are around 85-90%. **Table 4-1-1** below documents the sample size and response rates by the HRS cohort and year. The description of HRS in this section relies heavily on the documents available on the HRS website (https://hrs.isr.umich.edu/), please refer to the website for more information.

Table 4-1-1. HRS sample sizes and response rates by cohort and by year.

	1992/	**1994/**										
	1993	**1995**	**1996**	**1998**	**2000**	**2002**	**2004**	**2006**	**2008**	**2010**	**2012**	**2014**
Wave	1	2	3	4	5	6	7	8	9	10	11	12
Total Sample												
Interviewed	20,874	18,447	10,964	21,384	19,578	18,166	20,129	18,469	17,217	22,032	20,554	18,747
Response rate (%)	81.1	90.7	86.9	83.8	88.0	88.4	85.3	88.9	88.4	81.0	89.1	87.1
HRS												
Interviewed	12,652	11,420	10,964	10,584	10,044	9,724	9,362	8,879	8,493	7,904	7,395	6,624
Response rate (%)	81.6	89.4	86.9	86.7	85.4	86.6	86.4	88.6	88.6	88.6	89.6	87.9
AHEAD												
Interviewed	8,222	7,027		5,951	5,000	4,107	3,365	2,700	2,142	1,526	1,165	844
Response rate (%)	80.4	93.0		91.4	90.5	90.1	89.4	90.6	90.7	89.3	90.0	87.7
CODA												
Interviewed				2,320	2,124	1,951	1,777	1,618	1,454	1,255	1,112	903
Response rate (%)				72.5	92.3	91.2	90.1	91.4	90.4	89.0	90.8	88.7
War Baby												
Interviewed				2,529	2,410	2,384	2,295	2,237	2,165	2,138	2,065	1,939
Response rate (%)				69.9	90.9	90.6	87.9	88.1	87.0	87.4	88.0	89.0
EBB												
Interviewed							3,330	3,035	2,963	2,926	2,826	2,745
Response rate (%)							75.3	87.7	86.3	85.9	85.5	85.6
MBB												
Interviewed										3,283	3,121	2,982
Response rate (%)										68.8	89.6	84.9
Minority oversample												
Interviewed										3,000	2,870	2,710
Response rate (%)										66.2	90.9	88.4

HRS Tracker File. Cross-Wave Tracker File 2016 Tracker (Early, Version 3.0) contains one record for every person who was ever interviewed in any wave, which facilitates the use of HRS data within and across waves. For each person, it records basic demographic information, interview status, and if, when, and how an interview was conducted in each wave.

HRS Restricted Data. State-level geographic information for respondents interviewed in 1992 through 2016, matching the 2016 tracker file, was obtained from the HRS restricted data file—Cross-Wave Geographic Information (State) (1992-2016). This dataset contains geographic identifiers, especially the state of birth and state of residency.

RAND HRS Longitudinal File. The RAND Center for the Study of Aging produced a clean, user-friendly, and streamlined version of the HRS core interviews, with derived variables covering a very large number of measures (e.g., demographics, health, health insurance, out-of-pocket medical expenditure, and employment history); variables are named and derived consistently across waves. The RAND HRS Longitudinal File 2016 (v1) is used in this study.

RAND HRS Fat Files. As a complement to the RAND HRS Longitudinal File, the RAND Center also produced versions of the HRS "raw" data for each year from 1992 to 2016. Each file contains all the raw variables collected at the respondent or household level, except those from "Other Person" modules (i.e., data on children, siblings, household members, helpers, and transfers between respondents and their children). I used these files to extract information regarding self-reported health status and family financial situation when respondents were younger than 16.

4.1.2 Compulsory Schooling Laws

Compulsory schooling laws in the 1900s have been systematically compiled and analyzed by several authors and made publicly available. Most of these authors documented three features of the laws: continuation schooling laws, child labor laws, and child attendance laws. In this study, I rely on these papers as sources of compulsory schooling laws. For certain states and years missed by these sources, I extended the data series by collecting my own data from a variety of sources. Specifically, I use datasets from Lleras-Muney (2002) and Goldin & Katz (2003) for 48 states (excluding Alaska, Hawaii, and Washington D.C.) from 1910 to 1939, and the dataset from Acemoglu & Angrist (2000) for 48 states (excluding Alaska, and Hawaii) during 1940-1978, and for Washington D.C. during 1915-1978.

I also collected additional data to impute 1) compulsory schooling laws for Washington D.C. during 1910-1914, 2) missing continuation schooling laws for Washington D.C. during 1915-1978, 3) and missing continuation schooling laws for 48 states (excluding Alaska, and Hawaii) during 1940-1978. **Table 4-1-2** documents all data sources for compulsory schooling laws. I will explain each cited source in the table in greater detail in the following paragraphs.

Table 4-1-2. Sources of compulsory schooling laws, 1910-1978

Birth cohorts	Year at 14 age	States	Continuation schooling laws	Child labor laws	Child attendance laws	Required schooling
1896-1964	1910-1939	48 states (excluding Alaska, Hawaii, and Washington D.C.)	Lleras-Muney (2002), Goldin & Katz (2003)	Lleras-Muney (2002), Goldin & Katz (2003)	Lleras-Muney (2002), Goldin & Katz (2003)	
1896-1900	1910-1914	Washington D.C.	Y	Y	Y	
1911-1964	1915-1978	Washington D.C.	Y	Acemoglu & Angrist (2000)	Acemoglu & Angrist (2000)	
1926-1964	1940-1978	48 states (excluding Alaska, and Hawaii)	Y	Acemoglu & Angrist (2000)	Acemoglu & Angrist (2000)	
1905-1961	1919-1975	48 states (excluding Alaska, and Hawaii)				Stephens & Yang(2014)

Notes: Y indicates data were collected by my own from the following sources; for years without published reports, I carried forward the previous year's data:

1) US Office (Bureau) of Education. 1910. *Education Report*, 1910. (Annual Report of the Commissioner of Education). "Compulsory Education and Child-Labor Laws." Washington, D.C.: GPO.

2) US Department of the Interior, Bureau of Education. *Laws Relating to Compulsory Education*. Bulletin No. 20 by Ward W. Keesecker, US GPO 1929.

3) Alexander, K. & Jordan, K.F. (1973). *Legal Aspects of Educational Choice: Compulsory Attendance and Student Assignment*.

4) US Department of Labor, Division of Labor Standards (July 1946) *State Child-Labor Standards. A State-by-State summary of laws affecting the employment of minors under 18 years of age*.

5) US Department of Labor, Division of Labor Standards (Sep 1949) *State Child-Labor Standards. A State-by-State summary of laws affecting the employment of minors under 18 years of age*. Bulletin 114.

6) US Department of Labor, Division of Labor Standards (Apr 1952) *State Child-Labor Standards. A State-by-State summary of laws affecting the employment of minors under 18 years of age*. Bulletin 158.

7) US Department of Labor, Division of Labor Standards (Sep 1965) *State Child-Labor Standards. A State-by-State summary of laws affecting the employment of minors under 18 years of age*. Bulletin 158 (Revised 1965).

For the first time, Lleras-Muney (2002) analyzed the effectiveness of compulsory attendance and child labor laws from 1915 to 1939 on education attainment (Lleras-Muney 2002). The following attributes were documented and analyzed: the maximum age by which a child must enter school (Entrance Age), the minimum age at which a child can drop out of school (Dropout Age), the minimum education level required to be exempted from school (Education to Dropout), the minimum age required to obtain a work permit and leave school (Work Permit Age), the minimum education level required to obtain a work permit (Education to Work), and whether working children were required to go to school on a part-time basis (Continuation School). The dataset is available on the author's website (https://adriana-llerasmuney.squarespace.com/data/).

Goldin and Katz (2003) also compiled a similar dataset containing the same information as Lleras-Muney (2002). But, it goes back further to 1910, to match all of the cohorts to the school entry age laws in effect when the cohorts were younger than 14, like at age 6 (Goldin and Katz 2003). This dataset is available on the first author's website (https://scholar.harvard.edu/goldin/pages/data) as well.

Acemoglu and Angrist (2000) were probably the first to compile data regarding child attendance laws and child labor laws. Their data covers a broader period ranging from 1915 to 1978 (Acemoglu and Angrist 2000). It contains all the variables available in Lleras-Muney (2002) and Goldin and Katz (2003) except for the measure of continuation laws. The dataset is publicly available online (https://economics.mit.edu/faculty/acemoglu/data/aa2000).

Besides, Stephens and Yang (2014) further extended the previous coding of compulsory schooling laws. They constructed a new measure called "required schooling (RS)," by iterating through ages 6 to 17 to determine whether the child is required to attend school at that age based on the law in place that same year. As such, this measure accounts for any changes to the compulsory attendance and child labor laws that may occur during the child's school years. Changes in these laws were extracted by building upon previous sources (Acemoglu and Angrist 2000, Goldin and Katz 2003), and using a number of additional secondary sources as well as the original legislation found in state session laws. The dataset has been published along with the paper (Stephens Jr and Yang 2014) and contains measures for birth cohorts 1905-1961.

4.1.3 Quality of Schooling

Card and Krueger (1992) compiled a dataset measuring the quality of public schools for birth cohorts between 1920 and 1949, based on issues of the *Biennial Survey of Education* that contains the results of surveys conducted by the US Office of Education from 1918 to 1966. For each of these quality attributes, the authors create a single measure for each state of birth/year of the birth cohort by averaging the prevailing measures during the years in which that cohort aged 6 to 17. It contains information on three main characteristics: the ratio of enrolled students to instructional staff in the state (pupil/teacher ratio), the average length of the school term (term length), and average annual teacher salaries.

Moreover, to account for geographic differences in the living cost and in the level of alternative wages available to potential teachers, they normalized teacher wages in each state by the level of average wages in the state; regional wage rates for workers on federal construction projects were used for normalization prior to 1940, and average weekly earnings of employees covered by the social security system were used from 1940 onward. Additionally, they further removed the trend in average relative teacher salaries by dividing the relative teacher wage in each state by the national average of this ratio in the same year. Here and after, I refer to this measure as relative teacher wage.

Following the same procedure, Stephens and Yang (2014) extended the data series to birth cohorts from 1905 to 1959, using various editions of the *Digest of Educational Statistics*. Together with the paper, the dataset is available on the journal website.

4.2 Measurements

In this section, I shall describe the variables used for the attrition analysis and the measures available from the HRS that could be proxies for the above concepts in the conceptual framework. Note that, otherwise stated, all these measures are available from the RAND HRS Longitudinal file in which variables are named and derived consistently across waves. The description of variables below heavily relies on the RAND HRS Longitudinal File 2016 (v1) Document. The documentation also provides a detailed description of the methodology in imputing missing values in income and wealth. Importantly note that I recoded most of these variables based on their distributions in my analytic sample to make sure they are suitable for this study.

Outcomes

Attrition status. The outcome "wave-specific attrition status" includes three categories: response, non-response due to death (died), and non-response due to other reasons (non-response). The variable was recreated based on the variable "interview status" from the RAND HRS Longitudinal File. RAND constructed this variable by taking mortality status and in sample status from the HRS Tracker files.

Hospitalization. Hospitalization is used as a proxy for health care use. Hospitalizations have low demand elasticities (Manning et al. 1987), and as such, are more likely to reflect changes in health status and corresponding health care use compared to outpatient care. It also captures the demand for health care and access to hospitals. The wave specific hospitalization indicator is available for all waves and represents whether the respondent reports any overnight hospital stay since the last interview. In the first two waves, the recall window was 12 months. In other waves, it was two years. As an example, in 2010, the question was asked as "(Since R's LAST IW MONTH, YEAR/In the last two years), have you been a patient in a hospital overnight?"

Focal independent variable

I use two measures for educational attainment: years of education and categorical education. **Years of education** was provided by RAND HRS based on respondents' highest education in the HRS core interviews and years of schooling from the HRS tracker file. It is truncated at 17, ranging from 0 to 17. **Categorical education** includes four groups: less than high school, high school, some college, and college or above. The "high school" group consists of those with a high school diploma or General Educational Development (GED). Respondents who have a high school degree or GED and years of education over 12 were classified into the category of "some college". The categorical education measure is used to investigate the potential non-linear relationship between education and hospitalizations.

Covariates for the attrition analysis (the first study)

I included an extensive list of covariates in the attrition analysis, including demographics, health status, and socioeconomic status, as described below.

Female. Female is a dummy variable with 1 as female and 0 as male. it is recoded from the "Gender" variable in the HRS Tracker file. **US-born.** This variable is also from the Tracker file; 1 indicates the respondent was born in the US and 0 otherwise. **Age.** Age is a wave-specific variable reflecting the age at the beginning of the interview, which is calculated based on the respondent's birthdate and beginning interview date (available in RAND HRS). **Race.** It has three categories: White/Caucasian, Black/African American, and Other. **Hispanic.** An indicator variable for whether the respondent is Hispanic. **Marital status** is a wave-specific variable and encompasses four categories: married (married, married but spouse absent, and partnered), divorced (divorced, separated, divorced/separated), widowed, and never married.

Census region refers to the wave-specific residence of one of the four census regions (Northeast, Midwest, South, and West). **The number of people in HH**. It recodes the number of residents in the household, including the respondent and spouse, to a categorical variable that contains four groups: one, two, three, and four-plus. Similarly, the **number of living children** recodes the number of living children of the respondent and spouse or partner to four categories: no child, one/two children, three/four children, and five/more children.

The proxy interview indicates whether the interview was done by a proxy (1=Yes, 0=No). **Self-reported health** is the Respondents' self-rated health status, coded as 1 = Excellent, 2 = Very Good, 3 = Good, 4 = Fair, 5 = Poor. **Ever had severe disease** is set to 1 if respondent was ever diagnosed with any of the following diseases: a) cancer or a malignant tumor of any kind except skin cancer; b) heart attack, coronary heart disease, angina, congestive heart failure, or other heart problems; c) stroke or transient ischemic; and d) chronic lung disease except asthma such as chronic bronchitis or emphysema attack (TIA). Otherwise, it is set to 0. **Ever had mild disease** is coded to 1 if the respondent was ever diagnosed with any of the following conditions: a) high blood pressure or hypertension; b) diabetes or high blood sugar; and c) emotional, nervous, or psychiatric problems. **Body Mass Index (BMI).** BMI is weight divided by the square of height. Based on BMI values, I created a categorical variable with 1 denotes normal weight if BMI < 25; 2 denotes overweight if BMI is between 25 and 30; and 3 represents obesity if BMI is larger than 30.

Education. The categorical education described above was used in the attrition analysis. **House ownership** indicates whether the respondent and his/her husband/wife/partner own the house (1= Yes, 0 = No). Labor force status includes four categories; 1= Working, 2 = (partly) Retired, 3 = Disabled, and 4 = Outside labor force. **Individual earnings.** Individual earnings are generally the sum of respondents' wage/salary income, bonuses/overtime/commissions/tips, and other incomes. When individual earnings are missing, RAND imputed them in the RAND HRS Longitudinal File. I recoded the variable to a four-categorical variable by quintiles in each wave. **HH total income.** The household's total income includes the sum of respondent and spouse earnings, pensions, government transfers, and other incomes. I also recoded it to a categorical variable, including four quintiles.

Concept proxies for the relationship between education and hospitalizations (the second and third studies)

Opportunities and Constraints

Childhood Environment. To capture the concept of childhood environment, I used two measures as proxies.

1) *Parents' education.* HRS collects information on respondents' fathers' years of completed education and mothers' years of completed education. Given some respondents lived in single-parent families, I created a composite measure for this variable as the highest grade of completed education of the respondent's father or mother, which ranges from 0 to 17.

2) *Childhood family financial situation*. Starting from wave 4, HRS asks respondents to recall their family financial situation when they were a child. The question is "Now think about your family when you were growing up, from birth to age 16. Would you say your family during that time was pretty well off financially, about average, or poor?". I extracted this variable from the HRS Fat Files since it is not compiled in the RAND HRS Longitudinal File. I also recreated it as a dummy variable, which is set to 1 if the response is "poor" and 0 otherwise.

Race and Ethnicity. It is worth noting that race/ethnicity also captures individuals' other socioeconomic status during childhood, as well as the opportunities for schooling. HRS has two questions for race and ethnicity. One question is about the Hispanic type that includes four categories: Hispanic, Mexican; Hispanic, Other; Hispanic, type unknown; and Non-Hispanic. I coded it as a binary outcome (Hispanic versus Non-Hispanic) in this dissertation. Another question is about race. It includes three responses: White/Caucasian, Black or African American, and Other.

Individual Tastes and Preferences, Biological Factors

The concept of individual tastes and preferences are always hard to capture in empirical studies. In this dissertation, I use *Female, Race,* and *Hispanics* (described above) to reflect differences in personal taste and preference. Admittedly, there is no way that these three measures could fully capture these differences. For example, the variation in the "discount rate" across individuals is likely to be omitted. Due to the limited set of information available in HRS, I turned to the instrumental variables approach to overcome this limitation.

Child Health

HRS has a question about respondents' s childhood health status. The question asked in the HRS is "Consider your health while you were growing up before you were 16 years old. Would you say that your health during that time was excellent, very good, good, fair, or poor?". It is available starting from wave 4. I included this question as a proxy for the concept of child health as a categorical variable ranging from 1 for excellent to 5 for poor.

State-of-Birth Education and Health Investment

To capture education and health investment at the level of state-of-birth, I used *state-of-birth fixed effects* and *year of birth fixed effects*. State-of-birth dummies account for the determinants of education and health that differ across locations but are time-invariant. Year of birth is to hold constant those that vary uniformly across states over time. I also included a state-of-birth specific linear time trend to capture state-level changes in policies, investment, and other factors that affect both education and hospitalizations.

Construction of Instruments

The compilation of state-level compulsory education and labor laws includes six key variables:

1. Minimum age of compulsory schoolings (Entry age);
2. Maximum age of compulsory schooling (Dropout Age);
3. Education for exemption from maximum age rule (Education to Dropout);
4. Age at which youth can obtain a work permit (Work Permit Age);
5. Education required to receive a work permit (Education to Work);
6. Whether a state has mandatory continuation schools (Continuation Laws)

Based on these elements, I constructed some composite measures for compulsory schooling years. Aspects of child attendance laws and child labor laws used to construct these measures are those prevailing in the individual’s state when they were 14 years old, except that entry age is assigned based on laws in place at age 6. More specifically, Years of compulsory schooling required by child attendance laws (CA_{sct}) for those born in state s in year c and were 14 years old in year t is computed as

$$CA_{sct} = \min\,(Dropout\ Age_{sct} - Entry\ Age_{sc,t-8}, Education\ to\ Dropout_{sct})$$

Years of compulsory schooling required by child labor laws (CL_{sct}) for those born in state s in year c and were 14 years old in year t is computed as

$$CL_{sct} = \max\,(Work\ Permit\ Age_{sct} - Entry\ Age_{sc,t-8}, Education\ to\ Work_{sct})$$

Leave age (LA_{sct}) for those born in state s in year c and were 14 years old in year t is computed as below

$$LA_{sct} = \min\,(Dropout\ Age_{sct}, Work\ Permit\ Age_{sct})$$

I also adopted the “required schooling (RS_{sct})” constructed by Stephen and Young (2015) by iterating through ages 6 to 17 to determine whether the child is required to attend school at that age based on the law in place that same year. In addition, I created several dummy variables that attempt to capture the potential non-linear effects. The cutoff values were used to ensure there are sufficient observations per group. To save notations, I dropped the subscripts in the following equations.

$$\begin{gathered} CA6 = 1\ if\ CA \leq 6; 0\ otherwise \\ CA7 = 1\ if\ CA = 7; 0\ otherwise \\ CA8 = 1\ if\ CA = 8; 0\ otherwise \\ CA9 = 1\ if\ CA = 9; 0\ otherwise \\ CA10 = 1\ if\ CA \geq 10; 0\ otherwise \end{gathered}$$

$$\begin{gathered} CL6 = 1\ if\ CL \leq 6; 0\ otherwise \\ CL7 = 1\ if\ CL = 7; 0\ otherwise \\ CL8 = 1\ if\ CL = 8; 0\ otherwise \\ CL9 = 1\ if\ CL \geq 9; 0\ otherwise \end{gathered}$$

$$\begin{gathered} RS6 = 1\ if\ RS \leq 6; 0\ otherwise \\ RS7 = 1\ if\ RS = 7; 0\ otherwise \\ RS8 = 1\ if\ RS = 8; 0\ otherwise \\ RS9 = 1\ if\ RS \geq 9; 0\ otherwise \end{gathered}$$

The school quality measures (pupil/teacher ratio, term length, and relative teacher wage) are also considered as potential candidates for instruments in this dissertation.

4.3 Attrition Analysis

4.3.1 Motivation

HRS has been collecting longitudinal data for people aged 51 or older and their spouses of any age since 1992. Multiple cohorts have been included in the survey between 1992 and 2016. The feature of collecting repeated measures for HRS respondents allows us to track the evolution of socioeconomic and health outcomes over time. In this dissertation, for example, I am interested in the evolution of the likelihood of being hospitalized by educational levels as individuals age. Importantly, it also substantially increases the sample size and makes it possible for more sophisticated statistical analysis, such as an instrumental variables approach.

However, since HRS focuses on near-elderly population, non-random loss of follow-up due to death or other reasons could be a serious concern, especially for those analyses using the longitudinal feature—following respondents over time to study some patterns of behaviors or pooled data analyses. If those respondents dropped out of the survey had systematically different characteristics, it could lead to a biased inference of estimates. Specifically, since those having fewer years of schooling were more likely to leave the follow-up surveys due to earlier death or non-response, the relationship between education and health would be underestimated if analyses using the panel data failed to account for such attrition bias.

The HRS-provided weights cannot address attrition bias for longitudinal data analyses. The goal of sampling weights constructed by HRS is to make the HRS weighted sample representative of the community-dwelling US population over age 50 in a given cross-sectional survey year (Ofstedal et al. 2011). Yet, sample weights for HRS follow-up waves adjust for the non-response based on post-stratification of the sample to the Current Population Survey (for waves 1992-2004) or American Community Survey (for waves 2006 and subsequent waves). The post-stratification was based on age, gender, and race/ethnicity. Thus, HRS-provided weights would only restore the representativity of the sample if attrition was completely random or only driven by age, gender, and race/ethnicity. In addition, even dropout due to mortality imposes no threat to sample representativity of survivors after the weighting, it will cause severe concerns for longitudinal analysis. For example, the relationship between education and health care use may be contaminated due to that people with lower educational levels are more likely to die earlier. As such, these weights could not account for the attrition bias in causal inference analyses pooling available data from all survey cohorts over multiple waves.

There are limited studies examining the extent to which attrition affects the inference of estimates from a longitudinal analysis. Prior studies are confined to early HRS cohorts and focused more on the representativity of the remaining sample. Kapteyn et al. (2006) investigated the determinants of attrition status in 2002 using the original HRS cohort (those born 1931-1941). The authors found that attritors (dropout due to non-response) had different baseline characteristics (e.g., race/ethnicity, immigration status, education, health, etc.), those who skipped some of the intermediate waves were the most different from those who always remained in the sample in terms of baseline characteristics, and those who were dead had systematically lower socioeconomic status (Kapteyn et al. 2006). Another similar study distinguished attrition due to mortality (passive attrition) from attrition due to refusal and other reasons (active attrition) in 2002 and included both original HRS and AHEAD cohorts in the analyses. This study concludes that active attrition was not selective and statistically ignorable, but passive-active was probably selective (Cao and Hill 2005). However, results from these two studies might not be relevant now for at least two reasons. First, there is dramatic attrition for the two earliest birth cohorts in 2016 relative to 2002, largely due to mortality; respondents from the original HRS cohort aged 61 to 71 in 2002, but they were 75 to 85 in 2016. Second, several other cohorts were added to the survey every two years, which might have different patterns of attrition compared to HRS and AHEAD.

The *objective* of this analysis is to investigate what observed factors drove different types of attrition and set the stage for developing weights for accounting for potential attrition bias.

4.3.2 Study population

The study population was comprised of all respondents in the HRS survey from 1992 to 2016. Those included eligible respondents and their spouses from all the birth cohorts except the Later Baby Boomer cohort (born 1960 to 1965) that first entered the survey in 2016.

4.3.3 Study design

I used data from The RAND HRS Longitudinal File 2016 (V1) merged into the 2016 HRS Tracker File. The two datasets have been described in the above section.

Baseline characteristics and attrition status in 2016. I first explored the relationship between baseline characteristics and attrition status in 2016. Baseline characteristics were measured during the wave when they first entered the survey. Based on respondents' attrition pattern in the follow-up waves, I constructed two types of attrition status. 1) Always In. Respondents who have never skipped any follow-up surveys since they entered the study were considered as "always in." 2) Died. Respondents who dropped out of the survey due to death. 3) Non-response. Non-response represents those dropouts who did not come back to the survey until 2016 but were alive when they first left the study. 4) Ever-out. The last group refers to those who had missed at least one intermediate wave but reentered the survey at least once.

Determinants of between wave attrition. Given the long follow-up time of the survey, respondents' attrition can also be driven by events after the initial interview. Those events include those that are related to baseline characteristics. For instance, respondents with lower levels of educational attainment tended to die earlier and then drop out of the survey. Those events also include random shocks in respondents' lives, such as loss of wealth or diagnosis of urgent, life-threatening diseases. To capture the immediate effects of events between waves, I conducted a between wave analysis that looks at attrition between each wave. In this case, there are only three modes of attrition: Responded, Died, and Non-response.

4.3.4 Econometric model

Both observables and unobservables determine attrition. In this study, I primarily focus on how observed measures from the HRS affect attrition. The assumption for this analysis is that attrition is a random event conditional on observed variables included in the model. To account for unobservables in attrition bias, variables that trigger the attrition but have no effect on the outcome of interest—the exclusion restrictions—are required in models such as the Heckman sample selection model. However, since a majority of attrition is driven by mortality and the outcome of this study is health, it is virtually impossible to find such variables.

A cross-sectional analysis of baseline characteristics and attrition status in 2016. In the first analysis, a multinomial logit model was applied, which is specified as follows

$$y_i = g(\alpha + \beta x_i) + \epsilon_i$$

Where y_i denotes attrition status, including four categories, as stated above. The $g()$ function indicates a non-linear model. x_i includes Female, US-born, Age, Race, Hispanic, Marital status, Census region, Number of people in HH, Number of living children, Proxy interview, Self-reported health, Ever had severe disease, Ever had mild disease, Body mass index, Education, House ownership, Labor force status, Individual earnings, and HH total income, all defined at the baseline wave for each entry cohorts accordingly.

A dynamic model for between-wave attrition. In the second analysis, I investigated the between-wave attrition using a dynamic multinomial logit model. This analysis captures the immediate effects of time-varying variables in the previous wave, which are more likely to affect respondents' response status in this wave. The model is specified below.

$$y_{it} = g(\alpha + \beta x_i + \lambda \delta_{i,t-1}) + \epsilon_{it}$$

Where y_{it} denotes attrition status including three categories (responded, died, and non-response). The $g()$ function indicates a non-linear model. x_i represents time-constant variables containing Female, US-born, Race, Hispanic, and Education. $\delta_{i,t-1}$ includes one-wave lagged variables that include Age, Marital status, Census region, Number of people in HH, Number of living children, Proxy interview, Self-reported health, Ever had severe disease, Ever had mild disease, Body mass index, House ownership, Labor force status, Individual earnings, and HH total income.

For ease of interpretation, marginal effects, instead of relative risk ratio (RRR), are reported. The first concern is whether some categories of outcomes could be combined into one group. To test this, I conducted a Wald test to see whether none of the predictors significantly predict the odds of alternative A versus alternative B; if so, we should combine the two alternatives to improve efficiency. Another main concern for a multinomial model is the *Independence of Irrelevant Alternatives* (IIA) assumption. The IIA assumption requires that the odds of choosing A over B should not depend on other alternatives available. The two most commonly used tests are the Hausman-McFadden test (1984) and the Small-Hsiao test (1985). However, these two tests perform rather poorly and often provide conflicting information on the violation of IIA. Prior studies have concluded that the two tests are not useful and suggested the assumption should be held on a theoretical basis (Fry and Harris 1998, 1996, Cheng and Long 2007). Nonetheless, I presented the results from the two tests in the section of results.

4.3.5 Sensitivity analyses

Since the outcome of attrition is about the duration for which one respondent stayed in the sample, survival analysis is another suitable approach. However, survival analysis cannot generate the probability of being one mode of attrition without a strong assumption on the hazard function. As such, I considered survival analysis as one of the sensitivity analyses. I fitted the following survival models.

1) A Cox proportional hazards model for attrition due to death. In this model, I treated those non-responses as censorings and used the default Breslow method for tied failures. The Cox model assumes that every subject shares a common baseline hazard function, and the covariates multiplicatively shift the baseline hazard function. To check whether this proportional-hazards assumption holds, I tested whether the slope is zero in a generalized linear regression of the scaled Schoenfeld residuals on functions of linear time; Stata's $estat\ phtest$ implements this test (Cleves, Gould, and Marchenko 2016). Stata's $estat\ phtest$ provides test results for covariates individually and globally. Given a large number of covariates (denoted by k) included in the model, I used Bonferroni Correction to lower the significance criterion (alpha value) to account for the number of comparisons conducted; $corrected\ \alpha = \alpha/\ k.$

2) *A competing risks regression model.* To focus on the determinants of non-response specific attrition, I performed a competing risk regression model treating death as a competing risk for non-response. I used Stata's $stcrreg$ command to implement this model. However, the typical competing risks model does not allow for multiple failure times; I thus only included the first attrition due to non-response as the failure event and considered death as a competing risk.

Because $estat\ phtest$ is not supported after $stcrreg$, I used another method to test the proportional-subhazards assumption. Specifically, I added an interaction term between one covariate and log-transformed time (such as $female * \ln\ (time)$) to the model. If the coefficient on the interaction term is significant, then the proportional-subhazards assumption is violated. This procedure was repeated for all covariates included in the model.

3) Given the limitation of the competing risks model mentioned above, I fitted the Andersen-Gill model that could account for multiple failure times (Andersen and Gill 1982). However, this model does not distinguish attrition due to non-response from attrition due to death. I used Stata’s $estat\ phtest$ for testing the proportional-hazards assumption.

All above survival analyses controlled for Female, US-born, Age, Race, Hispanic, Marital status, Census region, Number of people in HH, Number of living children, Proxy interview, Self-reported health, Ever had severe disease, Ever had mild disease, Body mass index, Education, House ownership, Labor force status, Individual earnings, and HH total income, all defined at the baseline wave for each entry cohorts accordingly.

4.4 Education and Hospitalizations: A Longitudinal Analysis

4.4.1 Motivation

As discussed in the literature review section, there is limited evidence on the relationship between education and health care utilization. The objective of this analysis is to examine the association between educational attainment and the probability of ever being hospitalized in the past two years based on the conceptual framework in Chapter 3.

4.4.2 Study Population

This analysis included all observations (person-wave) from the HRS 1992 – 2016, which consists of both eligible respondents and their spouses. To account for state-of-birth into the regression, I merged the 2016 HRS restricted state identifier file to the 2016 HRS tracker file and the RAND HRS Longitudinal file 1992-2016 (v1). As such, the study population included those born in the United States (with valid state of birth information) and recruited in the HRS survey from 1992 to 2016.

4.4.3 Study design

It is a cohort study following respondents for multiple waves; I pooled all observations available in the dataset across 1992 to 2016 waves. I used the fully robust standard errors to adjust for arbitrary within-person correlations. The analytic approach accounted for attrition bias using inverse probability weighting and adjusted for the confounders identified from the conceptual framework.

4.4.4 Econometric model

Before jumping into regression models, I started with several exploratory analyses. Particularly, I explored how the probability of being hospitalized changes by educational levels as respondents age. I also investigated to what extent the attrition bias that people with lower levels of education were more likely to attrite is a severe issue if left uncorrected.

For ease of interpretation, I used a linear regression model to examine the relationship between educational attainment (less than high school, high school/some college, college and above) and the probability of being hospitalized.

To correct for attrition bias, I employed the inverse probability weights (IPW) framework (Wooldridge 2010). Specifically, a logistic regression model was first used to fit predictive models for between-wave attrition. In other words, an attrition probability at time t was estimated restricting attention to those units still in the sample at time $t-1$. For $t = 2, \dots T$, let $\hat{\pi}_{it}$ denote the fitted probabilities of attrition. Then the probability weights were constructed as $\widehat{p_{it}} = \hat{\pi}_{i2}\hat{\pi}_{i3} \dots \hat{\pi}_{iT}$. However, there are two distinct forms of attrition—attrition due to death (passive attrition) and attrition due to non-response (active attrition). Previous evidence suggests that factors influencing those two forms of attrition were different (Kapteyn et al. 2006), and it is important to distinguish them in the model constructing attrition weights (Marden et al. 2017, Weuve et al. 2012). Following the spirit of these studies, I constructed robust attrition weights as follows.

Inverse probability of survival at wave t (S_{it})

$$IPSW_i = \prod_{t=1}^{13} \frac{\Pr\ (S_{it}|X_i, S_{i,(t-1)} = 1, UC_{i(t-1)} = 1)}{\Pr\ (S_{it}|X_i, L_{i,(t-1)}, S_{i,(t-1)} = 1, UC_{i(t-1)} = 1)}$$

The denominator for $IPSW_i$ is the probability of being alive for individual i at wave t given time-constant variables, one-wave lagged time-varying factors, and conditional on that the respondent was alive and responded in the previous wave. The numerator differs from the denominator by only including time-constant variables, which aims to avoid undue weights (Weuve et al. 2012).

Inverse probability of uncensored in the outcome “hospitalization” at wave t (UC_{it})

$$IPUCW_i = \prod_{t=1}^{13} \frac{\Pr\ (UC_{it}|X_i, S_{i,t} = 1, UC_{i(t-1)} = 1)}{\Pr\ (UC_{it}|X_i, L_{i,(t-1)}, S_{i,t} = 1, UC_{i(t-1)} = 1)}$$

The denominator for $IPUCW_i$ is the probability of being uncensored/having a valid response to the outcome for individual i at wave t given time-constant variables, one-wave lagged time-varying factors, and conditional on the respondent was alive this wave and responded in the previous wave. The numerator differs from the denominator by only including time-constant variables, which aims to avoid undue weights.

Where X_i includes baseline variables (Female, race/ethnicity, place of birth including nine census regions, year of birth, and educational attainment) and $L_{i,(t-1)}$ represents time-varying variables (marital status, four census region of residence, number of people in a household, number of living children, whether proxy interview, self-reported health, ever had severe diseases, ever had mild diseases, body mass index, house ownership, individual earnings, and household total income).

The final weights were constructed as $Finalweight_i = \ IPSW_i * \ IPUCW_i$. For ease of interpretation, I used a linear probability model. That says, I estimated the main analytic model using Pooled Ordinary Least Square (POLS) as

$$y_{it} = \alpha + \beta * Edu_i + X_i\theta + \epsilon_{it}$$

X_i here denotes child health status, the family financial situation in childhood, race, ethnicity, gender, state-of-birth, and year-of-birth. Edu_i includes three categories: less than high school (as reference), high school/some college, and college or above. The final model was weighted by the final weights constructed above for correcting for attrition bias. Standard errors were clustered at the respondent levels to accommodate for within-subject correlation over time in the survey.

Inspired by the exploratory analyses, the effects of educational attainment on hospital admissions might attenuate after age 78. There are several plausible explanations. For example, those survived beyond age 78 might be healthier (selective survival) and individuals’ health status might decay to a level that requires treatment in hospitals regardless of their educational attainment. To check this hypothesis, I performed a stratified analysis by age 78.

4.4.5 Robustness checks

A wide range of sensitivity analyses has been conducted to check the robustness of estimates from the above model.

First, although the outcome—ever being hospitalized in the past two years—is binary, an ordinary least squares linear regression is very likely to be appropriate based on the Central Limit Theory in the large sample size. To verify this, I estimated a logistic regression model and reported the marginal effects to compare with those from the linear regression model.

Second, in this analysis, I developed the weights correcting attrition bias by multiplying the probability of attrition in each follow-up wave. However, as argued in the literature of Econometrics (Wooldridge 2010), it might be more straightforward and appropriate to treat attrition as an absorbing state; once respondents dropped out of the survey, we considered them as attritors for all the following waves. A sensitivity analysis was done by treating attrition as an absorbing state.

Third, as both hospitalization and death reflect a severe health status, I created another outcome as the “adverse health outcome” defined as whether respondents experienced hospitalization or death. In this analysis, I considered non-response as the only form of attrition and developed the weights based on $IPUCW_i$.

Fourth, similar to a prior study (Behrman et al. 2011), I constructed another relevant outcome—ever being hospitalized 2 years before death—to extract hospitalizations from those related to terminal conditions that led to imminent mortality. I estimated the final model among the overall observations and subgroups stratified by age 78.

Finally, to address missing values and include all respondents instead of only those complete cases, I employed the multiple imputations approach. Although there were missing values in variables predicting attrition weights, it did not lead to the loss of respondents due to the multiplicative feature in constructing those weights. There would be a loss of respondents if variables were missed at every wave, which was not the case in the HRS data.

4.5 Secondary Schooling and Hospitalizations: An Instrumental Variable Approach

4.5.1 Motivation

The previous chapter aims to show the correlation between education and hospitalizations. The association may be taken as causal under the assumption that all potential confounding factors not captured in the adjusted covariates are captured by the state-of-birth fixed effects, year-of-birth fixed effects, or state-of-birth specific linear trends. However, there is no guarantee that it is the case. Failing to account for these factors will then lead to omitted variables bias in the estimated education effect. For example, individuals' discount rates might be left uncontrolled in the model. If patient individuals invested more in both education and health, the coefficient from the prior association study would be overstated and biased away from zero.

This chapter examines whether education has a causal impact on hospitalization by taking advantage of a unique quasi-experiment that states implemented legislation for compulsory schooling and child labor. Changes in those laws across different states in the 1900s governed the ages at which children were required to attend and allowed to leave school. The extra years of completed education induced by these laws for respondents born from different states in different years were plausibly unrelated to other determinants of health. By using these laws as instruments, this section aims to uncover the causal effects of education on hospitalization.

4.5.2 Study design

It is a quasi-natural experiment that leverages the variation in compulsory years of schooling required by state attendance laws and child labor laws in the 1900s. If these laws forced students to stay in school longer than they would have chosen otherwise, then students from states or birth cohorts who exposed to more years of compulsory schooling were more likely to complete more years of education than their counterparts exposed to fewer years of compulsory schooling. In this way, it would help eliminate bias due to individuals' choice over different investment in education and health.

In this study, I matched each individual to the laws that were in place in their state-of-birth when they were 14 years old. I then used these laws as instruments for years of education. The idea is that these laws affect hospitalizations only through the channel of years of completed education; these laws were plausibly exogenous. If this is true, estimators from this approach will be immune to bias due to omitted variables, reverse causality, and other endogenous problems.

4.5.3 Study population

Given the prior evidence (Lleras-Muney 2002) that suggests compulsory schooling laws and child labor laws impacted only the lower end of the distribution of education among white persons, I restricted the analysis to all white persons who had an education of high school or less and were born in the 48 states and the District of Columbia (Hawaii and Alaska were not then part of the Union). Due to the availability of measures for compulsory schooling laws and quality of schooling (as described in the section of “Data Sources”), I further restricted the analysis to those born between 1905 and 1959.

4.5.4 Econometric model

Pooled Ordinary Least Squares (POLS)

I pooled all the data and specified the following OLS regression model as a benchmark. It is a linear probability model, as used in the previous association study, but using a restricted sample and continuous measure of education. I estimated the following equation:

$$Y_{itcs} = \alpha + \theta * E_{ics} + X_{ics}\beta + \mu_c + \gamma_s + l_{sc} + b_{ics} + \epsilon_{itcs}$$

Where, Y_{itcs} denotes the outcome (ever being hospitalized in the past two years) for individual i being to cohort c and born in state s. E_{ics} represents years of completed education for individual i. X_{ics} are time-invariant individual characteristics, including gender, parents’ years of completed education, health status when the individual was younger than 16, and family socioeconomic status when the individual was younger than 16. μ_c is a set of birth cohort dummies, γ_s is a set of state-of-birth dummies, and l_{sc} represents state-of-birth linear trends. b_{ics} is the individual-specific unobserved time constant effect, and ϵ_{itcs} is the idiosyncratic error term.

Consistent identification of the coefficient θ from the POLS requires at least the following assumption: conditional on X_{ics}, μ_c, and γ_s, E_{ics} is uncorrelated with b_{ics} or ϵ_{itcs}. However, this assumption is very likely to be violated due to issues such as omitted variables. For example, individual preferences or discount rates could be related to both educational attainment and outcomes.

Two-stage least squares panel data (P2SLS)

The econometric model can be written as

$$E_{ics} = \alpha + Z_{cs}\pi + X_{ics}\beta + \mu_c + \gamma_s + l_{sc} + b_{ics} + v_{itcs}$$

$$Y_{itcs} = \alpha + \theta * E_{ics} + X_{ics}\beta + \mu_c + \gamma_s + l_{sc} + b_{ics} + \epsilon_{itcs}$$

Where, Y_{itcs} denotes the one outcome (ever being hospitalized in the past two years) for individual i being to cohort c and born in state s. E_{ics} represents years of completed education for individual i. Z_{cs} include exogenous instruments. X_{ics} are time-invariant individual characteristics as above. μ_c is a set of birth cohort dummies, γ_s is a set of state-if-birth dummies. b_{ics} is the individual-specific unobserved time constant effect, and l_{sc} represents state-of-birth linear trends. v_{itcs} and ϵ_{itcs} are the idiosyncratic error terms.

In the first step, I modeled years of completed education as a function of excluded instruments (Z_{cs}) together with other control variables. In the second step, I then fitted the second equation by replacing years of completed education with corresponding predicted values from the first step. The variance-covariance matrix of the use of the predicted values were adjusted to produce consistent standard errors.

Instrumental Variables and Generalized Method of Moment (GMM-IV) with panel data

Alternatively, we can use the Generalized Method of Moments to compute the IV estimators (GMM-IV). Compared to P2SLS, the approach of estimating instrumental variable estimators using GMM yields consistent but more efficient estimators (Wooldridge 2010). One reason is that P2SLS needs to reduce the dimension of instruments to the same level as the primary regressor. In contrast, GMM-IV does not need to, which is important in this study since many instruments are available. Also in the presence of non-i.i.d assumption of errors, which is the case in this study, GMM estimators are more efficient than P2SLS estimators (Baum 2006).

In addition, the GMM-IV approach provides various tests for the validity of instruments, including statistics for under-identification, weak instruments, and overidentifying restrictions (Baum 2006). For example, it reports the Anderson-Rubin Wald test statistics for weak identification; a failure to reject the null hypothesis calls the relevance of instruments into question. It also reports the J statistic of Hansen (Hansen 1982) to test for the overidentifying restrictions; a rejection of the null hypothesis suggests that either the instruments are not truly exogenous or are being incorrectly excluded from the regression.

Sparse Models (LASSO-IV) for many instruments and many controls

When using either P2SLS or GMM-IV, one issue arises: which sets and forms of instruments to use? Prior research applies various sets of instruments with different researchers using a different set of instruments. For example, these instruments include dummies of continuation laws and dummies for years of compulsory schooling required by child labor laws (Lleras-Muney 2005), categories for years of compulsory schooling required by child labor laws and attendance laws (Acemoglu and Angrist 1999), categories for accumulative years of compulsory schooling (Stephens Jr and Yang 2014), dropout age (Oreopoulos 2006), and even state-level quality measures of schooling (Nguyen et al. 2016). However, IV estimators are sensitive to choices of instruments. Also, IV estimators with many instruments are largely biased based on 2SLS and have too small standard errors (Hansen, Hausman, and Newey 2008, Chao and Swanson 2005). More importantly, there is not a priori way to decide which instruments to use in these studies.

Another commonly criticized point for using compulsory schooling laws as instruments is the concern about weak instruments, especially when the state-specific trend is taken into consideration. Prior studies show that statistically significant causal estimates of education on a number of outcomes (e.g., mortality, wages, unemployment, and divorce) became insignificant when allowing the year of birth effects to vary across regions or states (Mazumder 2008, Stephens Jr and Yang 2014). It also lets researchers conclude that "the further use of this methodology may require even larger, and potentially unattainable, sample sizes in the US" (Fletcher 2015).

To overcome these two challenges, we should be wise in the selection of instruments and controls in using these laws as instruments. Recent development in the literature of econometrics has proposed to apply machine learning techniques, such as the Least Absolute Shrinkage and Selection Operator (LASSO) regression, to select instruments and controls in the framework of causal inference (Belloni et al. 2012, Belloni, Chernozhukov, and Hansen 2014, Chernozhukov, Hansen, and Spindler 2015). It helps to address the above two concerns and get consistent estimators by selecting the optimal set of instruments and the appropriate set of controls, given the assumption of approximate sparsity. That is, the conditional expectation of the endogenous variables given the instruments can be approximated well by a parsimonious, yet an unknown set of variables. The approximately sparse imposes a restriction that only some of all variables have non-zero associated coefficients and permits a non-zero approximation error. As such, it allows model selection mistakes in the LASSO regression. The sparsity assumption is violated when there are weak instruments; testing for weak instruments can be considered as a check for the validity of the assumption. This method has been recently applied in a study evaluating the Workplace Wellness Programs on health and medical spending (Jones, Molitor, and Reif 2019).

A variant of the LASSO estimator (Belloni et al. 2012, Belloni, Chernozhukov, and Hansen 2014) was used in selecting instruments and controls as

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |b_i| \gamma_j$$

Where λ is the “penalty level” and γ_j is the “penalty loadings.” The penalty loadings are estimated from the data to ensure the equivalence of coefficients estimates to a rescaling of $x_{i,j}$ and to address heteroskedasticity, clustering, and non-normality in model errors.

The algorithm for the “Post-Double-Selection (PDS)” methodology (Belloni et al. 2012, Belloni, Chernozhukov, and Hansen 2014, 2011) is as follows.

1. Estimate a LASSO regression with the focal independent variable (E_{ics}) as the dependent variable and all potential instruments (Z_{cs}) and controls ($X_{ics}, \mu_c, \gamma_s, l_{sc}, b_{ics}$) as regressors. Get a selected set of instruments and controls.
2. Estimate a LASSO regression with the outcome variable (Y_{itcs}) as the dependent variable and all control variables ($X_{ics}, \mu_c, \gamma_s, l_{sc}, b_{ics}$) as regressors. In the LASSO model, gender, childhood health, childhood family financial situation, and parents’ education were partialled out. Get a selected set of controls.
3. Estimate a LASSO regression with the focal independent variable (E_{ics}) as the dependent variable and all control variables ($X_{ics}, \mu_c, \gamma_s, l_{sc}, b_{ics}$) as regressors. In the LASSO model, gender, childhood health, childhood family financial situation, and parents’ education were partialled out. Get a selected set of controls.
4. Estimate a 2SLS regression using the selected set of instruments from step 1 and the union of selected sets of controls from step 2 and step 3. This produces a Post-LASSO IV estimator (LASSO-IV). The Post-LASSO estimator discards the Lasso coefficient estimates and refits the regression via ordinary least squares (OLS) to alleviate Lasso’s shrinkage bias.

Another closely related method is the “Post-Regularization” methodology (Chernozhukov, Hansen, and Spindler 2015), which uses selected variables to construct orthogonalized versions of the dependent variable (Y_{itcs}) and the focal independent variable (E_{ics}). I reported estimators based on this method as a robustness check. I used the “*pdslasso*” and “*ivlasso*” packages to compute these two IV estimators (Ahrens, Hansen, and Schaffer 2018).

Finally, for all the above models, to address the within-subject correlation and potential heterogeneity problems, I used fully robust standard errors that are robust to arbitrary correlation and arbitrary heterogeneity (Wooldridge 2010). To account for attrition bias in the HRS longitudinal survey, I employed the inverse probability weighting approach using those weights developed in Section 4.3.

4.5.5 Sensitivity Analyses

I conducted the following sensitivity analyses. First, given the evidence from the exploratory analyses that the effect of education of high school disappears after the respondent reaches age 78, I performed two separate subgroup analyses using the LASSO-IV method. One analysis used observations of respondents younger than 78, and another analysis used observations when respondents were older than 78.

Second, to further confirm whether the quality of schooling measures are valid instruments. I added the quality of schooling measures into the selection pool of instruments. I then employed the selected instruments and controls based on the LASSO-IV approach in a GMM-IV regression model. I tested whether the selected instruments are weak and satisfy the overidenfication restrictions.

Lastly, since I adopted the attrition weights developed in the association study in section 5.2 based on the all-race sample, these weights might not be suitable for the specific sample in this analysis (whites born in the continental US between 1905 and 1959). However, it should not be a serious concern since the majority of the HRS sample are whites. Nevertheless, I reconstructed these weights based on this sample and reran the main analyses.

Chapter 5. Results

This chapter reports the results for the three research topics: the attrition analysis, the longitudinal analysis of the association between education and hospitalizations, and the causal effect of secondary schooling on hospitalizations based on an instrumental variables approach.

5.1 Attrition Analyses of the Health and Retirement Study

Part A: Baseline characteristics and attrition status in 2016

Sample size

Figure 5-1-1 shows the sample size for this analysis. There are 43,216 respondents in the 2016 HRS Tracker file and 42,053 respondents in the RAND HRS Longitudinal File (1992 - 2016). The differences in the number of respondents are due to the fact that 1,163 respondents included in the tracker file are not part of the HRS core interviews. I further excluded respondents who did not enter the survey until 2016 since attrition is not a relevant issue for those respondents. Finally, the study included 35,231 unique respondents from the following entry cohorts: Initial HRS (12,651), AHEAD (8,116), CODA (2,320), WB (2,529), EBB (3,330), MBB (3,285), and the Minority oversample cohort (3,000).

Figure 5-1-1 illustrates that a total of 13,227 (37.5%) respondents never left the survey since they were recruited for the study (Always-in); 12,175 (34.6%) respondents, mostly from the initial HRS and AHEAD cohorts, have dropped out of the survey due to death; 4,390 (12.5%) respondents skipped at least one of the intermediate waves but reentered the survey afterward; the remaining 5,439 (15.4%) respondents left the survey permanently due to non-death reasons.

Missing values

As shown in **Table 5-1-1**, there are few missing values for the included baseline characteristics. The variables with the most missing values were the *Number of living children* and *Body Mass Index*; 235 (0.67%) respondents did not have valid information on the number of living children in the baseline survey, and 454 (1.29%) respondents had no values for Body Mass Index. However, only 2.2% of the studied sample were missing at least one of the included variables, which was very likely to be missing at random. The following analytic sample only included the 34,457 complete cases, which accounts for 97.8% of the studied sample.

Descriptive statistics

As shown in **Table 5-1-2**, the overall analytic sample were predominately US-born, above 51 years old, white, non-Hispanic, married, not living alone, having more than one kid, and being able to respond to the survey independently. The sample included more females than males (56.3% versus 43.7%) and more respondents from the South (40%). A majority of the sample were homeowners (76.9%), had no severe diseases (74.1%), and considered their health status as good, very good, or excellent (72.5%). Approximately half of the sample had an educational level of high-school or less (60.3%) and were in the labor force (48.6%). Respondents from the sample were nearly evenly divided among the distribution of total household income.

Always in. Respondents who responded to every follow-up survey were more likely to be female and racial/ethnic minorities. One possible explanation for the racial/ethnic minorities is that the

“Minority oversample” cohort entered late in HRS. Age is a significant predictor of attrition status in 2016. This group of respondents tends to be aged 51-61 in the baseline survey, which indicates that they were from the entry cohorts other than AHEAD. Respondents in this group were more likely to have better health status; 80% considered their self-reported health status as good, very good, or excellent, and only 15.5% of them ever had severe diseases. They also had a higher socioeconomic status; half of them had a degree higher than high school; most had a total household income above the median level of the overall sample and were still working in the labor market.

Died. Those who left the survey due to death were more likely to be older, being widowed, living alone, having a worse health status, having lower educational attainment, and having lower income. Notably, 67% of them were over 61 in the baseline survey, which is much higher than that in the overall sample (34%), and that in other groups; 24.9% of them were windowed, and 25.1% of them were living alone, which are nearly twice that in the overall sample. This group of respondents had a much higher prevalence of both severe (41.9%) and mild (55.7%) diseases and were more likely to report health status as fair or poor. Importantly, it shows that those with lower education and lower income were more likely to attrite due to death. Among this group, 71.1% (49.4% in the always-in group) had an education level of high school or less, and 67% (37.6% in the always-in group) had a household income less than the median level of the overall sample.

Non-response. The baseline characteristics of those in the non-response group are similar to the overall sample. It is reassuring and suggests those dropouts due to non-death reasons are not systematically different from the overall sample. Compared to the overall sample, this group comprises of more respondents who were whites (78.7% versus 75.9%), had no severe diseases (78.9% versus 74.1%), and were house owner (80.1% versus 76.9%).

Ever-out. The baseline characteristics of those in the ever-out group were similar to the overall sample as well. However, there are some important differences relative to those in the non-response group. For example, relative to those in the non-response group, those who reentered the survey were more likely to be immigrants (16.6% versus 13.9%), racial minorities (28.7% versus 21.3%), and Hispanics (14.5% versus 9.7%). They also tended to have lower educational attainment in comparison to those in the non-response group; 29.8% of respondents from the ever-out group had less than high-school degrees while 23.1% in the non-response group. They were also less likely to have no living children (7.2% versus 11.0%) and have severe diseases (18.1% versus 21.1%) relative to those in the non-response group.

Estimates from a cross-sectional multinomial logistic regression model

Table 5-1-3 documents the results from a multinomial logistic regression. Marginal effects are reported in the table. In the following paragraphs, I will describe the results by groups of respondents' characteristics: demographics, living arrangements, health status, and socioeconomic status. I shall focus on statistically significant effects.

Demographics. Women were 7.5 percentage points (pp) more likely to remain the sample, and 8.2 pp less likely to die than men. Those who were born in the United States were more likely to drop out due to death (5.5 pp) but were less likely to be non-responsive (3.6 pp) or ever-out (3.2 pp). Older respondents had a significantly higher probability of dropping out due to death and a lower chance of staying in all the follow-up surveys. Relative to those aged less than 51, those aged over 61 are 19.51 pp were more likely to attrite due to death, 13.96 pp less likely to remain in the sample all the time, and 6.6 pp less likely to skip intermediate waves and come back. Compared to whites, black respondents had a 1.7 pp (2.7 pp) lower probability of being always-in (non-responsive) but a 5.2 pp higher probability of being ever-out. Those from other racial groups had a 3.2 pp less likelihood of death and a 3.0 pp higher likelihood of being ever-out. Hispanics were more likely to stay in the sample (2.7 pp) and reenter the survey when they dropped out (4.3 pp), and less likely to attrite due to both death (-4.1 pp) and non-response (-2.8 pp). Marital status also matters for the attrition status. In comparison with being married, being divorced or widowed reduced the probability of remaining in the sample by 1.8 pp and 5.3 pp, respectively. Being widowed also increased the likelihood of death by 5.5 pp. Besides, the entry cohorts also had significant impacts on the attrition status, largely due to the cohort effect; respondents from the later entry cohort were more likely to stay and less likely to die or ever-out. It is worth noting that those from the EB and MBB cohorts were more likely to be a non-respondent (3.7 pp and 2.6 pp), relative to the initial HRS cohort.

Living arrangements. The region where respondents live was associated with attrition status. Relative to those living in the Northeast region, those living in Midwest were more inclined to be always-in (3.4 pp) and dropout due to death (2.0 pp) but less likely to be non-response (-3.8 pp) and ever-out (-1.6 pp). Those living in the West had similar patterns with those in the Midwest. However, those in the South were 3.5 pp more likely to attrite due to death and 3.5 pp less likely to be non-responses; HRS's oversample of Floridians might be a plausible explanation for this phenomenon. The number of people in the household seemed to have little effect on the attrition status; having 4 + people in the household was associated with a 1.9 pp reduction of dropping out due to death and a 2.1 pp increase of reentering the study after non-death dropout. However, the number of living children matters a lot for respondents' participation in the survey. Specifically, having three/four (five or more) children increased the likelihood of being always-in by 4.0 pp (4.3 pp), and decreased the probability of being non-response by 4.2 pp (6.3 pp), relative to those without a living child.

Health Status. Health status has a pronounced effect on respondents' participation in the longitudinal survey. The worse the self-reported health status, the less likely the respondents stayed in every wave, and the more likely they left due to death. Compared to those with excellent self-reported health, those with good, fair, and poor health saw a reduction in the probability of being always-in by 4.2 pp, 7.3 pp, and 10.6 pp, and an increase in the probability of death by 3.2 pp, 8.0 pp, and 13.0 pp, respectively. Also, relative to those with "excellent" health, individuals with "very good" health were more likely to be non-response (1.3 pp), and those with "poor" health were less likely to be ever-out (-2.4 pp). Similarly, those ever diagnosed with severe and mild diseases were less likely to remain in the sample all the time (-5.3 pp and -3.9 pp) and more likely to die (7.4 pp and 4.8 pp) during the follow-up surveys. In contrast, in comparison to respondents with a normal BMI, overweight respondents had a 2.1 pp higher likelihood of being "always-in" and a 1.5 pp lower likelihood of being "non-response." Obese respondents had a 3.5 pp higher probability of being "always-in" and 2.6 pp (1.1 pp) lower probability of being "non-response" ("ever-out").

Socioeconomic status. In general, socially advantaged respondents tended to be more contactable for each follow-up survey. Compared to respondents with less than high school degree, those with an educational level of high school, some college, and college or above had a higher probability of being always-in by 2.7 pp, 3.4 pp, and 6.8 pp, respectively. Having a higher educational level also reduced the likelihood of death, though the effect of some college was not statistically significant. Those with a college degree or above were less likely to be ever-out. House ownership had similar effects as educational attainment; it increased the probability of being always-in (2.1 pp) and non-response (2.0 pp) and decreased the probability of being dead (-3.0 pp) and ever-out (-1.1). Labor force participation also affects attrition status. For instance, being partly retired, disabled, and outside of the labor force increased the likelihood of attrition due to death by 5.0 pp, 6.2 pp, and 2.4 pp. Individual earnings had a positive impact on individuals' participation in each wave of the study (approximately 1.7 pp). At the same time, total household reduced the likelihood of death by 3.1 pp (4.1 pp) for those in the third (fourth) quintile of the income distribution.

Robustness checks

Table 5-1-4 documents the results from Wald tests for combining alternatives. If none of the independent variables significantly affect the odds of alternative A and alternative B, then the two alternatives could be combined to improve efficiency. The results from Wald tests show that none of the paired alternative could be combined. As shown in the descriptive analysis, respondents in the non-response group and the ever-out group are similar, which is confirmed by the smallest χ_2 statistics in **Table 5-1-4**. However, the corresponding Wald test rejects the null hypothesis that those two alternatives are indistinguishable.

Table 5-1-5 displays result from both the Hausman test and Small-Hsiao test for the Independence of Irrelevant Alternatives (IIA) assumption. The Hausman test suggests that the IIA assumption has been violated for the two alternatives: non-response and ever-out. However, results from the Small-Hsiao test indicate no violation of the IIA assumption. This conflictive evidence is not surprising, given previous work that concludes the two tests do not provide useful information for testing IIA (Fry and Harris 1998, 1996, Cheng and Long 2007). Combining the non-response and ever-out into one category did not help the IIA assumption, as shown in the bottom of **Table 5-1-5**. To further check the IIA assumption, I reran three multinomial logistic regression models by excluding “Died”, excluding “Ever-out”, and excluding both “Died” and “Ever-out” from the outcome alternatives. As shown in **Table 5-1-6**, the point estimates from those three models are virtually the same as the full model including all alternatives, and the significance level changes only for a limited set of variables. It suggests that the odds of being non-response versus being always-in is not affected by the presence or absence of other alternatives (Died and Ever-out). As such, the IIA violation is not violated in this analysis. Similar results were found in other scenarios focusing on the odds of died versus always-in and the odds of ever-out versus always-in (not shown but available upon request).

Moreover, as argued by previous researchers, the multinomial model should only be used in cases where the alternatives “can plausibly be assumed to be distinct and weighted independently in the eyes of each decision-maker (McFadden 1973, Long and Freese 2014).” In other words, a multinomial logistic regression model works best when alternatives are dissimilar and not substitute for each other. In this study, those who left the survey and never came back should have different degrees of willingness to participate in the survey than those who reentered the survey.

Part B: A dynamic model for between-wave attrition

This part examines respondents' wave-by-wave attrition status using the same dataset as the previous analysis on the 2016 attrition status. It investigates the determinants of attrition due to either death or non-response, conditional on that individuals responded in the last survey. As such, it makes use of all the observations (person-wave records) from the data.

Sample size

As shown in **Figure 5-1-2**, a total of 458,004 person-wave records, including 35,231 respondents for 13 waves, were available in the data. However, for those who entered the survey in later waves, the records in earlier waves for those respondents were not applicable; no valid information in these records. There are 99,207 of those not relevant records in the data. Among the remaining records, there were 231,689 observations that individuals responded, 88,990 records that document respondents passed away at that wave and 38,117 records with an indicator for non-response but still alive. Thus, in theory, a maximum number of 231,689 observations could be included for the between-wave attrition analysis, as explained in the figure. After excluding those with missing covariates, 210,674 complete person-wave cases were included in the final analytic sample.

Missing values

Similar to the above cross-sectional analysis of baseline characteristics and 2016 attrition status in Part A, there was a low percentage of missing values for the included variables. The two variables with the most missing values are “the number of people in the household” and “Body Mass Index.” There are 1.15% of observations with missing values in the number of living children, and 1.42% of observations with missing values in body mass index. The percentages of missing values in all other variables are less than 1%. The analytic sample included the 231,689 complete person-wave cases, which accounts for 90.93% of the studied sample.

Descriptive statistics

Table 5-1-8 documents the number of respondents who responded, died, and did not respond but were still alive (non-response) by each cohort and over time. It also indicates the wave when respondents from each entry cohorts were recruited for the study. As an example, for the initial HRS and AHEAD cohorts, we see a dramatic drop in the number of individuals who responded from wave 1 to wave 13, and a significant increase in the number of respondents died during the follow-up period. In wave 13, there were only 5,176 respondents (40.9%) from the HRS cohort, and 442 respondents (5.4%) from the AHEAD cohort responded to the survey. A large number of respondents from the later cohorts (e.g., WB, EBB, MBB) were still in the sample in wave 13.

To make the evolution of sample composition clearer, I visualized the numbers from **Table 5-1-8** in **Figure 5-1-3** in the form of a cumulative distribution. In general, death was the primary driver of dropping out for respondents from the earlier entry cohorts (HRS, AHEAD, and CODA), while non-response was the main reason why individuals left the study for the later entry cohorts (WB, EBB, MBB, and the Minority cohort). Another takeaway message from those figures is that attrition due to death or non-response resulted in a loss of many respondents, particularly among those from the early cohorts.

Estimates from a dynamic multinomial logistic regression model

As shown in **Table 5-1-9**, the effects of independent variables from a dynamic multinomial logistic regression model are qualitatively similar to those from the cross-sectional multinomial logistic regression model in part A but differ in magnitude. I will describe the results by groups of demographics, living arrangements, health status, and socioeconomic status, with an emphasis on statistically significant effects.

Demographics. Females were 3.3 pp more likely to always stay in the sample, and 2.9 pp (0.5 pp) less likely to attrite due to death (non-response). Those born in the USA had a 1.1 pp higher likelihood of death and 1.2 pp lower probability of leaving the survey for other reasons. Black respondents were 0.9 pp less likely to respond, while respondents from other racial/ethnic groups and Hispanics were less likely to die by 0.66 pp and 0.9 pp, respectively. Racial minorities had an approximately 1.0 pp higher probability of being non-response. As shown by results from the variables for age and entry cohorts, older respondents tended to leave the survey due to death, were less likely to always remain in the survey, and less likely to be non-response. Marital status in the previous wave had significant impacts on respondents' participation in follow-up surveys. Being divorced, being widowed, and never married in the last wave led to reductions in the probability of a response by 1.5 pp, 1.8 pp, and 1.0 pp, and increased in the likelihood of death by 1.0 pp, 2.1 pp, and 1.2, respectively.

Living arrangements. Compared to living in the Northeast, living in the Midwest, South, and West was associated with 1.1 pp, 0.8 pp, and 1.4 pp higher likelihood of responding, and 1.4 pp, 1.0 pp, and 1.4 pp lower probability of non-response. Although the number of people in the household had limited effects, the number of living children had substantial impacts on respondents' attrition status. The more living children the respondents had in the previous wave, the more likely they responded, and less likely to attrite. Specifically, those with 1-2 living children, 3-4 children, and 5 + children were 1.3 pp, 2.1 pp, and 2.3 more likely to respond, 0.7 pp, 0.1 pp, and 0.1 pp less likely to die, and 0.6 pp, 1.1 pp, and 1.4 pp less likely to not respond.

Health status. In general, better health status in the previous wave had a significant positive impact on the probability of responding and a negative impact on death. Relative to those with excellent self-reported health, those with poor, fair, and good health status had an 8.2 pp, 3.5 pp, and 1.4 pp lower likelihood of responding, largely due to death. Those with severe (mild) diseases were 3.7 pp (2.0 pp) less likely to respond, and 4.0 pp (1.6 pp) more likely to die. While respondents with severe diseases were also less likely to be non-response (-0.3 pp), but those with mild diseases were more likely to be non-response (3.7 pp). In contrast, those who were obese or overweight were more likely to respond (4.1 pp, and 2.7 pp, respectively), and less likely to attrite due to death (-3.1 pp, -2.4 pp) or non-response (-1.1 pp and -0.3 pp). Besides, whether the interview was conducted by a proxy had a substantial effect on attrition; the proxy interview was associated with a 10.1 pp reduction in the probability of responding, and a 5.5 pp (4.6 pp) increase in the likelihood of attrition due to death (non-response).

Socioeconomic status. After taking wealth and income in the previous wave into account, educational attainment had little impacts on attrition; those with a college degree or above were still more likely to respond (0.9 pp) relative to those with less than high school educational level.

House ownership was associated with a higher likelihood of response (1.5 pp) and a lower likelihood of death (-1.7 pp). Not working (either partly retired, disabled, or outside labor force) was related to a lower probability of responding and a higher probability of death, especially for those outsides of the labor force. Last but not least, those with higher total household income (individual earnings) in the previous wave in quintile 3 and quintile 4 were 0.5 pp (1.8 pp) and 0.8 pp (1.5 pp) more likely to respond, and less likely to die, relative to those from the lowest quintile of household income (individual earnings).

Robustness checks

The use of the cluster option to accommodate within-respondent correlation in the above dynamic multinomial logistic regression makes it difficult to conduct a Hausman or Small-Hsiao test for the IIA assumption. Instead, I employed a “Seemingly unrelated estimation (SUEST)” to test whether coefficients on independent variables from the full model are statistically significant from those from another model excluding one type of alternative. Specifically, I made two comparisons; one compared the odds of non-response versus response from the full model to those from the restricted model that excluded death from the outcome categories, another one compared the odds of died versus response between the full model and the restricted model excluding the alternative of non-response.

Results from the above two comparisons were statistically significant at the alpha = 0.05 level. It suggests the independent variables from the full model were statistically significant from the restricted model, which indicates violations of IIA assumption. However, those tests might be misleading, especially given that the large sample size makes even minor differences statistically significant. As shown in **Table 5-1-10**, the coefficients on covariates from the full and restricted model are quite similar; most of the estimates are different in the third decimal point. Also, in this analysis, the three alternatives (response, non-response, and death) are distinct and not substitutes for each other. As such, the IIA assumption should not be a concern for estimates from the dynamic multinomial logistic model.

Additionally, I performed another sensitivity analysis using a survival analysis model. **Table 5-1-11** reports results from the three survival models. Since the interpretation of estimates from survival models relies on Hazard Ratio (HR) or Subharzard Ratio (SHR), we cannot directly compare these estimates to those from the multinomial logit model. For example, the HR is 0.74 from the Andersen-Gill model. It indicates that, on average, the hazard of attrition due to either death or non-response for women was 26% lower than that for men, all else being equal. In the following paragraphs, I will describe whether results from survival analyses are consistent with those from the dynamic multinomial logistic regression qualitatively concerning both the direction and significance.

Estimates from the Andersen-Gill model versus estimates for "responded" from the dynamic model. In the Andersen-Gill model, a coefficient less than one means the respondents with the covariate were less likely to attrite due to either death or non-response. In other words, they were more likely to respond. In this way, results from the two models are fairly consistent in identifying variables that increase/decrease the probability of responding with respect to both the direction and significance with some minor exceptions. There are four variables with consistent directions but differ in the significance level: “Other” racial groups, age 51-61, age >61, and never married. The proportional-hazards assumption was satisfied for most of the independent variables.

Estimates from the Cox model for mortality versus estimates for “died” from the dynamic model. In the Cox model, the hazard ratio represents a higher hazard of death if it is larger than one. Similarly, the direction and significance level of coefficients from the two models match perfectly. The only exception was that the variable “some college” was not statistically significant in the Cox model but significant in the dynamic model, but both were in the same direction. The proportional-hazards assumption was satisfied for most of the independent variables.

Estimates from the competing risks model versus estimates for “non-response” from the dynamic model. A subhazard ratio larger than 1 from the competing risks model indicates a higher subhazard of being non-response. Since the competing risks model only considers the first event of non-response and ignores the later non-response cases, the coefficients match well in directions of those from the dynamic model but not in the significant levels. Specifically, 1) the following variables from the competing risks model are in the same direction but differ in significance level: Hispanic, Age >61, Divorced, Other regions, two people in the household, 4 + people in the household, fair self-reported health, poor self-reported health, quintile 3 and quintile 4 of individual earnings. 2) The coefficients for the following variables from the competing risks model are in the opposite direction: female, quintile 3 and quintile 4 of household total income. 3) Lastly, the coefficients on the following variables from the competing risks model are different in both direction and significance: ages 51-61, overweight, outside of labor force, and quintile 2 of individual earnings. Importantly note that the proportional-subhazards assumption was violated for many covariates. These estimates should be interpreted with caution. The purpose of this analysis is to uncover the determinants of non-response while taking mortality into consideration. Despite of the restriction to the first non-response, it should not impose severe concerns on the overall analyses. As shown in **Figure 5-1-3**, attrition due to non-response is only responsible for a small share of dropouts and likely to be negligible.

5.2 The Association Between Education and Hospitalizations

5.2.1 Sample size

As shown in **Figure 5-2-1**, a total of 193,772 person-wave observations (29,199 unique respondents) have included in the analytic sample this analysis.

More specifically, I first merged the 2016 HRS state identifier file into the 2016 HRS tracker file, which resulted in 43,216 matched subjects. I then linked these matched subjects to the RAND HRS Longitudinal File 1992 – 2016 (v1). There was only one respondent from the RAND HRS without a record from the HRS state identifier file. This linkage led to a dataset containing 42,052 unique individuals. I further excluded those who were born outside of US (5,775 subjects) or unaware of their birthplaces (40 subjects) or missing either state-of-birth or year of birth (785 subjects). The analytic sample included 35,451 respondents corresponding to 460,863 (35,451 persons * 13 waves) observations in theory in the longitudinal data. However, a large proportion of these observations were inapplicable including observations before respondents entered the survey and observations after respondents dropped out. After excluding those irrelevant records, there were a total of 215,724 observations.

As discussed in attrition analysis, there was a small percentage of missing values in covariates. The missing values were likely to be completely random. In the main analysis, I only included complete cases (29,199 persons, and 193,772 person-waves).

5.2.2 Descriptive analyses

Descriptive statistics of individual characteristics

Since the primary variable of interest is education, **Table 5-2-1** displays the descriptive statistics of respondents' characteristics by educational attainment. The overall sample were predominately white (77.1%), non-Hispanic (94.7 %), and with childhood health status as good or above (94.7%). More than half of the included respondents were female (56.8%) and rated their childhood financial situation as pretty well off or above the average (69.9%). The average years of schooling of their better-educated parents was 10.6 years.

Respondents differ in all characteristics across educational categories. Based on the descriptive statistics, it seems that male, white, non-Hispanic, self-rated child health, self-rated childhood family financial situation, and higher educated parents were associated with better educational attainment. The percentage of men among those with college or above (49.3%) was higher than that among those with less than high school (40.9%) and those with high school or some college (43.2%). A total of 83.9% of those with higher education (college degree or above) were white respondents, whereas only 77.8% of those with a high school degree or some college experience and 66.7% of those in the group of less than high school were whites. Correspondingly, racial/ethnic minorities were more likely to have a lower level of education. As an example, 27.6% of respondents among those having education less than high school were blacks, whereas only 12.7% of people with a college degree or above were blacks. Childhood health status seemed to play a big role in determining educational attainment. More than half (62.0%) of those with a higher education degree rated their health status in childhood as excellent, while only 39.4% of respondents among the less than high school group did so. Similar pattern was found for the childhood family financial situation and parents' years of schooling. For instance, in the "college or above" group, 79.6% rated their childhood family financial situation as pretty well off or above the average, and the average parents' years of schooling was 12.6. However, in the "less than high school" group, 55.1% considered childhood family financial situation as pretty well off or above the average, and the average of their parents' education was only 8 years.

Changes in the probability of hospitalizations by education.

To explore the relationship between educational attainment and hospital admissions, I plotted the probability of ever being hospitalized against age by educational attainment. The longitudinal feature of HRS also allows me to follow this relationship as respondents age. To ensure stable estimates for hospitalizations, I only kept those education-age cells with at least 100 observations. The graph on the left of **Figure 5-2-2** shows that those with an educational level of some college are indistinguishable from those with a high school degree. Thus, it makes more sense to combine these two groups into one, as shown on the right of **Figure 5-2-2**. In general, the probability of being hospitalized increases as respondents age regardless of their educational levels. Those with education less than high school had a persistently higher likelihood of being hospitalized than those with education beyond high school, whereas those with a college degree or above had the lowest likelihood. At age 50, the probability of being hospitalized for those with less than high school, high school/some college, and college or above is 21.0%, 15.8%, and 11.7%, respectively. Although those with education less than high school still had the highest likelihood of hospitalization (38.5%) at age 78, those with an education higher than high school had a similar probability of being hospitalized; 32.3% for those with high school/some college and 31.9% for those with college and above.

The education-hospitalization gradient is fairly stable before age 64, then starts decreasing, and becomes fuzzy after age 78. For instance, the education-hospitalization gradient between those with less than a high school degree and those with a college degree or above is, on average, 9.5 percentage points (pp) across ages 48-64, 7.0 pp across ages 65-77, and 4.1 pp between ages 78-93. The difference in the probability of being hospitalized between those having less than high school and those with high school/some college is decreasing over time; the gradient is 4.3 pp during ages 48-64, 3.1pp during ages 65-77, and only 0.6 pp after age 78.

The evolution of attrition up to wave 13 by education and by entry cohort.

Figure 5-2-3 visualizes the differential attrition caused by the fact that individuals with fewer years of completed education were more likely to attrite due to both death and non-response. If the attrition rate were the same for all educational groups, the distribution of educational attainment would not change over time as most respondents were over age 50. However, the graphs in the figure demonstrate that those with higher educational level were more likely to remain in the sample, which increased their share in the sample. This unequal attrition across educational groups would lead to systematic bias (attrition bias) in the relationship between education and hospitalizations.

The attrition bias related to differential attrition by education attainment tended to be remarkable for the early cohorts (Original HRS, AHEAD, and CODA), particularly for respondents from the AHEAD cohort. The bias seemed to be minor for the later cohorts: WB, EBB, MBB, and the Minority oversample cohort. Respondents from the later cohorts also have a higher level of education than their counterparts from the early cohorts. I shall describe changes in the education distribution in more detail for the three earliest cohorts: Original HRS, AHEAD, and CODA.

Since the AHEAD covers the oldest respondents, the attrition bias is the most obvious concern. During the baseline, 36.0% of the sample had education less than high school, 34.2% had a high school degree, 17.2% had some college experience, and 12.5% had a college degree or above. However, in 2016 (wave 13), only 21.6% (14.4 pp reduction) of the AHEAD remaining sample had an education less than high school, 39.9% (5.7 pp increase) had an education of high school, 21.6% (9.1 pp increase) had education of some college, and 16.9% (4.4 pp increase) had a college education or above.

For the initial HRS cohort, the percentage of respondents with an education level of “less than high school” decreased from 21.5% in wave 1 to 16.5% in wave 13. Over the same period, the percentage of “high school” slightly increased from 39.4% to 39.8%, the proportion of “some college” increased from 20.8% to 22.2%, and lastly the percentage of “college or above” increased from 18.2% to 21.5%.

For the CODA cohort, the percentage of respondents in the remaining sample having education less than high school dropped by 6.1 pp (from 21.9% at baseline in wave 4 to 15.7% in wave 13), whereas the that of those with high school education increased by 2.1 pp (from 37.3% to 39.45%), and the percentage of those with college or above education increased by 4.0 pp (from 20.1% to 24.1%).

5.2.3 Estimates from the longitudinal analysis of education and hospitalizations

Main results

Table 5-2-2 documents the results from the main regression models. Column (1) and Column (2) reports the results from the final econometric model (unweighted OLS) without accounting for attrition bias, whereas Column (2) additionally controls for state-of-birth-specific linear time trends that capture changes in state-level policies. Results are virtually the same between the two models, which suggests state linear trends have a limited impact on the relationship between education and hospitalization after accounting for the state of birth and year of birth fixed effects. It shows that, on average, having a college degree or above was significantly associated with a 5.84 pp (95% CI, -6.53 pp to -5.15 pp) lower probability of being hospitalized, and having education of high school or some college was related to 2.05 pp (95% CI, -2.60 pp to -1.50 pp) lower probability, compared to having an education less than high school, all else equal.

As shown in Column (3) and Column (4), the effects of education on hospitalization became larger after accounting for attrition bias using a weighted OLS regression. Similarly, results were robust to state linear trends; results without adjusting for state linear trends yielded slightly smaller estimates compared to those in Column (4). Finally, Column (4) documents the final results. Compared to those having education less than high school, those with an education of high school/some college had a 3.37 pp (95% CI, -3.93 pp to -2.80 pp) lower likelihood of being hospitalized, and those with a higher education degree had an 8.39 pp (95% CI, -9.10 pp to -7.67 pp) lower likelihood of hospitalizations. Both effects were statistically significant. We can see that estimates without accounting for attrition bias in HRS would underestimate the effects of education on hospitalizations. The effect size of underestimation is not statistically ignorable since estimates from models accounting for attrition bias were significantly larger than those from models ignoring attrition bias; the 95% confidence intervals from the two types of models do not overlap.

Results of Column (4) also show that women were 2.25 pp (statistically significant) less likely than men to be hospitalized. Relative to whites, blacks had a 1.60 pp higher probability of hospitalizations, and respondents from other racial/ethnic groups had a 2.85 higher probability of being hospitalized, which were statistically significant. Although Hispanics had about a 1.2 pp lower chance of getting hospitalized, it became indistinguishable from zero after correcting attrition bias. Moreover, the worse the childhood health status, the more likely respondents were hospitalized in later life. Specifically, compared to those having excellent child health, individuals whose child health status was poor, fair, good, and very good had a significantly higher probability of hospitalization by 10.73 pp, 6.93 pp, 3.00 pp, and 0.92 pp, respectively. Respondents who grew up in low-income families were also significantly more likely (1.37 pp) to get hospitalized. Besides, parents’ education had a minor and insignificant negative effect (-0.13 pp) on hospitalization, largely due to the inclusion of childhood family financial situation and childhood health status (see **Table 5-2-3**). After removing these two variables from the model, one extra year of parents' schooling was significantly associated with a 2.0 pp reduction in the likelihood of hospitalizations. Nonetheless, since the econometric model was built up considering educational attainment as the focal independent variable, coefficients on other covariates should be interpreted with caution.

As discussed in the conceptual framework, childhood health status would lead to a reverse causality problem if the relationship between child health and educational attainment were bidirectional. To investigate this, I excluded the variable "childhood health status" from the model and reran the model. As shown in Column 2 of **Table 5-2-3**, the effect of education on hospitalizations slightly increased by about a half percentage point and remained statistically significant. It is reassuring and indicates that the potential bias due to reverse causality is likely to be negligible.

Given that descriptive analyses suggest the effect of education on hospitalizations is larger during ages 48-64, then slightly decreases during ages 65 to 78, and then becomes vague after 78, I conducted a subgroup analysis stratified by these three periods. As shown in **Table 5-2-4**, the education effects were larger for respondents younger than 78, but there was no difference between those aged under 65 and those aged 65 to 75. Thus, I conclude that, among respondents younger than 78, having education beyond college resulted in a 10 pp reduction in the probability of hospitalizations, and having education of high school degree contributed to a 5 pp reduction in the probability. All these effects were statistically significant. However, after age 78, the probability of hospitalization for those having education of high school was not significantly different from that of those having education less than high school; the estimate was -0.96 pp and not significant. Having a higher education degree still significantly reduced the chance of hospitalization by 5.5 pp for those aged over 78.

Sensitivity analyses

I conducted several sensitivity analyses to check the robustness of the estimates. First, a logistic regression model was employed because the outcome—hospitalization—is a binary variable. It is worth noting that the logistic model did not converge once the state-of-birth linear trends were included in the model. However, given the limited impact of these linear trends on estimates, as shown in the main analyses, **Table 5-2-5** reports the results from a logistic regression without controlling for state linear trends. Results are consistent with the main results from a linear regression but slightly smaller in magnitude. For instance, the weighted logistic regression model shows that an educational level of “college or above” reduced the probability of hospitalizations by 8.21 pp, and an educational level of “high school or some college” reduced the probability by 2.89 pp, both of which were statistically significant.

Treating attrition as an absorbing state yielded similar results (**Table 5-2-6**). The effects of education on the adverse outcome (either hospitalized or died) were about one-percentage-point larger than those for hospitalizations only; the coefficient on “college or above” is -0.0911 and on “high school/some college” is -0.0394, all are statistically significant (**Table 5-2-7**). Extracting hospitalization from the terminal conditions that cause imminent death, **Table 5-2-8** reports the results of using hospitalization two years before death and by age subgroups. The estimates from the overall sample and the subsample for those aged over 78 are in line with the main results from the final model. However, the estimates for those under 78 are approximately one-percentage-point lower compared to those from the final model. More specifically, the effect of a college degree or above on hospitalization two years before death is -9.47 (-10.30 pp for the overall hospitalization), and the effect of education of high school or some college is -4.50 pp (about -5.0 pp for the overall hospitalization).

Finally, prior models restricting to complete cases dropped 6,252 (17.6%) respondents and 21,952 observations (10.2%) out of analyses. As described in **Figure 5-2-1**, complete case analyses only included 29,199 persons (193,772 person-years) out of 35,451 persons (215,724 person-years) in the analytic sample. To check whether missing values would lead to systematic bias, I performed an analysis using multiple imputations. As shown in **Table 5-2-9**, the unweighted estimates are -0.0216 for high school/some college and -0.0599 for college or above, and the weighted estimates for “high school/some college” and “college or above” are -0.0349 and -0.0851, respectively. All of these estimates are quite similar to those from the main results in terms of both magnitude and significance.

5.3 Secondary Schooling and Hospitalizations: Instrumental Variables Analysis

Sample size and descriptive statistics

The sample used in this analysis builds upon the final regression sample (denoted as “Sample A” in **Figure 5-3-1**) from the association study in Section 5.2. I then matched each individual to the compulsory schooling laws that were in place in their state-of-birth when they were 14 years old. In the process of matching, I excluded those born in Hawaii and Alaska since the two states were not then part of the Union. I also restricted the birth cohorts to 1905-1959 due to the availability of data in compulsory schooling laws and school quality measures, as illustrated in the “Data Source” section. It led to a sample of 30,898 persons and 207,172 person-years. I further restricted the sample to white respondents who had education less than or equal to 12 years and without missing values in all variables included in the regression. The final regression sample included 10,724 persons and 82,606 person-years.

Table 5-3-1 documents the summary statistics of the included sample. The overall average rate of hospitalization in the baseline was 17%. The respondents’ average years of education were 10.9, which reflects a skewed distribution towards the higher end. More than half of the respondents were female (58.9%), rated their childhood health status beyond “Very Good” (74.7%), and had an above-average or well-off family financial situation during childhood (65.7%). The average of their parents’ highest education was 9.3 years. Approximately half of them were born between 1921 and 1940.

For instruments, 63% of respondents were exposed to continuation laws. The average number of years required for respondents when they were 14 years based on compulsory attendance laws, child labor laws, and accumulative required schooling was 9.6, 7.8, and 8.2, respectively. Respondents had an average pupil-teacher ratio of 27.13, an average length of the school term of 175.5, and a group of teachers with $2928.29 average salary during the time when they were in school.

Description of compulsory schooling laws and quality of schooling

As shown in **Table 5-3-2**, the correlation matrix, years of compulsory schooling, constructed in different ways, were moderately correlated, and school quality measures were also moderately correlated. The years of compulsory schooling had a relatively low correlation with school quality measures. It is reassuring and suggests that the inclusion of all these variables in one regression model will not lead to severe multicollinearity issues. It also indicates that each of these constructed measures carries unique information about compulsory schooling and school quality measures.

Specifically, from the table, we see that years of compulsory schooling required by child labor laws had a relatively high correlation with those required by attendance laws (0.53), and dropout age (0.46). The correlation between schooling by child attendance laws and required schooling was as high as 0.57. Required schooling was also related to school quality measures with a correlation coefficient equal to about 0.37. In addition, the pupil-teacher ratio was negatively correlated with school length of term (r=-0.62) and teachers' wages (-0.52).

Table 5-3-3 addresses the concern that a large proportion of variations in these laws and quality measures could be explained by state-of-birth fixed effects, year of birth fixed effects, and in particular, the region or state trends. The first column displays the mean and standard deviation (SD) of the raw numbers. I then purged out state-of-birth fixed effects and year of birth fixed effects by computing the residuals from regressing raw numbers on these fixed effects. The remaining columns report the mean, SD, and percent reductions in SD compared to raw numbers, based on the computed residuals. I further added region/state linear trends into the regression models and calculated the same statistics in the remaining columns.

In general, variations in compulsory schooling laws were relatively robust to these fixed effects and the inclusion of region or state trends; there was still more than half of the variation left. For schooling by child labor laws and dropout age, only one-third of the variation was explained by these fixed effects and trends. Other aspects of compulsory laws still had about half of the variations left. However, there were almost no variations left for school quality measures after accounting for state linear trends. The variation of pupil-teacher ratio, length of the term, and teachers' wages dropped by 82.3%, 74.2%, and 87.4%, respectively. These results thus cast doubt on the validity of using school quality measures as instruments. Even for compulsory schooling laws, we are better to be wise in adding those fixed effects and trends in regression models.

Figure 5-3-2 shows the trends in the average years of compulsory schooling by educational laws overtime when respondents were 14 years old. In general, states required more years of education towards the end of the period. However, this was not always the case; there were some ups and downs over the period. Importantly, there were decreases in required schooling around the year 1939 (birth cohort 1925) and 1959 (birth cohort 1945). These non-linear trends might explain why there was still some variation left after accounting for state-specific linear trends, which laid the foundation for the use of these variables as instruments.

Effects of compulsory schooling laws on education

As preliminary evidence, I plotted the average education over laws or required years of schooling by-laws, by categories of educational attainment (see **Figure 5-3-3**). These graphs imply that these laws were only effective in improving education for those with lower education levels, particularly those having education less than high school. There was little effect on those with completed education as “some college” and “college or above.” For those in the lower end distribution of education, average education was higher for those in states that required more years of compulsory schooling, and higher within states, after laws forced more years of schooling. Besides, these effects were not linear; the magnitude decreased after years of compulsory schooling reach eight. Thus, we might need to include a non-linear transformation of these laws as instruments for education.

I then turned to regression models to add further controls and assess the strength of using them as instruments. The specification of the econometric model was the same as the first stage regression of the P2SLS weighted by attrition weights and clustered at respondent levels. I estimated several models varying in the addition of state, year fixed effects, and region, state trends. **Table 5-3-4** documents the results. Column 1 is the basic model without controlling for any state and year of birth effects, it shows that exposure to continuation laws was associated with 0.420 more years of education, one additional required year of schooling by child labor laws, attendance laws, and accumulative laws increased education by 0.086, 0.037, and 0.065, respectively. All of these coefficients were statistically significant. As adding state fixed effects (Column 2), year of birth effects (Column 3), state and year effects (Column 4), regional trends (Column 5), and state-level trends (Column 6), most of these coefficients gradually shrunk to near-zero and loss significance with some noteworthy exceptions. First, the inclusion of regional trends seemed to have a substantial effect on these coefficients. Second, the accumulative required schooling had a robust effect on years of education. More importantly, **Table 5-3-4** also documents the F statistics from a joint test of the compulsory schooling laws, which are commonly used to evaluate whether instruments are weak. The F statistic decreased from 26.08 to a very low level (0.59 in Column 5 and 1.55 in Column 6), suggesting an apparent weak instrument problem if trends were included in the regression model. This comes as no surprise, given the results in **Table 5-3-3**.

POLS results of the effect of education on hospitalization

As a benchmark, I pooled all the data and fitted an OLS regression. Note that the OLS estimator only indicates a correlation relationship. **Table 5-3-5** documents the results. I experimented with different model specifications with and without adjusting for attrition bias and state-level trends. The estimators are fairly robust. After controlling for state-of-birth fixed effects, year-of-birth fixed effects, and state-specific linear time trends, the results in Column 4 suggests that one more year of completed schooling was associated with about a 0.8 percentage point reduction in the probability of hospitalization, which was statistically significant. It is worth noting that a large majority of coefficients on state-of-birth dummies, year-of-birth dummies, and state-level trends were not statistically significant (not shown). Given there were few HRS respondents born from some small states, the inclusion of all these dummies might cause overfitting concerns. It thus motivated another analysis to select a parsimonious set of controls.

Assuming years of completed education is exogenous, I used the PDS methodology (step 2 and step 3) described in the Methods section to select an appropriate set of controls with a high correlation with both hospitalization and education. In this way, the omitted variables only have mild or near-zero correlation, and the omitted variable bias is thus negligible. **Table 5-3-6** reports the results without attrition weights in Column 1 and with these weights in Column 2. For the weighted analysis that accounts for attrition bias, only Texas was included for state-of-birth fixed effects. The year of birth fixed effects included 1913, 1918, 1924, 1942, 1949, 1955, and 1959. Note that no state-specific linear time trend was selected by the PDS method. The coefficients on years of schooling are similar to those from OLS regressions. It shows that one additional year of schooling was associated with about 1 percentage point decrease in the probability of hospitalization.

P2SLS results of the effect of education on hospitalization

While there is evidence for weak instruments in **Table 5-3-4**, it is unclear which sets of instruments are the optimal instruments, especially given the non-linear effects of laws on average education shown in **Figure 5-3-2**. To probe into this issue, I estimated five P2SLS regression models with a variety of instruments while controlling for individual characteristics, state-of-birth fixed effects, year-of-birth fixed effects, and state-specific linear time trends (see **Table 5-3-7**). Column 1, Column 2, and Column 3 report results from P2SLS using a previously used set of instruments. Those include CA7, CA8, CA9, CA10, CL7, CL8, CL9 (Acemoglu and Angrist 2000), Continuation laws, CL7, CL8, CL9 (Lleras-Muney 2005), and RS7, RS8, RS9 (Stephens Jr and Yang 2014). However, the F statistics on instruments from the above models are pretty small and indicate weak instrument problems (Bound, Jaeger, and Baker 1995). The effect of years of schooling from these models were not precise and had a wide confidence interval, which led to inconclusive results.

Column 4 shows the results when all potential instruments are included. The full set of instruments included continuation laws, years of compulsory schooling required by child labor laws (CL_{sct}), years of compulsory schooling required by child labor laws (CA_{sct}), leaving age (LA_{sct}), years of required schooling (RS_{sct}), CL7, CL8, CL9, CA7, CA8, CA9, CA10, RS7, RS8, and RS9. The use of the full set of instruments indeed improved the efficiency of the IV estimator, as indicated by the reduced standard error on educational attainment, though it remained insignificant. However, it has an even smaller F statistic on instruments, and the IV estimator is very likely to be biased.

I then applied the LASSO regression to select instruments but without selections in controls (Belloni et al. 2012). The selected instruments include continuation laws, CL_{St}, CA_{St}, CL7, CA7, and RS6. However, as seen in Column 5, this set of selected instruments was not optimal and had the lowest F statistic. This finding was not unexpected. Recall the results from **Table 5-3-3**, where there was a substantial loss of variation after accounting for state-specific linear trends. Thus, the LASSO regression was unable to choose the optimal instruments in this case. It provides further evidence justifying the necessity of selections on controls, particularly on fixed effects and trends.

LASSO-IV results of the effect of education on hospitalization

Finally, I selected both instruments and controls using the PDS methodology. The selected instruments included continuation laws, CL_{st}, CA_{st}, RS_{st}, and $CL8$. The selected state-of-birth fixed effects and cohort fixed effects, displayed in **Table 5-3-8**, included Texas, 1912, 1918, 1924, 1942, 1949, 1955, and 1959. The LASSO model did not pick up other fixed effects due to their limited empirical impacts on years of education and hospitalizations, though they were conceptually relevant. In other words, the PDS LASSO model led to a parsimonious model where Texas alone versus other states was sufficient to identify the state variation. This is part of the approximate sparsity assumption of the PDS LASSO method, which will be examined next. Those consistent results from a POLS regression model with all fixed effects (Column 4 of **Table 5-3-5**) and with selected fixed effects (Column 3 of **Table 5-3-8**) further confirm that it is sufficient to only include selected instruments and controls. Another important concern is the exclusion restriction assumption for instruments that instruments should only exhibit effects on hospitalizations through educational attainment, which will be tested after a GMM-IV model below.

Using the selected instruments and controls, the P2SLS estimator (Post-LASSO estimator) suggests that an increase in one year of schooling lowered the probability of two-year hospitalization by 6.5 percentage points, which was statistically significant. The IV estimator on education from the "Post-Regularization" methodology (using Post LASSO Orthogonalized variables) was -0.097, which was also statistically significant. The first stage F statistic is 20.61, far beyond the critical value for weak instruments. The Super Score test (Belloni et al. 2012) rejects weak identification and supports the sparsity assumption.

Another key assumption for instruments to be valid is that instruments affect outcome only through the focal independent variable. I performed a test for overidentifying restrictions. Due to the use of inverse probability weighting, it was not straightforward to do the test for P2SLS and Post LASSO-IV. Instead, I reestimated the IV approach based on a GMM method, which reports the Hansen J statistic for overidentifying restrictions. The GMM-IV estimator is virtually the same as the P2SLS estimator. The Hansen J statistic is 6.29 (p=0.1785), suggesting the instruments satisfy the overidentification restrictions. The GMM-IV approach also conducts an under-identification test; the Kleibergen-Paak rk LM statistic is 76.86 ($p < 0.001$) and thus the model is properly identified. In addition, GMM-IV tests for weak-instrument-robust inference using an Anderson-Rubin Wald test; the χ_2 statistic for this test is 40.88 ($p<0.001$) and provide consistent evidence that the set of instruments is not subject to weak-instrument bias.

The IV estimator -0.065 (95% CI: -0.091 to -0.039) is much larger than that from the OLS regression using the same set of controls in Column 3 (-0.011, 95% CI: -0.014 to -0.007).

Sensitivity Analyses

As shown in the previous association study, the effect of education on hospitalization is different at different stages of life. To probe into this hypothesis, I computed the Post-LASSO IV estimator by two age groups: less than 78 and higher/equal to 78 (see **Table 5-3-9**). Consistent with results from the association study, an additional year of schooling significantly lowered the probability of hospitalization by 6.6 percentage points when respondents were younger than 78. After age 78, the effect of education (here is secondary schooling) on hospitalization is distinguishable from zero. The selected instruments are documented below **Table 5-3-9**. The equations were properly identified, and instruments satisfy the overidentification restrictions.

The quality of school measures was excluded from the proper analyses due to concerns over its validity. Nonetheless, I did another sensitivity analysis by including them in the selection pool of instruments. **Table 5-3-10** reports the results. The results show that one additional year of schooling significantly reduces the likelihood of hospitalization by approximately 8 percentage points. However, the selected instruments (pupil-teacher ratio and length of term) do not satisfy the overidentification restrictions; the Hasen J statistic from the GMM-IV model is 33.56 ($p < 0.001$). It confirms again that school quality measures should not be used as instruments.

The last concern is about the attrition weights. The main analyses adopted attrition weights developed in section 5.2 of the association study based on the full sample. Given that white persons comprise the majority of the HRS sample, it should not add too much noise to the IV estimators. To address this concern, I reconstructed the weights based on the regression sample in the IV analysis and reestimated the IV regression models. I got similar results with a slight increase in the magnitude; one more year of schooling lowers the probability of two-year hospitalization by about 8 percentage points (see **Table 5-3-11**). There is no empirical evidence signaling the selected instruments are invalid.

Results Tables and Figures

Results: 5.1 Attrition Analyses of the Health and Retirement Study

Figure 5-1-1. Sample size flowchart for baseline characteristics and 2016 attrition status

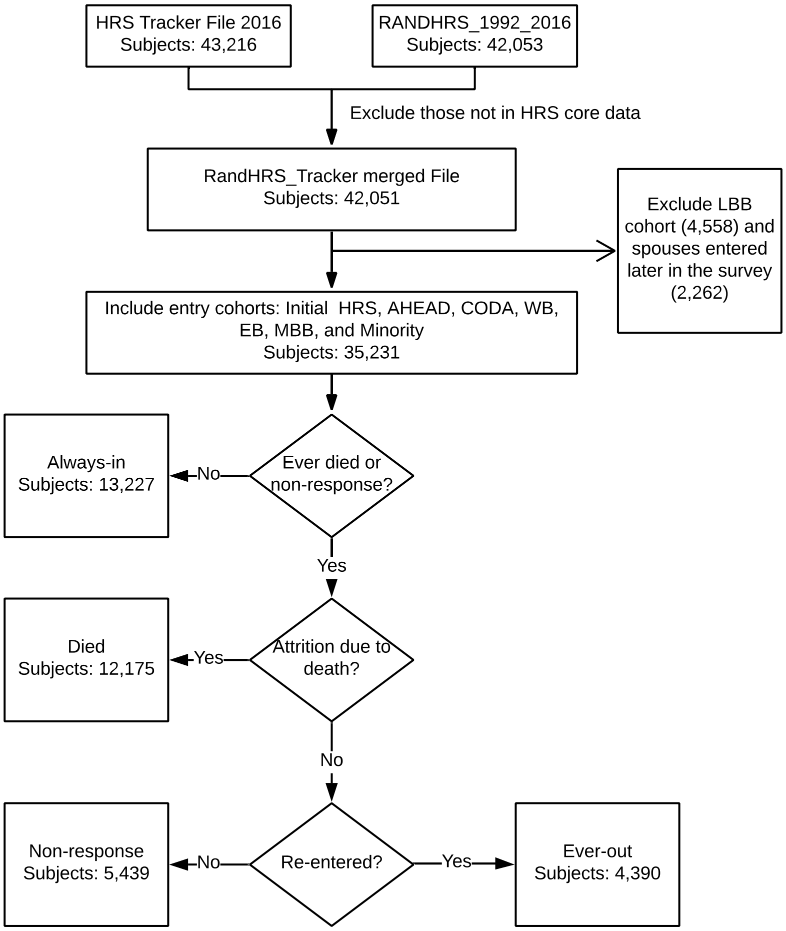


Table 5-1-1. Distribution of missing values in baseline characteristics

	Missing	Total	Percent Missing
Female	0	35,231	0.00
US born	1	35,231	0.00
Age	1	35,231	0.00
Race	61	35,231	0.17
Hispanic	34	35,231	0.10
Marital status	14	35,231	0.04
Census region	4	35,231	0.01
Number of people in HH	0	35,231	0.00
Number of kids alive	235	35,231	0.67
Proxy interview	0	35,231	0.00
Self-reported health	9	35,231	0.03
Ever had severe disease	0	35,231	0.00
Ever had mild disease	1	35,231	0.00
Body mass index	454	35,231	1.29
Educational attainment	1	35,231	0.00
House ownership	0	35,231	0.00
Labor force status	0	35,231	0.00
Individual earnings	0	35,231	0.00
HH total income	0	35,231	0.00
Any missing	774	35,231	2.20

Notes: The analytic sample only includes the 34,457 complete cases, which accounts for 97.8% of the studied sample.

Table 5-1-2. Descriptive statistics of respondents’ characteristics by 2016 attrition status

	HRS attrition status				
	Total (%)	Always in (%)	Died (%)	Non-Response (%)	Ever out (%)
Female					
Male	43.7	40.5	47.1	44.3	43.2
Female	56.3	59.5	52.9	55.7	56.8
US born					
Foreign born	12.1	13.6	8.0	13.9	16.6
US born	87.9	86.4	92.0	86.1	83.4
Age					
< 51	12.3	19.0	3.3	13.9	15.9
51-61	53.7	71.3	29.7	59.6	60.8
> 61	34.0	9.8	67.0	26.5	23.3
Race					
White	75.9	71.8	80.8	78.7	71.3
Black	17.9	19.6	16.3	14.6	21.3
Other	6.2	8.6	2.9	6.7	7.4
Hispanics					
Not Hispanic	89.5	86.7	93.6	90.3	85.5
Hispanic	10.5	13.3	6.4	9.7	14.5
Marital Status					
Married	71.0	76.7	62.3	73.8	75.4
Divorced	11.6	13.7	9.2	11.5	12.2
Widowed	13.0	4.5	24.9	9.7	9.1
Never married	4.3	5.1	3.6	5.0	3.3
Census region					
Northeast	17.6	16.2	16.9	21.0	19.3
Midwest	23.2	23.4	24.3	22.1	21.1
South	40.9	39.1	43.1	38.3	43.1
West	18.3	21.3	15.6	18.5	16.5
Num of people in HH					
1	17.2	11.9	25.1	15.8	13.0
2	45.5	41.3	50.9	45.8	42.9
3	18.4	22.1	13.4	19.5	19.9
4+	18.9	24.8	10.6	19.0	24.2
Num of kids alive					
No child	9.3	8.0	10.7	11.0	7.2
One/two kids	38.5	39.0	37.3	40.6	37.6
Three/four kids	33.6	35.7	31.5	33.3	33.2
Five/more kids	18.7	17.3	20.5	15.1	21.9
Proxy Interview					
Not proxy	94.0	97.4	91.9	92.8	91.7
Proxy	6.0	2.6	8.1	7.2	8.3
Self-reported health					
Excellent	17.0	21.3	11.1	18.7	18.9
Very good	26.4	30.3	20.6	29.5	26.8
Good	29.1	28.4	29.0	29.2	31.1
Fair	18.2	14.8	23.6	15.6	16.7
Poor	9.3	5.2	15.7	6.9	6.5
Ever had severe disease					
No severe diseases	74.1	84.5	58.1	78.9	81.9
Ever had severe diseases	25.9	15.5	41.9	21.1	18.1
Ever had mild diseases					
No mild diseases	53.4	58.8	44.3	57.2	58.3

Ever had mild diseases	46.6	41.2	55.7	42.8	41.7
Body Mass Index					
Normal	35.7	30.1	41.0	37.7	35.5
Overweight	38.7	38.4	37.9	39.5	40.6
Obese	25.6	31.5	21.1	22.9	23.8
Education					
Lt high-school	26.7	16.2	38.4	23.1	29.8
High school	33.6	33.2	32.7	35.4	34.9
Some college	21.4	25.9	16.6	21.4	21.1
College and above	18.3	24.7	12.3	20.0	14.2
House ownership					
Not house owner	23.1	21.7	25.6	19.9	24.0
House owner	76.9	78.3	74.4	80.1	76.0
Labor force status					
Working	48.6	66.4	24.0	54.4	57.3
(partly) Retired	28.5	16.8	46.5	22.6	20.1
Disabled	3.4	3.2	4.0	2.7	3.7
Outside labor force	19.5	13.6	25.6	20.2	18.9
Individual earnings					
Quintile 1	47.3	30.9	70.0	41.6	39.4
Quintile 2	7.5	7.0	6.4	7.8	11.7
Quintile 3	17.6	19.9	13.3	18.9	20.8
Quintile 4	27.7	42.1	10.3	31.7	28.1
HH total income					
Quintile 1	25.7	18.1	35.6	21.9	25.1
Quintile 2	25.0	19.5	31.4	23.4	25.3
Quintile 3	23.8	26.3	19.9	26.2	23.7
Quintile 4	25.6	36.0	13.0	28.5	25.9

Notes: Analytic sample includes 34,457 unique individuals with complete cases.

Table 5-1-3. Association between baseline characteristics and 2016 attrition status

	Always-in	Died	Non-response	Ever-out
Female	0.0751***	-0.0817***	0.0039	0.0027
	(0.0049)	(0.0046)	(0.0043)	(0.0040)
US born	0.0124	0.0554***	-0.0355***	-0.0323***
	(0.0081)	(0.0077)	(0.0077)	(0.0068)
Age (ref: <51)				
51-61	-0.0210**	0.0649***	-0.0158*	-0.0280***
	(0.0074)	(0.0085)	(0.0068)	(0.0069)
> 61	-0.1396***	0.1951***	0.0103	-0.0658***
	(0.0141)	(0.0139)	(0.0130)	(0.0120)
Race (ref: White)				
Black	-0.0171**	-0.0078	-0.0272***	0.0521***
	(0.0065)	(0.0061)	(0.0054)	(0.0059)
Other	-0.0084	-0.0318**	0.0104	0.0298***
	(0.0099)	(0.0110)	(0.0095)	(0.0086)
Hispanics	0.0267**	-0.0413***	-0.0283***	0.0429***
	(0.0091)	(0.0088)	(0.0072)	(0.0082)
Marital Status(ref: Married)				
Divorced	-0.0178*	0.0101	-0.0013	0.0089
	(0.0090)	(0.0089)	(0.0082)	(0.0076)
Widowed	-0.0533***	0.0545***	0.0051	-0.0064
	(0.0110)	(0.0090)	(0.0094)	(0.0079)
Never married	-0.0190	0.0230	0.0040	-0.0081
	(0.0133)	(0.0133)	(0.0119)	(0.0110)
Region (ref: Northeast)				
Midwest	0.0341***	0.0199**	-0.0379***	-0.0162**
	(0.0071)	(0.0064)	(0.0063)	(0.0057)
South	0.0045	0.0345***	-0.0350***	-0.0040
	(0.0064)	(0.0058)	(0.0058)	(0.0052)
West	0.0310***	0.0197**	-0.0302***	-0.0205***
	(0.0075)	(0.0070)	(0.0068)	(0.0060)
Num people in HH (ref: 1)				
2	0.0052	0.0022	-0.0098	0.0024
	(0.0093)	(0.0079)	(0.0082)	(0.0070)
3	0.0101	-0.0043	-0.0096	0.0039
	(0.0101)	(0.0089)	(0.0090)	(0.0077)
4+	0.0104	-0.0194*	-0.0119	0.0209**
	(0.0103)	(0.0093)	(0.0093)	(0.0081)
Num living children (ref: No living child)				
One/two children	0.0173	-0.0042	-0.0298***	0.0167*
	(0.0092)	(0.0083)	(0.0086)	(0.0072)
Three/four children	0.0400***	-0.0043	-0.0420***	0.0063
	(0.0095)	(0.0084)	(0.0088)	(0.0073)
Five/more children	0.0433***	0.0066	-0.0631***	0.0132
	(0.0103)	(0.0090)	(0.0093)	(0.0078)
Proxy interview	-0.1176***	0.0028	0.0612***	0.0536***
	(0.0104)	(0.0088)	(0.0100)	(0.0087)
Self-reported health (ref: Excellent)				
Very good	-0.0216**	0.0080	0.0132*	0.0005
	(0.0066)	(0.0066)	(0.0059)	(0.0054)
Good	-0.0423***	0.0318***	0.0051	0.0054
	(0.0069)	(0.0066)	(0.0060)	(0.0056)
Fair	-0.0727***	0.0803***	-0.0026	-0.0049
	(0.0084)	(0.0079)	(0.0073)	(0.0067)
Poor	-0.1058***	0.1297***	0.0004	-0.0243**
	(0.0110)	(0.0102)	(0.0097)	(0.0082)

Ever had severe diseases	-0.0534***	0.0738***	-0.0044	-0.0160***
	(0.0059)	(0.0053)	(0.0051)	(0.0046)
Ever had mild diseases	-0.0386***	0.0477***	-0.0025	-0.0066
	(0.0049)	(0.0044)	(0.0043)	(0.0039)
Body mass index (ref: Normal)				
Overweight	0.0209***	-0.0150**	-0.0064	0.0005
	(0.0053)	(0.0048)	(0.0047)	(0.0042)
Obese	0.0350***	0.0013	-0.0256***	-0.0107*
	(0.0060)	(0.0056)	(0.0052)	(0.0047)
Education (ref: Lt High School)				
High school	0.0265***	-0.0149**	-0.0027	-0.0090
	(0.0066)	(0.0055)	(0.0058)	(0.0051)
Some college	0.0336***	-0.0112	-0.0135*	-0.0089
	(0.0074)	(0.0066)	(0.0064)	(0.0059)
College and above	0.0677***	-0.0158*	-0.0114	-0.0406***
	(0.0082)	(0.0074)	(0.0071)	(0.0061)
House owner	0.0208***	-0.0295***	0.0200***	-0.0113*
	(0.0060)	(0.0056)	(0.0051)	(0.0049)
Labor force (ref: Working)				
(partly) Retired	-0.0068	0.0497***	-0.0170*	-0.0260***
	(0.0076)	(0.0075)	(0.0067)	(0.0062)
Disabled	-0.0267*	0.0624***	-0.0348**	-0.0008
	(0.0134)	(0.0129)	(0.0112)	(0.0115)
Outside labor force	-0.0395***	0.0242**	0.0239**	-0.0086
	(0.0077)	(0.0080)	(0.0074)	(0.0065)
Earning (ref: Quintile 1)				
Quintile 2	0.0183*	-0.0005	-0.0144	-0.0034
	(0.0093)	(0.0087)	(0.0083)	(0.0072)
Quintile 3	0.0158*	-0.0084	-0.0012	-0.0063
	(0.0075)	(0.0075)	(0.0068)	(0.0062)
Quintile 4	0.0187*	-0.0248**	0.0060	0.0002
	(0.0079)	(0.0086)	(0.0074)	(0.0069)
Total HH income (ref: Quintile 1)				
Quintile 2	-0.0036	-0.0088	0.0025	0.0099
	(0.0073)	(0.0062)	(0.0062)	(0.0054)
Quintile 3	0.0094	-0.0306***	0.0126	0.0085
	(0.0080)	(0.0073)	(0.0070)	(0.0062)
Quintile 4	0.0132	-0.0408***	0.0067	0.0208**
	(0.0090)	(0.0085)	(0.0079)	(0.0073)
Entry cohorts (ref: initial HRS)				
AHEAD	-0.1415***	0.1794***	-0.0451***	0.0072
	(0.0111)	(0.0150)	(0.0108)	(0.0138)
CODA	0.1059***	-0.0418**	-0.0553***	-0.0088
	(0.0174)	(0.0134)	(0.0111)	(0.0147)
WB	0.1563***	-0.1233***	-0.0067	-0.0263***
	(0.0104)	(0.0107)	(0.0090)	(0.0078)
EB	0.2409***	-0.2316***	0.0368***	-0.0462***
	(0.0103)	(0.0091)	(0.0093)	(0.0070)
MBB	0.3906***	-0.2976***	0.0262**	-0.1192***
	(0.0105)	(0.0079)	(0.0095)	(0.0054)
Minority	0.4481***	-0.3016***	-0.0167	-0.1298***
	(0.0102)	(0.0074)	(0.0092)	(0.0050)
Observations	34,457	34,457	34,457	34,457

Notes: Estimates are average marginal effects from the cross-sectional multinomial logit model that examines the baseline determinants of 2016 attrition status. * p<0.05, ** p < 0.01, ** p<0.001.

Table 5-1-4. Wald tests for combining alternatives in the cross-sectional multinomial logistic regression for baseline determinants of 2016 attrition.

	chi2	df	P>chi2
Always-in & Died	7751.187	47	<0.001
Always-in & Non-response	1942.848	47	<0.001
Always-in & Ever-out	2392.883	47	<0.001
Died & Non-response	3734.828	47	<0.001
Died & Ever-out	3257.016	47	<0.001
Non-response & Ever-out	644.377	47	<0.001

Notes: The null hypothesis for the Wald test is that all coefficients except intercepts associated with a given pair of alternatives are 0 (i.e., alternatives can be combined). A rejection of the hypothesis suggests that the two alternatives should not be combined.

Table 5-1-5. Hausman test for IIA in the cross-sectional multinomial logistic regression for baseline determinants of 2016 attrition.

	chi2	df	P>chi2
Hausman Test			
Always-in	-17.834	96	.
Died	65.193	96	0.993
Non-response	201.671	96	<0.001
Ever-out	122.05	96	0.038
Small-Hsiao Test			
Always-in	100.83	96	0.348
Died	112.762	96	0.116
Non-response	93.797	96	0.545
Ever-out	99.874	96	0.373
Hausman Test After Combing Non-Response and Ever-Out			
Always-in	25.698	48	0.997
Died	-28.538	48	.
Non-response & Ever-out	97.662	48	<0.001
Small-Hsiao Test After Combing Non-Response and Ever-Out			
Always-in	41.167	48	0.747
Died	54.26	48	0.248
Non-response & Ever-out	73.782	48	0.010

Notes: Significant p values indicate that the independence of irrelevant alternatives (IIA) assumption has been violated.

Table 5-1-6. IIA tests for non-response versus always-in based on the cross-sectional multinomial logistic regression

	(1) Full Model	(2) Exclude Died	(3) Exclude Ever-out	(4) Exclude Died and Ever-out
Non-response				
Female	0.749***	0.741***	0.750***	0.748***
US born	0.803***	0.776***	0.816***	0.789***
Age (ref: <51)				
51-61	0.995	0.994	1.000	0.997
> 61	1.878***	1.884***	1.895***	1.872***
Race (ref: White)				
Black	0.879***	0.877***	0.871***	0.860***
Other	1.077***	1.043*	1.087***	1.047*
Hispanic	0.731***	0.736***	0.722***	0.721***
Marital Status(ref: Married)				
Divorced	1.063**	1.076***	1.058**	1.074***
Widowed	1.295***	1.390***	1.296***	1.434***
Never married	1.111***	1.129***	1.095**	1.124***
Region (ref: Northeast)				
Midwest	0.714***	0.722***	0.717***	0.724***
South	0.815***	0.830***	0.807***	0.824***
West	0.759***	0.767***	0.765***	0.778***
Num people in HH (ref: 1)				
2	0.924***	0.932***	0.928***	0.945**
3	0.906***	0.919***	0.912***	0.933**
4+	0.885***	0.893***	0.894***	0.913***
Num living children (ref: No living child)				
One/two children	0.793***	0.787***	0.787***	0.787***
Three/four children	0.676***	0.668***	0.664***	0.650***
Five/more children	0.579***	0.566***	0.572***	0.556***
Proxy	2.237***	2.338***	2.218***	2.354***
Self-reported health (ref: Excellent)				
Very good	1.173***	1.168***	1.178***	1.172***
Good	1.217***	1.195***	1.243***	1.220***
Fair	1.330***	1.310***	1.350***	1.316***
Poor	1.586***	1.526***	1.621***	1.535***
Ever had severe diseases	1.230***	1.202***	1.222***	1.182***
Ever had mild diseases	1.161***	1.145***	1.167***	1.145***
Body mass index (ref: Normal)				
Overweight	0.884***	0.869***	0.883***	0.867***
Obese	0.748***	0.731***	0.751***	0.734***
Education (ref: Lt High School)				
High school	0.886***	0.901***	0.878***	0.887***
Some college	0.807***	0.821***	0.797***	0.803***
College and above	0.729***	0.740***	0.723***	0.727***
House owner	1.043**	1.040**	1.036**	1.041**
Labor force (ref: Working)				
(partly) Retired	0.941***	0.949**	0.955**	0.986
Disabled	0.891***	0.898***	0.893***	0.913**

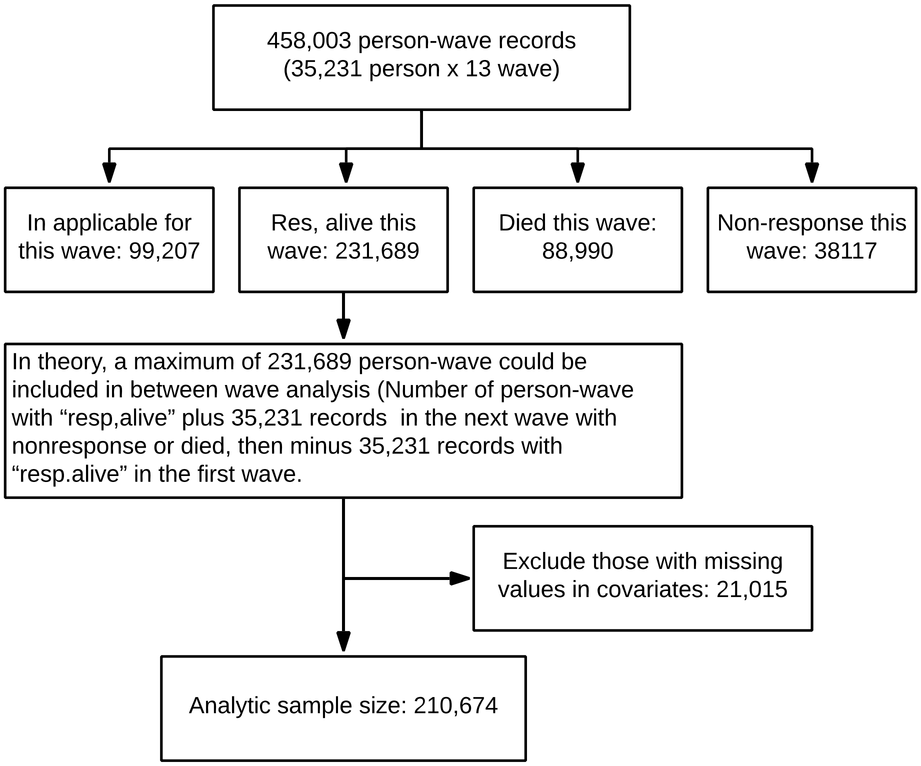
Outside labor force	1.344***	1.355***	1.320***	1.314***
Earning (ref: Quintile 1)				
Quintile 2	0.850***	0.862***	0.848***	0.856***
Quintile 3	0.934***	0.943***	0.929***	0.929***
Quintile 4	0.957*	0.953**	0.952**	0.942***
Total HH income (ref: Quintile 1)				
Quintile 2	1.024	1.030	1.026	1.036*
Quintile 3	1.030	1.042*	1.024	1.036*
Quintile 4	0.973	0.996	0.964	0.987
Entry cohorts (ref: initial HRS)				
AHEAD	1.673***	1.712***	1.651***	1.709***
CODA	0.477***	0.471***	0.471***	0.462***
WB	0.566***	0.570***	0.563***	0.566***
EB	0.559***	0.564***	0.554***	0.561***
MBB	0.385***	0.391***	0.384***	0.392***
Minority	0.273***	0.278***	0.272***	0.280***
N	447,941	291,590	391,794	235,443
Log Likelihood	-450141.411	-256288.118	-291913.446	-127499.268

Notes: Estimates are relative risk ratio (RRR) comparing those with 2016 attrition status as “non-response” to those as “always-in” from a multinomial logistic regression model.

Column 1 includes all four alternatives (always-in, died, non-response, and ever-out); Column 2 excluded “died” from the alternatives; Column 3 excluded “ever-out” from the alternatives; and Column 4 excluded both “died” and “ever-out” from the alternatives.

* $p<0.05$, ** $p < 0.01$, ** $p<0.001$.

Figure 5-1-2. Sample size flowchart for the between-wave attrition analysis


458,003 person-wave records
(35,231 person x 13 wave)
In applicable for this wave: 99,207
Res, alive this wave: 231,689
Died this wave: 88,990
Non-response this wave: 38117
In theory, a maximum of 231,689 person-wave could be included in between wave analysis (Number of person-wave with "resp,alive" plus 35,231 records in the next wave with nonresponse or died, then minus 35,231 records with "resp.alive" in the first wave.
Exclude those with missing values in covariates: 21,015
Analytic sample size: 210,674

Notes: This figure builds upon Figure 5-1-1.

Table 5-1-7. Distribution of missing values in observations (person-years) from 1992-2016

Variable	Missing	Total person-wave	Percent Missing
Female	0	231,689	0.00
US born	1	231,689	0.00
Race	220	231,689	0.09
Hispanic	101	231,689	0.04
Education	1	231,689	0.00
Age	2	231,689	0.00
Marital status	180	231,689	0.08
Census region	458	231,689	0.20
Num of people in HH	4	231,689	0.00
Num of living children	2,669	231,689	1.15
Proxy Interview	0	231,689	0.00
Self-reported health	156	231,689	0.07
Ever had severe disease	42	231,689	0.02
Ever had mild disease	20	231,689	0.01
Body mass index	3,282	231,689	1.42
House ownership	0	231,689	0.00
Labor force participation	0	231,689	0.00
HH total income	0	231,689	0.00
Individual earnings	0	231,689	0.00
Any missing	21,015	231,689	9.07

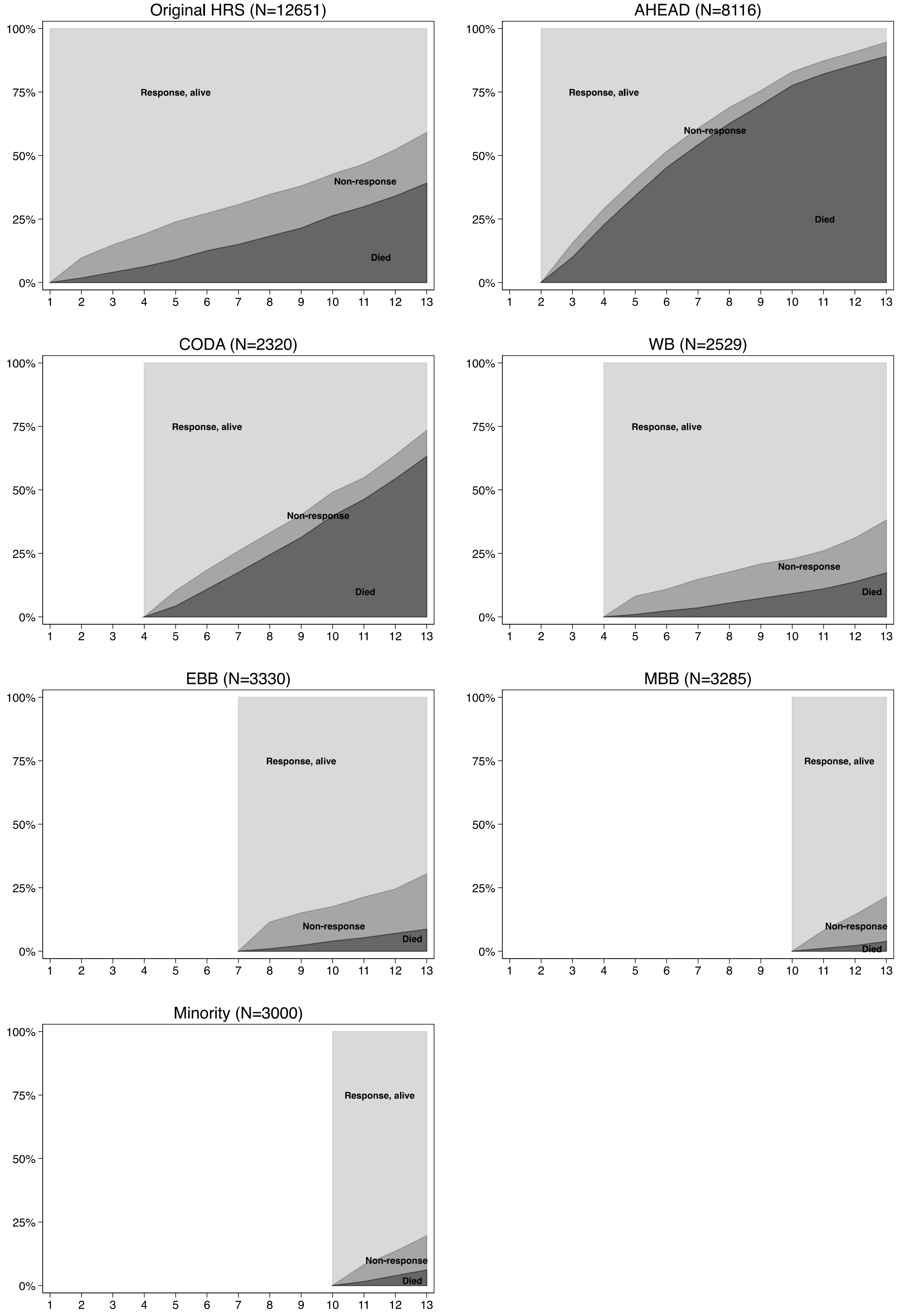
Notes: The analytic sample includes the 231,689 complete person-wave cases, which accounts for 90.93% of the studied sample.

Table 5-1-8. Sample size by interview response status in each wave, by entry cohort

		HRS	AHEAD	CODA	WB	EBB	MBB	Minority
wave 1	Resp, alive	12,651						
	Died	0						
	Non-response	0						
wave 2	Resp, alive	11,423	8,116					
	Died	226	0					
	Non-response	1,002	0					
wave 3	Resp, alive	10,775	6,859					
	Died	507	804					
	Non-response	1,369	453					
wave 4	Resp, alive	10,242	5,760	2,320	2,529			
	Died	789	1,840	0	0			
	Non-response	1,620	516	0	0			
wave 5	Resp, alive	9,630	4,817	2,082	2,323			
	Died	1,142	2,770	99	24			
	Non-response	1,879	529	139	182			
wave 6	Resp, alive	9,206	3,929	1,893	2,254			
	Died	1,583	3,669	252	60			
	Non-response	1,862	518	175	215			
wave 7	Resp, alive	8,773	3,189	1,719	2,155	3,330		
	Died	1,898	4,402	406	90	0		
	Non-response	1,980	525	195	284	0		
wave 8	Resp, alive	8,260	2,527	1,553	2,083	2,948		
	Died	2,311	5,081	567	138	32		
	Non-response	2,079	508	200	308	350		
wave 9	Resp, alive	7,843	1,986	1,389	2,002	2,827		
	Died	2,715	5,672	726	185	76		
	Non-response	2,092	458	205	342	427		
wave 10	Resp, alive	7,253	1,386	1,182	1,953	2,746	3,285	3,000
	Died	3,326	6,297	921	231	133	0	0
	Non-response	2,072	433	217	345	451	0	0
wave 11	Resp, alive	6,749	1,037	1,049	1,871	2,621	3,016	2,754
	Died	3,776	6,661	1,073	279	176	37	47
	Non-response	2,126	418	198	379	533	232	199
wave 12	Resp, alive	6,029	742	842	1,743	2,515	2,815	2,596
	Died	4,311	6,954	1,261	349	233	72	116
	Non-response	2,311	420	217	437	582	398	288
wave 13	Resp, alive	5,176	442	617	1,567	2,317	2,581	2,412
	Died	4,948	7,225	1,465	438	288	126	183
	Non-response	2,527	449	238	524	725	578	405

Notes: Wave 1 = 1992-1993, wave 2 = 1994/1995, wave 3= 1996, wave 4 = 1998, wave 5 = 2000, …, wave 13 = 2016,

Figure 5-1-3. Attrition patterns for all entry cohorts across survey waves, 1992-2016



Notes: the x-axis represents the number of waves; Wave 1 = 1992-1993, wave 2 = 1994/1995, wave 3= 1996, wave 4 = 1998, wave 5 = 2000, …, wave 13 = 2016. The blank areas in graphs indicate that the entry cohorts have not entered the study yet. The dark grey areas represent the share of those died, the medium grey denote the share of those not responded but alive, and the light grey areas represents the share of those who responded.

Table 5-1-9. Determinants for between-wave attrition

	Responded	Died	Non-response
Female	0.0333***	-0.0285***	-0.0049***
	(0.0016)	(0.0011)	(0.0013)
US born	0.0008	0.0113***	-0.0121***
	(0.0027)	(0.0016)	(0.0022)
Race (ref: White)			
Black	-0.0093***	-0.0006	0.0099***
	(0.0023)	(0.0014)	(0.0018)
Other	-0.0027	-0.0062*	0.0090**
	(0.0037)	(0.0027)	(0.0027)
Hispanic	0.0049	-0.0087***	0.0037
	(0.0030)	(0.0020)	(0.0024)
Education (ref: Lt High School)			
High school	0.0005	0.0015	-0.0020
	(0.0020)	(0.0012)	(0.0017)
Some college	-0.0014	0.0044**	-0.0031
	(0.0023)	(0.0015)	(0.0019)
College and above	0.0087***	0.0028	-0.0115***
	(0.0025)	(0.0017)	(0.0020)
Age (ref: <51)			
51-61	-0.0021	0.0107**	-0.0086**
	(0.0045)	(0.0038)	(0.0028)
> 61	-0.0153**	0.0303***	-0.0150***
	(0.0047)	(0.0038)	(0.0031)
Entry cohorts (ref: initial HRS)			
AHEAD	-0.0295***	0.0389***	-0.0093***
	(0.0021)	(0.0014)	(0.0017)
CODA	-0.0139***	0.0166***	-0.0028
	(0.0029)	(0.0017)	(0.0025)
WB	0.0127***	-0.0091***	-0.0036
	(0.0027)	(0.0019)	(0.0021)
EB	0.0028	-0.0132***	0.0104***
	(0.0032)	(0.0022)	(0.0024)
MBB	-0.0049	-0.0107**	0.0156***
	(0.0043)	(0.0033)	(0.0030)
Minority	0.0108**	-0.0115***	0.0007
	(0.0038)	(0.0028)	(0.0028)
One wave lagged variables			
Marital Status (ref: Married)			
Divorced	-0.0150***	0.0101***	0.0048*
	(0.0031)	(0.0021)	(0.0024)
Widowed	-0.0176***	0.0213***	-0.0037
	(0.0027)	(0.0018)	(0.0022)
Never married	-0.0104*	0.0118***	-0.0013
	(0.0046)	(0.0032)	(0.0033)
Region (ref: Northeast)			
Midwest	0.0111***	0.0028	-0.0138***
	(0.0023)	(0.0014)	(0.0019)
South	0.0083***	0.0021	-0.0103***
	(0.0021)	(0.0013)	(0.0017)
West	0.0136***	0.0008	-0.0144***
	(0.0024)	(0.0016)	(0.0019)
Other	0.0266	-0.0023	-0.0243*
	(0.0200)	(0.0158)	(0.0123)
Num people in HH (ref: 1)			
2	0.0015	0.0012	-0.0028
	(0.0025)	(0.0015)	(0.0021)
3	0.0000	0.0047*	-0.0048*
	(0.0029)	(0.0019)	(0.0023)
4+	0.0008	0.0025	-0.0033

	(0.0031)	(0.0021)	(0.0024)
Num living children (ref: No living child)			
One/two children	0.0133***	-0.0073***	-0.0060*
	(0.0032)	(0.0020)	(0.0025)
Three/four children	0.0212***	-0.0099***	-0.0114***
	(0.0032)	(0.0021)	(0.0026)
Five/more children	0.0231***	-0.0093***	-0.0139***
	(0.0034)	(0.0022)	(0.0027)
Proxy interview	-0.1006***	0.0548***	0.0458***
	(0.0034)	(0.0020)	(0.0028)
Self-reported health (ref: Excellent)			
Very good	0.0010	0.0009	-0.0018
	(0.0024)	(0.0017)	(0.0018)
Good	-0.0136***	0.0132***	0.0004
	(0.0025)	(0.0017)	(0.0019)
Fair	-0.0348***	0.0371***	-0.0023
	(0.0028)	(0.0020)	(0.0021)
Poor	-0.0828***	0.0843***	-0.0014
	(0.0036)	(0.0026)	(0.0026)
Ever had severe diseases	-0.0366***	0.0395***	-0.0029*
	(0.0016)	(0.0010)	(0.0013)
Ever had mild diseases	-0.0200***	0.0163***	0.0037**
	(0.0015)	(0.0010)	(0.0012)
Body mass index (ref: Normal)			
Overweight	0.0271***	-0.0244***	-0.0026*
	(0.0017)	(0.0012)	(0.0013)
Obese	0.0408***	-0.0307***	-0.0101***
	(0.0019)	(0.0013)	(0.0015)
House owner	0.0152***	-0.0165***	0.0013
	(0.0018)	(0.0013)	(0.0014)
Labor force (ref: Working)			
(partly) Retired	-0.0083***	0.0243***	-0.0161***
	(0.0025)	(0.0019)	(0.0018)
Disabled	-0.0056	0.0222***	-0.0166***
	(0.0041)	(0.0029)	(0.0030)
Outside labor force	-0.0363***	0.0303***	0.0060*
	(0.0032)	(0.0023)	(0.0024)
Total HH income(ref: Quintile 1)			
Quintile 2	0.0006	-0.0006	-0.0000
	(0.0020)	(0.0013)	(0.0016)
Quintile 3	0.0047*	-0.0029	-0.0018
	(0.0024)	(0.0016)	(0.0018)
Quintile 4	0.0079**	-0.0088***	0.0009
	(0.0028)	(0.0020)	(0.0022)
Earnings (ref: Quintile 1)			
Quintile 2	-0.0042	-0.0046	0.0088
	(0.0084)	(0.0075)	(0.0047)
Quintile 3	0.0176***	-0.0151***	-0.0025
	(0.0026)	(0.0020)	(0.0018)
Quintile 4	0.0147***	-0.0136***	-0.0011
	(0.0028)	(0.0023)	(0.0018)
Observations	210,674	210,674	210,674

Notes: Estimates are average marginal effects from a dynamic multinomial regression model that examines the determinants of between-wave attrition. * p<0.05, ** p < 0.01, ** p<0.001.

Table 5-1-10. IIA Tests for the dynamic multinomial logistic regression model

	Non-response versus Response		Died versus Response	
	Full model	Restricted model	Full model	Restricted model
Female	0.878***	0.881***	0.559***	0.558***
US born	0.811***	0.811***	1.267***	1.268***
Race (ref: White)				
Black	1.204***	1.206***	0.999	0.988
Other	1.177***	1.176***	0.884*	0.892
Hispanic	1.063	1.065	0.831***	0.824***
Education (ref: Lt High School)				
High school	0.965	0.966	1.030	1.028
Some college	0.948	0.950	1.092**	1.094**
College and above	0.792***	0.794***	1.046	1.049
Age (ref: <51)				
51-61	0.873**	0.874**	1.360*	1.356*
> 61	0.789***	0.789***	2.106***	2.096***
Marital Status (ref: Married)				
Divorced	1.109**	1.108*	1.250***	1.258***
Widowed	0.953	0.952	1.525***	1.529***
Never married	0.988	0.986	1.282***	1.284***
Region (ref: Northeast)				
Midwest	0.774***	0.774***	1.042	1.044
South	0.830***	0.829***	1.031	1.028
West	0.763***	0.762***	0.999	1.001
Other	0.602	0.601	0.925	0.896
Num people in HH (ref: 1)				
2	0.949	0.946	1.023	1.027
3	0.916*	0.913*	1.095*	1.102*
4 +	0.941	0.937	1.049	1.052
Num living children (ref: No living child)				
One/two children	0.890**	0.889**	0.860***	0.869***
Three/four children	0.799***	0.798***	0.810***	0.821***
Five/more children	0.758***	0.757***	0.818***	0.826***
Proxy interview	2.135***	2.126***	2.583***	2.536***
Self-reported health (ref: Excellent)				
Very Good	0.966	0.966	1.026	1.024
Good	1.023	1.024	1.449***	1.451***
Fair	0.998	0.999	2.342***	2.351***
Poor	1.074	1.072	4.511***	4.531***
Ever had severe diseases	0.991	0.990	2.320***	2.319***
Ever had mild diseases	1.097***	1.095***	1.432***	1.440***
Body mass index (ref: Normal)				
Overweight	0.925***	0.924***	0.618***	0.618***
Obese	0.786***	0.786***	0.526***	0.527***
House owner	1.006	1.009	0.723***	0.727***
Labor force (ref: Working)				
(partly) Retired	0.751***	0.753***	1.775***	1.766***
Disabled	0.740***	0.742***	1.697***	1.684***
Outside labor force	1.149***	1.147***	2.050***	2.051***
Total HH income(ref: Quintile 1)				
Quintile 2	0.998	1.000	0.988	0.988
Quintile 3	0.961	0.962	0.941	0.938
Quintile 4	1.006	1.008	0.829***	0.829***
Earning (ref: Quintile 1)				
Quintile 2	1.166	1.167	0.919	0.923
Quintile 3	0.935	0.935	0.710***	0.707***
Quintile 4	0.963	0.963	0.738***	0.735***
Entry cohorts (ref: initial HRS)				
AHEAD	0.863***	0.867***	2.052***	2.056***
CODA	0.967	0.966	1.425***	1.426***
WB	0.920*	0.920*	0.780***	0.784***

EB	1.184***	1.185***	0.699***	0.697***
MBB	1.291***	1.292***	0.763**	0.756**
Minority	0.999	1.001	0.730***	0.731***
Constant	0.165***	0.165***	0.009***	0.009***
N	210,674	197,721	210,694	199,249
Log Likelihood	-79,973	-42,791	-79,973	-36,469

Notes: Estimates are Relative Risk Ratio (RRR) from a dynamic multinomial logistic regression models comparing “non-response” versus “response”, and “died” versus “response”, respectively. The restricted model for the former comparison excludes “died” from the alternatives, whereas the restricted model latter comparison excludes “non-response” from the alternatives. * $p<0.05$, ** $p < 0.01$, ** $p<0.001$.

Table 5-1-11. Determinants of between-wave attrition based on survival analyses

	Andersen-Gill model		Cox model for mortality		Compete risks model	
	Hazard Ratio	PH test	Hazard Ratio	PH test	Subhazard Ratio (SHR)	var#lnt
Female	0.74***	0.088	0.60***	**0.001**	1.07*	**<0.001**
US born	0.98	**<0.001**	1.23***	0.473	0.81***	0.001
Race (ref: White)						
Black	1.08***	**<0.001**	1.00	0.005	1.19***	**<0.001**
Other	1.07*	0.036	0.90*	0.344	1.22**	0.007
Hispanic	0.95	0.279	0.83***	0.501	1.36***	**<0.001**
Education (ref: Lt High School)						
High school	0.98	0.806	0.98	0.500	1.03	**<0.001**
Some college	0.99	0.708	1.03	0.275	0.98	**<0.001**
College and above	0.89***	0.496	0.98	0.169	0.84**	**<0.001**
Age (ref: <51)						
51-61	0.89**	0.801	1.41**	0.165	1.09	0.392
> 61	0.94	0.795	1.92***	0.080	1.22**	0.013
Marital Status (ref: Married)						
Divorced	1.12***	0.229	1.19***	0.286	1.05	**<0.001**
Widowed	1.14***	0.610	1.29***	0.758	0.87	0.792
Never married	1.07	0.429	1.20**	0.872	0.89	0.741
Region (ref: Northeast)						
Midwest	0.91***	0.045	1.02	0.399	0.80***	0.862
South	0.93***	0.121	1.02	0.684	0.89**	0.366
West	0.88***	0.240	1.00	0.644	0.84***	0.499
Other	0.74	0.862	0.86	0.459	0.47	0.994
Num people in HH (ref: 1)						
2	0.99	0.449	1.03	0.451	0.87*	0.247
3	1.01	0.371	1.10**	0.968	0.79**	0.968
4+	1.02	0.980	1.06	0.055	0.86*	0.052
Num living children (ref: No living child)						
One/two children	0.90***	0.192	0.88***	0.739	0.75***	**<0.001**
Three/four children	0.83***	0.007	0.83***	0.414	0.71***	**<0.001**
Five/more children	0.81***	0.000	0.84***	0.093	0.66***	**<0.001**
Proxy	1.87***	0.217	1.76***	0.216	1.98***	**<0.001**
Self-reported health (ref: Excellent)						
Very good	0.96	0.311	1.02	0.637	1.00	0.141
Good	1.10***	0.032	1.43***	0.745	1.07	0.156
Fair	1.37***	**<0.001**	2.18***	0.342	0.86*	0.001
Poor	2.00***	**<0.001**	3.51***	0.164	0.69***	**<0.001**

Ever had severe diseases	1.39***	0.047	2.00***	0.099	0.90**	**<0.001**
Ever had mild diseases	1.14***	0.431	1.26***	0.055	1.14***	0.041
Body mass index (ref: Normal)						
Overweight	0.80***	0.206	0.69***	0.029	1.02	0.389
Obese	0.69***	0.118	0.61***	0.465	0.91*	0.227
House owner	0.89***	0.085	0.80***	0.690	0.93	**<0.001**
Labor force (ref: Working)						
(partly) Retired	0.90***	0.234	1.73***	0.381	0.78***	0.583
Disabled	0.88***	0.971	1.89***	0.236	0.63***	**<0.001**
Outside labor force	1.19***	**<0.001**	2.17***	**<0.001**	0.91	0.002
Total HH income(ref: Quintile 1)						
Quintile 2	0.99	0.504	0.97	0.999	1.05	**<0.001**
Quintile 3	0.95*	0.414	0.92**	0.392	1.02	**<0.001**
Quintile 4	0.95*	0.795	0.81***	0.617	1.11	**<0.001**
Earnings (ref: Quintile 1)						
Quintile 2	0.94	0.504	1.06	0.714	0.77**	**<0.001**
Quintile 3	0.85***	0.003	0.77***	0.004	0.85**	0.391
Quintile 4	0.94*	0.053	0.76***	0.046	0.88**	0.160
Observations	210,674		210,674		96,404	

Notes: The Andersen-Gill model does not distinguish non-response from death but accounts for multiple times of attrition due to either non-response or death. The Cox model for mortality treats death as failure and non-response as censoring. For these two models, Hazards Ratio and p values from a PH test are reported. PH test is a test for the proportional-hazards assumption based on the scaled Schoenfeld residuals, implemented by *estat phtest* in Stata. Significant p values indicate violations of the assumption.

The compete risks model considers the first non-response as failure and death as a competing risk. Sub-hazard Ratios are reported. P values on the interaction of each variable with log transformed time were displayed. Significant p values indicate violations of the assumption.

Bonferroni adjustment was used to correct the alpha level for multiple comparisons in testing for proportional-hazards assumption in the three models, as corrected alpha = 0.05 / 42 = 0.0012. The p values in bold indicate significance based on the corrected alpha level.

* p<0.05, ** p < 0.01, ** p<0.001.

Results 5.2: The Association Between Education and Hospitalizations

Figure 5-2-1. Sample size flowchart of the association study of education and hospitalization

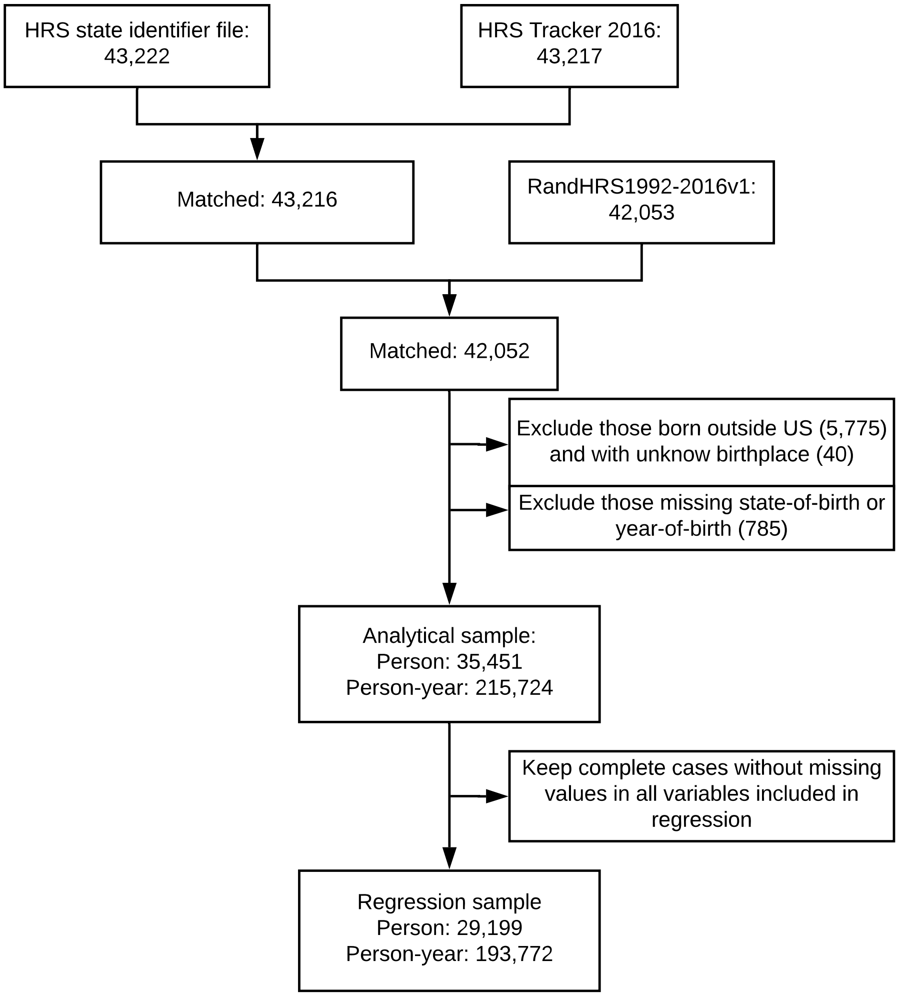
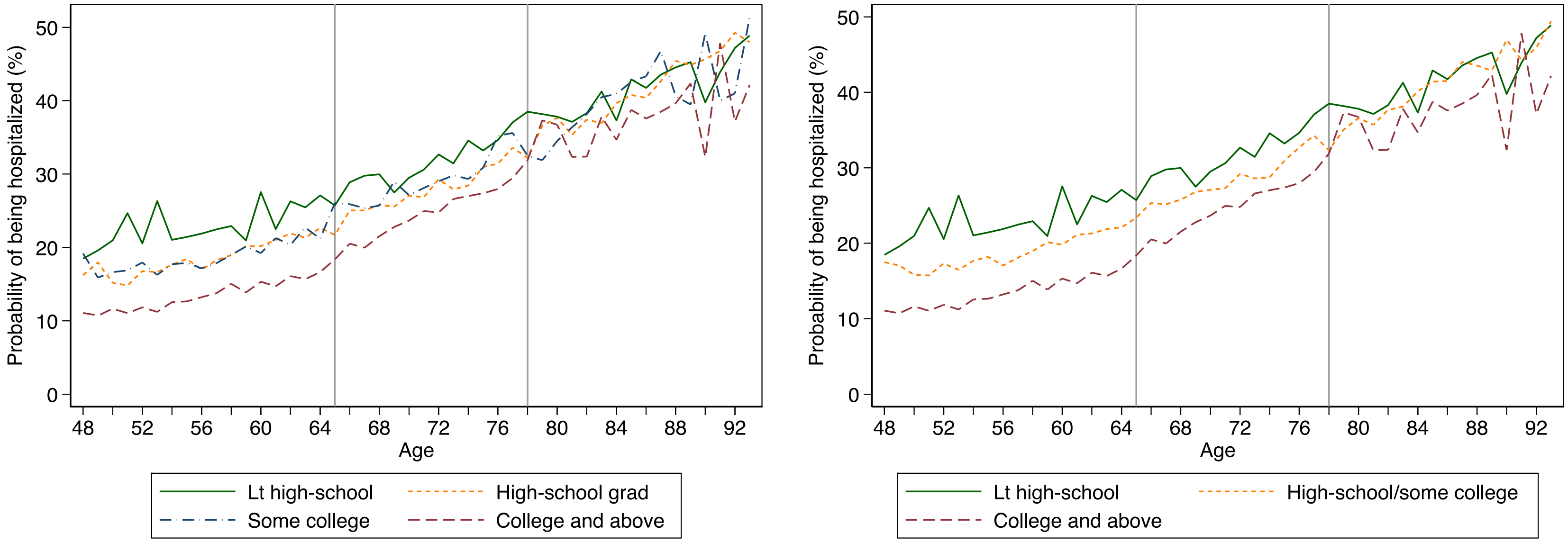

HRS state identifier file: 43,222
HRS Tracker 2016: 43,217
Matched: 43,216
RandHRS1992-2016v1: 42,053
Matched: 42,052
Exclude those born outside US (5,775) and with unknow birthplace (40)
Exclude those missing state-of-birth or year-of-birth (785)
Analytical sample: Person: 35,451 Person-year: 215,724
Keep complete cases without missing values in all variables included in regression
Regression sample Person: 29,199 Person-year: 193,772

Table 5-2-1. Descriptive statistics of respondents’ characteristics in the association study

	Lt High-school		High school/some college		College and above		Total	
	Freq	Mean	Freq	Mean	Freq	Mean	Freq	Mean
Gender (%)								
Male	2,332	43.2	7,138	40.9	3,138	49.3	12,608	43.2
Female	3,070	56.8	10,298	59.1	3,223	50.7	16,591	56.8
Race (%)								
White	3,602	66.7	13,559	77.8	5,338	83.9	22,499	77.1
Black	1,489	27.6	3,104	17.8	809	12.7	5,402	18.5
Other	311	5.8	773	4.4	214	3.4	1,298	4.4
Hispanics (%)								
Not Hispanic	4,860	90	16,600	95.2	6,186	97.2	27,646	94.7
Hispanic	542	10	836	4.8	175	2.8	1,553	5.3
Self-rated child health (%)								
Excellent	2,131	39.4	9,033	51.8	3,944	62.0	15,108	51.7
Very good	1,415	26.2	4,649	26.7	1,487	23.4	7,551	25.9
Good	1,313	24.3	2,666	15.3	669	10.5	4,648	15.9
Fair	388	7.2	841	4.8	212	3.3	1,441	4.9
Poor	155	2.9	247	1.4	49	0.8	451	1.5
Childhood family financial situation (%)								
Well off	2,975	55.1	12,365	70.9	5,065	79.6	20,405	69.9
Poor	2,427	44.9	5,071	29.1	1,296	20.4	8,794	30.1
Parents' highest education (mean)	5,402	8.0	17,436	10.6	6,361	12.6	29,199	10.6

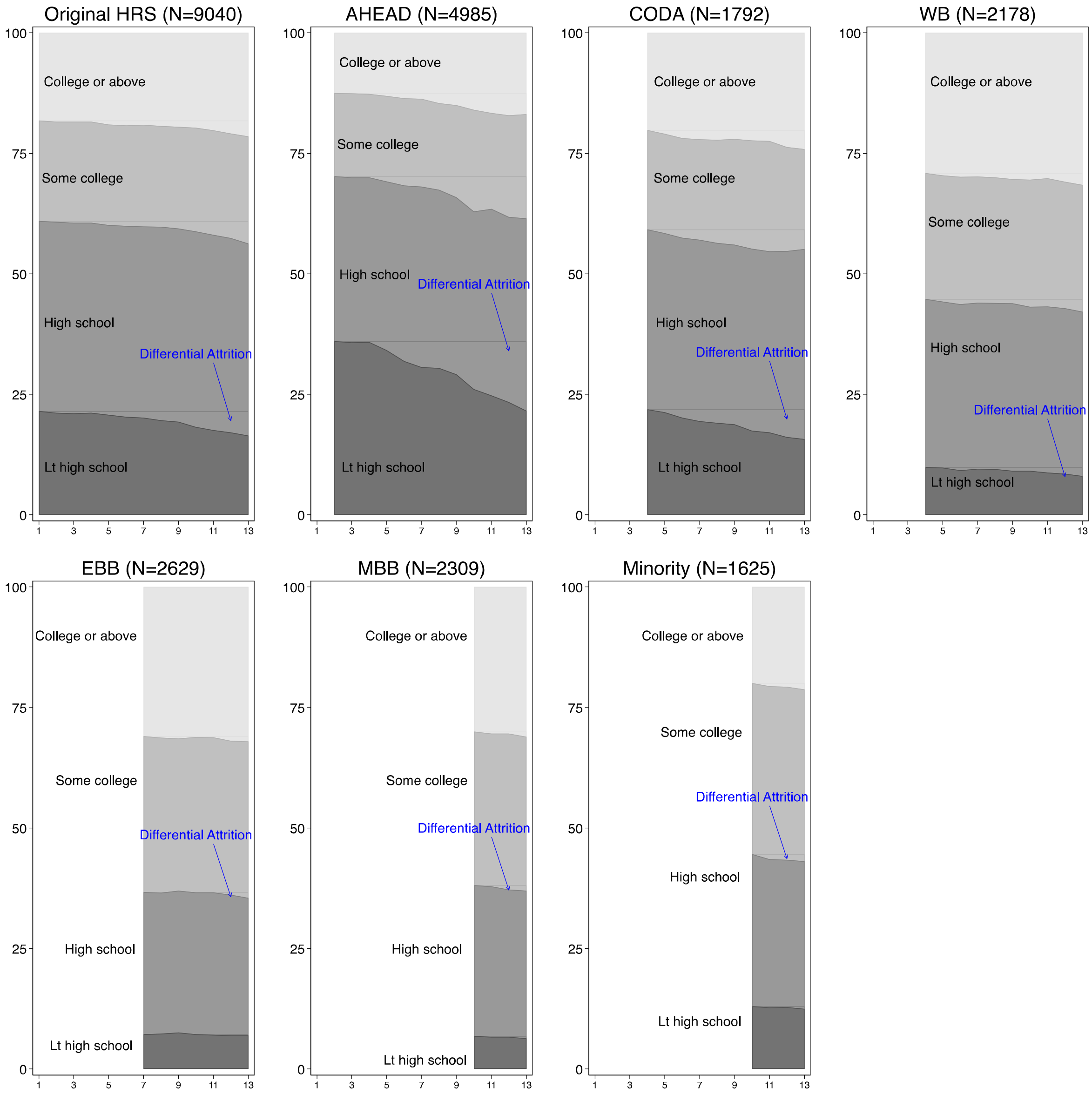
Notes: Estimates were based on those who were born in the United States and had no missing values in the variables listed in the table.

Figure 5-2-2. The probability of hospitalizations over age by educational attainment



Notes: Respondents aged 48 to 93 were included to make sure at least 100 observations available in each education-age cell. The average probability of being hospitalized for those with less than high school, high school/some college, and college or above is 22.7%, 18.4%, and 13.2% over ages 48-64; is 31.2%, 28.1%, and 24.2% over ages 65-77; is 41.6%, 41.0%, and 37.5% after age 78.

Figure 5-2-3. Differential attrition related to educational attainment, by entry cohort



Notes: N represents the number of unique subjects in the baseline. The x-axis represents the number of waves; Wave 1 = 1992-1993, wave 2 = 1994/1995, wave 3= 1996, wave 4 = 1998, wave 5 = 2000, …, wave 13 = 2016. The blank areas in graphs indicate that the entry cohorts have not entered the study yet.

Table 5-2-2. Results from OLS linear regression models of education and hospitalizations

	Unweighted OLS		Weighted OLS	
	(1)	(2)	(3)	(4)
Education (Ref: Lt High school)				
High school/some college	-0.0202***	-0.0205***	-0.0329***	-0.0337***
	(0.0043)	(0.0043)	(0.0062)	(0.0062)
College or above	-0.0584***	-0.0584***	-0.0833***	-0.0839***
	(0.0051)	(0.0052)	(0.0073)	(0.0072)
Female	-0.0173***	-0.0171***	-0.0226***	-0.0225***
	(0.0029)	(0.0029)	(0.0041)	(0.0040)
Race (Ref: White)				
Black	0.0100**	0.0098**	0.0164**	0.0160**
	(0.0047)	(0.0047)	(0.0068)	(0.0068)
Others	0.0317***	0.0309***	0.0303**	0.0285**
	(0.0095)	(0.0095)	(0.0122)	(0.0123)
Hispanics	-0.0146*	-0.0141*	-0.0125	-0.0116
	(0.0082)	(0.0083)	(0.0115)	(0.0116)
Self-rated child health (Ref: Excellent)				
Very good	0.0085**	0.0085**	0.0091*	0.0092*
	(0.0034)	(0.0034)	(0.0048)	(0.0048)
Good	0.0241***	0.0239***	0.0300***	0.0300***
	(0.0042)	(0.0042)	(0.0058)	(0.0058)
Fair	0.0638***	0.0634***	0.0700***	0.0693***
	(0.0074)	(0.0074)	(0.0093)	(0.0092)
Poor	0.1031***	0.1029***	0.1081***	0.1073***
	(0.0141)	(0.0141)	(0.0174)	(0.0174)
Childhood family financial situation (Ref: well-off)				
Poor	0.0095***	0.0095***	0.0136***	0.0137***
	(0.0032)	(0.0032)	(0.0046)	(0.0045)
Parents' education	-0.0007	-0.0007	-0.0013*	-0.0013*
	(0.0005)	(0.0005)	(0.0007)	(0.0007)
SOB	Yes	Yes	Yes	Yes
YOB	Yes	Yes	Yes	Yes
SOB#YOB	No	Yes	No	Yes
Constant	-0.1605***	0.0710	-0.1858***	-0.1486
	(0.0199)	(0.0901)	(0.0182)	(0.1185)
Observations	193,772	193,772	193,772	193,772
R-squared	0.0267	0.0271	0.0419	0.0426

Notes: Weighted OLS uses inverse probability weighting to account for attrition bias. SOB denotes State-of-Birth, YOB denotes Year-of-Birth, and SOB#YOB denotes state-of-birth specific linear time trends.
Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1

Table 5-2-3. Family SES and child health on the education-hospitalization relationship

	(1)	(2)	(3)
Education (Ref: Lt High school)			
High school/some college	-0.0342***	-0.0381***	-0.0407***
	(0.0062)	(0.0062)	(0.0060)
College or above	-0.0850***	-0.0911***	-0.0943***
	(0.0072)	(0.0072)	(0.0070)
Female	-0.0232***	-0.0209***	-0.0215***
	(0.0040)	(0.0040)	(0.0039)
Race (Ref: White)			
Black	0.0167**	0.0164**	0.0183***
	(0.0068)	(0.0067)	(0.0066)
Others	0.0319**	0.0294**	0.0345***
	(0.0124)	(0.0121)	(0.0122)
Hispanics	-0.0122	-0.0119	-0.0119
	(0.0117)	(0.0115)	(0.0114)
Self-rated child health (Ref: Excellent)			
Very good	0.0094**		
	(0.0048)		
Good	0.0310***		
	(0.0057)		
Fair	0.0724***		
	(0.0093)		
Poor	0.1098***		
	(0.0173)		
Childhood family financial situation (Ref: well-off)			
Poor		0.0176***	
		(0.0045)	
Parents' education	-0.0016**	-0.0015**	-0.0020***
	(0.0007)	(0.0007)	(0.0007)
SOB	Yes	Yes	Yes
YOB	Yes	Yes	Yes
SOB#YOB	Yes	Yes	Yes
Constant	-0.0980	-0.1908	0.3200
	(0.1168)	(0.1176)	(0.3446)
Observations	194,053	194,267	200,213
R-squared	0.0426	0.0407	0.0410

Notes: All estimators are from weighted OLS models that use inverse probability weighting to account for attrition bias. SOB denotes State-of-Birth, YOB denotes Year-of-Birth, and SOB#YOB denotes state-of-birth specific linear time trends. Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1

Table 5-2-4. The association between education and hospitalizations, stratified by age

	Younger than 64	Ages 65-77	Older than 78
Education (Ref: Lt High school)			
High school/some college	-0.0482***	-0.0507***	-0.0096
	(0.0086)	(0.0092)	(0.0112)
College or above	-0.1024***	-0.1032***	-0.0546***
	(0.0095)	(0.0109)	(0.0152)
Female	-0.0239***	-0.0361***	-0.0152*
	(0.0047)	(0.0062)	(0.0092)
Race (Ref: White)			
Black	0.0359***	-0.0136	0.0042
	(0.0074)	(0.0110)	(0.0185)
Others	0.0386***	0.0180	-0.0187
	(0.0127)	(0.0231)	(0.0395)
Hispanics	0.0029	-0.0556***	0.0085
	(0.0136)	(0.0185)	(0.0282)
Self-rated child health (Ref: Excellent)			
Very good	0.0207***	-0.0020	0.0070
	(0.0055)	(0.0074)	(0.0108)
Good	0.0376***	0.0196**	0.0348***
	(0.0069)	(0.0086)	(0.0127)
Fair	0.0849***	0.0612***	0.0492**
	(0.0112)	(0.0147)	(0.0198)
Poor	0.1538***	0.0704**	0.0752**
	(0.0217)	(0.0283)	(0.0292)
Childhood family financial situation (Ref: well-off)			
Poor	0.0155***	0.0163**	0.0036
	(0.0056)	(0.0068)	(0.0096)
Parents’ education	-0.0010	-0.0011	-0.0011
	(0.0008)	(0.0011)	(0.0019)
SOB	Yes	Yes	Yes
YOB	Yes	Yes	Yes
SOB#YOB	Yes	Yes	Yes
Constant	-0.3590***	-0.0066	-0.1393
	(0.0505)	(0.1690)	(0.1138)
Observations	93,125	65,058	35,589
R-squared	0.0242	0.0197	0.0194

Notes: All estimators are from weighted OLS models that use inverse probability weighting to account for attrition bias. SOB denotes State-of-Birth, YOB denotes Year-of-Birth, and SOB#YOB denotes state-of-birth specific linear time trends. Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1

Table 5-2-5. Results from logistic regression models of education and hospitalizations

	Unweighted Logit		Weighted Logit	
	OR	Marginal Effects	OR	Marginal Effects
Education (Ref: Lt High school)				
High school/some college	0.91***	-0.0173***	0.87***	-0.0289***
	(0.02)	(0.0041)	(0.02)	(0.0060)
College or above	0.73***	-0.0572***	0.66***	-0.0821***
	(0.02)	(0.0050)	(0.02)	(0.0071)
Female	0.91***	-0.0175***	0.89***	-0.0229***
	(0.01)	(0.0029)	(0.02)	(0.0041)
Race (Ref: White)				
Black	1.06**	0.0102*	1.09**	0.0169*
	(0.03)	(0.0048)	(0.04)	(0.0068)
Others	1.19***	0.0336***	1.17**	0.0322*
	(0.06)	(0.0100)	(0.07)	(0.0129)
Hispanics	0.92*	-0.0143	0.94	-0.0119
	(0.04)	(0.0083)	(0.06)	(0.0118)
Self-rated child health (Ref: Excellent)				
Very good	1.05***	0.0087**	1.05**	0.0095*
	(0.02)	(0.0034)	(0.03)	(0.0048)
Good	1.14***	0.0236***	1.16***	0.0294***
	(0.02)	(0.0041)	(0.03)	(0.0057)
Fair	1.39***	0.0637***	1.40***	0.0701***
	(0.05)	(0.0074)	(0.06)	(0.0092)
Poor	1.65***	0.1004***	1.63***	0.1050***
	(0.11)	(0.0139)	(0.12)	(0.0170)
Childhood family financial situation (Ref: well-off)				
Poor	1.05***	0.0091**	1.07***	0.0131**
	(0.02)	(0.0032)	(0.02)	(0.0044)
Parents' education	1.00*	-0.0009	0.99**	-0.0016*
	(0.00)	(0.0005)	(0.00)	(0.0007)
SOB	Yes		Yes	
YOB	Yes		Yes	
SOB#YOB	No		No	
Constant	0.21**		0.17***	
	(0.14)		(0.11)	
Observations	193,749		193,749	

Notes: Weighted logit models use inverse probability weighting to account for attrition bias. SOB denotes State-of-Birth, YOB denotes Year-of-Birth, and SOB#YOB denotes state-of-birth specific linear time trends.
Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1

Table 5-2-6. OLS regression treating attrition due to non-response as an absorbing state

	Unweighted OLS		Weighted OLS	
	(1)	(2)	(3)	(4)
Education (Ref: Lt High school)				
High school/some college	-0.0197***	-0.0201***	-0.0337***	-0.0344***
	(0.0045)	(0.0045)	(0.0064)	(0.0064)
College or above	-0.0558***	-0.0557***	-0.0818***	-0.0823***
	(0.0053)	(0.0053)	(0.0074)	(0.0074)
Female	-0.0170***	-0.0167***	-0.0224***	-0.0223***
	(0.0029)	(0.0030)	(0.0041)	(0.0041)
Race (Ref: White)				
Black	0.0088*	0.0085*	0.0167**	0.0163**
	(0.0049)	(0.0049)	(0.0071)	(0.0071)
Others	0.0299***	0.0291***	0.0270**	0.0256**
	(0.0098)	(0.0098)	(0.0121)	(0.0122)
Hispanics	-0.0196**	-0.0188**	-0.0182	-0.0171
	(0.0086)	(0.0086)	(0.0111)	(0.0112)
Self-rated child health (Ref: Excellent)				
Very good	0.0079**	0.0080**	0.0087*	0.0088*
	(0.0035)	(0.0035)	(0.0048)	(0.0048)
Good	0.0245***	0.0244***	0.0307***	0.0306***
	(0.0043)	(0.0043)	(0.0060)	(0.0059)
Fair	0.0646***	0.0643***	0.0701***	0.0696***
	(0.0076)	(0.0076)	(0.0095)	(0.0094)
Poor	0.1046***	0.1045***	0.1119***	0.1110***
	(0.0147)	(0.0147)	(0.0187)	(0.0187)
Childhood family financial situation (Ref: well-off)				
Poor	0.0105***	0.0105***	0.0135***	0.0138***
	(0.0033)	(0.0033)	(0.0047)	(0.0046)
Parents’ education	-0.0010*	-0.0009*	-0.0015**	-0.0015**
	(0.0005)	(0.0005)	(0.0007)	(0.0007)
SOB	Yes	Yes	Yes	Yes
YOB	Yes	Yes	Yes	Yes
SOB#YOB	No	Yes	No	No
Constant	-0.1912***	-0.4648***	-0.2035***	-0.6412***
	(0.0211)	(0.0930)	(0.0303)	(0.1204)
Observations	180,028	180,028	180,028	180,028
R-squared	0.0272	0.0276	0.0425	0.0433

Notes: Weighted OLS uses inverse probability weighting to account for attrition bias. SOB denotes State-of-Birth, YOB denotes Year-of-Birth, and SOB#YOB denotes state-of-birth specific linear time trends.
Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1

Table 5-2-7. The association between education and advserse health outcomes

	Unweighted OLS		Weighted OLS	
	(1)	(2)	(3)	(4)
Education (Ref: Lt High school)				
High school/some college	-0.0262***	-0.0268***	-0.0384***	-0.0394***
	(0.0044)	(0.0044)	(0.0062)	(0.0062)
College or above	-0.0658***	-0.0660***	-0.0903***	-0.0911***
	(0.0053)	(0.0053)	(0.0072)	(0.0072)
Female	-0.0252***	-0.0249***	-0.0302***	-0.0301***
	(0.0030)	(0.0030)	(0.0041)	(0.0040)
Race (Ref: White)				
Black	0.0115**	0.0111**	0.0190***	0.0185***
	(0.0050)	(0.0050)	(0.0071)	(0.0071)
Others	0.0326***	0.0320***	0.0296**	0.0284**
	(0.0098)	(0.0099)	(0.0122)	(0.0122)
Hispanics	-0.0227***	-0.0219**	-0.0222**	-0.0211*
	(0.0087)	(0.0088)	(0.0112)	(0.0113)
Self-rated child health (Ref: Excellent)				
Very good	0.0079**	0.0080**	0.0081*	0.0081*
	(0.0035)	(0.0035)	(0.0048)	(0.0047)
Good	0.0243***	0.0242***	0.0291***	0.0289***
	(0.0043)	(0.0043)	(0.0058)	(0.0058)
Fair	0.0648***	0.0646***	0.0693***	0.0687***
	(0.0076)	(0.0076)	(0.0094)	(0.0093)
Poor	0.1068***	0.1066***	0.1105***	0.1095***
	(0.0145)	(0.0144)	(0.0182)	(0.0182)
Childhood family financial situation (Ref: well-off)				
Poor	0.0095***	0.0096***	0.0122***	0.0124***
	(0.0033)	(0.0033)	(0.0046)	(0.0046)
Parents’ education	-0.0012**	-0.0012**	-0.0019***	-0.0018**
	(0.0005)	(0.0005)	(0.0007)	(0.0007)
SOB	Yes	Yes	Yes	Yes
YOB	Yes	Yes	Yes	Yes
SOB#YOB	No	Yes	No	No
Constant	-0.0292	-0.3717***	-0.0425	-0.5527***
	(0.0242)	(0.0940)	(0.0334)	(0.1210)
Observations	188,257	188,257	188,257	188,257
R-squared	0.0443	0.0447	0.0603	0.0609

Notes: Weighted OLS uses inverse probability weighting to account for attrition bias. SOB denotes State-of-Birth, YOB denotes Year-of-Birth, and SOB#YOB denotes state-of-birth specific linear time trends.
Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1

Table 5-2-8. Education and hospitalizations two years before death

	All	Younger than 78	Older than 78
Education (Ref: Lt High school)			
High school/some college	-0.0342***	-0.0450***	-0.0114
	(0.0063)	(0.0070)	(0.0122)
College or above	-0.0820***	-0.0947***	-0.0537***
	(0.0073)	(0.0079)	(0.0162)
Female	-0.0192***	-0.0255***	-0.0054
	(0.0041)	(0.0042)	(0.0100)
Race (Ref: White)			
Black	0.0155**	0.0179***	0.0019
	(0.0068)	(0.0069)	(0.0203)
Others	0.0268**	0.0337***	-0.0347
	(0.0125)	(0.0126)	(0.0424)
Hispanics	-0.0109	-0.0166	0.0187
	(0.0116)	(0.0119)	(0.0319)
Self-rated child health (Ref: Excellent)			
Very good	0.0092*	0.0101**	0.0086
	(0.0048)	(0.0049)	(0.0115)
Good	0.0302***	0.0286***	0.0378***
	(0.0058)	(0.0060)	(0.0135)
Fair	0.0638***	0.0709***	0.0318
	(0.0093)	(0.0101)	(0.0219)
Poor	0.1112***	0.1129***	0.1003***
	(0.0184)	(0.0193)	(0.0350)
Childhood family financial situation (Ref: well-off)			
Poor	0.0130***	0.0168***	-0.0030
	(0.0046)	(0.0048)	(0.0103)
Parents’ education	-0.0012*	-0.0011	-0.0008
	(0.0007)	(0.0007)	(0.0020)
SOB	Yes	Yes	Yes
YOB	Yes	Yes	Yes
SOB#YOB	Yes	Yes	Yes
Constant	-0.6812***	-0.4594***	-0.1359
	(0.1182)	(0.0723)	(0.1219)
Observations	184,466	154,180	30,286
R-squared	0.0361	0.0246	0.0213

Notes: Weighted OLS uses inverse probability weighting to account for attrition bias. SOB denotes State-of-Birth, YOB denotes Year-of-Birth, and SOB#YOB denotes state-of-birth specific linear time trends.
Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1

Table 5-2-9. Education and hospitalizations—multiple imputations

	Unweighted OLS	Weighted OLS
Education (Ref: Lt High school)		
High school/some college	-0.0217***	-0.0351***
	(0.0039)	(0.0055)
College or above	-0.0597***	-0.0853***
	(0.0048)	(0.0066)
Female	-0.0164***	-0.0223***
	(0.0027)	(0.0038)
Race (Ref: White)		
Black	0.0072*	0.0108*
	(0.0044)	(0.0061)
Others	0.0282***	0.0242**
	(0.0088)	(0.0111)
Hispanics	-0.0180**	-0.0178*
	(0.0078)	(0.0107)
Self-rated child health (Ref: Excellent)		
Very good	0.0095***	0.0109**
	(0.0032)	(0.0045)
Good	0.0243***	0.0293***
	(0.0039)	(0.0054)
Fair	0.0636***	0.0730***
	(0.0071)	(0.0092)
Poor	0.0970***	0.1036***
	(0.0130)	(0.0167)
Childhood family financial situation (Ref: well-off)		
Poor	0.0084***	0.0112***
	(0.0031)	(0.0043)
Parents' education	-0.0008	-0.0014**
	(0.0005)	(0.0007)
SOB	Yes	Yes
YOB	Yes	Yes
SOB#YOB	Yes	Yes
Constant	21.1384	16.3586
	(10.0242)	(10.7650)
Observations	215,724	215,724

Notes: Weighted OLS uses inverse probability weighting to account for attrition bias based on 20 multiple imputations. SOB denotes State-of-Birth, YOB denotes Year-of-Birth, and SOB#YOB denotes state-of-birth specific linear time trends.

Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1

Results 5.3: The causal effects of secondary schooling on hospitalizations

Figure 5-3-1. Sample flowchart for the instrumental variables analysis

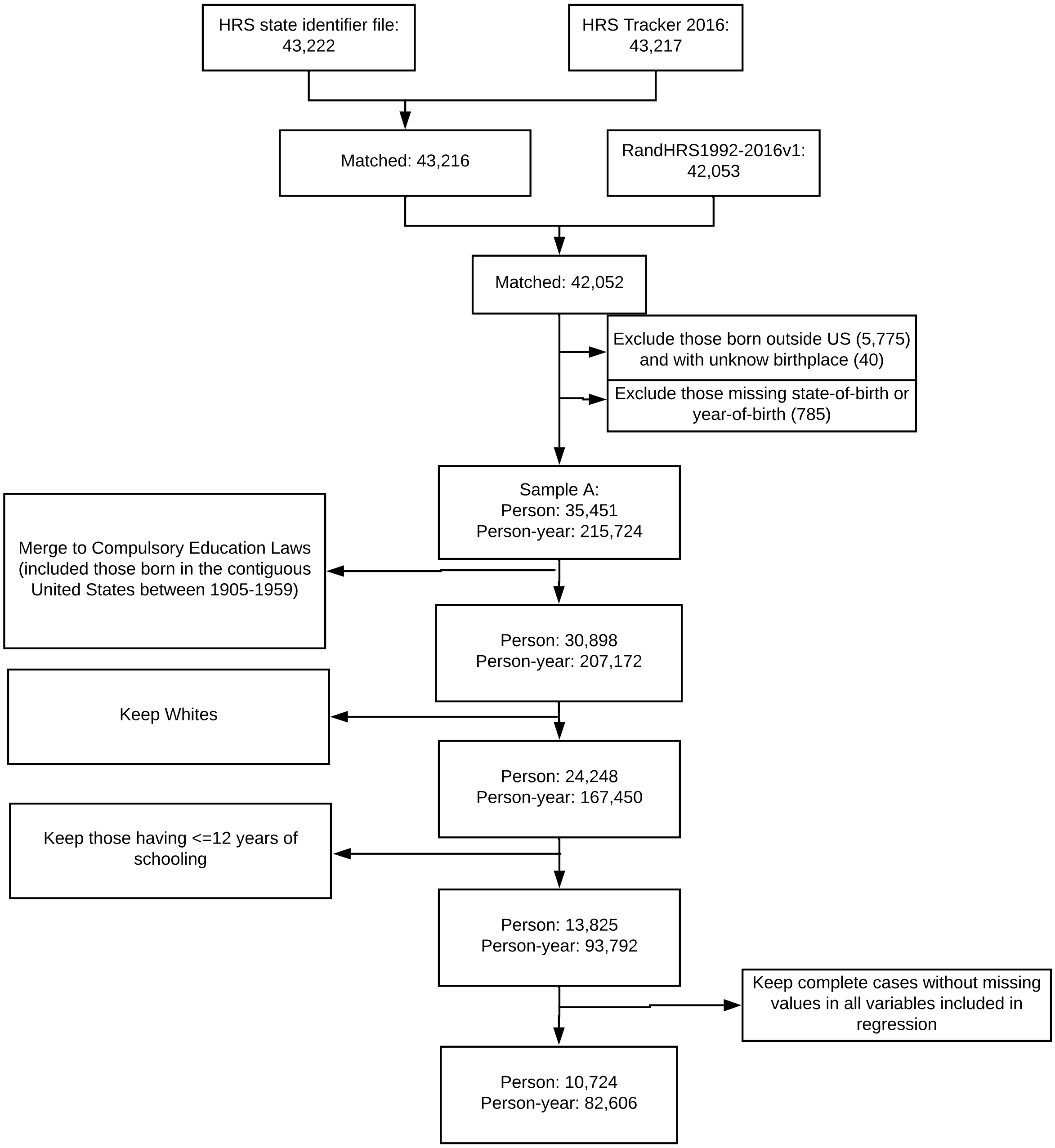

HRS state identifier file: 43,222
HRS Tracker 2016: 43,217
Matched: 43,216
RandHRS1992-2016v1: 42,053
Matched: 42,052
Exclude those born outside US (5,775) and with unknow birthplace (40)
Exclude those missing state-of-birth or year-of-birth (785)
Sample A: Person: 35,451 Person-year: 215,724
Merge to Compulsory Education Laws (included those born in the contiguous United States between 1905-1959)
Person: 30,898 Person-year: 207,172
Keep Whites
Person: 24,248 Person-year: 167,450
Keep those having <=12 years of schooling
Person: 13,825 Person-year: 93,792
Keep complete cases without missing values in all variables included in regression
Person: 10,724 Person-year: 82,606

Table 5-3-1. Summary statistics of individuals in the IV analysis

	Mean	Standard Deviation
The probability of hospitalizations	0.170	0.376
Years of completed education	10.881	1.969
Female	0.589	0.492
Self-rated child health		
Excellent	0.476	0.499
Very good	0.271	0.444
Good	0.187	0.390
Fair	0.050	0.217
Poor	0.017	0.130
Childhood family financial situation		
Well-off	0.657	0.475
Poor	0.343	0.475
Parents' years of completed education	9.321	2.910
Continuation laws	0.634	0.482
Years of schooling by child attendance laws	9.581	1.780
Years of schooling by child labor laws	7.804	1.183
Years of required schooling	8.194	1.194
Pupil-teacher ratio	27.125	4.349
Length of school tern	175.524	8.923
Teachers' average wage	2978.290	2195.698
Relative teachers' average wage	1.020	0.203
Region of Residence at baseline		
South	0.335	0.472
Midwest	0.358	0.479
Northeast	0.222	0.416
West	0.084	0.278
Year of birth		
1905 - 1910	0.026	0.160
1911 - 1920	0.151	0.358
1921 - 1930	0.220	0.414
1931 - 1940	0.298	0.458
1941 - 1950	0.184	0.387
1951 - 1959	0.120	0.325
N	10,724	

Notes: The analytic sample was restricted to White respondents who were born in the continental United States between 1905 and 1959.

Table 5-3-2. Correlation matrix of compulsory schooling laws and school resources

	Continuation laws	Schooling by child labor Laws	Schooling by child attendance Laws	Dropout age	Required schooling	Pupil teacher ratio	Length of term	Teachers' wage	Relative teachers' wage
Continuation laws	1								
Schooling by child labor Laws	0.1352	1							
Schooling by child attendance Laws	0.2619	0.5344	1						
Dropout age	0.1370	0.4560	0.1194	1					
Required schooling	0.1349	0.3384	0.5685	0.1537	1				
Pupil teacher ratio	-0.1091	-0.0633	-0.1288	-0.0107	-0.3787	1			
Length of term	0.3082	0.1062	0.0951	0.1720	0.3748	-0.6171	1		
Teachers' wage	0.1518	0.2109	0.1612	0.1919	0.2742	-0.5177	0.4302	1	
Relative teachers' wage	0.1494	0.0108	-0.0635	0.0467	0.0715	-0.0165	0.4555	0.1659	1

Notes: the correlation matrix was constructed based on state-level compulsory schooling laws and quality of school measures between 1919 and 1973 in 49 states (including District of Columbia); corresponding to birth cohort 1905-1959. The total number of observations is 2,695.

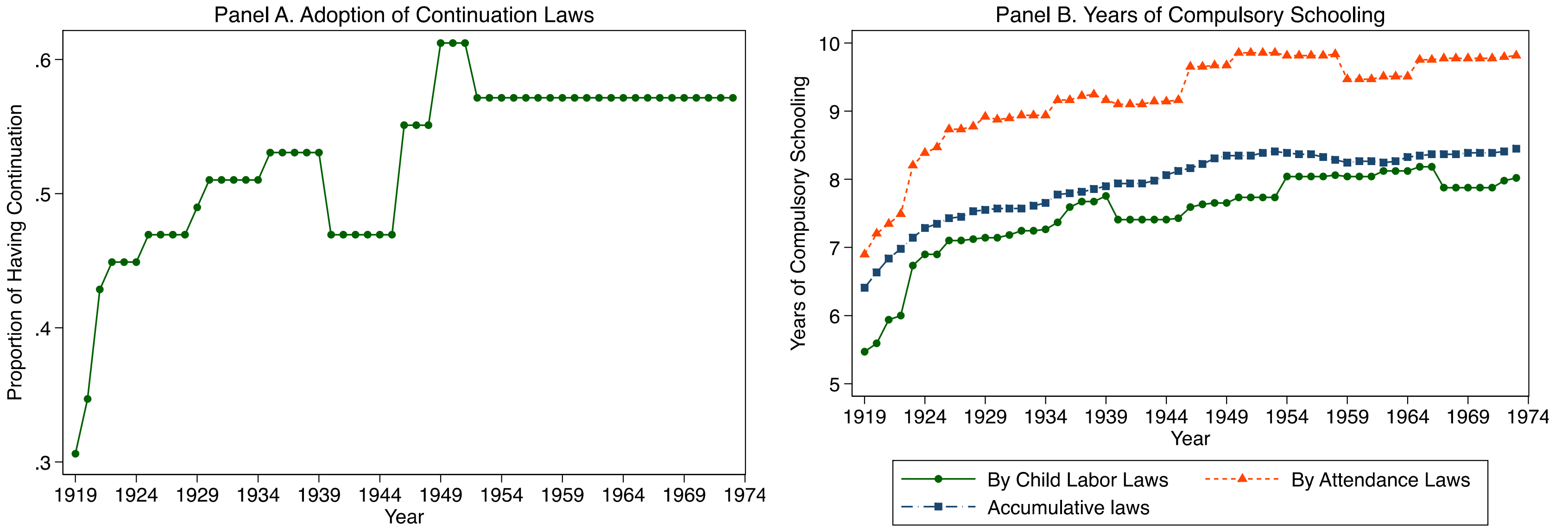
Table 5-3-3. Summary Statistics of compulsory schooling laws and school resources

	Raw	Control for State & Year FE		Control for Region Trends		Control for State Trends	
	Mean	Mean	Reductions in SD %	Mean	Reductions in SD %	Mean	Reductions in SD %
Continuation laws	0.5	0	-48.0	0	-50.0	0	-62.0
	(0.5)	(0.3)		(0.3)		(0.2)	
Schooling by child labor Laws	7.6	0	-14.7	0	-14.7	0	-26.6
	(1.4)	(1.2)		(1.2)		(1.1)	
Schooling by child attendance Laws	9.4	0	-26.4	0	-26.4	0	-46.6
	(1.9)	(1.4)		(1.4)		(1.0)	
Dropout age	14.4	0	-23.8	0	-25.0	0	-29.3
	(1.6)	(1.3)		(1.2)		(1.2)	
Required schooling	8.0	0	-29.9	0	-31.4	0	-49.6
	(1.4)	(1.0)		(0.9)		(0.7)	
Pupil teacher ratio	26.8	0	-60.3	0	-70.2	0	-82.3
	(5.5)	(2.2)		(1.6)		(1.0)	
Length of term	173.3	0	-40.9	0	-56.8	0	-74.2
	(11.1)	(6.6)		(4.8)		(2.9)	
Teachers' wage	2927.1	0	-87.4	0	-87.9	0	-94.7
	(2366.5)	(297.4)		(286.9)		(126.6)	
Relative teachers' wage	1.0	0	-47.4	0	-47.4	0	-63.2
	(0.2)	(0.1)		(0.1)		(0.1)	

Notes: Standard deviations in parentheses. The numbers for "control for State & Year FE" were the mean and standard deviation of predicted residuals from regressing laws and quality measures on state-of-birth and year-of-birth dummies. The numbers for "control for Region Trends" were the mean and standard deviation of predicted residuals from regressions with state-of-birth dummies, year-of-birth dummies, and region-specific linear time trends. Similarly, the numbers for "control for Region Trends" were the mean and standard deviation of predicted residuals from regressions with state-of-birth dummies, year-of-birth dummies, and state-of-birth-specific linear time trends.

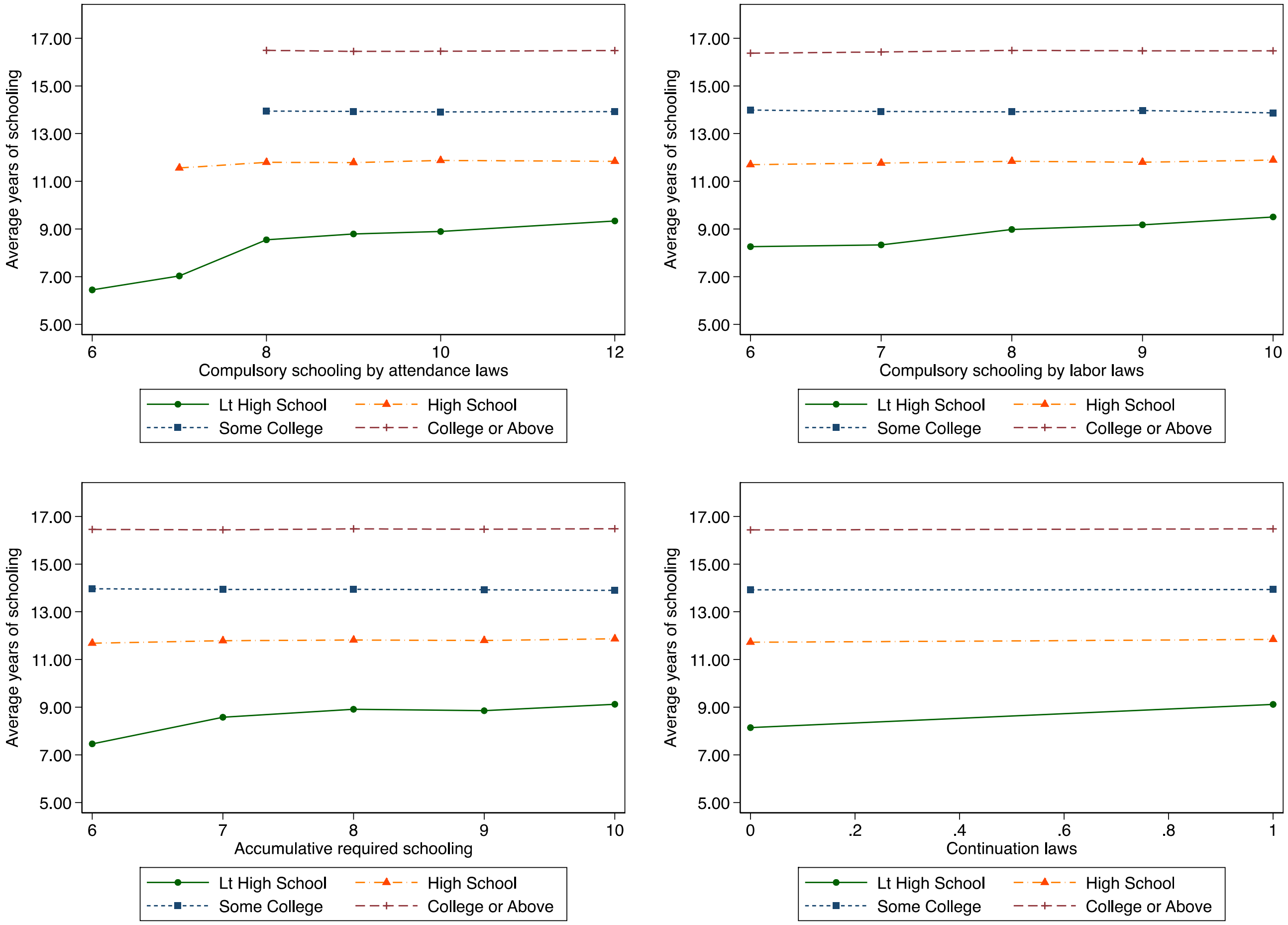
Data were state-level compulsory schooling laws and quality of school measures between 1919 and 1973 in 49 states (including District of Columbia); corresponding to birth cohort 1905-1959. The total number of observations is 2,695.

Figure 5-3-2. Trends in compulsory schooling laws from 1919 to 1973



Notes: Shown are aggregate average of state-level compulsory schooling laws and quality of school measures between 1919 and 1973 in 49 states (including District of Columbia). Since I matched each individual to the laws that were in place in their state-of-birth when they were 14 years old, these trends correspond to birth cohort 1905-1959. The total number of observations is 2,695 including 49 states for 55 years.

Figure 5-3-3. Average years of schooling and compulsory years, by educational categories



Notes: These graphs were based on the analytic sample that includes 10,724 unique respondents. Shown are the aggregate average of years of completed education by compulsory schooling and educational categories. To ensure stable estimates, only those “compulsory schooling” X “educational category” cells with 100 observations were included.

Table 5-3-4. Effects of compulsory schooling laws on education

	(1)	(2)	(3)	(4)	(5)	(6)
Continuation laws	0.420***	0.426***	0.197**	0.145	0.041	0.132
	(0.061)	(0.061)	(0.093)	(0.093)	(0.096)	(0.115)
Compulsory schooling by labor laws	0.086***	0.067**	0.059	0.020	-0.045	-0.029
	(0.028)	(0.028)	(0.036)	(0.036)	(0.036)	(0.037)
Compulsory schooling by attendance laws	0.037*	0.033	0.017	0.007	-0.008	-0.041
	(0.021)	(0.020)	(0.031)	(0.030)	(0.030)	(0.034)
Accumulative required schooling	0.065**	0.027	0.185***	0.092*	0.050	0.126**
	(0.033)	(0.034)	(0.047)	(0.048)	(0.047)	(0.063)
Female	0.339***	0.358***	0.367***	0.383***	0.386***	0.377***
	(0.060)	(0.058)	(0.057)	(0.056)	(0.055)	(0.054)
Childhood Health (Ref: Excellent)						
Very Good	-0.091	-0.074	-0.083	-0.069	-0.071	-0.068
	(0.061)	(0.059)	(0.060)	(0.058)	(0.058)	(0.057)
Good	-0.472***	-0.448***	-0.457***	-0.431***	-0.424***	-0.418***
	(0.077)	(0.077)	(0.075)	(0.075)	(0.074)	(0.074)
Fair	-0.427***	-0.416***	-0.418***	-0.408***	-0.406***	-0.399***
	(0.142)	(0.139)	(0.133)	(0.131)	(0.130)	(0.131)
Poor	-0.749**	-0.731**	-0.678**	-0.663**	-0.635**	-0.641**
	(0.302)	(0.292)	(0.280)	(0.269)	(0.268)	(0.268)
Childhood Family SES (Ref: Well off)						
Poor	-0.486***	-0.480***	-0.467***	-0.456***	-0.455***	-0.458***
	(0.060)	(0.059)	(0.058)	(0.057)	(0.057)	(0.056)
Parents' highest education	0.183***	0.172***	0.152***	0.137***	0.143***	0.144***
	(0.011)	(0.012)	(0.011)	(0.011)	(0.011)	(0.011)
Constant	7.453***	7.314***	6.786***	7.221***	-68.776***	-104.874
	(0.282)	(0.517)	(0.449)	(0.642)	(15.356)	(64.893)
Observations	82606	82606	82606	82606	82606	82606
F-statistic on instrument	26.08	20.9	10.8	2.24	.59	1.55
Year of Birth	No	Yes	No	Yes	Yes	Yes
State of Birth	No	No	Yes	Yes	Yes	Yes
Region linear trends	No	No	No	No	Yes	No
State linear trends	No	No	No	No	No	Yes

Notes: All estimators are from weighted OLS models that use inverse probability weighting to account for attrition bias. Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1

Table 5-3-5. Effect of years of schooling on hospitalizations—OLS results

	Unweighted OLS		Weighted OLS	
	(1)	(2)	(3)	(4)
Years of Schooling	-0.007***	-0.007***	-0.007***	-0.008***
	(0.001)	(0.001)	(0.002)	(0.002)
Female	-0.008*	-0.008*	-0.011*	-0.011*
	(0.005)	(0.005)	(0.006)	(0.006)
Childhood Health (Ref: Excellent)				
Very Good	0.008	0.008	0.012	0.012
	(0.005)	(0.005)	(0.008)	(0.008)
Good	0.019***	0.018***	0.027***	0.026***
	(0.006)	(0.006)	(0.008)	(0.008)
Fair	0.074***	0.075***	0.097***	0.097***
	(0.011)	(0.011)	(0.014)	(0.014)
Poor	0.113***	0.112***	0.122***	0.122***
	(0.022)	(0.022)	(0.024)	(0.024)
Childhood Family SES (Ref: Well off)				
Poor	0.013***	0.012**	0.014**	0.013*
	(0.005)	(0.005)	(0.007)	(0.007)
Parents' highest education	0.001	0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.489***	3.485	0.527***	4.082
	(0.052)	(5.228)	(0.075)	(5.831)
Observations	82606	82606	82606	82606
State of Birth	Yes	Yes	Yes	Yes
Year of Birth	Yes	Yes	Yes	Yes
State linear trends	No	Yes	No	Yes

Notes: Weighted OLS uses inverse probability weighting to account for attrition bias. Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1

Table 5-3-6. Effect of education on hospitalizations—PDS LASSO with selection on controls

	(1) Unweighted PDS	(2) Weighted PDS
Years of Schooling	-0.010***	-0.011***
	(0.001)	(0.002)
Female	-0.007	-0.003
	(0.005)	(0.007)
Childhood Health (Ref: Excellent)		
Very Good	0.012**	0.021***
	(0.005)	(0.008)
Good	0.024***	0.033***
	(0.006)	(0.009)
Fair	0.071***	0.093***
	(0.011)	(0.015)
Poor	0.115***	0.129***
	(0.022)	(0.024)
Childhood Family SES (Ref: Well off)		
Poor	0.016***	0.017**
	(0.005)	(0.007)
Parents' highest education	-0.004***	-0.007***
	(0.001)	(0.001)
Selected state-of-birth fixed effects		
Texas	-0.034***	-0.046***
	(0.011)	(0.014)
Virginia	-0.016	
	(0.015)	
Selected year-of-birth fixed effects		
1909	0.089***	
	(0.033)	
1912	0.125***	0.145***
	(0.023)	(0.035)
1913	0.086***	
	(0.023)	
1915	0.109***	
	(0.020)	
1918	0.090***	0.119***
	(0.017)	(0.023)
1921	0.099***	
	(0.017)	
1924		0.107***
		(0.024)
1942		-0.079***
		(0.016)
1949		-0.096***
		(0.019)
1955		-0.112***
		(0.022)
1959		-0.119***
		(0.022)
Constant	0.401***	0.482***
	(0.017)	(0.022)
Observations	82,606	82,606

Notes: Weighted OLS uses inverse probability weighting to account for attrition bias. Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1. The LASSO regression only selected a subset of state and year fixed effects among all fixed effects and state-specific linear time trends. All individual-level characteristics were partialled out from the LASSO regression.

Table 5-3-7. Effect of years of secondary schooling on hospitalizations—P2SLS results

	(1)	(2)	(3)	(4)	(5)
	AA	ALM	SY	FULL	SEL
Years of Schooling	0.020	-0.012	0.011	0.010	-0.008
	(0.033)	(0.042)	(0.045)	(0.025)	(0.042)
Female	-0.021	-0.009	-0.018	-0.018	-0.011
	(0.014)	(0.017)	(0.018)	(0.012)	(0.017)
Childhood Health (Ref: Excellent)					
Very Good	0.014*	0.012	0.013	0.013*	0.012
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Good	0.038**	0.024	0.034	0.034**	0.026
	(0.016)	(0.019)	(0.021)	(0.014)	(0.020)
Fair	0.108***	0.096***	0.105***	0.105***	0.097***
	(0.019)	(0.021)	(0.023)	(0.017)	(0.021)
Poor	0.139***	0.119***	0.134***	0.133***	0.121***
	(0.035)	(0.036)	(0.039)	(0.031)	(0.037)
Childhood Family SES (Ref: Well off)					
Poor	0.025	0.011	0.021	0.021	0.013
	(0.017)	(0.021)	(0.021)	(0.013)	(0.021)
Parents' highest education	-0.005	-0.000	-0.003	-0.003	-0.001
	(0.005)	(0.006)	(0.007)	(0.004)	(0.006)
State of Birth	Yes	Yes	Yes	Yes	Yes
Year of Birth	Yes	Yes	Yes	Yes	Yes
State linear trends	Yes	Yes	Yes	Yes	Yes
Constant	7.233	3.562	6.234	6.157	4.027
	(6.718)	(7.422)	(7.616)	(6.307)	(7.446)
Observations	82,606	82,606	82,606	82,606	82,606
F statistics on instruments	1.58	1.78	2.97	1.53	1.26

Notes: All models are from a pooled two-stage least square (P2SLS) method with inverse probability weighting to account for attrition bias. Different sets of instruments are included in the models. **Column 1** includes CA7, CA8, CA9, CA10, CL7, CL8, CL9 (Acemoglu and Angrist 2000) as instruments; **Column 2** uses continuation laws, CL7, CL8, CL9 (Lleras-Muney 2005); **Column 3** uses RS7, RS8, RS9 (Stephens Jr and Yang 2014), **Column 4** includes all relevant instruments: include continuation laws, years of compulsory schooling required by child labor laws (CL_{st}), years of compulsory schooling required by child labor laws (CA_{st}), Leaving age (LA_{st}), years of required schooling (RS_{sct}), CL7, CL8, CL9, CA7, CA8, CA9, CA10, RS7, RS8, and RS9; and **Column 5** includes a set of instrument selected by a LASSO regression of education on the full set of instruments.
Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1.

Table 5-3-8. Effect of years of secondary schooling on hospitalizations—LASSO-IV results

	(1) LASSO-IV	(2) GMM-IV	(3) OLS
Years of Schooling	-0.065***	-0.064***	-0.011***
	(0.013)	(0.013)	(0.002)
Female	0.015*	0.015*	-0.003
	(0.009)	(0.008)	(0.007)
Childhood Health (Ref: Excellent)			
Very Good	0.016*	0.017*	0.021***
	(0.009)	(0.009)	(0.008)
Good	0.007	0.009	0.033***
	(0.012)	(0.012)	(0.009)
Fair	0.068***	0.069***	0.093***
	(0.018)	(0.017)	(0.015)
Poor	0.085***	0.085***	0.129***
	(0.027)	(0.027)	(0.024)
Childhood Family SES (Ref: Well off)			
Poor	-0.010	-0.010	0.017**
	(0.010)	(0.010)	(0.007)
Parents' highest education	0.003	0.002	-0.007***
	(0.003)	(0.003)	(0.001)
Selected state-of-birth fixed effects			
Texas	-0.116***	-0.115***	-0.046***
	(0.024)	(0.024)	(0.014)
Selected year-of-birth fixed effects			
1912	0.089**	0.092**	0.145***
	(0.042)	(0.042)	(0.035)
1918	0.113***	0.118***	0.119***
	(0.029)	(0.029)	(0.023)
1924	0.057*	0.061**	0.107***
	(0.029)	(0.029)	(0.024)
1942	-0.066***	-0.065***	-0.079***
	(0.016)	(0.016)	(0.016)
1949	-0.089***	-0.088***	-0.096***
	(0.020)	(0.020)	(0.019)
1955	-0.088***	-0.087***	-0.112***
	(0.024)	(0.024)	(0.022)
1959	-0.107***	-0.106***	-0.119***
	(0.022)	(0.022)	(0.022)
Constant	0.997***		0.482***
	(0.128)		(0.022)
Observations	82,606	82,606	82,606
F statistics on instruments	20.61	NA	NA

Notes: Column 1 reports P2SLS estimators using selected instruments and controls from the PDS method. The selected set of instruments include continuation laws, years of compulsory schooling required by child labor laws (CL_{sct}), years of compulsory schooling required by child labor laws (CA_{sct}), years of required schooling (RS_{sct}), and whether $CL_{sct} = 8$ (CL8). Selected set of controls are displayed in the table; all individual-level characteristics were partialled out from the LASSO regression. Column 2 documents GMM-IV estimators using selected instruments and controls. Column 3 reports OLS estimators of regressing hospitalization on selected controls.

Post-estimation of the GMM-IV model: Anderson-Rubin Wald test for Weak-instrument-robust inference shows that Chi-sq(5)= 40.88 (p<0.001) rejecting that weak instruments. Kleibergen-Paap rk LM statistic for the under-identification test is 76.86 (p<0.001) suggesting the model is identified. Hansen J statistic is 6.291 (p =0.1785), which shows no evidence of violations of exclusion restrictions.

All models apply inverse probability weighting to account for attrition bias. Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1.

Table 5-3-9. Effect of education on hospitalizations—LASSO-IV results, by age

	Younger than 78		Older than 78	
	LASSO-IV	GMM-IV	LASSO-IV	GMM-IV
Years of Schooling	-0.066**	-0.066**	0.001	0.001
	(0.028)	(0.028)	(0.011)	(0.011)
Female	0.001	0.001	-0.006	-0.007
	(0.012)	(0.012)	(0.013)	(0.013)
Childhood Health (Ref: Excellent)				
Very Good	0.016*	0.016*	0.010	0.010
	(0.009)	(0.009)	(0.015)	(0.015)
Good	0.009	0.009	0.019	0.019
	(0.016)	(0.016)	(0.019)	(0.019)
Fair	0.068***	0.068***	0.104***	0.102***
	(0.021)	(0.021)	(0.026)	(0.026)
Poor	0.082**	0.082**	0.116***	0.118***
	(0.035)	(0.035)	(0.036)	(0.036)
Childhood Family SES (Ref: Well off)				
Poor	0.000	0.000	0.008	0.009
	(0.013)	(0.013)	(0.015)	(0.015)
Parents' highest education	0.006	0.006	-0.003	-0.003
	(0.005)	(0.005)	(0.004)	(0.004)
Selected state-of-birth fixed effects				
Texas	-0.111***	-0.111***		
	(0.037)	(0.037)		
Constant	0.953***		0.458***	
	(0.264)		(0.104)	
Observations	63,728	63,728	18,878	18,878
F statistic on instruments	45.34		32.02	
Kleibergen-Paap rk LM statistic		22.6***		72.0***
Hansen J statistic		NA		3.8

Notes: Selected instruments for the “younger than 78” group includes continuation laws, whereas selected instruments for the “older than 78” group includes continuation laws, years of compulsory schooling required by child labor laws (CA_{sct}), years of required schooling (RS_{sct}), and whether $CL_{sct} = 8$ (CL8).

Kleibergen-Paap rk LM statistics for the under-identification test suggest all models are identified. The non-significance of Hansen J statistics suggests no evidence of violations of exclusion restrictions.

All models apply inverse probability weighting to account for attrition bias. Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1.

Table 5-3-10. Effect of education on hospitalizations—LASSO IV with quality of schooling

	(1) LASSO-IV	(2) GMM-IV
Years of Schooling	-0.078***	-0.083***
	(0.011)	(0.011)
Female	0.019**	0.020**
	(0.009)	(0.009)
Childhood Health (Ref: Excellent)		
Very Good	0.015	0.013
	(0.009)	(0.009)
Good	0.001	-0.000
	(0.012)	(0.012)
Fair	0.063***	0.058***
	(0.018)	(0.018)
Poor	0.075***	0.079***
	(0.028)	(0.028)
Childhood Family SES (Ref: Well off)		
Poor	-0.017*	-0.021**
	(0.010)	(0.010)
Parents' highest education	0.005**	0.006**
	(0.002)	(0.002)
Selected state-of-birth fixed effects and trends		
Texas	3.988	4.275
	(2.933)	(2.933)
Texas X Year	-0.002	-0.002
	(0.002)	(0.002)
Selected year-of-birth fixed effects		
1912	0.075*	0.077*
	(0.043)	(0.043)
1918	0.108***	0.110***
	(0.031)	(0.031)
1924	0.045	0.042
	(0.030)	(0.030)
1942	-0.062***	-0.057***
	(0.017)	(0.017)
1949	-0.086***	-0.080***
	(0.020)	(0.020)
1955	-0.079***	-0.073***
	(0.024)	(0.024)
1959	-0.100***	-0.094***
	(0.022)	(0.022)
Constant	1.115***	
	(0.109)	
Observations	82606	82606
F statistic on instruments	67.77	NA
Kleibergen-Paap rk LM statistic		88.9***
Hansen J statistic		33.6***

Notes: Selected instruments includes pupil-teacher ratio and length of school term.

Kleibergen-Paap rk LM statistics for the under-identification test suggest all models are identified. The significance of Hansen J statistics suggests violations of exclusion restrictions.

All models apply inverse probability weighting to account for attrition bias. Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1.

Table 5-3-11. Effect of education on hospitalizations—LASSO IV with weights from whites

	(1) LASSO-IV	(2) GMM-IV
Years of Schooling	-0.079***	-0.078***
	(0.017)	(0.016)
Female	0.020**	0.021**
	(0.010)	(0.010)
Childhood Health (Ref: Excellent)		
Very Good	0.016*	0.018*
	(0.010)	(0.010)
Good	0.002	0.004
	(0.014)	(0.014)
Fair	0.065***	0.067***
	(0.020)	(0.019)
Poor	0.073**	0.075**
	(0.031)	(0.031)
Childhood Family SES (Ref: Well off)		
Poor	-0.016	-0.016
	(0.011)	(0.011)
Parents' highest education	0.005	0.005
	(0.003)	(0.003)
Selected state-of-birth fixed effects		
Texas	-0.127***	-0.128***
	(0.027)	(0.027)
Selected year-of-birth fixed effects		
1918	0.121***	0.124***
	(0.032)	(0.032)
1924	0.044	0.048
	(0.033)	(0.033)
1942	-0.066***	-0.065***
	(0.018)	(0.018)
1949	-0.091***	-0.091***
	(0.020)	(0.020)
1955	-0.088***	-0.087***
	(0.025)	(0.025)
1959	-0.108***	-0.107***
	(0.023)	(0.023)
Constant	1.133***	
	(0.157)	
Observations	82606	82606
F statistic on instruments	21.81	NA
Kleibergen-Paap rk LM statistic		64.3***
Hansen J statistic		2.8

Notes: Selected instruments includes continuation laws, years of compulsory schooling required by child labor laws (CL_{st}), years of compulsory schooling required by child labor laws (CA_{st}), and years of required schooling (RS_{sct}). Kleibergen-Paap rk LM statistics for the under-identification test suggest all models are identified. The non-significance of Hansen J statistics provides no evidence of violations of exclusion restrictions.

All models apply inverse probability weighting to account for attrition bias with weights constructed based on white persons only. Standard errors are clustered at individual levels. *** p<0.01, ** p<0.05, * p<0.1.

Chapter 6. Discussion

6.1 Summary and Interpretation of Findings

6.1.1 Attrition Study

Results from the attrition analyses support the hypothesis in Section 3.2 that those with lower socioeconomic status and worse health status are more likely to drop out of the survey.

First, baseline characteristics are significant predictors of respondents' attrition status in 2016. Individuals who were female, white, Hispanic, married, and who had more living children were more likely to remain in the survey and respond in every follow-up wave. Lower health status led to both a lower likelihood of retention and a higher probability of death regardless of whether health is self-reported or measured by diagnosed health conditions. Socioeconomic status, particularly education and income, had significant influences on attrition; those having higher educational attainment and higher income were more likely to respond to the survey at each wave.

Second, using one wave lagged measure for time-varying variables such as income in a dynamic model yields similar results relative to using the baseline measures in terms of both magnitude and significance. It indicates that the influences of random shocks on these variables were limited. The coefficients on demographic variables are also similar. However, there are some notable changes in the effect of educational attainment; only "college or above" is significantly linked to a higher probability of remaining in the sample. It is not surprising because of the lagged income, as an important pathway running from education to outcomes, better captures the effect of education than the baseline income measures.

Importantly note that many variables have different impacts on death and non-response. For instance, those born in the US were more likely to attrite due to death but less likely to attrite due to non-response. Also, those having education of college and above had a higher probability of death, though not statistically significant, but a lower probability of non-response. These findings are important as it suggests that drivers of attrition due to death and non-response are different. As such, in constructing weights for adjustment of attrition bias, we should model the probability of survival and response separately.

Results from this study are generally consistent with the analysis of the 2002 attrition status of respondents from the original HRS cohort (Kapteyn et al. 2006), or based on both HRS and AHEAD cohort (Cao and Hill 2005), but with some important exceptions. In particular, my results demonstrated that more years of schooling significantly reduced the probability of attrition due to death, but Kapteyn et al. (2006) found mixed and insignificant estimates on educational attainment. This might because a few numbers of respondents died in 2002 compared to 2016 (as shown in **Figure 5-1-3**), and the attrition bias related to education is not a serious concern (see **Figure 5-2-3**). Consistent with another study focusing on racial/ethnic differences in 2008 attrition status, my results also showed that race/ethnicity played an important role in respondents' attrition status in 2016.

6.1.2 Association Study

There are three important findings from the association study. For the overall sample, higher educational attainment is associated with a lower probability of being hospitalized. Those with a college or above degree saw a larger effect (-8.4 pp) than those with high school degrees (-3.4 pp), relative to their counterparts who had education less than high school.

Attrition bias matters for the estimates of education; the effect of education on hospitalization would be underestimated for both "college or above" and "high school/some college" if attrition bias was ignored. Estimates from models accounting for attrition bias are significantly larger than those from models without controlling for attrition bias.

Age modifies the relationship between education and hospitalizations, particularly for the effect of "high school or some college." Before age 78, both an educational level of "high school/some college" and of "college or above" had a significantly negative effect on hospitalizations with the larger effect size for the latter. However, after 78, all of these effects decreased. The probability of hospitalizations among those having an education level of high school was no longer distinguishable from that of those having education less than high school. The impact of a degree as college or above on hospitalization, though attenuated, remains statistically significant. These results are consistent with the majority of education-health literature and support the age-as-levelers hypothesis, which posits that health is more age-dependent at older ages than at younger ages (House et al. 1994, Lynch 2003). Thus, education gradients in health should be larger at younger ages. For example, Elo and Preston (1996) show the largest effect of education on mortality occur among persons of working ages. Similarly, my results also show an attenuated effect of education on hospitalizations, but those with a college degree or above still saw a significantly decreased probability of being hospitalized. Another reason for the decreased education differentials in health is selective mortality that the surviving population at older ages consist of robust persons with both low and high levels of education (Beckett 2000). My results also support this explanation as the education gap in hospitalizations increased after accounting for attrition due to death or non-response. However, it is unable to fully account for the bias due to selective survival because we only included a limited set of observables in constructing attrition weights.

6.1.3 Causal Effects

I found a substantive negative causal effect of secondary schooling on hospitalizations. One additional year of secondary schooling lowers the probability of two-year hospitalizations by 6.5 percentage points. This effect only appeared to be large and significant (-6.6 percentage points) when respondents were younger than 78. After that, it becomes indistinguishable from zero (0.1 percentage points, p = 0.909), consistent with the age-as-levelers hypothesis.

The IV estimator (-0.065, 95% CI: -0.091 to -0.039) is significantly larger than the OLS estimator (-0.011, 95% CI: -0.014 to -0.007). This could be explained by at least two reasons. First, the OLS estimator is biased, large due to omitted confounding variables. Second, the OLS estimator represents the Average Treatment Effect (ATE) while the IV estimator reflects the Local Average Treatment Effect (LATE). The LATE measures the average causal effects among the compliers rather than the general population. In this specific study, it represents the causal effect of schooling on hospitalizations among those who would drop out of school if these compulsory schooling laws were then not in place.

The IV estimator is larger than those from previous studies, largely due to the difference in the way how hospitalizations were measured and the research methods. My IV estimator (6.5 pp, $p<0.01$) on the probability of two-year hospitalizations is more than twice larger than (2.7, $p>0.05$) that from a similar study on the probability of annual hospitalization (Mazumder 2008). The Mazumder (2008) study, based on data from the Survey of Income and Program Participation panel data, uses a set of dummy variables of years of compulsory schooling required by child labor laws as instruments. The smaller estimate from his study might be due to the inclusion of those older than 78 years old and those having education higher than high school. Besides, the inclusion of state-specific linear time trends could explain the loss of significance in his study.

My finding that more years of education are causally linked to a lower probability of hospitalization is roughly consistent with a study using a similar method (Arendt 2008); the author finds that having an education beyond primary schooling reduces the probability of hospitalization in a given year by 1.9 pp, which is statistically significant, for women, but no significant impact for men. But my finding contrasts with the result of a within-twins study design that fails to detect any significant effects of schooling on hospitalizations (Behrman et al. 2011).

6.2 Implications for Research and Policy

Understanding the causal impact of educational attainment on health care use is important to design the most effective social and health policies related to educational attainment. This study provides a systematic analysis of the relationship between educational attainment and hospitalizations using a longitudinal dataset. It leverages rigorous quantitative methods and shows a substantial and significant education effect on reducing hospitalizations. This dissertation should contribute to research and policymaking in several ways.

My study highlights the attrition bias in the HRS matters and sets up a framework to account for it. It should set the stage for future longitudinal analysis based on HRS and other panel surveys. More specifically, I found that determinants of attrition due to death are different from those of attrition due to non-response. As such, these two modes of attrition should be treated separately. Since those with lower socioeconomic status were less likely to stay in the follow-up surveys, empirical studies would underestimate the effect of those variables on health outcomes if attrition bias left uncontrolled. The framework used in this study should advance the relevant health services research based on longitudinal data.

Although it is well-established that education is one of the greatest correlates of health, there is still substantial uncertainty as to what extent this relationship is causal. Prior studies that leverage school reforms or educational legislations to uncover the causal effects of education on health provided imprecise estimates and thus cannot reach an agreement. In particular, those studies in the United States employing compulsory schooling laws as instruments to address the endogeneity of education were constrained by weak instruments, especially when state-specific linear time trends were accounted for. My study overcomes this issue by using a consistent but more efficient approach that applies a LASSO regression model to select both the optimal instruments and the parsimonious set of controls. This innovation should guide further research in evaluating the effect of education on other outcomes, such as health expenditure and longevity, and, more generally, guide those studies involving many instruments and many controls. Also, as more data are available for researchers in health sectors and other industries, this method could also be applied to identify the most relevant set of control variables and reduce the dimension of big data (e.g., electronic health records, and administrative data) in health policy and services research. For example, we could use the method of selecting a parsimonious set of health measures from the electronic health data to study the impact of certain policy reforms or clinical procedures on health and health care utilization outcomes.

In this study, I investigate the gross effect of educational attainment on hospitalizations. I did not adjust for potential mediators, such as income, occupation, and wealth. That said, the estimates from this study indicate the total effect of education on hospitalization, not the effect net of mediating pathways. The result that educational attainment has a large effect on hospitalizations contributes to the growing literature on social determinants of health. Educational attainment is highly related to other socioeconomic factors such as income, wealth, and occupation, thus results from this study could indicate how social factors could be used as policy levers to improve health and reduce health care costs. Moreover, hospital care is expensive and represents the largest share of overall health care costs in the US health care system. As health policymakers and researchers are seeking solutions to reduce health care costs, and health disparities attached to socioeconomic status, policy reforms that address social determinants of health could be an effective option. However, such policy reforms are struggling with limited rigorous evidence on the relationship between social factors and health care utilization. Results from this study should inform policymakers that providing more health care resources to those with less educational attainment might be an effective means to reduce health disparities. For instance, in the current health care system, there are many means-tested programs primarily based on income, such as Medicaid and the Supplemental Nutrition Assistance Program (SNAP), which have contributed to reducing socioeconomic disparities in health. This study provides evidence that education also matters for health, which could also be considered in addition to income in the policymaking process. Compared to income, education can be measured more accurately and is relatively stable for those in the middle- and older ages. For example, we could consider more investment in safety net providers that may have less educational attainment or targeting more resources to areas with lower educational attainment. In a broader context, it also suggests that investment in the educational system could be a more cost-effective way to reduce intense health care use and health care costs compared to increased expenditures in the health care system.

My study also provides rigorous evidence for current policy reforms that considering integrating social factors into the health care delivery system. One notable example is the current ongoing value-based payment reforms that aim to shift the focus of care from quantity to quality by financially penalizing or rewarding health care providers based on their patients' health outcomes. However, since socially disadvantaged patients, such as those having less educational attainment, are often concentrated among a subset of providers, the quality of care of those providers would be underestimated if patients' characteristics were not appropriately adjusted. A report from National Academies of Science, Engineering, Medicine concluded that incorporating social factors in the Medicare payment schemes would have great implications on quality improvements and cost control. It also highlighted the absence of rigorous empirical evidence (National Academies of Sciences and Medicine 2017). My results should inform these risk adjustment models from at least two aspects. First, educational attainment matters for health and health care. Given the large and significant causal effect of education on hospitalizations, educational attainment should be considered as an important social factor in the risk adjustment model for value-based payment schemes. Admittedly, educational attainment is not available in the Medicare and Medicaid claims data, which precludes such practice. Current studies have considered area-level measures of educational attainment as proxies to improve the risk adjustment, but more research is needed to examine whether they are good proxies for individual-level education. More importantly, as the majority of the Medicare enrollees are aged 65 or above, educational attainment could be collected in the enrollment stage. As my study shows that those with fewer years of schooling had a higher probability of hospitalizations, adding education into the enrollment will facilitate more effective risk adjustment for payment reforms and more targeted resource allocations. Second, since my results suggest a decreased education effect on hospitalization after age 78, such risk adjustment models that consider including educational attainment will achieve better predictive ability if they are stratified by whether patients reach ages 78.

6.3 Limitations

Several limitations of this study need to be acknowledged.

First, for the attrition analysis, the list of variables used in predicting attrition is limited by available individual measures available in the HRS data. Although I improved the prior studies by including an expanded set of predictors, it is possible that we left out some other unobserved and unmeasured factors that could lead to attrition. The inverse probability weighting approach applied in this study to account for attrition bias relies on the assumption missing at random or selection on observables. That says conditional on a set of variables attrition should be a random event, which is an untestable assumption.

Second, similar to other studies based on self-reported survey data, estimators from this study might suffer from recall bias. The questions about hospitalization were based on respondents' experience in the past two years. Although hospitalization should be a memorable event, it is uncertain to what extent reporting errors in hospitalizations influence the estimators in the study. It should not be a concern in the IV analysis, as the measurement errors in hospitalizations and education could be addressed by valid instruments. Moreover, respondents were asked to recall their health status and family financial situation during their childhoods. Given the majority of respondents were elderly adults, recall bias could be a concern for these two variables. Similarly, the IV estimators should be immune to these biases. Since it is rare to follow individuals over the life course, these variables are not even available in many other health surveys. But these variables are worth collecting as most of health and economic outcomes have their roots in childhood health and living conditions.

Third, the association study demonstrates a substantial and significant correlation between more years of schooling and a lower probability of hospitalizations. However, correlation does not mean causation. There are several important variables driving both education and hospitalization that might be missed from the association study. Those variables, for example, include time preference, personality traits, and intelligence. The compulsory schooling laws used as instruments in this study only had impacts on years of secondary schooling, and as such, the IV estimators from this dissertation do not apply to years of education beyond high school or some college.

Fourth, the quality of schooling, though it is important for health outcomes, is not considered in this study. Respondents with the same years of education but from schools with different resources should have different levels of returns to education. It is an important area for future research but beyond the scope of this study. Because of this, my results on the education effect should be interpreted as respondents from schools with an average level of quality. Although I included state-level measures, they tend to be fully captured by state-specific linear time trends and are very likely to be endogenous.

Lastly, the generalizability of results from this study is quite limited, largely due to pooling data from all the cohorts and waves. The emphasis on improving the internal validity of education effect on hospitalizations comes at the cost of external validity. I pooled all observations available in the HRS to maximize the sample size to make sure robust estimates for the IV approach. Thus, the analytic sample includes not only eligible respondents but also their spouses. Nonetheless, a majority of the sample are elderly adults; results from this study could provide informative evidence on the elderly population in general. Moreover, compulsory schooling laws were intended to increase the educational level of those who would otherwise drop out of high school if there were no laws in place. Thus, it makes it difficult to generalize the IV estimators to the general population.

6.4 Conclusions and Future Research

In summary, I found that those with greater educational attainment had a lower probability of two-year hospitalizations compared to those with lower levels of educational attainment. The education-hospitalization relationship would be underestimated if attrition was not accounted for. The association between education and hospitalization is monotonic; the correlation of higher educational levels with reduced hospitalizations is larger than that of a high school degree. Moreover, age modifies the relationship. After age 78, the effect of having an education level of high school or some college on hospitalizations becomes indistinguishable from zero, but the effect of higher educational levels remains significant. Importantly, the IV results provide evidence that suggests that years of secondary schooling have a large causal effect on reducing the probability of hospitalizations.

Future research could estimate the causal relationship between educational attainment on later-life objective health measures and health care use. For example, it is of considerable policy implications to look at how educational attainment could reduce potentially avoidable hospitalizations—hospitalizations that could have been avoided because the disease or symptoms could have been prevented or treated outside of an inpatient hospital setting. It is also essential to directly examine whether those with a higher level of education spend more or less in medical care. Besides, such studies are also quite relevant to the on-going value-based payment reforms, particularly for Medicare. Future research could look at how incorporating individual-level educational attainment change health care providers' performance matrices and the related bonuses and/or penalties.

Future research should also investigate potential reporting bias in health status and health care utilization by different socioeconomic groups. Due to the lack of socioeconomic measures in administrative data, policymakers and researchers rely on surveys to monitor population health and health disparities across different socioeconomic groups. However, potential reporting bias due to varying levels of awareness and of willingness to report in health conditions and health care utilization may distort the validity of estimates from surveys, and therefore, mislead the policymaking decisions. Future research is warranted to investigate the reporting bias and to what extent it affects health disparity studies.

It is also of great policy relevance for identifying potential leverages that help narrow socioeconomic disparities by examining the mechanisms through which education affects health and health care use. For example, if the effects of higher education levels on decreased hospitalizations are through a healthier lifestyle or better access to health care providers, then health promotion and insurance expansion may help narrow the education gradients in hospitalizations. For future research, it is also crucial to investigate the effects of the quality of education on health and medical care use. Returns to education on health are related to the quality of schooling. Understanding the role of quality of schooling in the education-health relationship could help us make more effective policies in education investment.

References

Acemoglu, Daron, and Joshua Angrist. 1999. How large are the social returns to education? Evidence from compulsory schooling laws. *NBER Working Paper* 7444.

Acemoglu, Daron, and Joshua Angrist. 2000. "How large are human-capital externalities? Evidence from compulsory schooling laws." *NBER Macroeconomics Annual* 15:9-59.

Acevedo-Garcia, Dolores, Theresa L Osypuk, Nancy McArdle, and David R Williams. 2008. "Toward a policy-relevant analysis of geographic and racial/ethnic disparities in child health." *Health Affairs* 27 (2):321-333.

Adams, Scott J. 2002. "Educational attainment and health: Evidence from a sample of older adults." *Education Economics* 10 (1):97-109.

Ahrens, A., C.B. Hansen, and M.E. Schaffer. 2018. "pdslasso and ivlasso: Programs for post-selection and post-regularization OLS or IV estimation and inference." Accessed on 20 December 2019. http://ideas.repec.org/c/boc/bocode/s458459.html.

Albouy, Valerie, and Laurent Lequien. 2009. "Does compulsory education lower mortality?" *Journal of Health Economics* 28 (1):155-168.

Amin, Vikesh, Jere R Behrman, and Tim D Spector. 2013. "Does more schooling improve health outcomes and health related behaviors? Evidence from UK twins." *Economics of Education Review* 35:134-148.

Andersen, Per Kragh, and Richard D Gill. 1982. "Cox's regression model for counting processes: a large sample study." *The Annals of Statistics*:1100-1120.

Angrist, Joshua D, and Alan B Krueger. 1992. "The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples." *Journal of the American Statistical Association* 87 (418):328-336.

Angrist, Joshua D, and Alan B Krueger. 1999. "Empirical strategies in labor economics." In *Handbook of Labor Economics*, 1277-1366. Elsevier.

Arbaje, Alicia I, Jennifer L Wolff, Qilu Yu, Neil R Powe, Gerard F Anderson, and Chad Boult. 2008. "Postdischarge environmental and socioeconomic factors and the likelihood of early hospital readmission among community-dwelling Medicare beneficiaries." *The Gerontologist* 48 (4):495-504.

Arendt, J. N. 2008. "In sickness and in health - Till education do us part: Education effects on hospitalization." *Economics of Education Review* 27 (2):161-172.

Arendt, Jacob Nielsen. 2005. "Does education cause better health? A panel data analysis using school reforms for identification." *Economics of Education Review* 24 (2):149-160.

Ashenfelter, Orley, William J Collins, and Albert Yoon. 2006. "Evaluating the role of Brown v. Board of Education in school equalization, desegregation, and the income of African Americans." *American Law and Economics Review* 8 (2):213-248.

Avendano, Mauricio, Hendrik Jürges, and Johan P Mackenbach. 2009. "Educational level and changes in health across Europe: longitudinal results from SHARE." *Journal of European Social Policy* 19 (4):301-316.

Bailey, Martha J, and Susan M Dynarski. 2011. Gains and gaps: Changing inequality in US college entry and completion. *NBER Working Paper* 17633.

Baum, C.F. 2006. An Introduction to Modern Econometrics Using Stata: Taylor & Francis.

Beal, Anne C. 2004. "Policies to reduce racial and ethnic disparities in child health and health care." *Health Affairs* 23 (5):171-179.

Becker, Gary S, and Casey B Mulligan. 1997. "The endogenous determination of time preference." *The Quarterly Journal of Economics* 112 (3):729-758.

Becker, Gary Stanley. 1967. Human capital and the personal distribution of income: An analytical approach: Institute of Public Administration.

Beckett, Megan. 2000. "Converging health inequalities in later life-an artifact of mortality selection?" *Journal of Health and Social Behavior*:106-119.

Behrman, Jere R, Hans-Peter Kohler, Vibeke Myrup Jensen, Dorthe Pedersen, Inge Petersen, Paul Bingley, and Kaare Christensen. 2011. "Does more schooling reduce hospitalization and delay mortality? New evidence based on Danish twins." *Demography* 48 (4):1347-1375.

Behrman, Jere R, and Mark R Rosenzweig. 2004. "Returns to birthweight." *Review of Economics and Statistics* 86 (2):586-601.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. 2012. "Sparse models and methods for optimal instruments with an application to eminent domain." *Econometrica* 80 (6):2369-2429.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2011. "Inference for high-dimensional sparse econometric models." *arXiv preprint arXiv:1201.0220*.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-dimensional methods and inference on structural and treatment effects." *Journal of Economic Perspectives* 28 (2):29-50.

Ben-Porath, Yoram. 1967. "The production of human capital and the life cycle of earnings." *Journal of Political Economy* 75 (4, Part 1):352-365.

Berkman, Nancy D, Stacey L Sheridan, Katrina E Donahue, David J Halpern, and Karen Crotty. 2011. "Low health literacy and health outcomes: an updated systematic review." *Annals of Internal Medicine* 155 (2):97-107.

Black, Dan A, Yu-Chieh Hsu, and Lowell J Taylor. 2015. "The effect of early-life education on later-life mortality." *Journal of Health Economics* 44:1-9.

Black, Sandra E, Paul J Devereux, and Kjell G Salvanes. 2007. "From the cradle to the labor market? The effect of birth weight on adult outcomes." *The Quarterly Journal of Economics* 122 (1):409-439.

Bolin, Kristian, Lena Jacobson, and Björn Lindgren. 2001. "The family as the health producer—when spouses are Nash-bargainers." *Journal of Health Economics* 20 (3):349-362.

Bolin, Kristian, Lena Jacobson, and Björn Lindgren. 2002a. "The demand for health and health investments in Sweden 1980/81, 1988/89, and 1996/97." In *Individual Decisions for Health*, pp. 109-128. Routledge.

Bolin, Kristian, Lena Jacobson, and Björn Lindgren. 2002b. "Employer investments in employee health: Implications for the family as health producer." *Journal of Health Economics* 21 (4):563-583.

Bolin, Kristian, Lena Jacobson, and Björn Lindgren. 2002c. "The family as the health producer—when spouses act strategically." *Journal of Health Economics* 21 (3):475-495.

Bound, John, David A Jaeger, and Regina M Baker. 1995. "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak." *Journal of the American Statistical Association* 90 (430):443-450.

Braakmann, Nils. 2011. "The causal relationship between education, health and health related behaviour: Evidence from a natural experiment in England." *Journal of Health Economics* 30 (4):753-763.

Brittain, John, and Callie Kozlak. 2007. "Racial disparities in educational opportunities in the United States." *Seattle Journal for Social Justice* 6 (2):11.

Buckles, Kasey, Andreas Hagemann, Ofer Malamud, Melinda Morrill, and Abigail Wozniak. 2016. "The effect of college education on mortality." *Journal of Health Economics* 50:99-114.

Cao, Hongdao, and Daniel H Hill. 2005. Active versus Passive Sample Attrition: The Health and Retirement Study. Ann Arbor, MI: University of Michigan.

Card, David, Ciprian Domnisoru, and Lowell Taylor. 2018. The intergenerational transmission of human capital: Evidence from the golden age of upward mobility. *NBER Working Paper* 25000.

Card, David, and Alan B Krueger. 1992. "Does school quality matter? Returns to education and the characteristics of public schools in the United States." *Journal of Political Economy* 100 (1):1-40.

Case, A., D. Lubotsky, and C. Paxson. 2002. "Economic Status and Health in Childhood: The Origins of the Gradient." *American Economic Review* 92 (5):1308-34.

Case, Anne, Angela Fertig, and Christina Paxson. 2005. "The lasting impact of childhood health and circumstance." *Journal of Health Economics* 24 (2):365-389.

Chao, John C, and Norman R Swanson. 2005. "Consistent estimation with a large number of weak instruments." *Econometrica* 73 (5):1673-1692.

Cheng, Simon, and J Scott Long. 2007. "Testing for IIA in the multinomial logit model." *Sociological Methods & Research* 35 (4):583-600.

Chernozhukov, Victor, Christian Hansen, and Martin Spindler. 2015. "Post-selection and post-regularization inference in linear models with many controls and instruments." *American Economic Review* 105 (5):486-90.

Chetty, Raj, John N Friedman, and Jonah E Rockoff. 2014. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *American Economic Review* 104 (9):2633-79.

Clark, Damon, and Heather Royer. 2013. "The Effect of Education on Adult Mortality and Health: Evidence from Britain." *American Economic Review* 103 (6):2087-2120.

Cleves, M.A., W. Gould, and Y.V. Marchenko. 2016. An Introduction to Survival Analysis Using Stata: Stata Press.

Collins, William J, and Robert A Margo. 2006. "Historical perspectives on racial differences in schooling in the United States." *Handbook of the Economics of Education* 1:107-154.

Currie, Janet. 2009. "Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development." *Journal of Economic Literature* 47 (1):87-122.

Cutler, David M, and Adriana Lleras-Muney. 2010. "Understanding differences in health behaviors by education." *Journal of Health Economics* 29 (1):1-28.

Cutler, David M, Adriana Lleras-Muney, and Tom Vogl. 2008. Socioeconomic status and health: dimensions and mechanisms. *NBER Working Paper* 14333.

Cutler, David M., and Adriana Lleras-Muney. 2006. "Education and Health: Evaluating Theories and Evidence." *NBER Working Paper* 12352.

Davies, Neil M, Matt Dickson, George Davey Smith, Gerard van den Berg, and Frank Windmeijer. 2016. "The causal effects of education on health, mortality, cognition, well-being, and income in the UK Biobank." *bioRxiv*:074815.

De Walque, Damien. 2004. "Education, information, and smoking decisions: evidence from smoking histories, 1940-2000." *Working Paper* 3362, Word Bank, Washington.

De Walque, Damien. 2007. "Does education affect smoking behaviors?: Evidence using the Vietnam draft as an instrument for college education." *Journal of Health Economics* 26 (5):877-895.

Deary, Ian J, Steve Strand, Pauline Smith, and Cres Fernandes. 2007. "Intelligence and educational achievement." *Intelligence* 35 (1):13-21.

Deaton, Angus, and Christina Paxson. 2001a. "Mortality, Income, and Income Inequality Over Time in Britain and the United States." *NBER Working Paper* 8534.

Deaton, Angus S, and Christina Paxson. 2001b. "Mortality, education, income, and inequality among American cohorts." In *Themes in the Economics of Aging*, 129-170. University of Chicago Press.

Doorslaer, EKA van. 1987. "Health, knowledge and the demand for medical care: an econometric analysis." Maastricht University.

Dudovitz, Rebecca N, Paul J Chung, Sarah Reber, David Kennedy, Joan S Tucker, Steve Shoptaw, Kulwant K Dosanjh, and Mitchell D Wong. 2018. "Assessment of exposure to high-performing schools and risk of adolescent substance use: a natural experiment." *JAMA Pediatrics* 172 (12):1135-1144.

Dudovitz, Rebecca N, Bergen B Nelson, Tumaini R Coker, Christopher Biely, Ning Li, Lynne C Wu, and Paul J Chung. 2016. "Long-term health implications of school quality." *Social Science & Medicine* 158:1-7.

Ehrlich, Isaac, and Hiroyuki Chuma. 1990. "A Model of the Demand for Longevity and the Value of Life Extension." *Journal of Political Economy* 98 (4):761-782.

Elo, Irma T, and Samuel H Preston. 1996. "Educational differentials in mortality: United States, 1979–1985." *Social Science & Medicine* 42 (1):47-57.

Erbsland, Manfred, Walter Ried, and Volker Ulrich. 1995. "Health, health care, and the environment. Econometric evidence from German micro data." *Health Economics* 4 (3):169-182.

Ettner, Susan L. 1996. "New evidence on the relationship between income and health." *Journal of Health Economics* 15 (1):67-85.

Fiscella, Kevin, Peter Franks, Marthe R Gold, and Carolyn M Clancy. 2000. "Inequality in quality: addressing socioeconomic, racial, and ethnic disparities in health care." *JAMA* 283 (19):2579-2584.

Fischer, Martin, Martin Karlsson, and Therese Nilsson. 2013. "Effects of compulsory schooling on mortality: evidence from Sweden." *International Journal of Environmental Research and Public Health* 10 (8):3596-3618.

Fletcher, Jason M. 2015. "New evidence of the effects of education on health in the US: Compulsory schooling laws revisited." *Social Science & Medicine* 127:101-107.

Fletcher, Jason M, and David E Frisvold. 2009. "Higher education and health investments: does more schooling affect preventive health care use?" *Journal of Human Capital* 3 (2):144-176.

Friis, Karina, Mathias Lasgaard, Gillian Rowlands, Richard H Osborne, and Helle T Maindal. 2016. "Health literacy mediates the relationship between educational attainment and health behavior: a Danish population-based study." *Journal of Health Communication* 21 (sup2):54-60.

Frisvold, David, and Ezra Golberstein. 2011. "School quality and the education–health relationship: Evidence from Blacks in segregated schools." *Journal of Health Economics* 30 (6):1232-1245.

Fry, Tim RL, and Mark N Harris. 1996. "A Monte Carlo study of tests for the independence of irrelevant alternatives property." *Transportation Research Part B: Methodological* 30 (1):19-30.

Fry, Tim RL, and Mark N Harris. 1998. "Testing for independence of irrelevant alternatives: some empirical results." *Sociological Methods & Research* 26 (3):401-423.

Fuchs, Victor R. 1980. "Time Preference and Health: An Exploratory Study." *NBER Working Paper* 539.

Fujiwara, Takeo, and Ichiro Kawachi. 2009. "Is education causally related to better health? A twin fixed-effect study in the USA." *International Journal of Epidemiology* 38 (5):1310-1322.

Galama, Titus J, Adriana Lleras-Muney, and Hans van Kippersluis. 2018. The Effect of Education on Health and Mortality: A Review of Experimental and Quasi-Experimental Evidence. *NBER Working Paper* 24225.

Galea, Sandro, Melissa Tracy, Katherine J Hoggatt, Charles DiMaggio, and Adam Karpati. 2011. "Estimated deaths attributable to social factors in the United States." *American Journal of Public Health* 101 (8):1456-1465.

Gilleskie, Donna B, and Amy L Harrison. 1998. "The effect of endogenous health inputs on the relationship between health and education." *Economics of Education Review* 17 (3):279-295.

Glied, Sherry, and Adriana Lleras-Muney. 2008. "Technological innovation and inequality in health." *Demography* 45 (3):741-761.

Glymour, M Maria, Ichiro Kawachi, Christopher S Jencks, and Lisa F Berkman. 2008. "Does childhood schooling affect old age memory or mental status? Using state schooling laws as natural experiments." *Journal of Epidemiology & Community Health* 62 (6):532-537.

Goldin, Claudia, and Lawrence Katz. 2003. "Mass Secondary Schooling and the State." *NBER Working Paper* 10075.

Goldman, Dana P, and James P Smith. 2002. "Can patient self-management help explain the SES health gradient?" *Proceedings of the National Academy of Sciences* 99 (16):10929-10934.

Gottfredson, Linda S, and Ian J Deary. 2004. "Intelligence predicts health and longevity, but why?" *Current Directions in Psychological Science* 13 (1):1-4.

Grimard, Franque, and Daniel Parent. 2007. "Education and smoking: Were Vietnam war draft avoiders also more likely to avoid smoking?" *Journal of Health Economics* 26 (5):896-926.

Grossman, Michael. 1972a. "The demand for health: a theoretical and empirical investigation." *NBER Books*.

Grossman, Michael. 1972b. "On the concept of health capital and the demand for health." *Journal of Political Economy* 80 (2):223-255.

Grossman, Michael. 1976. "The correlation between health and schooling." In *Household Production and Consumption*, 147-224. NBER.

Grossman, Michael. 2000. "The human capital model." In *Handbook of Health Economics*, 347-408. Elsevier.

Grossman, Michael. 2006. "Education and nonmarket outcomes." *Handbook of the Economics of Education* 1:577-633.

Grossman, Michael, and Robert Kaestner. 1997. "Effects of Education on Health." In *The Social Benefits of Education*, edited by Jere R. Behrman and Nevzer Stacey, 69-124. University of Michigan Press.

Gundersen, Craig, Brent Kreider, and John Pepper. 2012. "The impact of the National School Lunch Program on child health: A nonparametric bounds analysis." *Journal of Econometrics* 166 (1):79-91.

Hamad, Rita, Holly Elser, Duy C Tran, David H Rehkopf, and Steven N Goodman. 2018. "How and why studies disagree about the effects of education on health: A systematic review and meta-analysis of studies of compulsory schooling laws." *Social Science & Medicine* 212:168-178.

Hammond, Cathie. 2002. "What is it about education that makes us healthy? Exploring the education-health connection." *International Journal of Lifelong Education* 21 (6):551-571.

Hansen, Christian, Jerry Hausman, and Whitney Newey. 2008. "Estimation with many instrumental variables." *Journal of Business & Economic Statistics* 26 (4):398-422.

Hansen, Lars Peter. 1982. "Large sample properties of generalized method of moments estimators." *Econometrica: Journal of the Econometric Society*:1029-1054.

Heckman, James J, Jora Stixrud, and Sergio Urzua. 2006. "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior." *Journal of Labor Economics* 24 (3):411-482.

House, James S, James M Lepkowski, Ann M Kinney, Richard P Mero, Ronald C Kessler, and A Regula Herzog. 1994. "The social stratification of aging and health." *Journal of Health and Social Behavior*:213-234.

Hurd, Michael, and Arie Kapteyn. 2003. "Health, Wealth, and the Role of Institutions." *Journal of Human Resources* 38 (2):386-415.

Idler, Ellen L, and Stanislav V Kasl. 1995. "Self-ratings of health: do they also predict change in functional ability?" *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 50 (6):S344-S353.

Iloabuchi, Tochukwu C, Deming Mi, Wanzhu Tu, and Steven R Counsell. 2014. "Risk Factors for Early Hospital Readmission in Low-Income Elderly Adults." *Journal of the American Geriatrics Society* 62 (3):489-494.

Institute of Medicine. 2004. Health Literacy: A Prescription to End Confusion. edited by L. Nielsen-Bohlman, A. M. Panzer and D. A. Kindig. Washington (DC): National Academies Press (US)

Ippolito, R. 2002. Health, education and investment behaviour in the family. Law and Economics Working paper series 03-04. School of Law, George Mason University, Arlington, Virginia.

Jacobson, Lena. 2000. "The family as producer of health—an extended Grossman model." *Journal of Health Economics* 19 (5):611-637.

Jansen, Tessa, Jany Rademakers, Geeke Waverijn, Robert Verheij, Richard Osborne, and Monique Heijmans. 2018. "The role of health literacy in explaining the association between educational attainment and the use of out-of-hours primary care services in chronically ill people: a survey study." *BMC Health Services Research* 18 (1):394.

Jasti, Harish, Eric M Mortensen, David Scott Obrosky, Wishwa N Kapoor, and Michael J Fine. 2008. "Causes and risk factors for rehospitalization of patients hospitalized with community-acquired pneumonia." *Clinical Infectious Diseases* 46 (4):550-556.

Johnson, Rucker C. 2011. Long-run impacts of school desegregation & school quality on adult attainments. *NBER Working Paper* 16664.

Jones, Damon, David Molitor, and Julian Reif. 2019. "What do workplace wellness programs do? Evidence from the Illinois workplace wellness study." *The Quarterly Journal of Economics* 134 (4):1747-1791.

Jürges, Hendrik, Steffen Reinhold, and Martin Salm. 2011. "Does schooling affect health behavior? Evidence from the educational expansion in Western Germany." *Economics of Education Review* 30 (5):862-872.

Kapteyn, Arie, Pierre-Carl Michaud, James P Smith, and Arthur Van Soest. 2006. "Effects of attrition and non-response in the Health and Retirement Study." Santa Monica, CA: RAND Corporation, 2006. https://www.rand.org/pubs/working_papers/WR407.html.

Kemptner, Daniel, Hendrik Jürges, and Steffen Reinhold. 2011. "Changes in compulsory schooling and the causal effect of education on health: Evidence from Germany." *Journal of Health Economics* 30 (2):340-354.

Kenkel, Donald, Dean Lillard, and Alan Mathios. 2006. "The roles of high school completion and GED receipt in smoking and obesity." *Journal of Labor Economics* 24 (3):635-660.

Kenkel, Donald S. 1991. "Health behavior, health knowledge, and schooling." *Journal of Political Economy* 99 (2):287-305.

Kenkel, Donald S. 1994. "The demand for preventive medical care." *Applied Economics* 26 (4):313-325.

Kitagawa, Evelyn M, and Philip M Hauser. 1973. "Differential mortality in the United States: A study in socioeconomic epidemiology." Cambridge, Harvard University Press.

Lager, Anton Carl Jonas, and Jenny Torssander. 2012. "Causal effect of education on mortality in a quasi-experiment on 1.2 million Swedes." *Proceedings of the National Academy of Sciences* 109 (22):8461-8466.

Laporte, Audrey, and Brian S Ferguson. 2007. "Investment in health when health is stochastic." *Journal of Population Economics* 20 (2):423-444.

Leigh, J Paul. 1983. "Direct and indirect effects of education on health." *Social Science & Medicine* 17 (4):227-234.

Leigh, J Paul. 1985. "An empirical analysis of self-reported, work-limiting disability." *Medical Care*:310-319.

Liljas, Bengt. 1998. "The demand for health with uncertainty and insurance." *Journal of Health Economics* 17 (2):153-170.

Lleras-Muney, Adriana. 2002. "Were compulsory attendance and child labor laws effective? An analysis from 1915 to 1939." *The Journal of Law and Economics* 45 (2):401-435.

Lleras-Muney, Adriana. 2005. "The relationship between education and adult mortality in the United States." *The Review of Economic Studies* 72 (1):189-221.

Lleras-Muney, Adriana, and Frank R Lichtenberg. 2005. "Are the more educated more likely to use new drugs?" *Annales d'Économie et de Statistique*:671-696.

Long, J.S., and J. Freese. 2014. Regression Models for Categorical Dependent Variables Using Stata, Third Edition: Stata Press.

Low, M David, Barbara J Low, Elizabeth R Baumler, and Phuong T Huynh. 2005. "Can education policy be health policy? Implications of research on the social determinants of health." *Journal of Health Politics, Policy and Law* 30 (6):1131-1162.

Lundborg, Petter. 2013. "The health returns to schooling—what can we learn from twins?" *Journal of Population Economics* 26 (2):673-701.

Lundborg, Petter, Carl Hampus Lyttkens, and Paul Nystedt. 2012. "Human capital and longevity: Evidence from 50,000 twins." *Unpublished Manuscript, University of York, Health, Econonometrics and Data Group*.

Lundborg, Petter, Martin Nordin, and Dan Olof Rooth. 2018. "The intergenerational transmission of human capital: the role of skills and health." *Journal of Population Economics* 31 (4):1035-1065.

Lynch, Scott M. 2003. "Cohort and life-course patterns in the relationship between education and health: A hierarchical approach." *Demography* 40 (2):309-331.

Malamud, Ofer, Andreea Mitrut, and Cristian Pop-Eleches. 2018. "The Effect of Education on Mortality and Health: Evidence from a Schooling Expansion in Romania." *NBER Working Paper* 24341.

Manning, Willard G, Joseph P Newhouse, Naihua Duan, Emmett B Keeler, and Arleen Leibowitz. 1987. "Health insurance and the demand for medical care: evidence from a randomized experiment." *American Economic Review*:251-277.

Marden, J. R., E. J. Tchetgen Tchetgen, I. Kawachi, and M. M. Glymour. 2017. "Contribution of Socioeconomic Status at 3 Life-Course Periods to Late-Life Memory Function and

Decline: Early and Late Predictors of Dementia Risk." *American Journal of Epidemiology* 186 (7):805-814.

Mazumder, Bhashkar. 2008. "Does education improve health? A reexamination of the evidence from compulsory schooling laws." Federal Reserve Bank of Chicago Economic Perspectives Q2: 2- 16.

Mazumder, Bhaskar. 2012. "The effects of education on health and mortality." *Nordic Economic Policy Review* 1 (2012):261-301.

McFadden, Daniel. 1973. "Conditional logit analysis of qualitative choice behavior." In *Frontiers of Econometrics*, ed. by P. Zarembka. New York: Academic Press.

Meghir, Costas, Mårten Palme, and Emilia Simeonova. 2012. Education and mortality: Evidence from a social experiment. *NBER Working Paper* 17932.

Meghir, Costas, Mårten Palme, and Emilia Simeonova. 2013. Education, cognition and health: Evidence from a social experiment. *NBER Working Paper* 19002.

Mehta, Neil K, Hedwig Lee, and Kelly R Ylitalo. 2013. "Child health in the United States: recent trends in racial/ethnic disparities." *Social Science & Medicine* 95:6-15.

Mokdad, Ali H, James S Marks, Donna F Stroup, and Julie L Gerberding. 2004. "Actual causes of death in the United States, 2000." *JAMA* 291 (10):1238-1245.

Montez, Jennifer Karas, Robert A Hummer, and Mark D Hayward. 2012. "Educational attainment and adult mortality in the United States: A systematic analysis of functional form." *Demography* 49 (1):315-336.

Muurinen, Jaana-Marja. 1982. "An economic model of health behaviour: with empirical applications to Finnish health survey data." University of York.

Nansel, Tonja R, Mary D Overpeck, Denise L Haynie, W June Ruan, and Peter C Scheidt. 2003. "Relationships between bullying and violence among US youth." *Archives of Pediatrics & Adolescent Medicine* 157 (4):348-353.

Nansel, Tonja R, Mary Overpeck, Ramani S Pilla, W June Ruan, Bruce Simons-Morton, and Peter Scheidt. 2001. "Bullying behaviors among US youth: Prevalence and association with psychosocial adjustment." *JAMA* 285 (16):2094-2100.

National Academies of Sciences, Engineering, and Medicine. 2017. Accounting for social risk factors in Medicare payment: National Academies Press.

Nguyen, Thu T, Eric J Tchetgen Tchetgen, Ichiro Kawachi, Stephen E Gilman, Stefan Walter, Sze Y Liu, Jennifer J Manly, and M Maria Glymour. 2016. "Instrumental variable approaches to identifying the causal effect of educational attainment on dementia risk." *Annals of Epidemiology* 26 (1):71-76. e3.

Nutbeam, Don. 2008. "The evolving concept of health literacy." *Social Science & Medicine* 67 (12):2072-2078.

Ofstedal, MB, DR Weir, KT Chen, and J Wagner. 2011. Updates to HRS Sample Weights. Ann Arbor, Michigan: Institute for Social Research, University of Michigan.

Oreopoulos, Philip. 2006. "Estimating average and local average treatment effects of education when compulsory schooling laws really matter." *American Economic Review* 96 (1):152-175.

Oreopoulos, Philip, Marianne E Page, and Ann Huff Stevens. 2006. "The intergenerational effects of compulsory schooling." *Journal of Labor Economics* 24 (4):729-760.

Pischke, Jörn-Steffen, and Till Von Wachter. 2008. "Zero returns to compulsory schooling in Germany: Evidence and interpretation." *The Review of Economics and Statistics* 90 (3):592-598.

Rasu, Rafia S, Walter Agbor Bawa, Richard Suminski, Kathleen Snella, and Bradley Warady. 2015. "Health literacy impact on national healthcare utilization and expenditure." *International Journal of Health Policy and Management* 4 (11):747.

Reinhold, Steffen, and Hendrik Jürges. 2010. "Secondary school fees and the causal effect of schooling on health behavior." *Health Economics* 19 (8):994-1001.

Rosen, Sherwin, and Paul Taubman. 1982. "Some socioeconomic determinants of mortality." *Economics of Health Care*:255-71.

Rosenzweig, Mark R, and T Paul Schultz. 1982. "The behavior of mothers as inputs to child health: the determinants of birth weight, gestation, and rate of fetal growth." In *Economic Aspects of Health*, 53-92. University of Chicago Press.

Ross, Catherine E, and John Mirowsky. 1999. "Refining the association between education and health: the effects of quantity, credential, and selectivity." *Demography* 36 (4):445-460.

Sambamoorthi, Usha, and Donna D McAlpine. 2003. "Racial, ethnic, socioeconomic, and access disparities in the use of preventive services among women." *Preventive Medicine* 37 (5):475-484.

Sander, William. 1995a. "Schooling and quitting smoking." *The Review of Economics and Statistics*:191-199.

Sander, William. 1995b. "Schooling and smoking." *Economics of Education Review* 14 (1):23-33.

Sapolsky, Robert M. 2005. "The influence of social hierarchy on primate health." *Science* 308 (5722):648-652.

Sickles, Robin C, and Paul Taubman. 1986. "An analysis of the health and retirement status of the elderly." *Econometrica: Journal of the Econometric Society*:1339-1356.

Stephens Jr, Melvin, and Dou-Yan Yang. 2014. "Compulsory education and the benefits of schooling." *American Economic Review* 104 (6):1777-92.

Strauss, John, and Duncan Thomas. 1998. "Health, Nutrition, and Economic Development." *Journal of Economic Literature* 36 (2):766-817.

Taubman, Paul, and Sherwin Rosen. 1980. Healthiness, education, and marital status. *NBER Working Paper* 611.

Van Der Pol, Marjon. 2011. "Health, education and time preference." *Health Economics* 20 (8):917-929.

Van Kippersluis, Hans, Owen O'Donnell, and Eddy Van Doorslaer. 2011. "Long-run returns to education does schooling lead to an extended old age?" *Journal of Human Resources* 46 (4):695-721.

Vaughn, Michael G, Christopher P Salas-Wright, and Brandy R Maynard. 2014. "Dropping out of school and chronic disease in the United States." *Journal of Public Health* 22 (3):265-270.

Vlassoff, Carol. 2007. "Gender differences in determinants and consequences of health and illness." *Journal of Health, Population, and Nutrition* 25 (1):47.

Wagstaff, Adam. 1986. "The demand for health: some new empirical evidence." *Journal of Health Economics* 5 (3):195-233.

Wagstaff, Adam. 1993. "The demand for health: an empirical reformulation of the Grossman model." *Health Economics* 2 (2):189-198.

Weuve, J., E. J. Tchetgen Tchetgen, M. M. Glymour, T. L. Beck, N. T. Aggarwal, R. S. Wilson, D. A. Evans, and C. F. Mendes de Leon. 2012. "Accounting for bias due to selective attrition: the example of smoking and cognitive decline." *Epidemiology* 23 (1):119-28.

Winkleby, Marilyn A, Darius E Jatulis, Erica Frank, and Stephen P Fortmann. 1992. "Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease." *American Journal of Public Health* 82 (6):816-820.

Wooldridge, Jeffrey M. 2010. Econometric analysis of cross section and panel data: MIT press.

Zajacova, Anna, and Robert A Hummer. 2009. "Gender differences in education effects on all-cause mortality for white and black adults in the United States." *Social Science & Medicine* 69 (4):529-537.