Helsø, Anne-Line Koch

Doctoral Thesis

Labor supply and earnings: In old age, in bad health, and across generations

PhD Series, No. 200

Provided in Cooperation with:
University of Copenhagen, Department of Economics

Suggested Citation: Helsø, Anne-Line Koch (2019) : Labor supply and earnings: In old age, in bad health, and across generations, PhD Series, No. 200, University of Copenhagen, Department of Economics, Copenhagen

This Version is available at:
https://hdl.handle.net/10419/240549

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

PhD Thesis

Anne-Line Koch Helsø

Labor Supply and Earnings:
In old age, in bad health, and across generations

Supervisors: Bertel Schjerning and Thomas H. Jørgensen

Date of submission: August 31 2018

Contents

English Introduction

The three largest items on the list of public expenditures in Denmark (and many other countries) are retirement, health, and education. These are also the topics of interest in the three Chapters of this dissertation. While the chapters are quite different from each other, they share three common denominators. First, as indicated by the dissertation title, they all revolve around the labor supply and earnings of individuals. Second, they seek to shed light on some highly policy-relevant questions. Last, but not least, all chapters are based on the exceptional Danish register data. The three chapters are self-contained with separate abstracts, bibliographies, and appendices.

In **Chapter 1: "Pension Wealth, Preference Heterogeneity and the Impact of Retirement Policies,"** we study the labor supply decision of seniors, and how financial incentives affect their retirement decisions. We propose and estimate a novel structural retirement model with leisure preference heterogeneity using Danish register data with unique information about individual private retirement wealth. We study the extent to which larger private retirement savings dampens the effect of retirement reforms that target an increase in labor supply. Our findings suggest that the effect of raising the normal retirement age by one year is more than halved for individuals who hold private retirement savings equivalent to at least four years of earnings.

Chapter 2: "The Economic Impact of Healthcare Quality" studies the labor supply and earnings costs of hospitalizations and evaluate the quality of treatment based on its ability to mitigate the labor market consequences of a given diagnosis. We measure substantial heterogeneity in treatment quality across hospitals, with a four percentage points difference in lost earnings between the best and worst hospital, all else equal. We also document a significant decline in the labor cost of hospitalizations over time and find that the average post-hospitalization reduction in labor supply has declined by 13.6 percentage points from 1998 to 2012.

Several studies have found evidence of much higher intergenerational income mobility in the Scandinavian countries compared to the U.S. Often researchers attribute these differences to the social and educational policies in the Scandinavian countries, which equalize opportunities for children. In **Chapter 3: "Intergenerational Income Mobility in Denmark and the U.S."**, I present a novel cross-country comparison of intergenerational income mobility in Denmark and the U.S. Unlike existing studies, I rely on high-quality administrative data for both countries. I find that Denmark is 50-100% more mobile than the U.S. I contrast my findings to the existing literature, which finds larger cross-country differences, and show that my results are more robust to sample selection and measurement error biases.

Dansk Introduktion

De tre største udgiftsposter på de offentlige finanser i Danmark (og mange andre lande) er pension, sundhed og uddannelse, og disse emner udgør hver især hovedtemaerne i denne afhandlings tre kapitler. Selvom kapitlerne er ganske forskellige fra hinanden, så deler de tre fællesnævnere. For det første, som titlen også indikerer, drejer de sig alle om individers arbejdskraft og indtjening. For det andet søger de at kaste lys på nogle politisk vigtige spørgsmål. Sidst, men ikke mindst, er alle kapitler baseret på dansk registerdata. Kapitlerne fremstår selvstændigt, med separate resuméer, bibliografier og bilag.

I **Kapitel 1: "Pension Wealth, Preference Heterogeneity and the Impact of Retirement Policies"** studerer vi seniorers arbejdsudbud, og hvordan økonomiske incitamenter påvirker deres beslutning om tilbagetrækning fra arbejdsmarkedet. Vi foreslår og estimerer en ny strukturel pensionsmodel med heterogene fritidspræferencer på danske registerdata, med unik information om private pensionsopsparinger. Vi undersøger, i hvilket omfang større private pensionsopsparinger dæmper effekten af pensionsreformer, der er målrettet et øget arbejdsudbud. Vores resultater viser, at virkningen af at hæve folkepensionsalderen med ét år er mere end halveret for personer, der har private pensionsopsparinger svarende til fire års arbejdsindkomst eller mere.

Kapitel 2: "The Economic Impact of Healthcare Quality" undersøger de arbejdsrelaterede tab, som er forbundet med hospitalsindlæggelser, herunder lønnedgang og tab i arbejdsudbud. Vi evaluerer kvaliteten hospitalsbehandlingerne ud fra deres evne til at mindske de arbejdsmarkedsrelaterede konsekvenser af en given diagnose. Vi måler betydelig heterogenitet i behandlingskvalitet på tværs af hospitaler, med en forskel på fire procentpoint i tabt indtjening mellem det bedste og dårligste hospital, alt andet lige. Vi dokumenterer også et betydeligt fald i lønomkostningerne ved hospitalsindlæggelser over tid, hvor vi finder, at det gennemsnitlige arbejdstab efter en indlæggelse er faldet med 13,6 procentpoint fra 1998 til 2012.

Flere undersøgelser har konkluderet, at der er meget højere inter-generationel indkomstmobilitet i de skandinaviske lande sammenlignet med USA. Ofte peger forskere på de skandinaviske landes uddannelsessystemer og socialpolitikker som forklarende faktorer, da disse udligner børns muligheder. I **Kapitel 3: "Intergenerational Income Mobility in Denmark and the U.S."** præsenterer jeg en ny sammenligning af inter-generationel indkomstmobilitet i Danmark og USA. I modsætning til eksisterende studier bruger jeg høj-kvalitets administrativ data for begge lande. Jeg finder, at Danmark er 50-100 % mere mobil end USA. Jeg sammenligner mine resultater med den eksisterende litteratur, som finder større forskelle på tværs af landene, og viser, at mine resultater er mere robuste over for bias fra målefejl og dataselektion.

Acknowledgements

With a background in math-econ, it took me quite a while to figure out what economics is all about. For a long time, I thought that economics was about money - and how to make as much as possible. It was only when I became a student researcher at the economic research institution DREAM, that I learned how economics is really about people, and how we make our choices in life.

Thanks to this epiphany, I got enrolled in the Ph.D. program at the Economics Department at Copenhagen University. I, therefore, owe a high debt of gratitude to my colleagues at DREAM who introduced me to the exciting world of register data and micro-econometrics. Not least to Peter Stephensen, who – in his way of phrasing – is also the Godfather of this thesis. I also thank DREAM for financing half of this dissertation, and for hosting me during the last three years.

I am also very thankful for my two supervisors, Bertel Schjerning and Thomas H. Jørgensen, who have guided me throughout the project and along the winding roads of the academic world. I am extremely glad that they enabled my visit at Stanford University, which came to be a turning point in my dissertation work.

I am grateful beyond words to Raj Chetty and the Stanford Opportunity Lab for hosting me, and for welcoming me in their research group. Their exciting work on equality of opportunities inspired me to write the 3rd chapter of this thesis about intergenerational mobility in Denmark and the U.S. I am also deeply indebted to Pablo Mitnik at the Stanford Center on Poverty & Inequality and grateful for the entire afternoons, which he kindly spent with me discussing this project. In this regard, I also wish to thank the Fulbright Foundation for enabling the visit with their generous financial support.

This thesis would not have been possible without the active engagement of my co-authors who have all shown incredible amounts of patience and enthusiasm throughout our work. I thank all for their collaboration, which I hope will last for many years to come. Also a special thanks to the numerous research assistants at the Danish Ministry of Finance who have spent a lot of time to compute the data used in Chapter 1.

A lot of excellent research has come out of the Danish register data, which is unique in many different aspects, and which make up the very foundation of my dissertation. I therefore also owe a great deal of gratitude to the many employees at Statistics Denmark and elsewhere, who enable us researchers to access this treasure of information.

Last, but not least, I wish to thank my husband, Morten, for his unconditional support throughout the last three years, and for always helping me put things in perspective.

Anne-Line Koch Helsø

Copenhagen, August 2018

Chapter 1

Pension Wealth, Preference Heterogeneity and the Impact of Retirement Policies

Pension Wealth, Preference Heterogeneity and the Impact of Retirement Policies

Søren Arnberg*, Anne-Line Koch Helsø** and Peter Philip Stephensen***

August 31, 2018

*Danish Ministry of Finance

**Department of Economics, University of Copenhagen

***Danish Institute for Economic Modeling and Forecasting, DREAM

Abstract

We propose and estimate a novel structural retirement model with leisure preference heterogeneity using high-quality Danish register data with unique information about individual private pension wealth. We apply a non-parametric estimation technique to measure the heterogeneity in leisure preferences, and our estimates suggest that leisure preferences are widely distributed among the population. We use the model to study the extent to which increasing private retirement savings dampens the effect of retirement reforms targeted to increase labor supply. Our findings suggest that the effects of an increase in the normal retirement age by one year increases the average retirement age by 0.2 year if individuals hold zero or low wealth. The corresponding effect is more than halved once individuals hold savings equivalent to four or more years of pre-retirement earnings.

Keywords: Structural retirement model; Labor supply of seniors; Preference heterogeneity; Pension wealth; Policy evaluation

JEL Classifications: C51, C63, J14, J22, J26

1 Introduction

In virtually all developed countries, policy-makers seek ways to reform their pension system to ease the fiscal burden of increasing life expectancy. Many countries have done so by cutting pension benefits or postponing statutory retirement ages. At the same time, most OECD countries have introduced preferential tax treatments of pension savings to encourage people to save for retirement. As a result, private pension savings have increased substantially and will continue to do so OECD. [2016].

Increased private funding of pension payments reduce the need for government retirement transfers, and thereby help reduce the fiscal vulnerability of governments. However, larger private retirement savings also provide retirees with more flexibility to decide their own retirement age independent of the statutory retirement ages defined by the public retirement plans. As such, increased private retirement savings could also counteract the policies targeted to raise retirement ages, such as postponed statutory retirement ages or decreased pension benefits. This paper aims to explore the extent to which the increase in private pension wealth affects the effectiveness of such retirement policies.

To organize the empirical analyses, we propose a novel structural retirement model of senior worker's retirement decision with heterogeneous leisure preferences, attrition, and improved health across generations[1]. The basic assumption underlying structural economic models is that people's decisions are formed by their circumstances and preferences combined. In real life, however, we often observe that agents with identical circumstances behave differently, due to the self-evident fact that people are different and have different preferences. Allowing for preference heterogeneity, therefore, contributes with an important extra dimension to structural models, making them more flexible and realistic. One of the main contributions of this paper is that we propose and estimate a new retirement model where we allow for individuals to have different preferences for leisure when determining their optimal retirement age. We also show that the assumed preference heterogeneity significantly improves the model fit and alters the policy experiment conclusions compared to a model where all agents are assumed to have the same leisure preferences.

We estimate our model on full population Danish register data from 1996-2016, where we consider the retirement decisions of birth cohorts 1942-1954. Our data contain information about each person's wealth in both retirement and non-retirement accounts, as well as highly detailed income information.

The Danish setting of our model has four main advantages: First, we have access to high-quality administrative information about the full population of Danes. Our data include detailed information about each senior's pension and non-pension wealth, earnings, transfer

[1] Our model is a development of the former model versions presented in Arnberg and Stephensen [2015] and Helsø [2015]

income and transfer eligibility which are mainly 3rd-party reported. This unique data set is particularly suitable for studying the effect of financial incentives in retirement, where both earnings, non-pension wealth and private pension savings are extremely important factors. Compared to previous studies which are often based on the HRS (Health and Retirement Survey) for the U.S., our data contain a many more observations, which are also of higher quality since they are mainly 3rd-party reported. Second, there were several sharp pension reforms in Denmark, targeting higher retirement ages, which provide close to exogenous variation in retirement incentives across wealth groups. We use one of these reforms to externally validate our model fit. Third, the mandatory retirement savings in Denmark offset other savings to a very limited extent (e.g. Chetty et al. [2014] and Arnberg and Barslund [2014]), suggesting that pension wealth is almost exogenous. As some groups have faced mandatory savings schemes longer than others have, there is a significant dispersion in the level of private retirement savings of senior workers in Denmark. Since mandatory labor market pension saving schemes were expanded during the 1980s and 1990s, private pension savings have increased substantially, and will continue to do so during the next decades. In 2017, 85% of all workers contributed about 12-18% of their monthly wages to defined contribution private pension schemes making Denmark the OECD country with the largest ratio of assets to GDP in funded private pension arrangements (OECD. [2016]). Fourth, since private retirement plans in Denmark are predominantly DC (defined contribution) plans, we're able to compute the future retirement income streams (as a function of the chosen retirement age) of almost everyone with very high accuracy.

Descriptive evidence suggests that financial incentives seem to be of high importance for the retirement decision of Danish seniors, who often retire immediately after meeting some eligibility criteria, and whose retirement patterns vary quite a lot with the level of private retirement wealth. We also observe that individuals respond to changes in eligibility criteria: when the statutory early retirement pension (ERP) age was increased from age 60 for cohort 1953 to 61 for cohort 1954 (2nd half), the corresponding drop in labor force participation at age 60 shifted almost entirely to age 61. As a result, the labor force participation at age 60 of cohort 1954 (2nd half) was 28 percentage points larger compared to cohort 1953. Furthermore, a means testing in ERP benefits with respect to private pension wealth can be avoided if retirement is postponed to at least two years after the ERP age. Our descriptive findings show that ERP eligible individuals with large retirement savings, who benefit the most from this rule, are also more likely to postpone their retirement age to 62.

Our proposed model shares many similarities with both the Option Value model and the Dynamic Programming model. The main difference lies in our assumption that agents, at the age of 57, have perfect foresight about their future income, with the only uncertainty being the timing of their death. This also implies that we don't account for any health or health cost related uncertainty. As such, our model mirrors the policy regime in Denmark

and many European countries, with free healthcare, guaranteed retirement benefits, and generous unemployment benefits. The main contribution of our model is that it allows for leisure preference heterogeneity which we estimate using a non-parametric estimation technique, which is a fixed-grid version of the Expectation Maximization (EM) algorithm as proposed by Train [2007]. We find substantial variation of leisure preferences within gender- and education specific groups. Our model fits the data well with reasonable parameter values. More importantly, our proposed model is able to fit the retirement response to a policy change in an external validation setting. We also find that out model performs much better compared to a similar model without preference heterogeneity, suggesting that the leisure preference heterogeneity assumption adds a significant contribution to the model. Previous studies have emphasized and studied the impact of heterogeneous leisure preferences on retirement behavior, see Gustman and Steinmeier [2005] and French and Jones [2011], but we are - to our knowledge - the first to propose and estimate a model with flexible and non-parametric heterogeneous leisure preferences.

Our work builds upon an extensive literature which studies the effects of financial incentives on senior workers' labor supply decision. One strand of the literature models and estimates the retirement decision within a structural model framework and simulates the effects of various policy experiments. Examples are the Option Value model by Stock and Wise [1990] and Dynamic Programming models as devised by Rust and Phelan [1997]. In a dynamic life-cycle model, French [2005] finds that a 20% reduction in social security income raises the average retirement age of U.S. workers by three months. French and Jones [2011] find that an increase in the social security eligibility age in the US from 65 to 67 increase retirement ages by less than one month. Gustman and Steinmeier [2005] find that the effect of a two-year increase in the social security early eligibility age in the U.S. is two months. For Denmark, Bingley et al. [2004] find that a three-year increase in both early- and normal retirement ages delays retirement by 1.3 years. Another strand of the literature examines the ex-post evaluations of actual policy changes in reduced-form studies. Compared to the structural ex-ante simulation exercises, these studies often find larger effects of an increase in statutory retirement ages (Blundell et al. [2016]). Mastrobuoni [2009] finds that the mean retirement age of U.S. cohorts increased by one month when the normal retirement age was increased by two months. Another example is Lalive and Staubli [2014] who find that a one year increase in the normal retirement age of Swiss women delayed their labor market exit by as much as 7.9 months.

Mastrobuoni [2009] argues that many structural models underestimate the effects of an increase of the NRA because of measurement error (due to lack of precise data on retirement incentives) and because they do not account for social norms related to the NRA. Our model includes a control for social norms related to the NRA, and we are convinced that our high-quality administrative data enables very precise calculations of the economic retirement

incentives. Compared to previous structural models, our estimates are also in the higher end, however lower compared to most reduced form evidence. However, since the estimated magnitudes of policy effects very much depends on the country-specific retirement systems, direct comparisons are not possible.

We analyze the effects of different policy experiments and show how these vary for different levels of private retirement savings. First, we simulate a base-line experiment in which we abolish the early retirement program. As the early retirement program is drastically being phased out in Denmark, the baseline experiment mimics the retirement decision of future generations. We then contrast the retirement decisions in our baseline experiment to three additional experiments.

The first experiment is an increase in the normal retirement age by one year. We find that individuals with zero private retirement savings delay their retirement with 0.15-0.2 years once the statutory retirement age increases from 65 to 66. For individuals with private retirement wealth equivalent to four or more years of pre-retirement earnings, the same response is less than half the size.

The second experiment decreases the old age pension benefits by five percent, and for this experiment, we estimate a decline in the expected retirement age of roughly 0.08-0.09 years for those with no retirement savings and 0.07-0.06 years for those with large retirement savings. As such, the effect of a decrease in retirement benefits is much more stable across levels of retirement wealth as compared to an increase in the NRA.

The third experiment introduces a reduction in the old age pensions' means testing with respect to private pension annuities when individuals retire one or more years after the statutory NRA. While low-wealth individuals are almost unaffected by this experiment, individuals holding large pension savings delay their retirement by up to 0.1 years.

In summary, our experiments show that the size of individual retirement savings can have important implications for the effect of different retirement reforms. For an increase in the NRA, we find a particularly large and negative effect of private retirement wealth on the reform's ability to increase labor supply. It is possible, however, to reverse this effect - e.g. in a reform which mimics the two-year rule of the ERP scheme.

The paper is structured as follows: In Section 2, we describe the institutional settings defining the Danish retirement system. In Section 3, we describe the data and how we compute the income components which are not directly observed in the data, and in Section 4 we motivate our analysis with some descriptive figures of the data. In Section 5 we present our structural retirement model and Section 6 specifies how we estimate the model, including a detailed description of the fixed-grid version of the EM algorithm which we use to estimate the heterogeneous leisure preference parameters. Our estimation results are shown in Section 7, and in Section 8 we show how our model fit the data, together with an external validation test of the model. In Section 9, we run and evaluate three different policy experiments, and

in Section 10, we compare our proposed model to a simpler version without leisure preference heterogeneity. Finally, Section 11 concludes.

2 Danish Institutional Settings

The following Section outlines the main components of the Danish retirement system, which mainly consists of three elements: the *Early Retirement Pension*, ERP (efterløn), the *Old Age Pension*, OAP (folkepension) and *private pension savings* which supplement the ERP/OAP benefits.[2] The Section begins with a description of the policy rules which applied to birth cohorts 1942-1952 and ends with a description of how these rules are going to change for future retirees.

Early Retirement Pension (ERP)

The ERP is a voluntary scheme in which participants pay a quarterly membership fee (1.122 DKr in 2004) for at least 10 years to obtain eligibility. ERP benefits apply from age 60 until the normal retirement age of 65. ERP payouts resemble the level of unemployment benefits (about 200,000 DKr in 2018), and thereby it contains elements of both a funded and unfunded plan, where the government finances roughly 70% (Jørgensen [2014]), making it a quite attractive scheme. 69% of the cohort born in 1942 (92% of those in the labor force) were eligible for ERP benefits. Following a reform in 2011 which made the ERP program less lucrative, many workers have opted out of the ERP-scheme.

ERP benefits are means tested with respect to earnings, private pension payments and the accumulated amount of private pension savings. The *two-year rule* within the ERP program introduces a financial incentive for ERP members to postpone retirement further until age 62. If retirement is postponed by at least two years from age 60 to 62, the ERP benefit rate is increased from 91 to 100% of the UI benefit rate, and ERP benefits are no longer means tested with respect to accumulated private pension wealth. As such, the two-year incentive rule is especially relevant to those who hold large amounts of accumulated private retirement savings. The exact means testing in ERP benefits with respect to private retirement savings depends on their size, type (life annuity, rate- or capital pension) and category (employer or employee administrated) - see Section 2 for a further description. For a person with the most frequent type of savings (employer administrated life annuity), who have saved an amount corresponding to a lifelong annual payment of 50,000 DKr at retirement age 60, his ERP benefits at the 91% UI benefit rate are reduced by roughly 10% if retires before age 62, whether or not he decides to initiate his private pension payments. If he retires after age 62

[2]Another important element is the disability pension system. However, as we only consider voluntary retirement in the present analysis, disability pensioners are excluded from our analysis.

and delays the initiation of his private pension payments until age 65, he will receive the full amount of the full ERP benefits at the 100% rate of UI benefits. The exact means testing rules of the ERP scheme with respect to private retirement savings are listed in Table 3. Had his accumulated life annuity savings corresponded to annual payments of 200,000 DKr (at retirement age 60), his ERP benefits would be reduced by roughly 50% if he retired before age 62 - a reduction which could be avoided if postponed his retirement age to or after age 62. When an ERP eligible individual fulfills the two-year rule but postpones his retirement even further, he receives one "portion" tax-free premium of 10,000 DKr (2004-level) for every four months of full-time work until the NRA at 65.

Old Age Pension (OAP)

The Danish OAP system is available to all Danish citizens aged 65 or above and is a fully government financed pay-as-you-go program available to all Danish citizens. It consists of a baseline amount (grundbeløb) and a supplement (pensionstillæg). The baseline annual amount (58.776 DKr in 2004) applies to all and is only means tested with respect to concurrent labor market income. The size of the OAP supplement varies by marital status (26.208 DKr in 2004 for married and 56.148 DKr for singles) and is also means tested by concurrent earnings. It is also further means tested by private pension payments to the recipient and the labor market status an earnings of his/her spouse. Unlike the ERP, the OAP is *not* means tested with respect to the accumulated amount of private pension savings. Seniors with poor financial circumstances can apply for additional benefits, e.g., housing allowances. For each year an individual postpones his/her retirement age beyond the NRA age at 65, his/her future OAP rate increases with 6% (with a maximum of 10 years).

Private Retirement Savings

Most jobs in Denmark are covered by collective agreements which often include an employer administrated contribution plan in which a fixed proportion of the monthly wages are paid into a defined contribution (DC) pension savings plan. These contribution plans are mandatory to all workers covered by the collective agreements, which counts about 85% of workers in 2018. The contribution rates are between 12-18 percent of the monthly wages (in 2018), where blue-collar workers typically contribute about 12 percent and white-collar workers between 15-18 percent. While the mandatory contribution plans started already in the 1950s for some high-skilled groups of workers, the contribution plans for large groups took form during the 1980s and 1990s, resulting in relatively large variation in private pension wealth for future cohorts of seniors. As such, private retirement savings have increased steadily during the last 3-4 decades, and continue to do so until the 2050s-2060s when most retirees will have the saved for retirement during their entire careers. In addition to the mandatory

employer administrated contribution plans, individuals can also voluntary save money for their retirement in employee administrated pension accounts, but these types of retirement savings only make up a small fraction of total private retirement savings.

The two main categories (employer and employee administrated savings accounts) each consists of three types of retirement savings: 1) life annuities (livrente), 2) term pension (ratepension) and 3) capital pension (kapitalpension). *Life annuities* guarantee a monthly payment from retirement until death, according to the accumulated amount. *Term pensions* are also paid out as annuities, but their duration only lasts between 10 and 25 years and can be initiated no later than age 77. C*apital pensions are* pension balances with no requirements on installment and are usually paid as a lump sump, no longer than 15 years after retirement.

As in most other OECD countries, the tax treatment of retirement savings provides a tax advantage when people save for retirement. While returns of regular non-retirement savings are taxed by approximately 33%, retirement savings are taxed by only 15.3%. However, private pension savings are means tested in public pension payments (ERP and OAP), which increases the effective taxation.

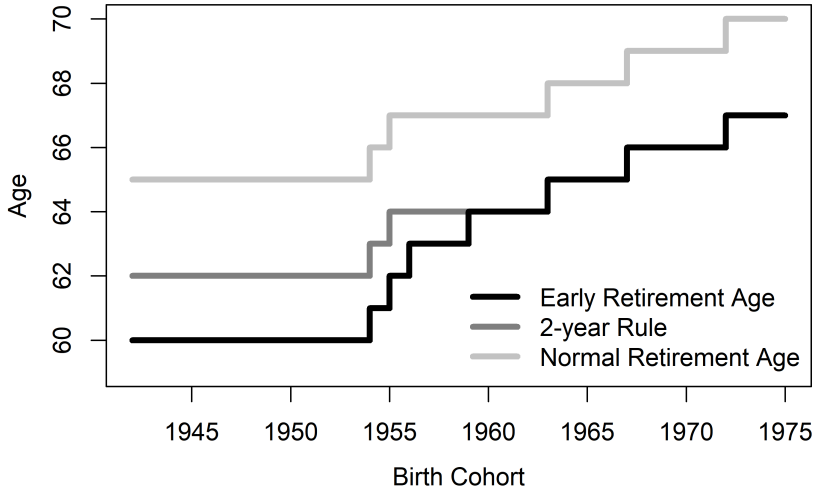
Trends in statutory retirement ages

Figure 1 shows how the statutory retirement ages change across cohorts. Both the ERA and the NRA increase over time, and the length of the ERP Program decreases from five to three years, with a gradual abolition of the two-year rule. For individuals born later than 1960, the ERP scheme is three years long, and there is no two-year rule. Whereas the full ERP amount is increased to 100% of unemployment benefits, there is no reduction in the means testing of the ERP benefit with respect to private pension wealth, as were the case with the two-year rule for earlier cohorts. On the contrary, the means testing is further sharpened for these cohorts, from 60 to 80%, and the lower-limit allowance is removed. Due to larger reductions in ERP benefits by private retirement savings, the scheme is now much more unattractive for most workers, and the share of ERP members has dropped drastically. For the cohort born in 1970 or later, only 12% are members of the ERP-scheme, compared to 69% in the 1942 cohort.

3 Data

We base our analysis on administrative register data available from Statistics Denmark, covering the full population of Danes (approximately 5.5 million people) in 1996-2016 who were born between 1942-1954. Our data includes detailed yearly information about a wide range of variables: earnings, pension savings, liquid wealth, transfer income, non-pension wealth, employment status, demographics, etc., which are mainly 3rd-party reported. For

Figure 1: Statutory Retirement Ages by Birth Cohort



The figure shows the statutory eligibility ages for the different retirement schemes for the different cohorts. The statutory ages for cohorts born after 1967 have not yet been approved by the Danish parliament, but are expected to develop according to the increasing life expediencies.

each person, income streams at age $a \in \{58, ..., 120\}$ is defined for each of the possible retirement ages, $r \in \{58, ..., 72\}$. While parts of the income streams can be directly observed in the data, we compute all future and counter-factual income streams based on each person's observed benefit entitlements, retirement savings and earnings history. Given the richness of our data, we are able to compute the income streams with a high level of detail and accuracy. In the following, we describe how we observe/compute each of the financial components, how we measure the actual retirement ages, and how we define our sample:

Income

Earnings: W observe annual labor earnings directly in data for both wage earners and self-employed up until retirement, and we assume that the counterfactual earnings equal the average of the latest three observed earnings prior to retirement, with a wage inflation of 3.2%. We also include unemployment insurance benefits in our earnings definition. Our binary retirement assumption results in individuals having zero earnings in retirement. As our sample only includes individuals with full-time employment at age 54, we assume that all individuals receive 0 transfers, excluding unemployment insurance benefits, before retirement.

Early Retirement Pension (ERP): We observe if an individual is entitled to receive early retirement benefits or not at age 59, and we compute the ERP benefit rates according to the rules described in Section 2. We assume that all ERP eligible individuals are full-time insured, that they are eligible from age 60, and thereby meet the two-year-rule requirement at age

62. Appendix Table 3 outlines in detail how the ERP benefit is means tested with respect to different types of retirement savings. Individuals who retire after age 62, and who is thereby fulfilling the two-year rule, are assumed to postpone any individual pension payments until age 65 in order to avoid any reductions in the ERP benefits. If a person fulfills the two-year rule but delays retirement, he will receive a tax-free bonus of approximately 40,000 DKr for each year of postponement.

Old Age Pension (OAP) We model the OAP benefit payments according to the rules described in Section 2. We assume that all individuals older than 65 are entitled to receive the OAP. We assume that individuals only choose to receive OAP after their retirement. As we assume 0 earnings in retirement, all individuals receive the full OAP baseline amount. The amount of supplemental OAP depends on the cohabiting partner's labor market status and earnings and is reduced with the recipient's level of total income (including private pension payouts). We consider the latest observed labor market status and earnings of each person's partner and assume that individuals do not change their partner after 2018. If the partner's retirement age is unobserved in data, we assume it to be 65. We further assume that the partner doesn't die. While these assumptions might seem coarse, recall that the partner status only influences our model through the supplemental OAP benefit rate. As for the remaining transfer income types, future benefit rates are assumed to grow with an assumed inflation rate of 3.2%.

The exact retirement age of the spouse is not necessarily observed. In that case, it is set to the default retirement age, 65. The latest observed salary of the spouse is extrapolated with wage inflation $\pi_{wage} = 3.2\%$ until the actual or assumed retirement. Some individuals lose their spouse and/or get a new spouse during the age interval of 60 to 67. You can argue both against and in favor of including observed changes in partner status when computing the future income streams. Whether individuals are able to predict divorces, the death of a partner, meeting a new partner, etc. is a delicate matter. We include all observed changes in partner status in accordance with the perfect foresight assumption.

Other Retirement Benefits that we also compute and include in our income definition include housing benefits ("boligsikring") and older check ("ældrecheck"). All future transfer income is assumed to grow with an assumed inflation rate of 3.2%.

Taxes: We apply the actual tax rules which applied up until, and including 2018. For later years, we apply the 2018 rules where we let the limit amounts for the different progressive tax-levels increase with the wage inflation of 3.2%. As such, we implicitly assume that individuals also have perfect foresight with respect to the different tax reforms which took place between 1996 and 1918.

Retirement Wealth and Retirement Income

We observe the accumulated amounts and types of pension wealth for all individuals when they are 59.5 years old. These are 3rd party reported by the financial institutions, and are therefore highly reliable. We also observe what types of pension savings each person hold, why we are able to compute the payment schemes and corresponding benefit deductions with high accuracy. Some pension savings are observed as the deposited amount and others as the annual commitment given retirement at age 60. We assume that individuals save for their retirement as long as they work. If the actual contribution rate for an individual is unobserved, we assume a contribution rate equal to the average of observed contribution rates from when the individual was 54 years old, until his/hers observed retirement age.

Capital Pension: We assume that Capital Pensions (CP) are paid as lump sums during the first year of retirement. We assume that the capital pension deposit is grow with the annual interest rate $i_d = 4.75\%$ in the period prior to retirement. All interest gains on retirement savings are taxed with the so-called PAL tax, $\tau_{PAL} = 15.3\%$. CP_{59} is observed directly in the data, and the subsequent years are computed as

$$CP_a = CP_{a-1} * (1 + i_d(1 - \tau_{PAL})) + \Delta CP_{a-1} \text{ for } a \leq r$$

where ΔCP_{a-1} denotes the contributed amount to the capital pension at age $a - 1$

Term Pension: We assume that all terms pension payments are equally distributed through annuities of 10 years, such that the payment size equals 10% of the deposited amount at the retirement age, growing with interests $1 + i_r(1 - \tau_{PAL})$ each year. Payments start at the year of retirement with exemption of ERP eligible individuals who retire at age 63 or 64. They are assumed to postpone the payments until age 65 to avoid reductions in ERP benefits.

Life annuities: We observe life annuities both as total deposited value (LA^{TOT}) and as annual commitments given retirement at age 60 (LA^{PAY}). Assuming that the deposited values at all times should equal the present value of future annuity payments, we get the following correspondence between the two figures:

$$LA_a^{PAY} = \frac{LA_a^{TOT}}{\sum_{i=a}^{100}(1 - \mu_i) \times \left(\frac{1+i_d(1-\tau_{PAL})}{1+\pi_{wage}}\right)^{-(i-a+1)}}$$

The total committed amount, LA^{TOT}, is assumed to follow same development as the capital pension such that

$$LA_a^{TOT} = LA_{a-1}^{TOT} * (1 + i_d(1 - \tau_{PAL})) + \Delta LA_{a-1}^{TOT} \text{ for } a \leq r$$

where ΔLA_{a-1}^{TOT} denotes the amount contributed to the life annuity savings at age $a -$

1. LA_a^{TOT} denotes the age a fixed-price value of the total commitment and μ_i the death probability at age i. The interests gained on the deposited value are assumed to equal the interest rate on deposits $i_d = 4.75\%$. We let π_{wage} denote the wage inflation, set to growth (g=1.5%) times price inflation $\pi_{price} = 1.75\%$, such that $\pi_{wage} = 1.015 * 1.0175 = 1.032 = 3.2\%$.

Retirement definition

To identify the chosen retirement ages of our sample, we also consider the DREAM-register contains weekly information about all public transfer payments made to each person, including ERP and OAP benefits. We combine these weekly observations with annual earnings and income observations to measure people's chosen retirement ages. We define an individual to be retired if at least one of the below statements are true: 1) Receives ERP benefits, 2) When the yearly salary is less than half of the pre-retirement salary for two years in a row. We define the pre-retirement salary as the average annual salary observed at ages 55-57. We prefer this relative income threshold rather than an absolute threshold to account for differences in worker productivity. Retirement is assumed to be an absorbing and binary state, and as such, it suffers from two limitations. First, some people may return to employment after retirement. Second, some people may still have little employment even though their yearly wages "permanently" become less than half of their initial wage.

Sample Definition

Covering the period from 1996 to 2016, we observe individuals born in 1942 to 1954 from when they are 55 years old. As such, we observe cohort 1942 until they are 74 years old, while we only observe cohort 1952 until they are 64 years old. The estimation sample includes cohorts 1942-1952, whereas the remaining cohorts are used for out- of sample validation and forecasts. We restrict the analysis to cover all individuals who at age 57 face an actual retirement decision (i.e. not retired people), which amounts to all individuals who are observed with earnings above 90.000 DKr (2001-amount), and individuals on disability- and transition benefits together with civil servants[3] are excluded. Table 1 outlines how the data size decreases when the different groups are discarded from the sample. We end up with roughly 53% of the population for cohort 1942, and the coverage increases over the cohorts such that our sample includes 69% of the population born in 1953 - mostly due to increased labor force participation at age 57 (increasing from 63 to 74%). Individuals who are not present in data throughout all 18 consecutive years (age 58 to 72) are not discarded from the analysis. Appendix D describes how data censoring is handled throughout the estimation process.

[3]Civil servants refer to those eligible for the Danish equivalence of a defined benefit plan in the US ("tjenestemandspension").

Table 1: Sample Selection

Cohort	Initial sample Size	Excl. not in labor force	Excl. diasb. pens.	Excl. civil serv.	Share of initial sample
1942	65998	41609	40294	33675	0.53
1943	69767	45126	43663	36217	0.56
1944	75475	50155	48430	40320	0.58
1945	78956	53921	51970	43050	0.60
1946	81348	56962	54906	45408	0.61
1947	78495	57022	54975	45958	0.64
1948	73537	54041	52072	43107	0.65
1949	69714	50872	48863	40424	0.65
1950	69843	51275	49331	41026	0.67
1951	67743	49944	48051	39952	0.66
1952	68742	50732	48973	40947	0.67
1953	69930	51836	50100	42651	0.69

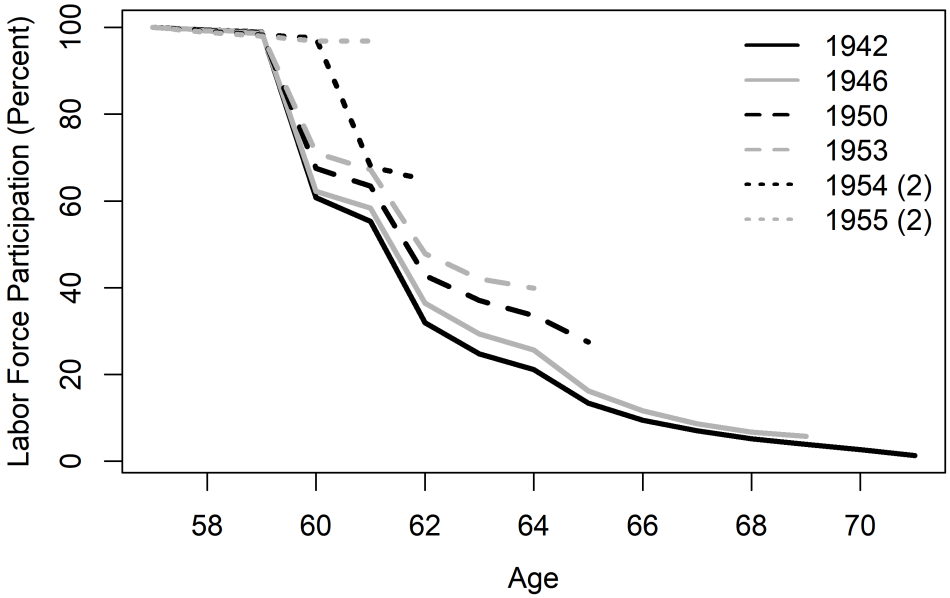
The table shows how the size of our main sample varies after we exclude individuals who are not facing a "standard" retirement decision, by cohort. The initial sample includes the full population of Danes who were alive and living in Denmark at age 57. The Column "Excl. not in labor force" shows the sample size once everyone who is not working (defined by annual earnings larger than 90.000 DKr (2001-value)) are excluded. The Column "Excl. disab. pens." shows the remaining sample size once also those who retire on disability insurance after age 57 are excluded. The Column "Excl. civil serv." excludes those individuals who ever paid into a "civil servant" retirement account. We exclude these individuals as we're not able to properly observe their private retirement entitlements (today, the civil servant arrangement is much smaller in scale).

4 Descriptive Statistics

4.1 Trends in retirement

Figure 2 shows the cohort-specific labor force participation rates of our main sample as specified in Table 1. The figure shows the cohort-specific developments in labor force participation from age 57 to 71. For cohorts 1942-1953, we see that employment shares drop at the ERP age (60), at age 62 where the two-year-rule is activated and again - however smaller in size - at the normal retirement age 65. For the cohort born in 1954 (2nd half), we see that the drop in employment share is shifted to the new ERP eligibility age 61. At age 60, the labor force participation of cohort 1954 (2nd half) was 28 percentage points larger compared to cohort 1953.

Figure 2: Labor Force Participation by Cohort, Estimation Sample



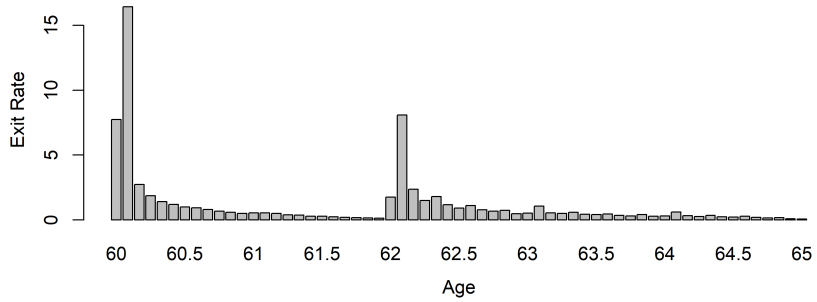
The figure plots the labor force participation rate at ages 57-71 for different cohorts for the subgroup of the population who were in the labor force at age 57. (2) indicates that the plot only covers those born in the second half of the year. Based on own calculations on full-population Danish register data, using the retirement age definition described in Section 3.

Figure 2 shows an upward parallel shift in employment shares across cohorts, meaning that the younger cohorts tend to retire later, even though the statutory retirement ages are unchanged for the cohorts 1942-1953.

Figure 3 depicts the retirement ages of individuals who retire on the early retirement scheme with a monthly precision for the birth cohorts 1942-45. Most individuals retire immediately after their birthday from which they become eligible for either ERP (age 60) or from which they can benefit from the two-year rule (62). The fact that individuals tend to retire at these favorable ages, and immediately after their birthdays, suggests that financial incentives are essential when they time their retirement age. It could also suggest that social norms and/or reference point effects are at stake.

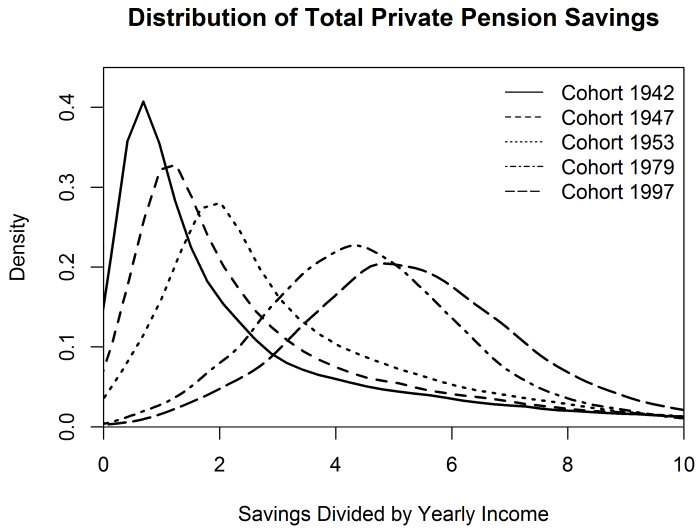
Figure 4 illustrates how the retirement wealth of seniors workers in Denmark increases over time. The figure shows the observed/predicted distribution of accumulated retirement savings for different cohorts, measured 7 years before the NRA. It clearly shows that the level of retirement savings are increasing over time, reflecting the maturation of the mandatory contribution plans, which were introduced in the 1990s for most workers. As such, younger cohorts have accumulated their pension wealth during more years. An increase in expected NRAs also explain some of the increase in retirement savings.

Figure 3: Exact retirement ages, cohort 1942-1943



The histogram shows the distribution of the precise retirement age of the seniors born in 1942-1943, who retire on Early Retirement Pension. Data are from the DREAM database (100% of the population) where we can identify the exact week in which an individual starts to collect ERP benefits.

Figure 4: Evolution of private pension savings across cohorts

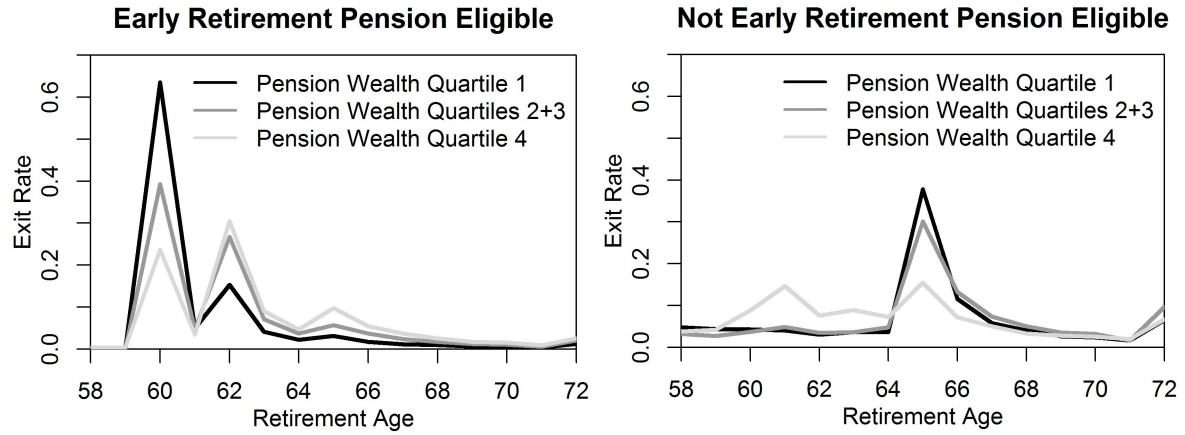


This figure shows the distribution of accumulated retirement savings of seniors who are working 7 years prior to their respective NRA, as measured by annual income for different cohorts. Retirement savings are measured six years prior to the NRA, divided by average annual income 7-10 years prior to the NRA. The distributions for cohort 1942, 1947 and 1953 are calculated using administrative register data, while the distributions for cohorts 1979 and 1997 are calculated using simulated data from the Danish Ministry of Finance's micro simulation model MFI. MFI models the incomes and pensions savings of future generations, assuming that individuals will continue to contribute to mandatory pension saving schemes with current contribution rates. Thus, the model captures the maturation of private pension wealth.

4.2 Lessons from the ERP-scheme

Figure 5 illustrates the exit rate distribution across ages for individuals with different levels of retirement wealth. The right figure shows the exit rates of people who are not entitled to ERP-benefits. As expected we see that those with large pension wealth are more likely to self-finance an earlier retirement prior to the NRA. Approximately 40% of those who are not members of the ERP program decide to self-finance an earlier retirement age, and this share increases with the level of retirement wealth. Different means testing and eligibility rules apply to the ERP-scheme and the old age pension as described in Section 2. Where the early retirement scheme induces individuals with large private pension savings to postpone their retirement until age 62. The exit rates in Figure 5 show that individuals respond to these incentives. In the group of individuals entitled to early retirement benefits, those with high private pension wealth retire later than those with smaller private pension savings. Thus, the two-year rule of the ERP-scheme reverses the negative retirement effect of private pension wealth.

Figure 5: Exit rates for different levels of pension wealth, ERP and Non-ERP eligible



The two figures plot the exit rates (share who exit the labor force) at ages 58-72 for our main sample (see Table 1), cohorts 1942-1944. The quartiles of the retirement savings distribution are computed within cohorts.

Of course, earnings levels also explain some of the heterogeneity in retirement behavior. To check that the findings in Figure 5 also apply to individuals with similar income levels, we show similar plots which are split by earnings quartile. Appendix Figure 17 confirms that high pension wealth individuals who are ERP-eligible tend to retire later than the low pension wealth ERP eligible individuals - conditional on them being in the same income quartile. For those who are not eligible for ERP benefits, we also observe that the high-wealth individuals tend to retire earlier than low-wealth individuals, conditional on them being in the same earnings quartile, see Appendix Figure 18. The described patterns hold across different gender- and education groups.

5 The Model

We model retirement as an absorbing state with the binary option of either working full-time or being retired at a given age. The model is derived in a random utility model framework with N utility-maximizing agents who at age $a = 57$ decides when to retire, $r \in \{58, ..., 72\}$. Evaluated at age 57, the utility of retiring at age r consists of a deterministic and a stochastic component:

$$U(r) = V(r) + \sigma\epsilon_r \tag{1}$$

$V(r)$ denotes the present value of discounted lifetime utility of consumption obtained from retirement age r. The *retirement age* dependent unobserved heterogeneity ϵ_r is assumed i.i.d. extreme value distributed with scaling parameter σ. ϵ_r is known to the agent, but not to the econometrician, and reflects individual preferences not captured by the model, such as health, behavior of the spouse, etc. The scale parameter σ determines the weight put on the deterministic relative to the stochastic part of the utility function. If $\sigma = 0$, economic incentives alone rule the retirement decision. Large values of σ, on the other hand, implies that the retirement decision, from the econometrician's perspective, is random.

We assume that agents have perfect foresight with respect to their future financial situation - they know with certainty their future salaries, interest rates, pension payments, benefit entitlements, etc. The only uncertainty is with respect to death, where age- and gender-specific death probabilities are known to the individuals. As such, we don't account for any income, health or health-cost related uncertainty. Such assumptions are less critical in a Danish setting with free universal health care and generous unemployment benefits, and these assumptions collapse the retirement decision problem into a discrete choice problem between the different retirement ages $r \in \{58, ..., 72\}$.

We assume CRRA utility with relative risk parameter ρ. $V(r)$ is given by

$$V(r) = \sum_{a=58}^{120} \frac{\left(\gamma\left(r, a|k, \alpha\right) c_a\right)^{1-\rho}}{1-\rho} D_a \tag{2}$$

where D_a is the mortality-adjusted discount factor:

$$D_a \equiv \beta^{a-57} \prod_{s=58}^{a} \left(1 - \mu_s\right).$$

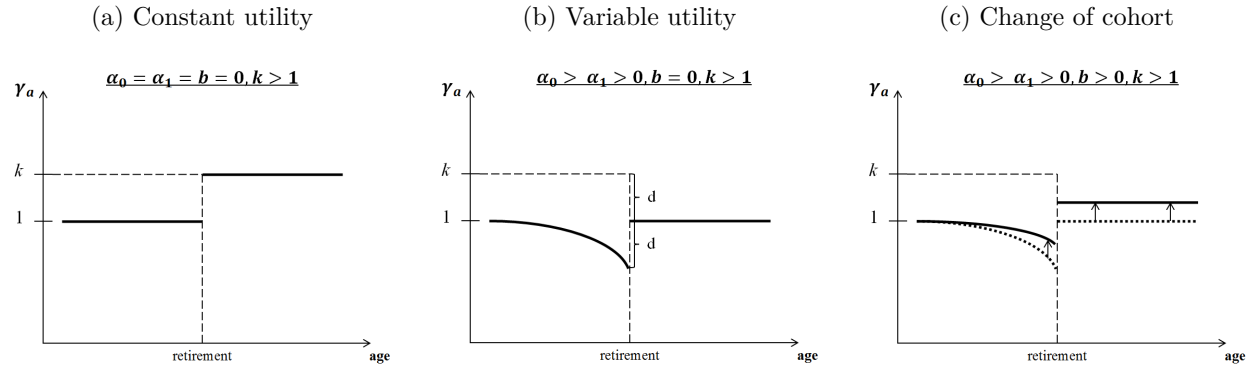
Here, μ_s denotes the death probability at age s, and β denotes the usual discount factor. c_a denotes the agent's chosen level of consumption in period a, given the optimal consumption problem described below. Consumption c_a is scaled with the age- and retirement-age dependent function $\gamma\left(r, a|k, \alpha\right)$, whose shape is defined by the leisure preference parameter k, and α which we will refer to as the "attrition parameter". The functional form of$\gamma\left(r, a|k, \alpha\right)$is

given by

$$\gamma\left(r,a|k,\alpha\right)=\begin{cases}e^{-\alpha a^{2}} & \text{for } a<r\\ k\cdot e^{-\alpha r^{2}} & \text{for } a\geq r\end{cases}$$

γ scales the utility of consumption, and figure 6 illustrates how its shape depends on the parameters α and k.

Figure 6: The parameter $\gamma_a\left(r\right)$



The leisure preference parameter k measures the utility effect of retirement as it scales the utility of consumption after retirement compared to while working. As such, our k parameter is similar to that in the option value proposed by Stock and Wise [1990]. If k is 1.1, the consumer values one unit of consumption 1.1 times higher once retired. We allow k to vary for each person, and estimate its population distribution. α decreases the utility level after retirement permanently - and more, the longer the retirement is postponed. Therefore, we think of α as capturing some kind of "attrition" related to working. Once retired, the utility level remains constant at the decreased value. We allow the attrition parameter to depend linearly on the cohort b, in order to capture an increase in health status across generations:

$$\alpha=\alpha_0+b\alpha_1$$

Agents with wealth endowment W_a, income y_a, and consumption c_a at age a face the budget restriction

$$W_a=\left(1+\left(1-\tau\right)i_a\right)W_{a-1}+y_a\left(r\right)-c_a, \tag{3}$$

where τ denotes a tax on capital income and i_a denotes the age-dependent and actuarial fair interest rate, dependent on the risk-free interest rate i and the death probability at age a, μ_a:

$$1+i_a=\frac{1+i}{1-\mu_a} \tag{4}$$

The future income stream $y_a(r)$, is assumed to be known by the individual with certainty for each possible r. It is determined by the observed earnings level of the individual, his/her accumulated retirement savings, benefit entitlements etc., as described in Section 2

Under these assumptions, standard calculus yields a closed-form solution to the optimal consumption problem, which we show in Appendix B. An agent who maximizes the utility function stated in 1 for a given retirement age r, and given the budget restriction in equation 3 and interest rate in equation 4, will have the *indirect utility function:*

$$V(r) = \frac{\left(\frac{W_{57}+H(r)}{P(r)}\right)^{1-\rho}}{1-\rho} \tag{5}$$

where W_{57} is wealth at the beginning of age 58. $H(r)$ and $P(r)$ are given by

$$H(r) = \sum_{a=58}^{120} y_a(r) R_a$$

$$P(r) = \left[\sum_{a=58}^{120} \gamma_a(r)^{\frac{1-\rho}{\rho}} D_a^{\frac{1}{\rho}} R_a^{\frac{\rho-1}{\rho}}\right]^{\frac{\rho}{\rho-1}}$$

where R_a is defined as

$$R_a = \prod_{s=58}^{a} \frac{1}{1+(1-\tau)\, i_a}$$

As such, $H(r)$ measures the future discounted income at the beginning of age 58 for chosen retirement age r, and $P(r)$ is the corresponding CES price index of future discounting factors. In a similar model, Helsø [2015] also models the consumption-savings decision separately, relaxing the assumption of a symmetrical interest rate function. Whereas this addition complicates the computation of the indirect utility function, it did not contribute significantly with respect to the model's performance.

As per our assumption that ϵ_r is iid. extreme value distributed, the probability of retirement at age r is computed by the logit specification

$$Prob(r) = \frac{exp(V(r))}{\sum_{r'=55}^{72} exp(V(r'))} \tag{6}$$

We observe a significant number of individuals who choose to retire at the statutory old age pension retirement age, despite relatively weak financial incentives at this specific retirement age. A common explanation is that the NRA provide a focal/reference point for decision-making and shapes the social norms about what is the appropriate retirement age. This effect has been studied in e.g. Behaghel and Blau [2012] and Seibold [2017]. To include such a focal point effect, we estimate a slightly different version of the model, in which the

indirect utility is given by

$$\overline{V}(r) = V(r) + d_{65}1_{[r=65]} \tag{7}$$

where d_{65} measures a jump in utility if the chosen retirement age is exactly equal to the statutory old age pension age, which is 65 years in the estimation sample.[4]

Consequently, we slightly alter the utility equation described in equation 1 such that

$$U(r) = \overline{V}(r) + \sigma\epsilon_r = V(r) + d_{65}1_{[r=65]} + \sigma\epsilon_r \tag{8}$$

The probability of retirement basically depend on three types of variables: 1) preference parameters that are assumed to be identical for all consumers $\Theta = (\alpha_0, \alpha_1, \beta, \sigma, \rho, d_{65}^0, d_{65}^1)$, 2) the heterogeneous leisure preference parameter k, and 3) variables that characterize the financial situation of each individual, $x^j = (V(58), .., V(72))$. The indirect utility function given by (5) and (7) is therefore a function of these variables, and we can write the retirement probability as

$$Prob(r|\Theta, x, k) = \frac{exp\left(\overline{V}(r|\Theta, x, k)\right)}{\sum_{r'=58}^{72} exp\left(\overline{V}(r'|\Theta, x, k)\right)} \tag{9}$$

6 Estimation

For each agent $j \in 1, ...n$, we observe(r^j, x^j), where r^j is person j's actual retirement age, and $x_j = \left(W_{57}^j, H(58)^j, .., H(72)^j\right)$ comprise his/her financial data (see Section 5 equation 5). We estimate the homogeneous model parameters $\Theta = (\alpha, \beta, \sigma, \rho, d_{65})$ together with the population distribution of the heterogeneous preference parameter, $p(k)$, in a nested 2-step procedure: first we assume some value of the homogeneous model parameters, $\hat{\Theta}$. Given $\hat{\Theta}$ and the observed data, (r^j, x^j), we're then able to compute the corresponding population distribution of the leisure preference parameter $p(k|\hat{\Theta})$. We do this using a fixed-grid version of the Expectation Maximization (EM) algorithm as described in the following. Given the model parameters $(\hat{\Theta}, p(k|\hat{\Theta}))$, we can compute the likelihood function, which is the joint probability of all observed retirement ages:

$$L\left(\hat{\Theta}, p(k|\hat{\Theta})|x, r\right) = \prod_{j=1}^{n} Prob(r^j|\hat{\Theta}, x^j, p(k|\hat{\Theta}))$$

Then we consider a new set of model parameters $\hat{\hat{\Theta}}$, and repeat the 2-step procedure. We continue to do so, until we have obtained the maximum likelihood. We use the non-linear

[4]We allow the value of d_{65} to vary between individuals born before and after 1946 to account for a temporary tax-reduction policy ("skattenedslaget") which made it more attractive for individuals born in 1946-1952 to retire on or after the statutory retirement age.

optimization routine NLM in R to search through the parameter space of Θ.

We compute $p(k)$ using the fixed-grid version of the Expectation Maximization (EM) algorithm as proposed in Train [2008] Section 6. The EM algorithm was originally a procedure developed for dealing with missing data, proposed by Dempster et al. [1977]. With this method, we estimate the population distribution of the preference parameter $p(k)$ as a discrete distribution, where we, for a fixed grid of k-values, $k_{min}, ..., k_{max}$, estimate the probability mass (or share) in each grid point, $p(k_{min}), ..., p(k_{max})$. As an alternative to the fixed-grid version of the EM-algorithm, one could also estimate $p(k)$ numerically, using the method proposed by Fox et al. [2016]

Assume that we know the homogeneous model parameters Θ. The probability that individual j retires at age r^j with financial circumstances x^j, conditional on a given $k' \in \{k_{min}, ..., k_{max}\}$, is given by the logit specification

$$Prob\left(r^j|\Theta, x^j, k'\right) = \frac{exp\left(\overline{V}(r^j|\Theta, x^j, k')\right)}{\sum_{r'=58}^{72} exp\left(\overline{V}\left(r'|\Theta, x^j, k'\right)\right)} \tag{10}$$

where $\bar{V}(r)$ denotes the indirect utility of retiring at age r given in equation 7. Applying Bayes' rule to r^j and k in the above equation, we're able to deduct the probability that individual j has preference parameter k

$$Prob(k'|\Theta, x^j, r^j) = \frac{Prob\left(r^j|\Theta, x^j, k'\right) p(k'|\Theta)}{Prob(r^j|\Theta, x^j)} \tag{11}$$

We compute the denominator in equation 10 by marginalizing out the conditional variable k in equation 11 using numerical integration:

$$Prob(r^j|\Theta, x^j) = \sum_{k''=k_{min}}^{k_{max}} Prob\left(r^j|\Theta, x^j, k''\right) p(k''|\Theta)$$

However, as we don't know $p(k')$, we cannot directly compute 10. Instead, we rely on the EM-algorithm, which enables us to derive $Prob(k'|r^j, x^j, \Theta)$ by an iterative recursion. Let i denote the iteration. We start with some initial guess of the probability mass of k': $p(k')^0$. Then we repeatedly update its value by the formula:

$$p(k'|\Theta)^{i+1} = \frac{1}{n} \sum_{j=1}^{n} \frac{Prob\left(r^j|\Theta, x^j, k'\right) p(k'|\Theta)^i}{Prob\left(r^j|\Theta, x^j\right)} \tag{12}$$

The EM recursion is repeated until convergence. In practice, we assume an even-spaced k-grid, where we find the optimal distance using cross-validation. Standard errors are bootstrapped as suggested by Train [2008]. A step-by-step description of f the estimation algorithm is included in the following text box.

Estimation Algorithm

1. Observe the actual retirement ages $r^j$and financial situations $x^j$of individuals $j = 1, .., n$
2. Set the fixed grid points of k within some boundary values, k_{min} and k_{max} and assume some initial distribution $p(k)^0$ on the defined interval (e.g. a flat uniform distribution)
3. Consider some set of homogeneous model parameters $\hat{\Theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{d_{65}})$
   - 3(a) For each person and each $k' \in \{k_{min}, ..., k_{max}\}$, compute $Prob(r^j|\Theta, x^j, k')$
   - 3(b) For each person and each $k' \in \{k_{min}, ..., k_{max}\}$, compute $Prob(k'|\Theta, x^j, r^j)^i = \frac{Prob(r^j|\hat{\Theta},x^j,k')p(k'|\Theta)^{i-1}}{Prob(r^j|\Theta,x^j)} = \frac{Prob\left(r^j|\Theta,x^j,k'\right)p(k'|\Theta)^{i-1}}{\sum_{k''=k_{min}}^{k_{max}} Prob(r^j|\Theta,x^j,k'')p(k''|\Theta)^{i-1}}$
   - 3(c) For $k' \in \{k_{min}, ..., k_{max}\}$, update the population distribution such that $p(k'|\hat{\Theta})^i = \frac{1}{n}\sum_{j=1}^{n} Prob(k'|r^j, x^j, \Theta)^i$ for $k' \in \{k_{min}, ..., k_{max}\}$
   - 3(d) Repeat step 3(b)-3(c) until convergence of $p(k|\hat{\Theta})$
4. Compute log likelihood $L\left(\Theta, \hat{p(k|}\hat{)}|x,\Theta\right)$
5. Return to step 3 and consider a new guess of parameters $\hat{\hat{\Theta}}$, until likelihood function is maximized

7 Estimation Results

We use a random 33% sample of the full-population sample described in Table 1 to estimate the model separately for eight gender- and education specific groups, covering the cohorts born in 1942 to 1952. Since our data end in 2016, we observe cohort 1942 until they are 73 year old, while we only observe cohort 1952 until age 62. Other individuals might also die before we observe their chosen retirement age. For the individuals where we don't observe their retirement, we will use the information that they didn't retire at the observed retirement ages. Appendix Section D elaborates on how we deal with this censoring. The homogeneous parameters $\Theta = (\alpha, \beta, \sigma, \rho, d_{65})$ are estimated in an outer log likelihood optimization search routine, while for each guess of $\hat{\Theta}$, the corresponding distribution of k is estimated in a nested iterative fixed point algorithm described in Section 6. We let $k_{min} = 0.05$ and $k_{max} = 3.05$ with a step size of 0.1. Table 2 presents the maximum-likelihood estimated parameters $\Theta = (\alpha, \beta, \sigma, \rho, d_{65})$.

Table 2: Estimated Homogeneous Model Parameters (Std. Errors in Parentheses)

Men	**N**	α_0	α_1	β	σ	ρ	$\mathbf{d^0_{65}}$	$\mathbf{d^1_{65}}$
Unskilled	22462	0.0110 (0.0004)	-0.000163 (0.000010)	0.944 (0.010)	0.0356 (0.0297)	1.053 (0.036)	0.057 (0.002)	0.017 (0.002)
Vocational	39890	0.0133 (0.0003)	-0.000197 (0.000008)	0.940 (0.004)	0.0984 (0.016)	0.981 (0.011)	0.143 (0.001)	0.048 (0.001)
Short-Medium Tertiary	13084	0.0128 (0.0003)	-0.000153 (0.000008)	0.943 (0.006)	0.1260 (0.0439)	0.990 (0.019)	0.165 (0.003)	0.072 (0.002)
Long Tertiary	6528	0.0100 (0.0004)	-0.000151 (0.000021)	0.968 (0.006)	0.1650 (0.0622)	1.016 (0.028)	0.160 (0.005)	0.116 (0.004)
Women	**N**	α_0	α_1	β	σ	ρ	$\mathbf{d^0_{65}}$	$\mathbf{d^1_{65}}$
Unskilled	25016	0.0079 (0.0002)	-0.000063 (0.000009)	0.961 (0.003)	0.0158 (0.0056)	1.134 (0.027)	0.033 (0.001)	0.012 (0.001)
Vocational	33361	0.0093 (0.0003)	-0.000064 (0.000008)	0.950 (0.006)	0.0395 (0.0136)	1.057 (0.024)	0.0681 (0.003)	0.020 (0.002)
Short-Medium Tertiary	17210	0.0135 (0.0003)	-0.000129 (0.000002)	0.910 (0.003)	0.0284 (0.1906)	1.073 (0.027)	0.0205 (0.000)	0.007 (0.001)
Long Tertiary	3247	0.0085 (0.0003)	-0.00008 (0.000002)	0.963 (0.004)	0.0434 (0.0147)	1.13 (0.028)	0.0610 (0.000)	0.027 (0.002)

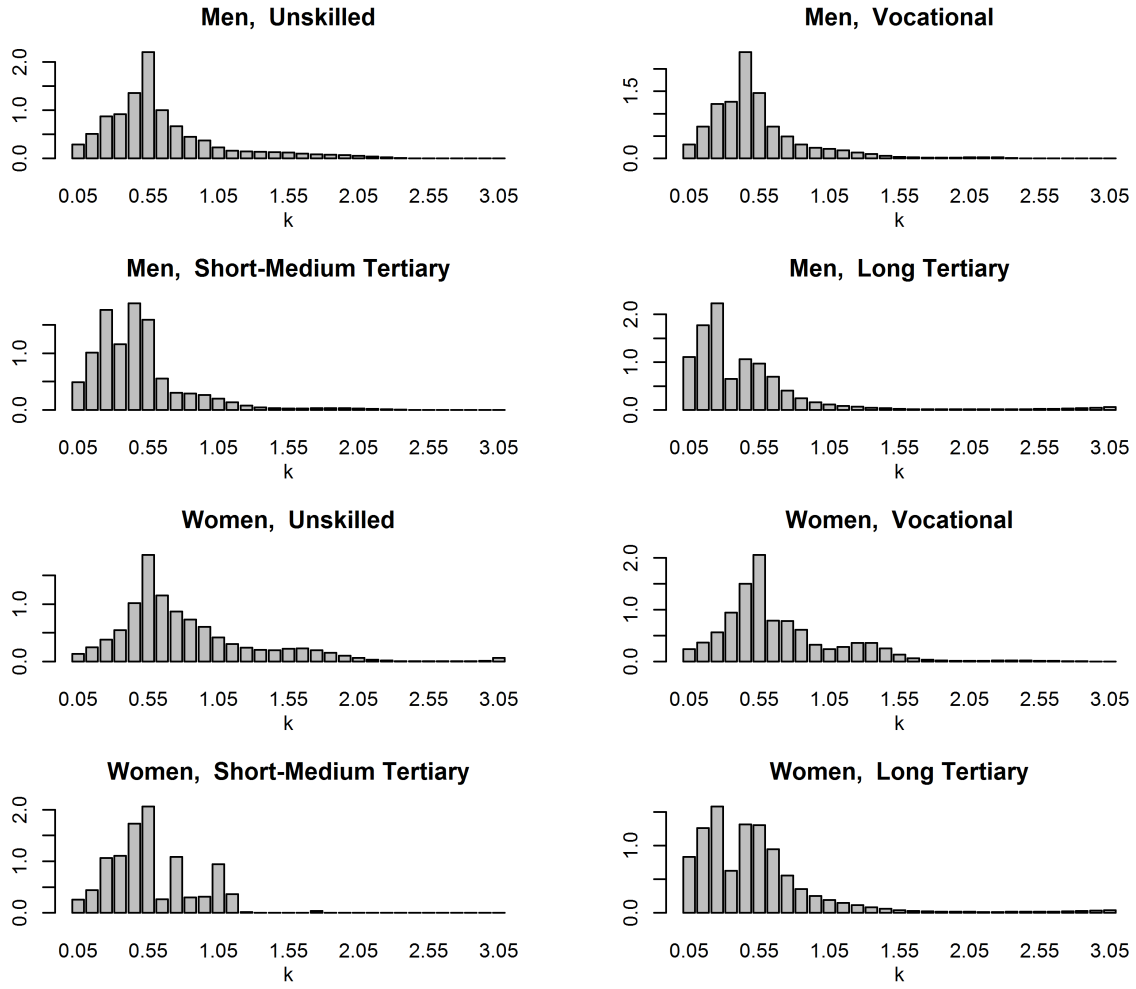
The estimated values of the attrition parameter α_0 are largest for males, and their magnitudes indicate that attrition is playing a key role in our specified model. For men born in cohort 1942 with vocational education, consumption will be down-scaled by a factor of 0.75 if retirement is postponed to age 62 and 0.47 if retirement is postponed to age 65, and this decline in utility persists throughout life. For women with vocational training, consumption is down-scaled by a factor of 0.84 if retirement is postponed to age 62 and 0.63 when delayed to age 65. The negative values of α_1 which specifies a cohort trend in attrition, is largest for men, suggest that attrition levels are decreasing for younger cohorts and that the gender gap in attrition is shrinking over time. The estimated discount factors, β, lie within a reasonable range from 0.937-0.973. Relative risk aversion coefficients, ρ, are close to 1 for all groups, and slightly larger for women, indicating that individuals are moderately risk averse. This is in line with the findings in Chetty [2006]. It is in line with what has been found

in similar studies, e.g. Rust and Phelan [1997] (1.07) and Blau and Gilleskie [2003] (0.95), but is smaller than other studies, e.g. French [2005] (2.2-5.1) and French and Jones [2011] (7.49). σ measures the scale of the extreme value type I error terms, and thereby weights the deterministic and stochastic components of our model as specified in equation 8. As such, a small value of σ indicates that the financial incentives are important for the retirement decision. However, the size of σ also reflects the overall variation of the deterministic part of the utility $\overline{V}(r)$, as specified in equation 8, which do not only depend on the remaining homogeneous model parameters, but also varies with k, and as such its interpretation is not straightforward.

As expected, the social norm dummy for cohorts 1942-46 (d_{65}^0) is larger than the social norm for the cohorts 1947-1952 (d_{65}^1). This is due to the temporary tax-reform "skattened-slaget", which made it more attractive for those born in 1947-1952 to retire on or later than the NRA. While the incentive did not lead to later retirement ages for these cohorts (few people understood this incentive), the financial incentives alone can explain a larger share of the individuals retiring at the NRA. The size of these social norm dummies should be interpreted in relation to the estimated scale parameter of the error term in the utility function, specified in equation 8, σ. The estimated parameter values of d_{65}^0 are roughly 1-2 times larger than σ, implying that the utility jump which an agent receives if he/she retires at the NRA (which we interpret as the magnitude of the social norms) is slightly larger than the overall variation of the agent's stochastic utility component, ϵ.

The homogeneous parameter estimates should be interpreted in the context of the estimated distributions of the heterogeneous leisure preference parameter k, which are plotted in Figure 7. The estimated distributions indicates that leisure preferences are widely distributed among the population. Recall that a low k implies a late retirement and vice versa. Appendix Figure 19 show, for four different types of individuals, how their expected retirement age varies across the k-distribution, and the plots clearly indicate that the different values of k have a large impact on explaining the retirement ages predicted by the model.

Figure 7: Estimated distributions of k



8 Model Fit

8.1 In-sample Fit

The primary objective of our model is to predict the retirement behavior of seniors so that it can be used for policy analysis and forecasts. This, of course, requires that the model predictions fit the data reasonably well. One of the main advantages of our model approach is that we, for each person, compute an individual-specific distribution of the k-parameter. For an in-sample policy experiment, we can use these individual-specific preferences. However, to calculate these individual-specific k-distributions, we need to observe the actual retirement ages. If we want to use our model on a new set of individuals for whom we don't observe a chosen retirement age (e.g., in a forecast), we have to rely on the education- and gender-specific population distributions of p(k) depicted in Figure 7. As such, we can test the fit of our model in two different ways, using either individual-specific k-distributions or the

population distribution.

For each person, we compute the average probability distribution over the possible retirement ages $r \in \{58, 59, ..., 72\}$. If we know individual j's actual retirement age r^j, we can calculate the probability that individual j retires at age r as: Θ, x^j, k'

$$\begin{aligned} P(r|\Theta, x^j, r^j) &= \int_{\mathcal{K}} P\left(r|k, \hat{\Theta}, x^j\right) P(k|r^j, \hat{\Theta}, x^j) dk \\ &= \int_{\mathcal{K}} P\left(r|k, \hat{\Theta}, x^j\right) \frac{P\left(r^j|k, \hat{\Theta}, x^j\right) p(k)}{P(r^j, \hat{\Theta}, \theta)} dk \end{aligned}$$

where we compute the integral numerically. However, if we don't know the actual retirement ages of the individuals, we instead use the estimated population distribution of $p(k)$ to compute the probability that individual j retires at age r as:

$$P(r|\Theta, x^j) = \int_{\mathcal{K}} P\left(r|k, \hat{\Theta}, x^j\right) p(k) dk$$

Let N_r denote the expected number of individuals who retire at age r. We sum the probability distribution over retirement ages for all individuals to get the expected number of retirements across the retirement ages:

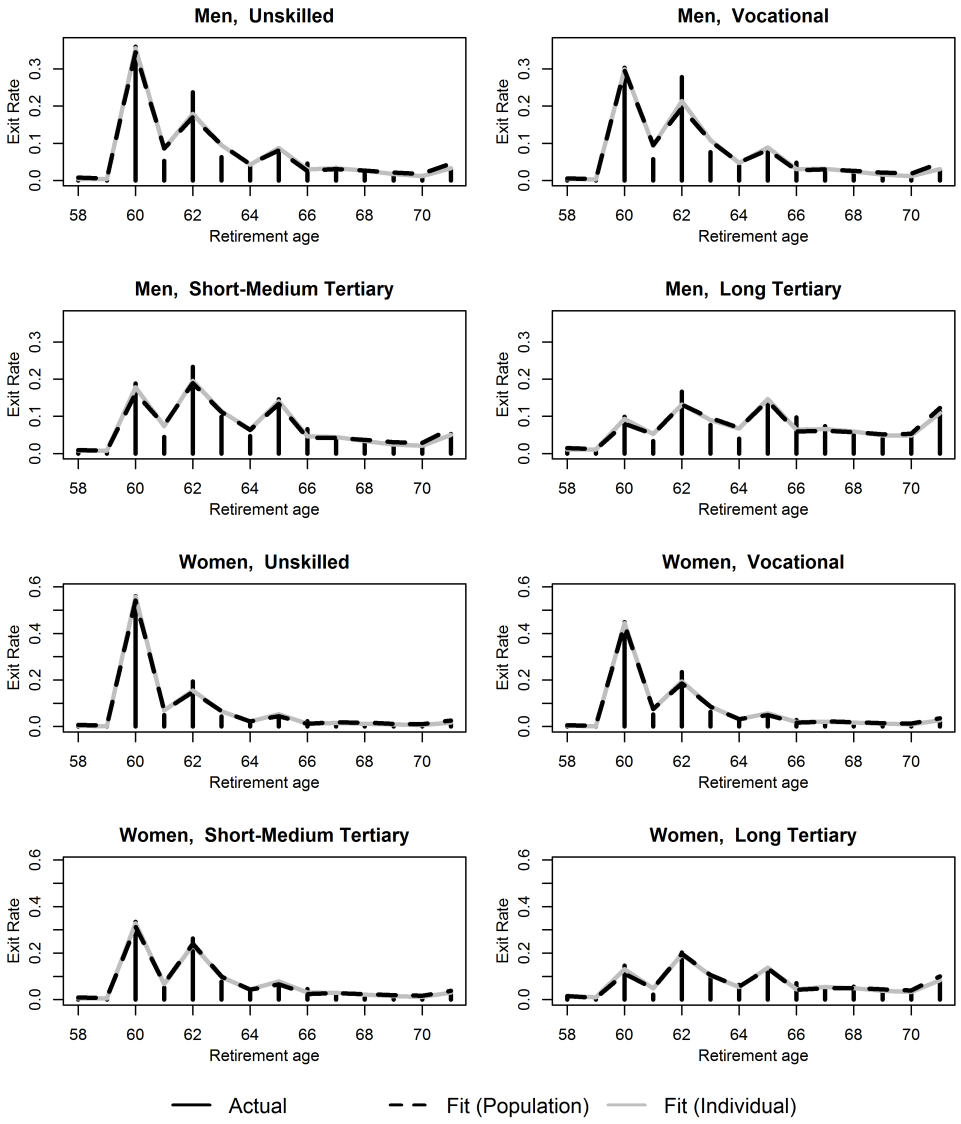
$$N_r^{individual} = \sum_{j=1}^{n} P(r|\hat{\Theta}, x^j, r^j),$$

$$N_r^{population} = \sum_{j=1}^{n} P(r|\hat{\Theta}, x^j)$$

The graphs in Figure 8 show, for each gender- and education group, the model prediction fit for the cohorts 1942-1944, where we observe retirements up until age 72.[5] The graphs plot the actual and predicted retirement distributions, relying both on both individual-specific and the aggregate population distributions of the leisure preference parameter k. We see that both predictions fit the data very well, with very little difference in estimated exit rates. The models slightly under-predict the exit rates at age 62, most pronounced for low-educated men, with a slight corresponding over-prediction of exit rates at ages 61 and 63. There is substantial heterogeneity in retirement age distributions across the different education and group where women and individuals with shorter education retire earlier. While the model fit is slightly better for women and groups with longer education, our model performs well across all groups, even for the predictions that rely only on the population distributions of k.

[5]For the younger cohorts, we don't observe all of the possible retirement ages, but these cohorts fit the data equally well.

Figure 8: Predicted vs. Actual Exit Rates



The graphs display the actual and predicted exit rates into retirement for birth cohorts 1942-1944 for whom we observe retirement exit rates from age 58-71. The vertical solid black lines plot the actual exit rates. The dashed black lines plot our model's predicted retirement distribution when we use only the population distribution of k. The solid gray lines show the predicted retirement distribution modeled using the estimated individual-specific k-distributions.

Our model also fits data well across the distribution of private pension wealth. In Appendix F, Figures 20 and 21 show the model fits similar to that in Figure, but for the subset of individuals with pension wealth in the first and fourth qualities of the pension wealth distribution within each estimation group. These figures confirm the findings in Figure 5, which is that individuals respond to the financial incentives within the ERP scheme: Individuals holding low levels of pension wealth are more likely to retire at age 62, while individuals

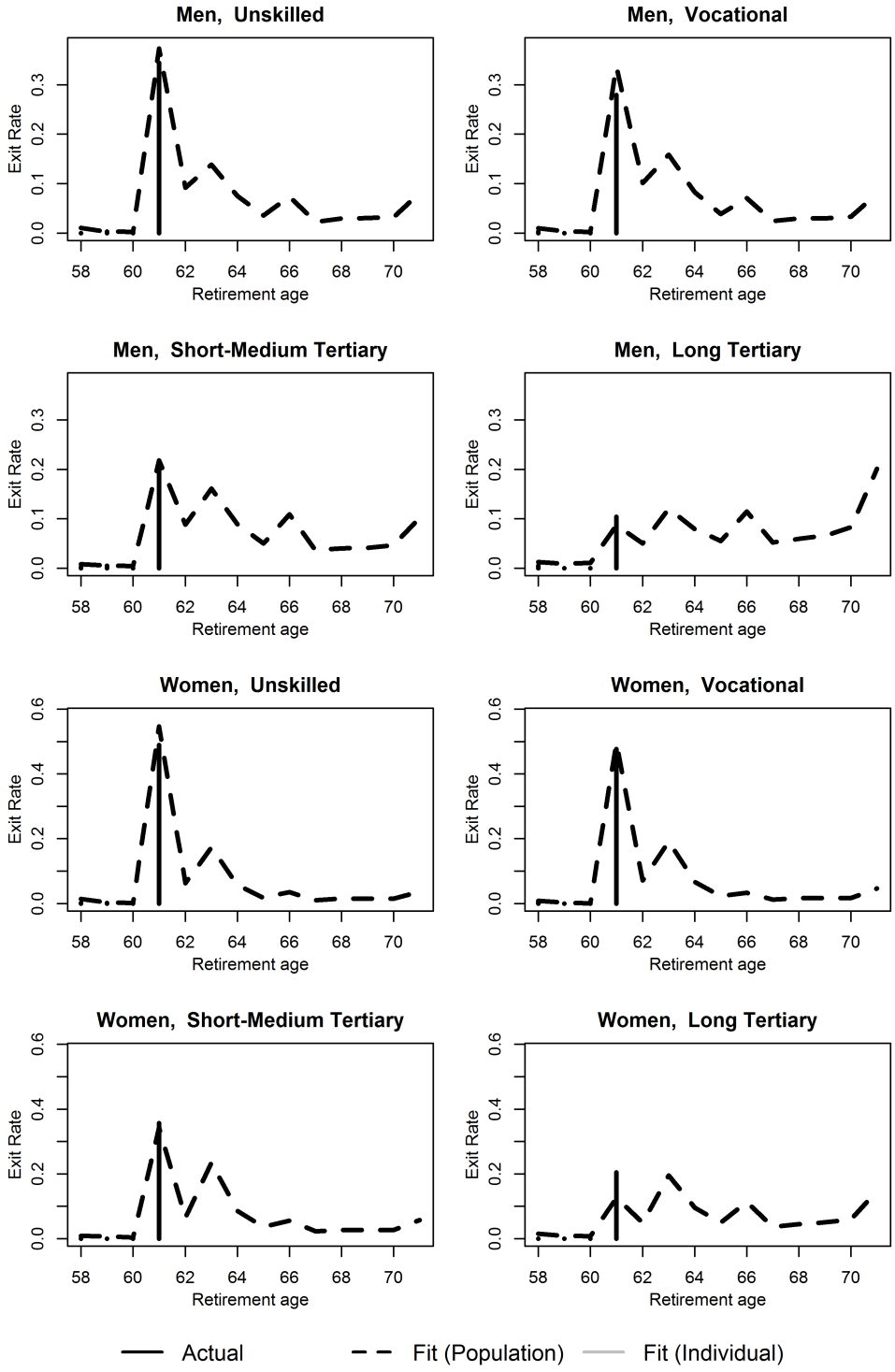
with more pension wealth to a larger extent postpone their retirement to age 62, so they can benefit from a smaller means-testing in their ERP benefits with respect to their pension wealth. Our model provides very reasonable fits for both the low- and high pension wealth groups, and as such we believe that our model captures the heterogeneity in retirement ages across the different levels of pension wealth. Appendix Figure 22 shows the model fit for the subgroup of individuals who are not members of the ERP program. The figure shows that the model prediction based on individual-level k distributions provide very good fits of the actual retirement age distribution, whereas the predictions that use the population distribution of k tend to slightly over-predict the exit rates at earlier retirement ages (61-64) and under-predict the exit rates at later retirement ages (66-71). This is driven by the fact that the subgroup of individuals who are not members of the ERP scheme, counting only 10% of the estimation sample, on average has lower levels of the leisure preference parameter k.

8.2 External Validation

We perform an out-of-sample validation exercise to test if our model can predict the retirement decision of persons not included in the model estimation, and to whom different retirement rules apply. We estimated the model on the cohorts born in 1942-1952, and all of these face an early retirement age of 60 and a statutory old age pension age of 65. Now we use our estimated model to predict the retirement decision of individuals born in the second half of 1954, facing an early retirement age of 61 and a statutory old age pension age of 66. We change the NRA dummy accordingly, and thereby we assume that the focal point/social norm effect follows the policy rule change.[6] As our data end in 2016, we can observe the share of individuals who retire before or on the early retirement age, 61. As we can only observe the chosen retirement ages at age 61 for individuals within the Early Retirement program (a self-financed retirement requires an extra year of observation), individuals who are not members of the Early Retirement scheme are excluded from the validation sample. Figure 9 shows the observed and predicted exit rates in the validation sample. The model does a good job predicting the exit rates at age 61 of the validation sample. For the groups with long tertiary education, the model fit is less precise. However, these predictions are more uncertain as they are based on few observations - 192 men and 128 women with long tertiary educations.

[6]The 1954 cohort was 51 years old when the reform was adopted

Figure 9: Predicted vs. actual exit rates, external validataion test on cohort 1954



The graphs display the actual and predicted exit rates into retirement for individuals born in the second half of 1954, who participate in the Early Retirement program. ´The vertical solid black lines plot the actual retirement age distribution. The dashed black lines plot our model's predicted retirement distribution when we use only the population distribution of k. The solid gray lines show the predicted retirement distribution modeled using the estimated individual-specific k-distributions.

9 Experiments

In this section, we present one baseline experiment in which we abolish the early retirement program, followed by three experiments where additional policy changes are added on top of the abolition of the early retirement program. We do this to mimic the retirement scene for future cohorts, where the large majority of individuals won't be eligible for ERP benefits. As such, the results of the main experiments will also reflect the prediction made by the baseline experiment. We run experiments within an experiment to get results which are more relevant for future policy making. All experiments are counterfactual by nature, as we conduct them on the estimation sample which includes the cohorts 1942-1952.

We will compare the estimated policy responses for individuals who hold different levels of retirement wealth. While earnings and non-retirement wealth are also important factors, these will not be the main focus of this analysis. We find that our conclusions are very similar once we restrict the analysis to only include average earning individuals. We also briefly show that non-retirement/liquid savings share many similarities with retirement savings in how it affects the response to the policy reforms.

A unique feature of our model is that we, for all persons, estimate an individual specific distribution of the leisure preference parameter k, which is retrieved from the person's observed retirement age and financial circumstances. As such, the model allows us to include these individual-specific k-distributions when we do our counter-factual experiments. However, we find that the results of our policy experiments are similar when we use the estimated population distribution of k, rather than the individual-specific distributions.

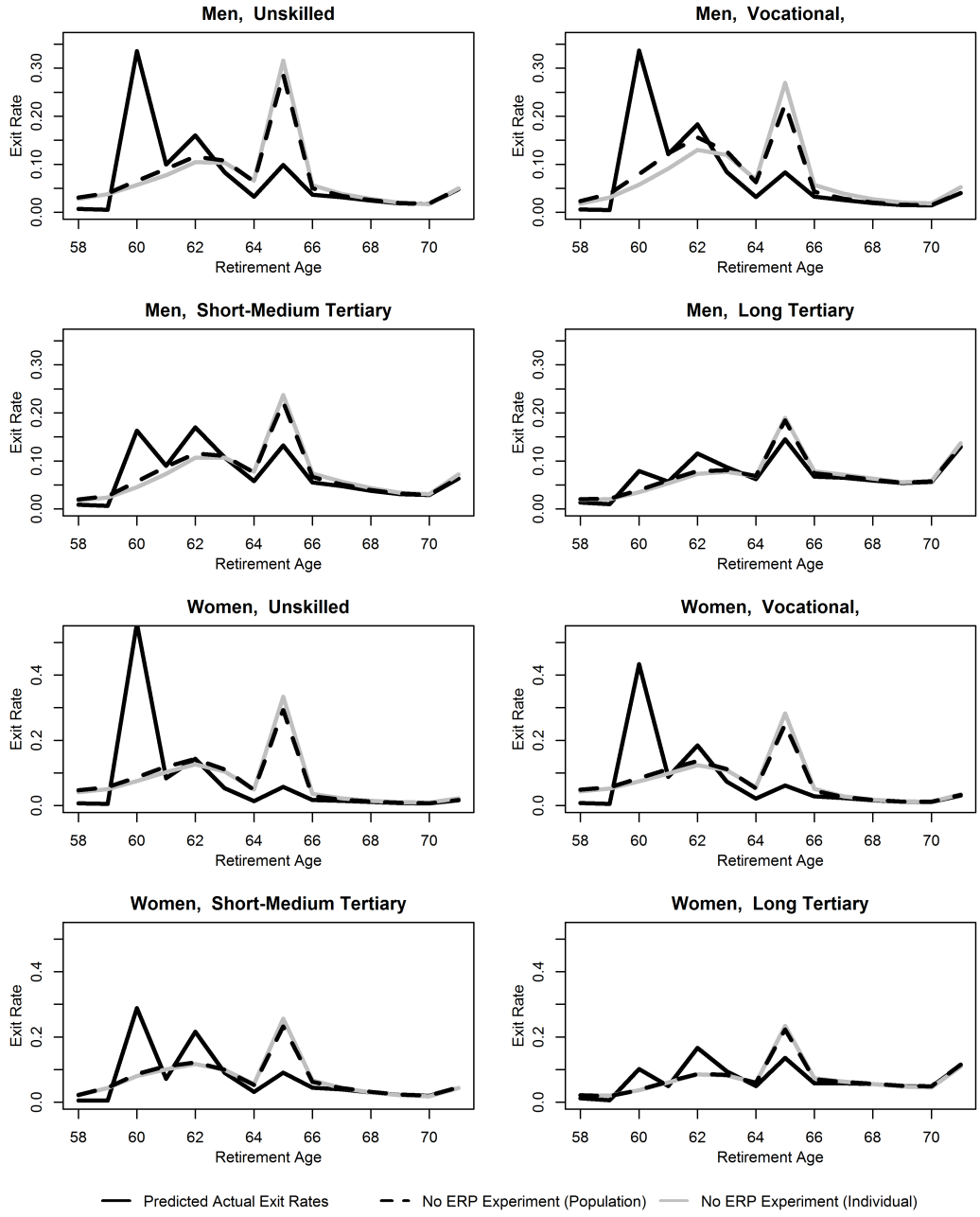
9.1 Baseline Experiment

First, we simulate a baseline experiment in which we abolish the early retirement program and thereby mimic the retirement setting of future cohorts. Whereas the ERP scheme DKr covers the majority of individuals, recent changes to the program which have made it significantly less attractive combined with an opt-out option have led to a drastic drop in individuals entitled to receive ERP benefits. In Appendix Figure 22, we showed that our model also fit the retirement ages of individuals who are not members of the ERP scheme - the fit was best when we used individual k distributions compared to the population distribution, but this distinction becomes less important when we consider the entire sample, and not just the selective group of individuals not entitled to ERP benefits.

Figure 10 shows the predicted exit rates for each retirement age when the Early Retirement Pension program is abolished, together with the in-sample model fit of the actual retirement rules of cohorts 1942-1952, where approximately 85% of the sample were entitled to receive ERP benefits starting at age 60. We see that the abolition of the ERP plan leads to a stark

increase in retirement ages according to our model. The peak in the exit rate at the ERP eligibility age (60) disappears, and the spike in the exit rate at the normal retirement age (65) drastically increases. The shift in exit rate spikes from 60 and 62 to 65 is strongest for those with lower education levels as expected since these groups were more inclined to retire on early retirement benefits. We also observe that the exit spike at age 62 (caused by the two-year rule) becomes a more smooth hump-shape in exit rates before the NRA. These are the individuals who decide to self-finance their retirement prior to when they become eligible for retirement benefits at age 65. The model predicts that 35-60% of seniors will choose to self-finance an earlier retirement, with largest shares for those with less education. While this share might seem very large, recall that we did observe a large proportion (approximately 40%) of those not eligible for ERP benefits retire prior to the NRA, see Figure 5. On average, our model predicts that individuals in the different gender- and education groups postpone their retirement by 0.5-1.5 years once the ERP scheme is abolished. The effect decreases when individuals hold large amounts of liquid and retirement wealth in an almost linear fashion, and the estimated effect is only 0.25-0.5 years for individuals with retirement savings equivalent to 9 or more years of pre-retirement income. This decline is illustrated in Appendix Figure 23, which also shows that the average effect of the reform is much larger for those with no or little education, and slightly larger for women compared to men.

Figure 10: Retirement Effect of Baseline Experiment: No ERP scheme



The solid lines depict the model's predicted exit rate distributions for the entire estimation sample consisting of cohorts 1942-1952 (these are similar to the model predictions shown in Figure 8, but include more cohorts). The dashed black lines plot our model's predicted retirement distribution in the baseline experiment, where the ERP scheme is abolished. The solid gray lines show the predicted retirement distribution modeled using the estimated individual-specific k-distributions.

9.2 Postponing the NRA

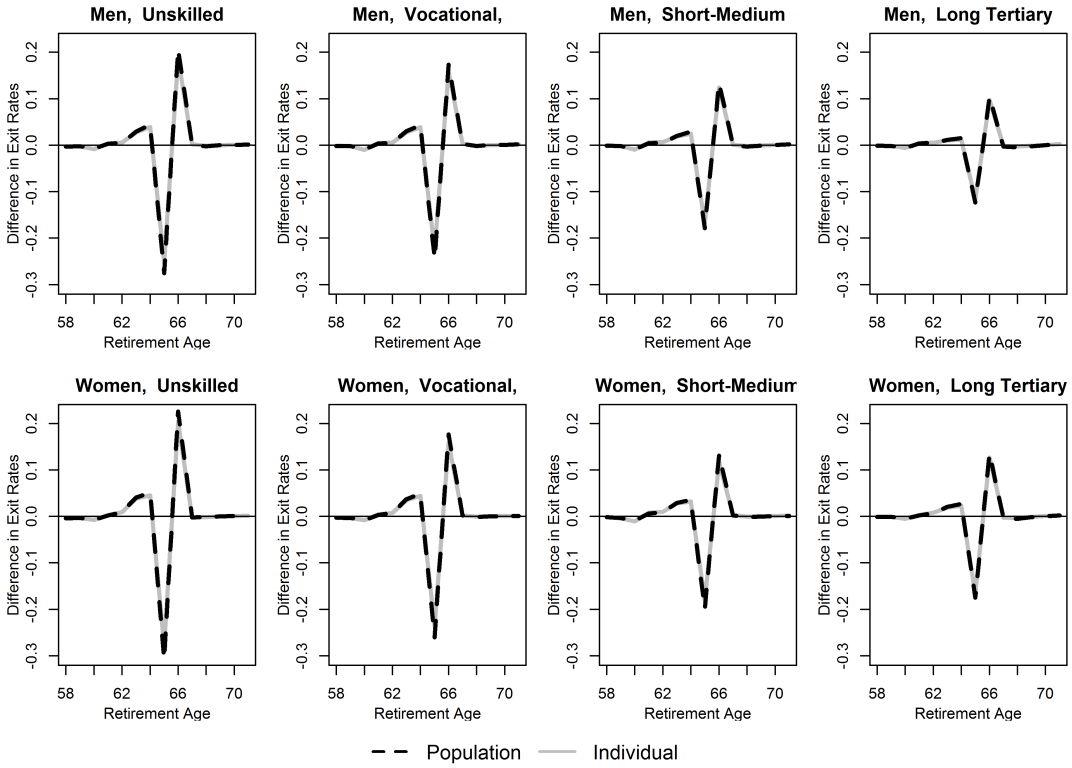
As described in Section 2, the Old Age Pension (OAP) eligibility age has been indexed to the life expectancy of future cohorts in Denmark. Most other countries, including the U.S. and the U.K., have also increased normal retirement ages (NRAs) to improve their fiscal sustainability. When the NRA is postponed, governments will have to pay retirement benefits to the retirees for fewer years, but more importantly, an increase of the NRA also induces individuals to work longer, and thereby the labor force (and tax base) is increased. Understanding how an increase of the NRA affects the labor supply decision of seniors is, therefore, of very high importance

An increase of the NRA corresponds to a decrease in the present value of discounted future OAP benefit payments, and does not impose strong financial incentives for individuals to respond to this change.[7] However, as we assume that the social norm/focal point dummies also shift from age 65 to 66, we expect that the increase of the NRA will have a large impact on retirement ages.

Figure 11 shows, for each gender- and education group born in 1942-1952, the expected change in exit rates when the NRA is raised from age 65 to 66 within the baseline experiment where there is no ERP scheme. The figure shows that an increase of the NRA leads to a stark drop in the exit rate at the old NRA, and a corresponding increase in the exit rates of the new NRA. If we compare the estimated changes to the exit rates at age 65 in the baseline experiment, we see that the NRA spike is almost entirely shifted from age 65 to 66. However, a small fraction of individuals who would have retired at the normal retirement age (65) before the increase of the NRA choose to retire even earlier after the increase. Postponing the retirement age has almost no effect for the exit rates prior to age 62, but exit rates at 63 and 64 increase slightly by a few percentage points. This increase is driven by those individuals who no longer prefer to postpone their retirement until they reach the normal retirement age to gain the social norm/focal point jump in utility, as they would now have to wait one year longer. These are predominantly individuals with high estimated k values. Consequently, the spike at the new OAP age, 66, is slightly smaller compared to when the NRA was 65. As expected, the increase of the NRA does not affect the exit rates at ages 66 and above.

[7] For individuals who are credit constrained, an increase of the NRA would impose strong financial incentives, but as our model assumptions imply that individuals can always borrow money against future income, this effect is not going to be captured by our model.

Figure 11: Retirement Effect of One Year Increase of the NRA



The solid lines depict the model's predicted exit rate distribution for the entire estimation sample consisting of cohorts 1942-1952. The dashed black lines plot our model's predicted retirement distribution in the baseline experiment, where the ERP scheme is abolished. The solid gray lines show the predicted retirement distribution when we use the estimated individual-specific k-distributions.

As such, an increase of the NRA age from 65 to 66 make more individuals retire at the new normal retirement age 66, which increases the average retirement age. In fact, close to everyone who retired at the old NRA at 65 also retires at the new NRA at 66. However, this effect is offset by the increase in the number of individuals who no longer wants to wait to retire at the normal retirement age, and instead decide to finance an earlier retirement, resulting in increased exit rates at ages 63 and 64. Overall, an increase of the NRA from 65 to 66 increases the expected retirement ages of the groups by 0.06-0.15 years, with largest effects for those with low education, and larger effects for men.[8] This number should be interpreted in the context of the baseline experiment, which predicts that 35-60% of seniors will chose to self-finance an earlier retirement once the ERP scheme is abolished.

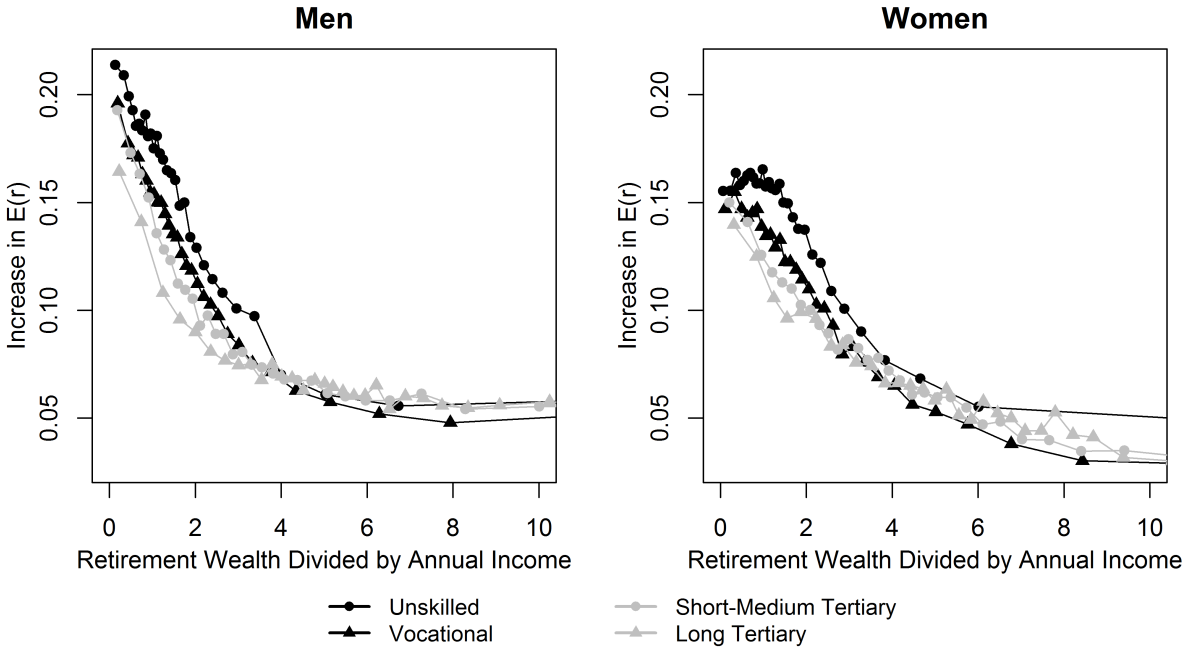
As self-financing an earlier retirement age is costly, one would expect individuals with

[8]Whereas the predicted changes in exit rate distributions are slightly different in an corresponding model without social norm dummies at the NRA (changes are more "smooth" compared to Figure 11), the aggregate effect as measured by increase in expected retirement age is very similar, also across different levels of private retirement savings.

larger private pension wealth to be more likely to do so. Figure 12 shows how the increase in the expected retirement age, following a one-year increase of the NRA, vary for different levels of retirement wealth. Here we see a decline in the effect of postponing the retirement age of more than 75% when we compare the low retirement wealth individuals to those with large retirement savings. The decline stagnates at retirement savings levels equivalent to approximately four years of annual pre-retirement earnings. The figure also suggests that, conditional on the level of retirement savings, the expected effect of the increase is more significant for those with higher levels of education and slightly larger for men than women.

Of course, seniors can also finance an earlier retirement age using their liquid/non-retirement savings. Appendix Figure 24 shows that the slopes equivalent to those depicted in Figure 12 are even steeper for individuals who hold zero or negative liquid wealth. For individuals with zero liquid and retirement wealth, our model predicts that an increase of the NRA from 65 to 66 will result in an increase in expected retirement age by 0.25 years for men and 0.2 for women. For individuals holding large non-retirement wealth, exceeding 1.5 million DKr, the slope becomes flat - this is expected, given that these individuals don't have to rely on retirement savings to finance their early retirement.

Figure 12: Effect of one year increase in NRA by accumulated pension wealth



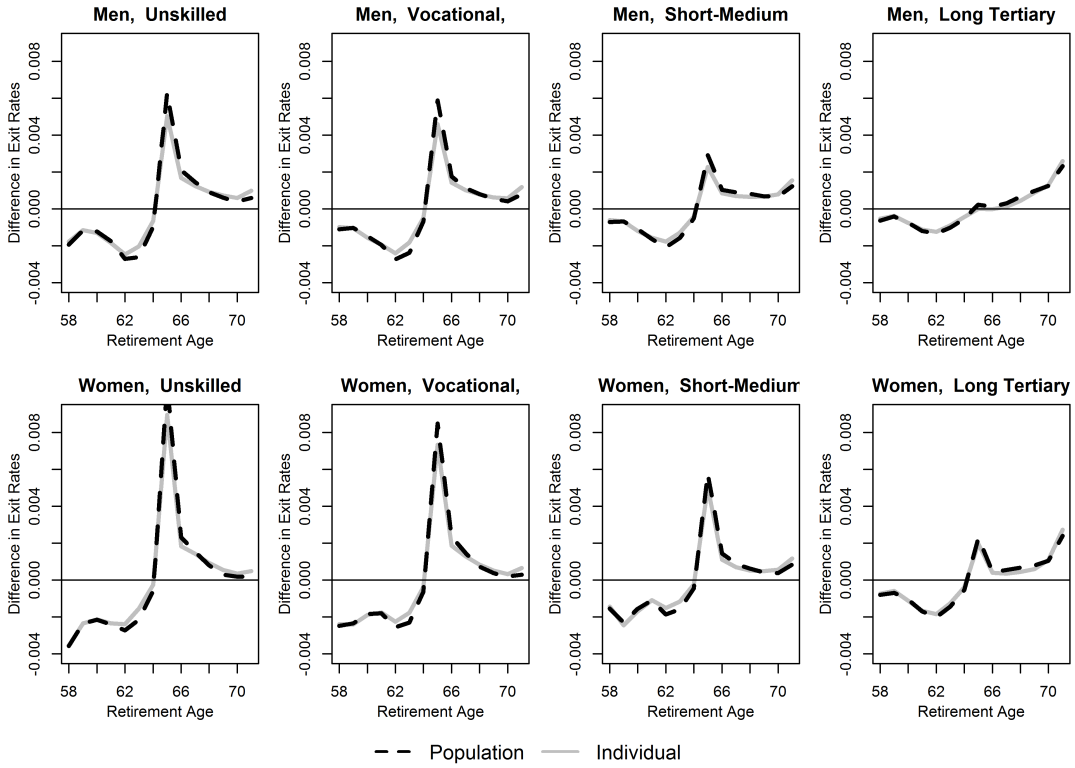
For each gender- and education specific group, the figure shows how our model predicts that an increase of the NRA from age 65 to 66 would have affected the expected retirement age of cohorts 1942-1952, given that the early retirement scheme was abolished. The effect is measured by years, as the change in expected retirement rates.

9.3 Reduced OAP benefit rates

Another strategy that countries use to improve the fiscal challenges of increased life expectancy is to decrease old age pension benefit rates. Effectively, such a policy is equivalent to a negative wealth shock to individuals similar to the one year increase of the NRA. Figure 15 plots, for each gender- and education group born in 1942-1952, the expected exit rates in the baseline experiment (solid gray line) with no ERP scheme together with the exit rates under an extended experiment (dashed black line) in which OAP benefits are reduced by five percent. As opposed to a one-year increase of the NRA, this experiment does not involve a change in the social norms/focal point of the retirement decision. This explains why the effect of the two types of reform differs. When OAP benefits are reduced by 5 percent, our model suggests that fewer individuals will retire prior to the NRA, and more individuals will retire after. The predicted changes in exit rates are small, but are distributed across all retirement ages - as there is a consistent drop at the retirement ages 58-65 and a consistent increase at ages 66-72, the accumulated effect is quite significant. Overall, the reduction in OAP payments increases the expected retirement ages of the groups by 0.05-0.09 years, with largest effect for those with low education, and larger effects for women compared to men.[9] Compared to an increase in the NRA as shown in the previous experiment, a 5 percent reduction does not lead to a stark drop in exit rates at age 65. This is because a reduction in OAP benefits does not to affect the credit-constrained individuals to the same extent as an increase in the NRA.

[9]Results are similar in a model with no NRA social norm dummy, with the only difference being a slightly smaller spike at age 65 and a larger estimated increase in exit rates at age 66

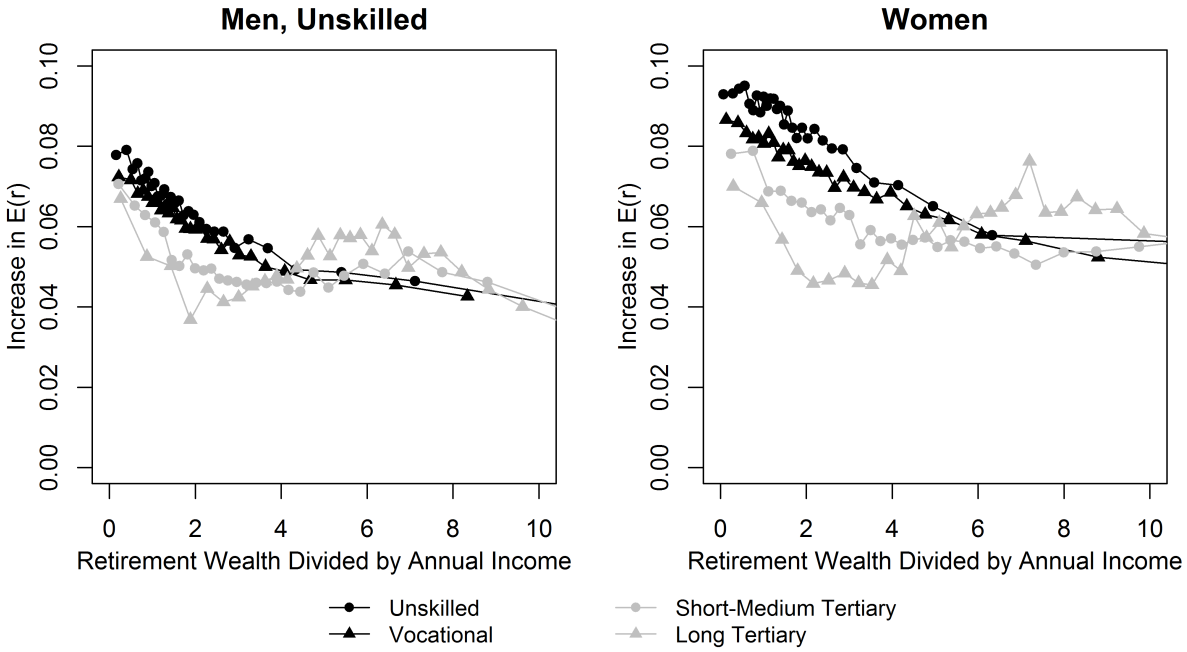
Figure 13: Retirement Effect of a five percent reduction of OAP-benefit rates



The figures plot, for each gender- and education specific group, the differences in predicted exit rates between the baseline experiment where the ERP scheme is abolished and an experiment in which the OAP benefit rates are reduced by 5%. The solid gray lines plot the change when we use individual-specific k-distributions, and the dashed black lines show the difference when the predictions are based on the population distribution of k.

Similar to the previous experiments, the estimated increase in the expected retirement age declines with increased levels of pension wealth, as pictured in figure 14. This means that individuals with larger retirement savings respond less to the policy change compared to those with low levels of accumulated retirement wealth. However, the decline in the policy response for increasing levels of retirement savings is much less steep when OAP benefits are reduced compared to when the NRA is increased. When the benefit rates are decreased by 5 percent, the model predicts that individuals with low levels of retirement savings will increase their retirement age by roughly 0.07 years for men and 0.09 years for women. The effect declines slightly to 0.05 years for men and 0.06 years for women who hold private retirement savings equivalent to four or more years of income at age 58. Thus, the two reforms (increasing the NRA by one year and decreasing OAP benefits by 5%) have roughly similar effects on the average retirement age for high-wealth individuals.

Figure 14: Retirement Effect of reduced OAP benefit rates by Pension Wealth



For each gender- and education specific group, the figure shows how our model predicts that an increase of the NRA from age 65 to 66 would have affected the expected retirement age of cohorts 1942-1952, given that the early retirement scheme was abolished. The effect is measured by years, as the change in expected retirement rates.

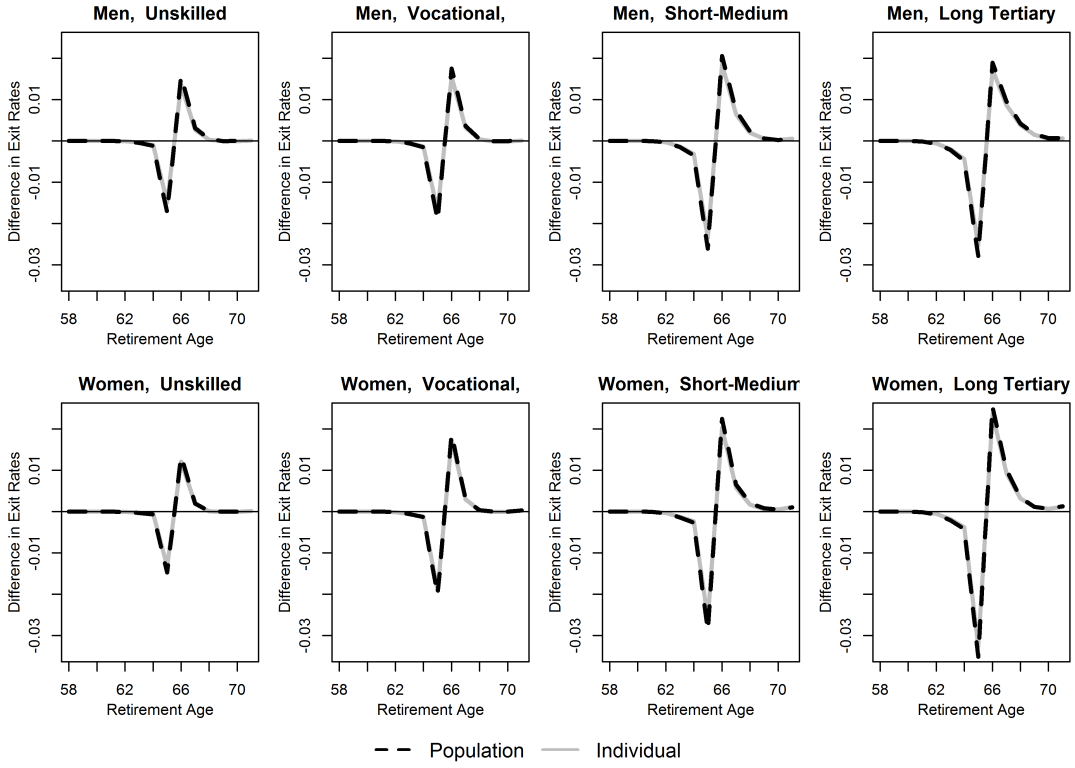
9.4 Reduced OAP means testing

As described in Section 4.2 (and in particular Figure 5), historical data suggests that the opportunity to avoid means testing of private pension wealth induced ERP eligible persons with pension wealth to postpone their retirement. In this subsection, we investigate the effect of a third policy, also targeted increased retirement ages, which is similar to the two-year rule of the Early Retirement Program. The suggested policy reduce the means testing of OAP benefits with respect to private pension wealth if retirement is postponed until at least age 66. Figure 15 shows, for each gender- and education group, the expected exit rates in the baseline experiment (solid gray line) with no ERP scheme together with the exit rates under an extended experiment (dashed black line) in which means testing of the OAP supplement with respect to private pension payments is avoided for three years if retirement is postponed to age 66 years or later (that is one later than the statutory NRA).[10] This experiment targets those who hold larger amounts of pension wealth, which mainly applies to workers with short-medium tertiary and long tertiary education. For these groups, the spike in retirement at

[10]Private pension payments above ≈70.000 DKr are deducted from the pension supplement with a rate of 31 percent for singles and 16 percent for couples

age 65 is reduced a little and moved to 66 years. Overall, the average expected retirement age is increased by 0.05-0.10 years. For unskilled and skilled workers, the effect is very small, because they have little pension wealth.[11] The average expected retirement age increases by roughly 0.01 years.[12]

Figure 15: Retirement Effect of reduced OAP means testing of private pension payments



The figures plot, for each gender- and education specific group, the differences in predicted exit rates between the baseline experiment where the ERP scheme is abolished and an experiment in which the ERP scheme is abolished and the means testing in OAP benefits by private pension payments is reduced when retirement is postponed to after the NRA. The solid gray lines plot the change when we use individual-specific k-distributions, and the dashed black lines show the difference when the predictions are based on the population distribution of k.

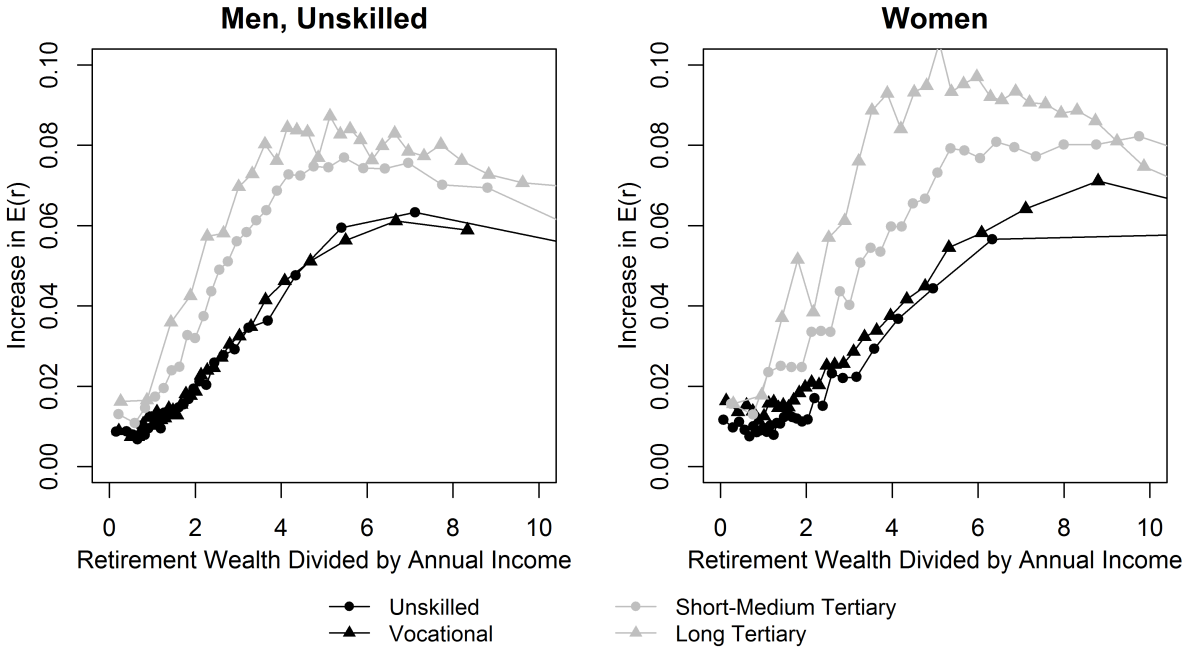
We would expect that individuals with large private pension wealth are more likely to respond to the experiment and postpone their retirement until age 66 or later, in order to avoid a means testing in their OAP benefits. As expected, Figure 16 shows an increasing trend in the predicted policy response, meaning that individuals holding large retirement savings respond stronger to the reform, compared those with less accumulated retirement

[11] Mandatory pension savings plans was introduced for most groups of unskilled and skilled workers in the 1990s with contribution rates below 5 percent until 1998.

[12] Results are similar in a model with no OAP social norm dummy.

savings. The predicted policy response is close to 0% for those holding low or no wealth, and increases to 0.05-0.09 for those holding retirement savings equivalent to six or more years of retirement savings. The effects of pension wealth are larger for workers with medium and tertiary educations.

Figure 16: Retirement Effect of Means Testing Relief



For each gender- and education specific group, the figure shows how our model predicts that an increase of the NRA from age 65 to 66 would have affected the expected retirement age of cohorts 1942-1952, given that the early retirement scheme was abolished. The effect is measured by years, as the change in expected retirement rates.

10 A simpler model without leisure preference heterogeneity

The main contribution of our proposed model is that we allow for different individuals to have different preferences for leisure. If we instead assumed that the leisure preference parameter k was a homogeneous parameter, our model would be very similar to a simplified dynamic programming model, where agents have perfect foresight in all respects except when they die. To find out whether the estimation of heterogeneous leisure preference parameters was worth the trouble, we estimate and test a similar model with a homogeneous k, and show how the model fit and experiment results differ compared to when we allow for leisure preference heterogeneity. The results are shown in Appendix J. Table 4 shows the estimated

parameters in the less flexible model, where the estimated point estimates of k range from 0.9 to 1.3. Compared to the model with leisure preference heterogeneity, agents discount future consumption higher (β coefficients are between 0.8 and 0.86), are more risk averse (ρ between 1.6 and 2) and face lower attrition rates (α_0 between 0.005-0.009). As expected, the less flexibility of the homogeneous k specification of the model causes the log likelihood value to decrease from an average of -1.67 per person to -1.80 per person.

Figure 25 shows the model fit of the simpler model. As expected, the simpler model does not fit the data as well, and it predicts an almost flat and steady decrease in exit rates following age 60, however with a spike in exit rates at the NRA 65, caused by the social norm dummy. As such, the model does not seem to capture the spikes in exit rates at age 62, caused by the two-year of the ERP scheme. When we run the external validation test on the simple model, where we use the estimated model to predict the retirement decision of individuals born in the second half of 1954, facing an early retirement age of 61 and a statutory old age pension age of 66 (compared to 60 and 65 in the estimation sample), the model under-predicts the exit rate at age 61 by 30-50% for the different gender- and education groups. Consequently, we should not have much faith in the predictions of this simpler model.

For illustrative purposes, we also conduct the baseline experiment in which the ERP scheme is abolished for the simpler model without leisure preference heterogeneity, and show how the simpler model predicts that a one year increase of the NRA affects retirement ages. Figure 27 shows the actual predicted exit rates together with the predicted exit rates in the baseline experiment where the ERP scheme is abolished. The exit rates predicted by the simple model in the baseline experiment are almost flat across the ages 58-71, with a small curvature. Due to the social norm dummy, also this model predicts a large spike in the exit rates at the NRA, however smaller compared to those predicted by the model with leisure preference heterogeneity. The overall predicted increase in expected retirement ages following an abolition of the ERP scheme is slightly larger in the simpler model (0.5-2 years) compared to the model with leisure preference heterogeneity (0.5-1.5 years).

We also compare the predicted response to an increase of the NRA by the two models. The predicted changes in exit rates after an increase of the NRA from age 65 to 66 are shown in Figure 28. Similarly, the simpler model also predicts that the

in the exit rate shifts from the old to the new retirement age. The difference between the model predictions is that the simpler model does not predict that anyone decides to retire earlier, at ages 63 and 64, once the NRA is increased. This is mainly due to the fact that the simpler model have smaller estimated attrition parameters, α, compared to the model with leisure preference parameters. As such, the simpler model suggest that the cost of remaining in the labor force for one more year is less costly, and as a result, the majority of those who retired at the old NRA will shift to the new NRA after the increase. The model with

preference heterogeneity estimates that attrition is much more important, with α parameters which are close to twice as large compared to those estimated by the simple model. The model with leisure preference heterogeneity is, therefore, suggesting that it is costly for individuals to retire late, which is offset by a large share of individuals having low values of k, making it costly to retire early as well. When individuals have different leisure preferences, they also respond differently to the policy changes. Whereas individuals with low leisure preferences are inclined to retire at the new NRA, individuals with large preferences for leisure are more inclined to choose to retire even earlier than before the increase in NRA, at ages 64 and 63.

Even though the simpler model predicts that a smaller share of individuals will respond to an increase of the NRA, the expected increase in average retirement age is still almost twice as large, as the model does not predict that some individuals will respond to the reform by decreasing their retirement age. Whereas the model with leisure preference heterogeneity predicted an overall increase in retirement ages of 0.06-0.15 years, with largest effects for those with low education, and larger effects for men, the simpler model predicts corresponding increases of the order 0.16-0.35 years.

Figure 29 shows how the simple model's predicted increase in expected retirement ages, following a one-year increase of the NRA, vary for different levels of retirement wealth. We see a drastic decline in the effect of postponing the retirement age of roughly 50% from 0.3 years to 0.15 years when we compare the low wealth individuals to those with larger retirement wealth. The decline stagnates at savings levels of around four years of annual income prior to retirement. As such, the decrease of roughly 50% for large vs. small retirement savings predicted by the simple model is lower compared to the decrease predicted by the model with preference heterogeneity, where the effect decreased by roughly 75%.

11 Conclusion

We propose a novel structural retirement model of senior worker's retirement decision, with heterogeneous leisure preferences, attrition, and improved health across generations. We estimate the model using high-quality Danish register data from 1996-2016, where we consider the retirement decisions of birth cohorts 1942-1954. We apply a non-parametric estimation technique to measure the heterogeneity in leisure preferences, and our estimates suggest that leisure preferences are widely distributed among the population. We find that the model fits the data well with reasonable parameter values. More importantly, we also find that our model provides decent predictions of an observed retirement response to an increase in the statutory retirement age in an external validation setting. Compared to a similar model without heterogeneous leisure preferences, we find that our proposed model performs significantly better - both in terms of in-sample fit and external validation of the model.

We use our model to study the extent to which the increase in private pension wealth -

providing individuals with more flexibility to choose their retirement age - impacts the effect of retirement reforms designed to increase the labor supply of seniors. We conduct several counter-factual experiments on the cohorts 1942-1952 which explore what their retirement decision would have been if different rules applied. First, we simulate a baseline experiment in which we abolish the ERP scheme. As the ERP scheme is being phased out, the baseline experiment mimics the retirement decision of future generations. We then contrast the retirement decisions in our baseline experiment to three additional experiments.

The first experiment increases the normal retirement age by one year from age 65 to 66. We find that individuals with zero private retirement savings delay their retirement with roughly 0.2 years for men and 0.15 years for women. For individuals with private retirement wealth equivalent to four or more years of pre-retirement earnings, the effect is reduced by roughly 75%. The second experiment reduces the old age pension benefits by five percent, and for this experiment, we estimate a decline in the expected retirement age of roughly 0.08-0.09 years for those with no retirement savings and 0.05-0.06 years for those with large retirement savings. As such, the effect of a decrease in retirement benefits is much more stable across different levels of retirement wealth as compared to an increase in the NRA. This is because a reduction in benefits is less consequential to credit-constrained individuals compared to a rise in the retirement age. We also consider a third policy experiment which introduces a means testing relief in the OAP which is similar to the two-year rule in the ERP-scheme: individuals can avoid a means testing of their old age pension with respect to their private pension income for three years if they retire one or more years after the statutory NRA. While low-wealth individuals are almost unaffected by this experiment, the expected retirement age increases by 0.06-0.10 years for individuals with large retirement savings. As such, this experiment - similar to the two-year rule of the retirement age - reverses the negative effect of pension wealth on retirement ages.

To summarize, our experiments find that the size of individual retirement savings can have important implications for the effect of different retirement reforms. For an increase in the NRA, we find a particularly large and negative effect of private retirement wealth on the reform's ability to increase labor supply. It is possible, however, to reverse this effect, e.g. in a reform which mimics the two-year rule of the ERP scheme.

References

Søren Arnberg and Mikkel Barslund. The crowding-out effect of mandatory labour market pension schemes on private savings: Evidence from Renters in Denmark. 2014.

Søren Arnberg and Peter Stephensen. Økonomiske incitamenter, nedslidning og tilbagetrækning - Semi-parametrisk estimation af heterogenitet og aldersbetingede ønsker om tilbagetrækning. *DREAM working paper*, 2015.

Luc Behaghel and David M Blau. Framing social security reform: Behavioral responses to changes in the full retirement age. *American Economic Journal: Economic Policy*, 4(4): 41–67, 2012.

Paul Bingley, Nabanita Gupta, and Peder Pedersen. The impact of incentives on retirement in Denmark. *Review of Economic Studies (2005) 723, pp.395-427*, 2004.

David M Blau and Donna B Gilleskie. The role of retiree health insurance in the employment behavior of older men. Technical report, National Bureau of Economic Research, 2003.

Richard Blundell, Eric French, and Gemma Tetlow. Retirement incentives and labor supply. In *Handbook of the economics of population aging*, volume 1, pages 457–566. Elsevier, 2016.

Raj Chetty. A new method of estimating risk aversion. *American Economic Review*, 96(5): 1821–1834, 2006.

Raj Chetty, John N Friedman, Søren Leth-Petersen, Torben Heien Nielsen, and Tore Olsen. Active vs. passive decisions and crowd-out in retirement savings accounts: Evidence from Denmark. *The Quarterly Journal of Economics*, 129(3):1141–1219, 2014.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Jeremy T Fox, Kyoo il Kim, and Chenyu Yang. A simple nonparametric approach to estimating the distribution of random coefficients in structural models. *Journal of Econometrics*, 195(2):236–254, 2016.

Eric French. The effects of health, wealth, and wages on labour supply and retirement behaviour. *Review of Economic Studies (2005) 723, pp.395-427*, 2005.

Eric French and John Bailey Jones. The effects of health insurance and self-insurance on retirement behavior. *Econometrica 79 (3)*, pages 693–732, 2011.

Alan L Gustman and Thomas L Steinmeier. The social security early entitlement age in a structural model of retirement and wealth. *Journal of public Economics*, 89(2-3):441–463, 2005.

Anne-Line Koch Helsø. Modelling retirement with heterogeneity - a semi-parametric estimation with push- and pull effects. Master's thesis, University of Copenhagen, 2015.

Thomas Jørgensen. *Leisure Complementarities in Retirement*. PhD thesis, University of Copenhagen, 2014.

Rafael Lalive and Stefan Staubli. How does raising women's full retirement age affect labor supply, income, and mortality. *Evidence from Switzerland*, 2014.

Giovanni Mastrobuoni. Labor supply effects of the recent social security benefit cuts: Empirical estimates using cohort discontinuities. *Journal of public Economics*, 93(11-12): 1224–1233, 2009.

OECD. *OECD PENSIONS OUTLOOK 2016*. ORGANIZATION FOR ECONOMIC, 2016.

John Rust and Christopher Phelan. How social security and medicare affect retirement in a world with incomplete markets. *Econometrica 65(4), pp. 781-831*, 1997.

Arthur Seibold. Statutory ages as reference points for retirement: Evidence from germany. 2017.

James Stock and David Wise. Pensions, the option value of work and retirement. *Econometrica, Vol. 58, Issue 5 pp.1151-1180*, 1990.

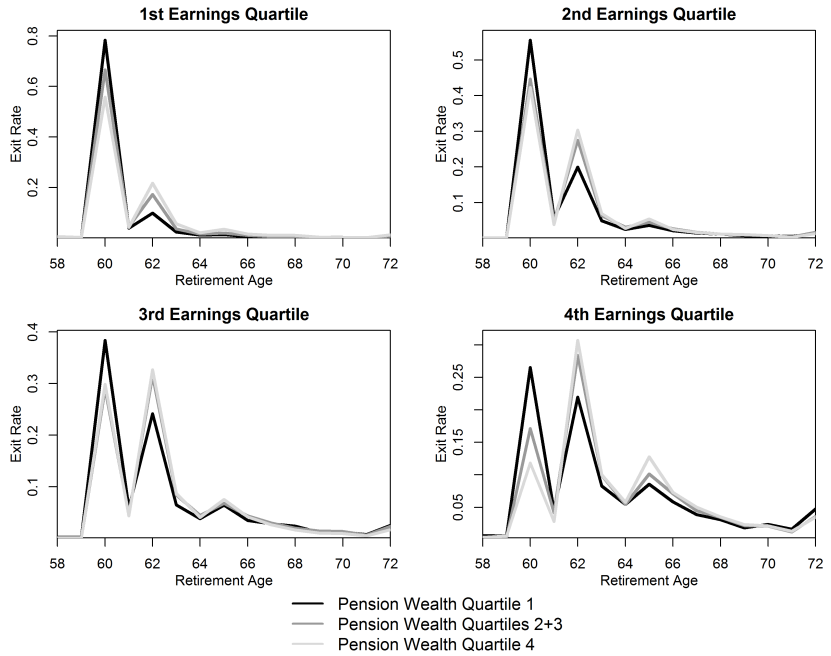
Kenneth Train. A recursive estimator for random coefficient models. *Working Paper, University of California*, 2007.

Kenneth Train. Em algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modeling, 1(1), pp. 40-69*, 2008.

A Appendices

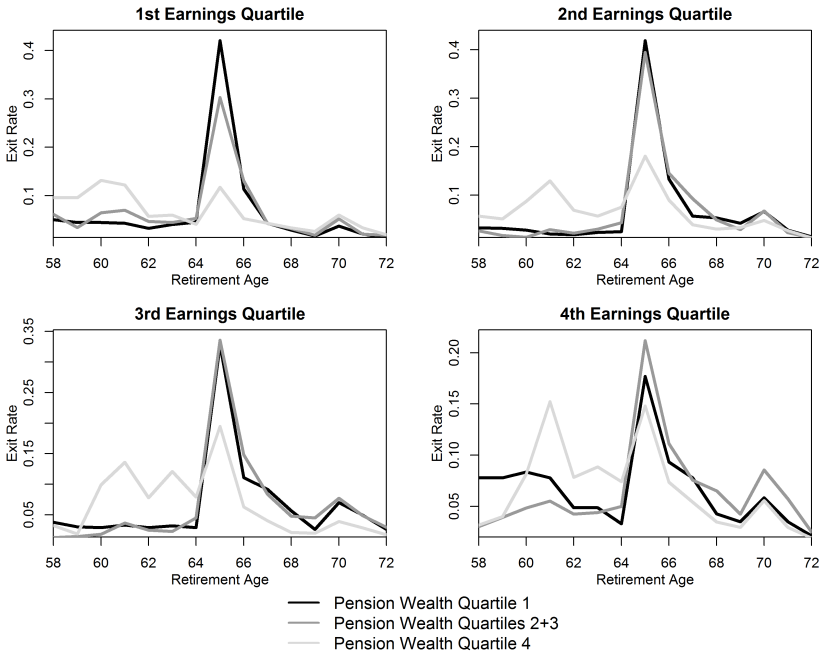
A Exit rates for ERP- and non ERP eligible, split on retirement wealth and earnings

Figure 17: Retirement by Pension Wealth and Earnings Quartile, ERP-eligible



The four figures plot the exit rates (share who exit the labor force) at ages 58-72 for individuals born in 1942-1944, who are included in our main sample (see Table 1), and who are eligible to retire on Early Retirement Pension benefits. The samples are split in pre-retirement earnings quartiles, and again in pension wealth quartiles, where quartile 2 and 3 are grouped together.

Figure 18: Retirement by Pension Wealth and Earnings Quartile, not ERP-eligible



The four figures plot the exit rates (share who exit the labor force) at ages 58-72 for individuals born in 1942-1944, who are included in our main sample (see Table 1), and who are not eligible to retire on Early Retirement Pension benefits. The samples are split in pre-retirement earnings quartiles, and again in pension wealth quartiles, where quartile 2 and 3 are grouped together.

B Closed-form solution for optimal consumption

We would like to derive the indirect utility function of the consumer for a *given* retirement age r. The consumer is faced with the problem defined by (1), (2) and (3) repeated below:

$$U(r) = V(r) + \epsilon_r \tag{13}$$

$$V(r) = \sum_{a=58}^{120} \frac{(\gamma_a(r)c_a)^{1-\rho}}{1-\rho} D_a \tag{14}$$

$$W_a = (1 + (1-\tau)\, i_a) W_{a-1} + y_a\,(r) - c_a \tag{15}$$

Given a retirement age r, ϵ_r is known by the consumer. According to (13) the consumer can then maximize $U\,(r)$ just by maximizing $V\,(r)$. We will prove that maximizing $V\,(r)$ given by (14) under the budget restriction, (15) yields the indirect utility function:

$$V(r) = \frac{\left(\frac{W_{57}+H(r)}{P(r)}\right)^{1-\rho}}{1-\rho}$$

where $H(r)$ and $P(r)$ are given by

$$H\,(r) = \sum_{a=58}^{120} y_a\,(r)\, R_a$$

$$P\,(r) = \left[\sum_{a=58}^{120} \gamma_a(r)^{\frac{1-\rho}{\rho}} D_a^{\frac{1}{\rho}} R_a^{\frac{\rho-1}{\rho}}\right]^{\frac{\rho}{\rho-1}}$$

Here R_a is defined as

$$R_a = \prod_{s=58}^{a} \frac{1}{1+(1-\tau)\, i_a}$$

Proof: We will start by demonstrating that we can transform the utility function (14) to a CES function. Define $E = 1/\rho$ and the transformation $T(V) = (\frac{1-\rho}{\phi} V)^{\frac{1}{1-\rho}}$. We have that $T'(V) = \frac{1}{\phi}\left(\frac{1-\rho}{\phi} V\right)^{\frac{\rho}{1-\rho}}$ and that $T'(V) > 0$ iff. $(1-\rho)V > 0$. Observe that from (14):

$$\hat{V}(r) \equiv T\,(V(r)) = \left[\sum_{a=58}^{120} (\gamma_a(r)c_a)^{\frac{E-1}{E}} D_a\right]^{\frac{E}{E-1}} \tag{16}$$

This is a CES utility function with elasticity of substitution E. This transformation is only OK for $T'(V) > 0$. But this is always the case: assume $1-\rho < 0$. Then form (14) $V < 0$, such that $(1-\rho)\, V > 0$. Assume $1-\rho > 0$. Then from (14) $V > 0$, such that $(1-\rho)\, V > 0$. Due to continuity, it will also be the case for $1-\rho = 0$.

The flow condition given in the budget restriction in (15) can be re-written as a stock condition:

$$\sum_{a=1}^{A} c_a R_a = W_{57} + H(r) \tag{17}$$

where R_a and $H(r)$ are defined as above. Maximizing 16 given 17 we get (as a standard result for CES-functions) that

$$c_a = \left(\gamma_a(r)^{\frac{E-1}{E}} D_a\right)^E \left(\frac{R_a}{P\left(r\right)}\right)^{-E} \frac{W_{57} + H(r)}{P(r)} \tag{18}$$

Where

$$\begin{aligned} P(r) &= \left[\sum_{a=58}^{120} \left(\gamma_a(r)^{\frac{E-1}{E}} D_a\right)^E R_a^{1-E}\right]^{\frac{1}{1-E}} \\ &= \left[\sum_{a=58}^{120} \gamma_a(r)^{\frac{1-\rho}{\rho}} D_a^{\frac{1}{\rho}} R_a^{\frac{\rho-1}{\rho}}\right]^{\frac{\rho}{\rho-1}} \end{aligned}$$

Inserting (18) in (16) we arrive at the indirect utility function

$$\hat{V}(r) = \frac{W_{57} + H(r)}{P(r)}$$

We then have that

$$V(r) = T^{-1}(\hat{V}(r)) = \frac{\hat{V}(r)^{1-\rho}}{1-\rho} = \frac{\left(\frac{W_{57}+H(r)}{P(r)}\right)^{1-\rho}}{1-\rho}.$$

C Means-testing of the Early Retirement Pension (ERP)

Table 3: Annual Means Testing of ERP based on individual pension savings

		Retirement Before 62		**Retirement After 62**	
		Employer Adm.	*Employee Adm.*	*Employer Adm.*	*Employee Adm.*
Life Annuities	*NP*	60% of (80% of RP - BA)	60% of (80% of RP - BA)	0	0
	P	50% of AP	60% of (80% of RP - BA)	58% of AP	0
Term Pension	*NP*	60% of (5% of RD - BA)	60% of (5% of RD - BA)	0	0
	P	50% of AP	60% of (80% of RP - BA)	58% of AP	0
Capital Pension		60% of (5% of RD - BA)	60% of (5% of RD - BA)	0	0

Note: **NP** = No Payments Made, **P** = With Payments, **AP** = Actual Payment (annual), **RP** = Reported Payment (annual), **RD** = Reported Deposited amount of total savings, **BA** = Basic Allowance (11.500 DKr in 2004) - can only be used once.

D Censored Observations

Recall our model assumptions about the retirement decision: the retirement age is decided upon at the beginning of age $a = 58$ where the agent has perfect foresight wrt. all future income. As long as we observe individual j at age 57, we're able to to compute $x_j = \left(W_{57}^j, H\left(58\right)^j, .., H\left(72\right)^j\right)$ (see Section 5). Given the financial information available at age 57, we can simulate all future income streams for ages $a \in \{58, ..., 120\}$ given retirement age $r^j \in \{58, ..., 71\}$. As such, we can compute the choice set of all individuals despite the fact that we don't observe their chosen retirement age. Some individuals die or emigrate before retiring. For others, the data ends before a retirement age is observed - this could be the case for the youngest cohorts included in the analysis. However, we can still use the information that the individuals did not retire during the years where we observe them.

The probability of not retiring prior to the last observed age a_d is given by:

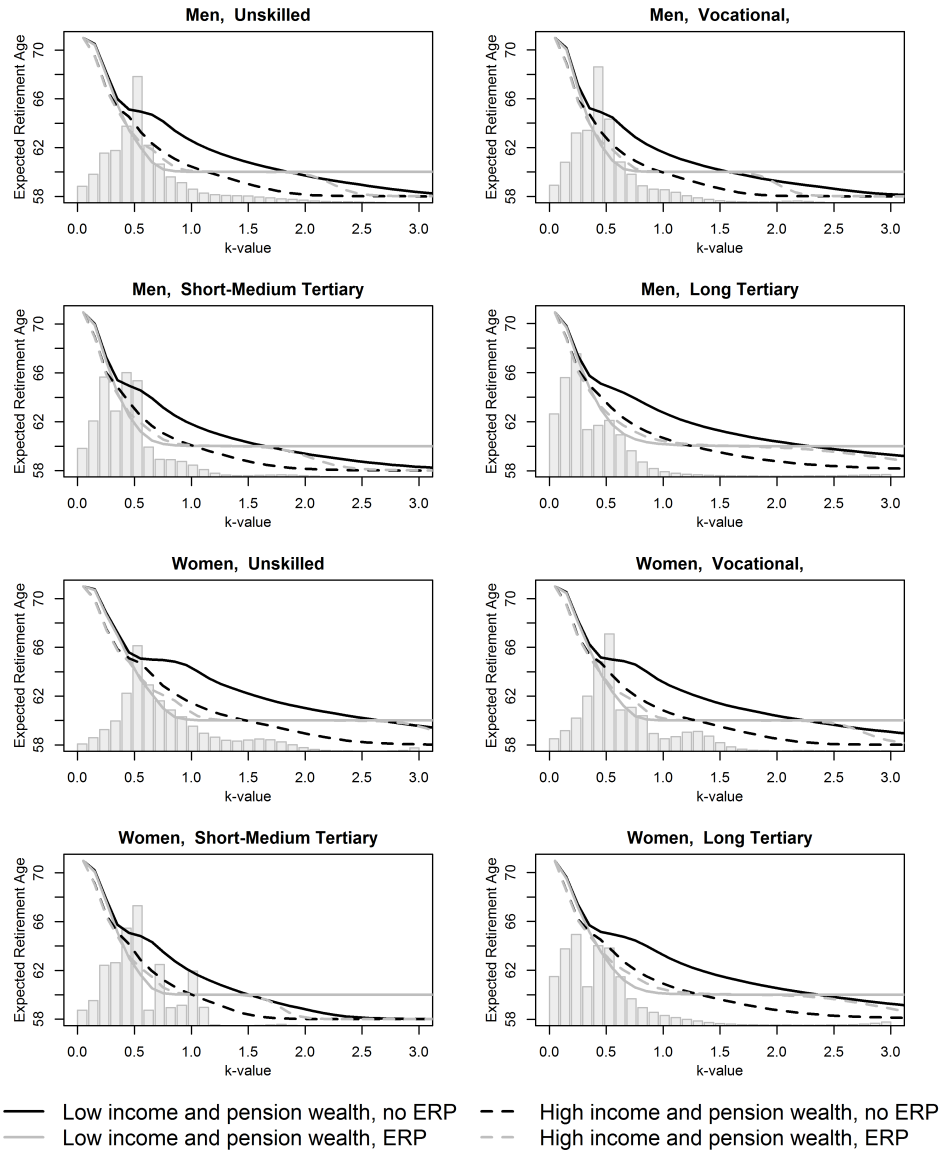
$$Prob_d\left(a_d|\Theta, x^j, k'\right) = \left(1 - \sum_{r=58}^{a_d} Prob\left(r|\Theta, x^j, k'\right)\right)$$

Consequentially, the probability of an observation depends on whether or not the retirement age is observed, $a_d \leq r$. As such, the probability that individual j either retire at age r^jor don't retire prior to age a_d^j is given by

$$Prob\left(r^j|\Theta, x^j, k', a_d^j\right) = \begin{cases} \frac{exp\left(\overline{V}(r^j|\Theta, x^j, k')\right)}{\sum_{r'=58}^{72} exp\left(\overline{V}(r'|\Theta, x^j, k')\right)} & \text{if } a_d^j > r^j \\ \left(1 - \sum_{a=a1}^{a_d} \frac{exp\left(\overline{V}(r|\Theta, x^j, k')\right)}{\sum_{r'=58}^{72} exp\left(\overline{V}(r'|\Theta, x^j, k')\right)}\right) & \text{if } a_d^j \leq r^j \end{cases}$$

E The effect of the leisure preference parameter k on expected retirement age

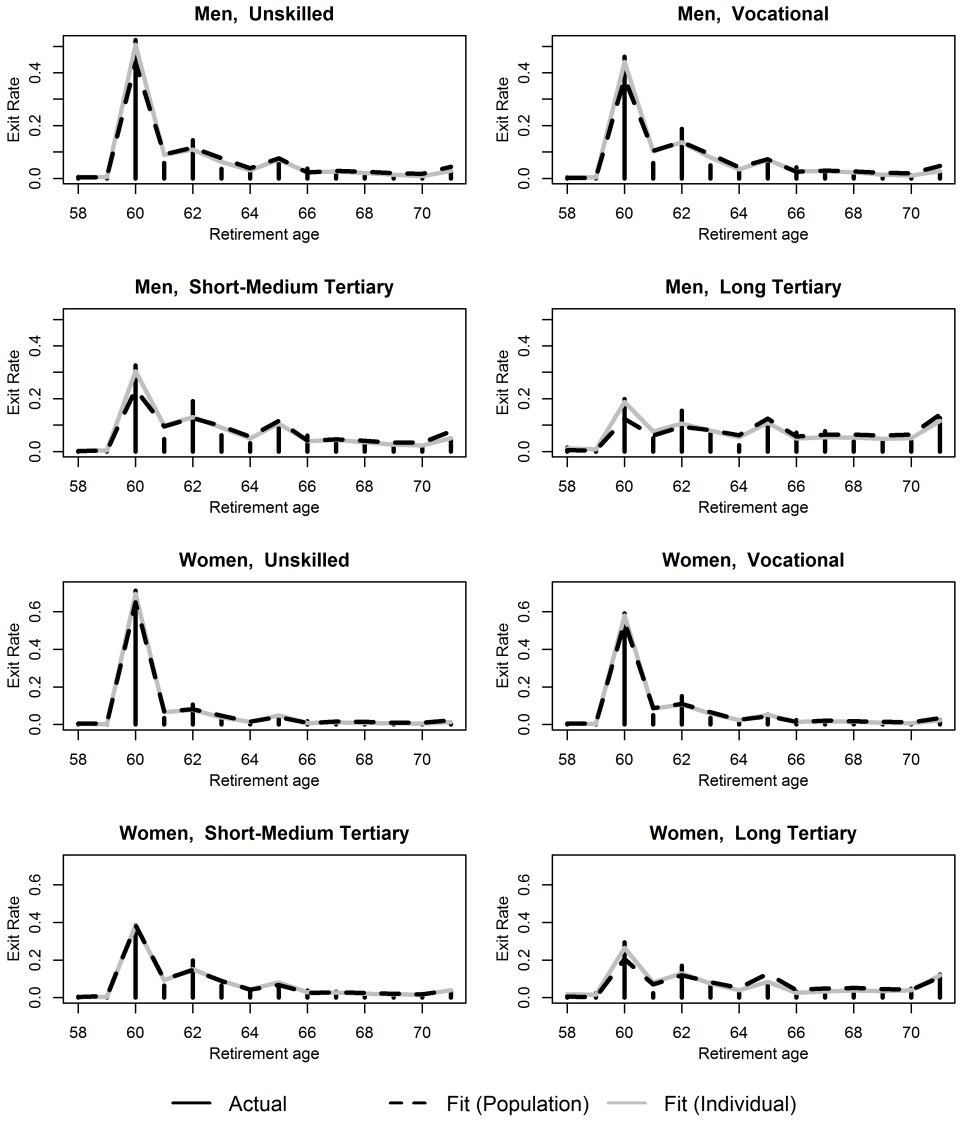
Figure 19: Predicted vs. Actual Exit Rates - Low Pension Wealth



The figures show how different values of k affects the expected retirement of four different types of individuals for each of the estimated gender- and education groups. We use the estimated homogeneous parameter estimates as presented in Table 2, and compute the expected retirement age predicted by the model for different values of k. The figures also show the estimated k-distribution histograms. Here, low income is set to 250,000 DKr at age 58, and a low pension wealth is set to 0 DKr. High income equals 500,000 DKr and a high pension wealth equals 500,000 DKr.

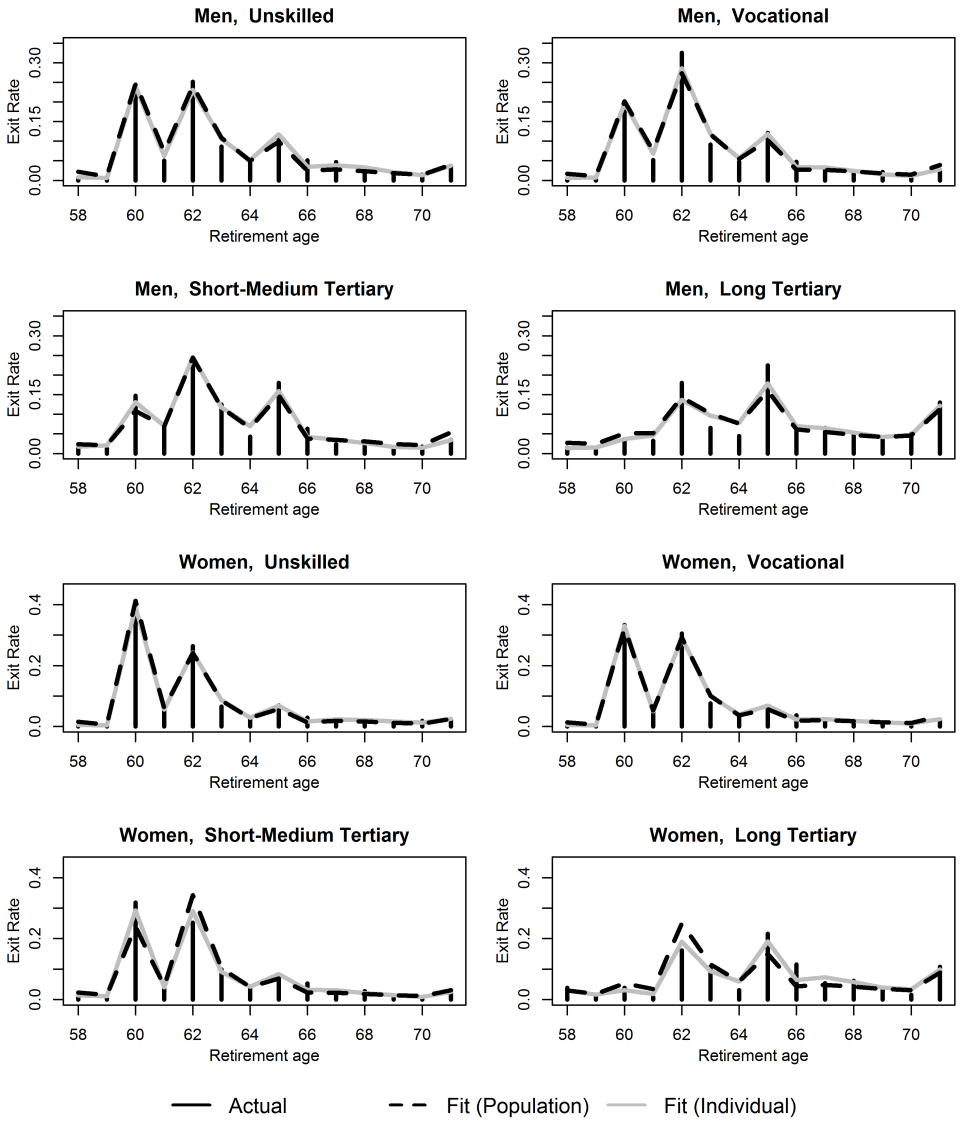
F Model fit for different levels of private pension wealth

Figure 20: Predicted vs. Actual Exit Rates - Low Pension Wealth



The graphs display the actual and predicted exit rates into retirement for birth cohorts 1942-1944 with private pension wealth in the 1st quartile of the private pension wealth distribution within the group. The vertical black lines plot the actual retirement age distribution. The dashed black lines plot our model's predicted retirement distribution when we use only the population distribution of k. The solid gray lines show the predicted retirement distribution modeled using the estimated individual-specific k-distributions.

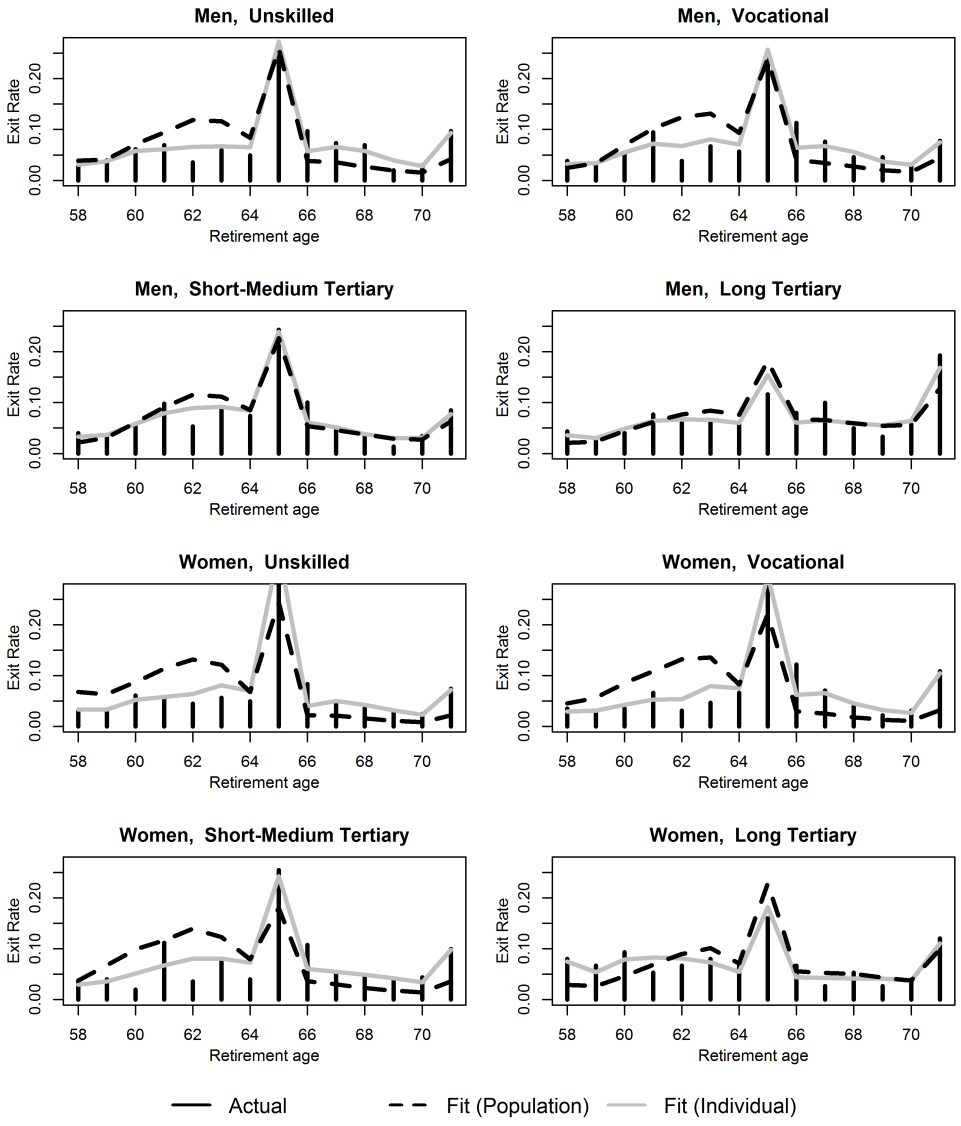
Figure 21: Predicted vs. Actual Exit Rates - High Pension Wealth



The graphs display the actual and predicted exit rates into retirement for birth cohorts 1942-1944 with private pension wealth in the 4th quartile of the private pension wealth distribution within the group. The vertical black lines plot the actual retirement age distribution. The dashed black lines plot our model's predicted retirement distribution when we use only the population distribution of k. The solid gray lines show the predicted retirement distribution modeled using the estimated individual-specific k-distributions.

G Model fit for individuals not entitled to early retirement benefits

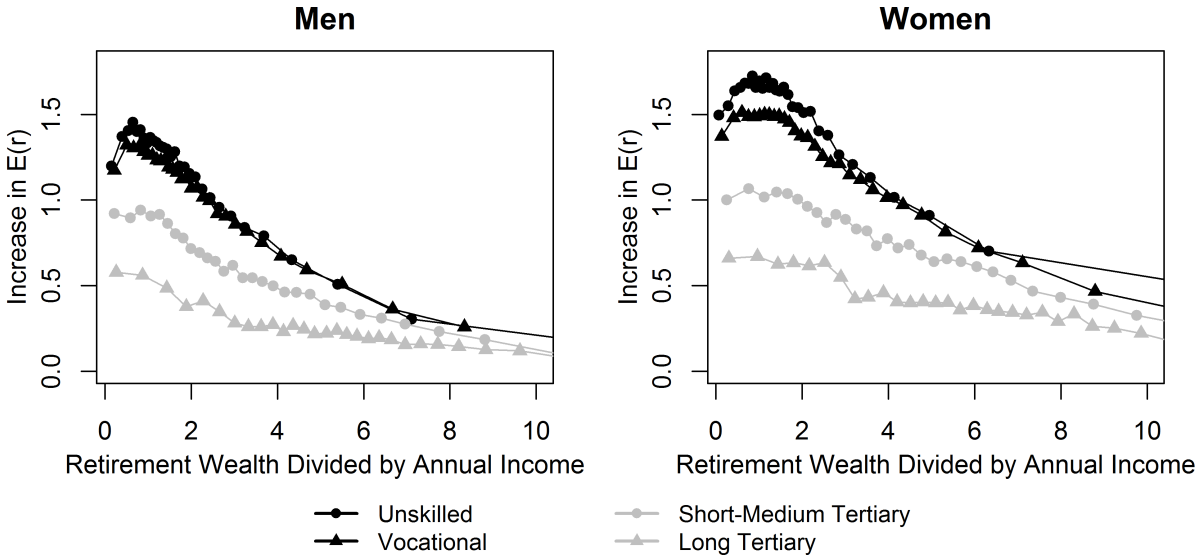
Figure 22: Predicted vs. Actual Exit Rates - Not ERP eligible



The graphs display the actual and predicted exit rates into retirement for birth cohorts 1942-1944 who are not members of the Early Retirement Program. The vertical black lines plot the actual retirement age distribution. The dashed black lines plot our model's predicted retirement distribution when we use only the population distribution of k. The solid gray lines show the predicted retirement distribution modeled using the estimated individual-specific k-distributions.

H Effect of baseline experiment across different levels of private retirement wealth

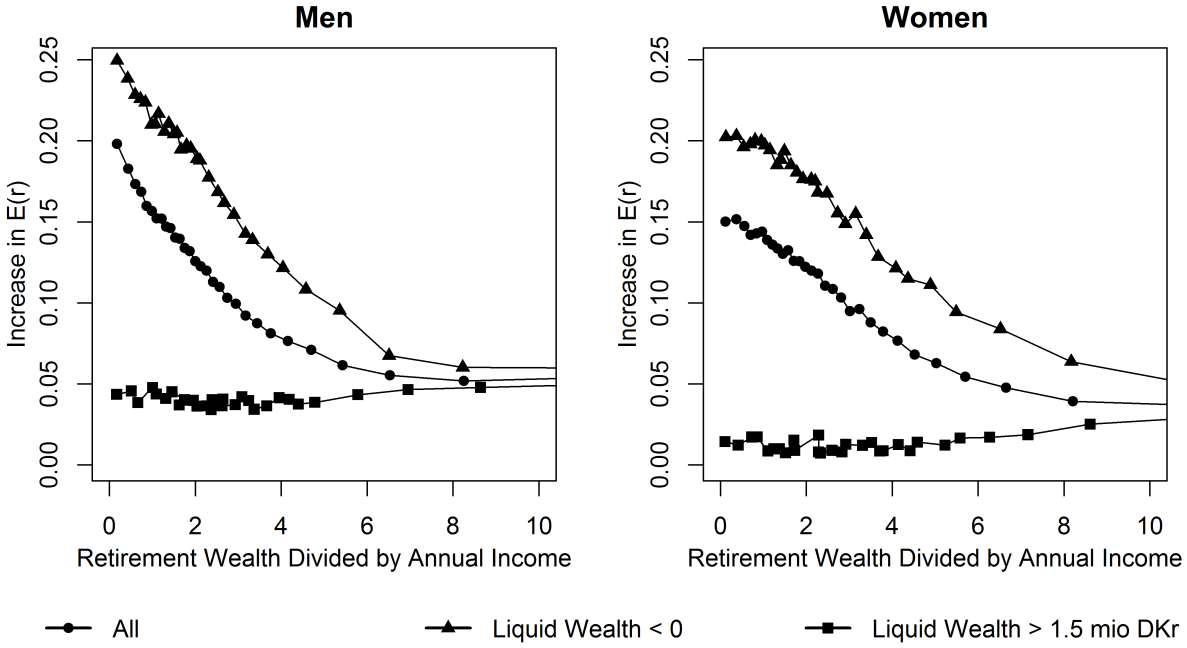
Figure 23: Effect of one year increase in NRA by accumulated pension wealth



For each gender- and education specific group, the figure shows how our model predicts that abolishing the ERP scheme would have affected the expected retirement age of cohorts 1942-1952, given that the early retirement scheme was abolished. The effect is measured by years, as the change in expected retirement rates.

I How non-retirement wealth affects the predicted response to an increase of the NRA

Figure 24: Effect of one year increase in NRA by accumulated pension wealth



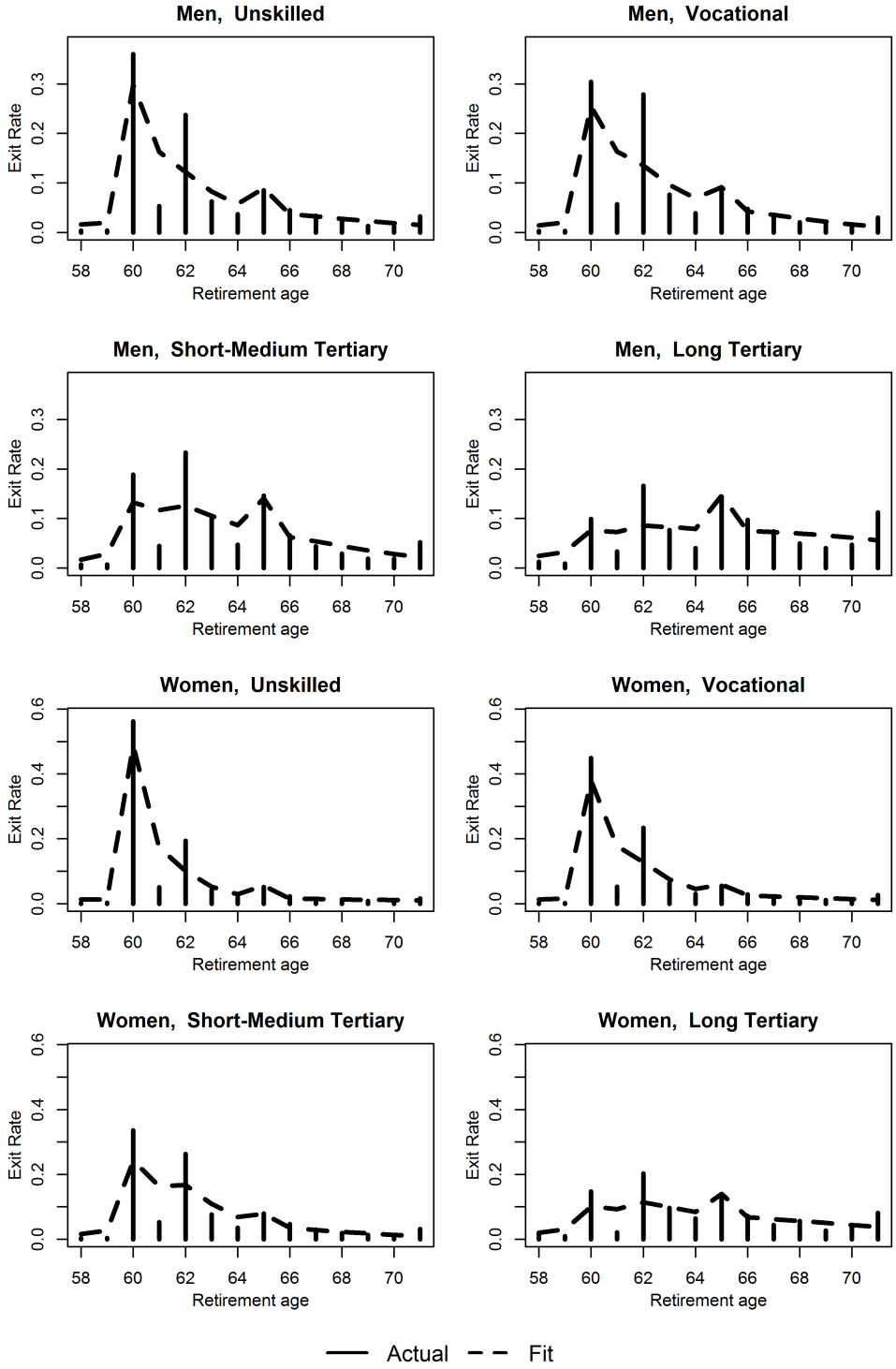
The figure plots the predicted effect of an increase of the NRA from age 65 to 66 on the expected retirement age of cohorts 1942-1952, assuming that the early retirement scheme was abolished. Split by gender, the black dots depicts the estimated effect for everyone by level of pension wealth, the black triangles depict the effect for everyone holding zero or negative wealth, and the black squares show the effect for the subsample holding more than 1.5 mio. DKr in non-retirement/liquid wealth.

J Model results for homogeneous leisure preferences, k

Table 4: Estimated Homogeneous Model Parameters (without leisure preference heterogeneity)

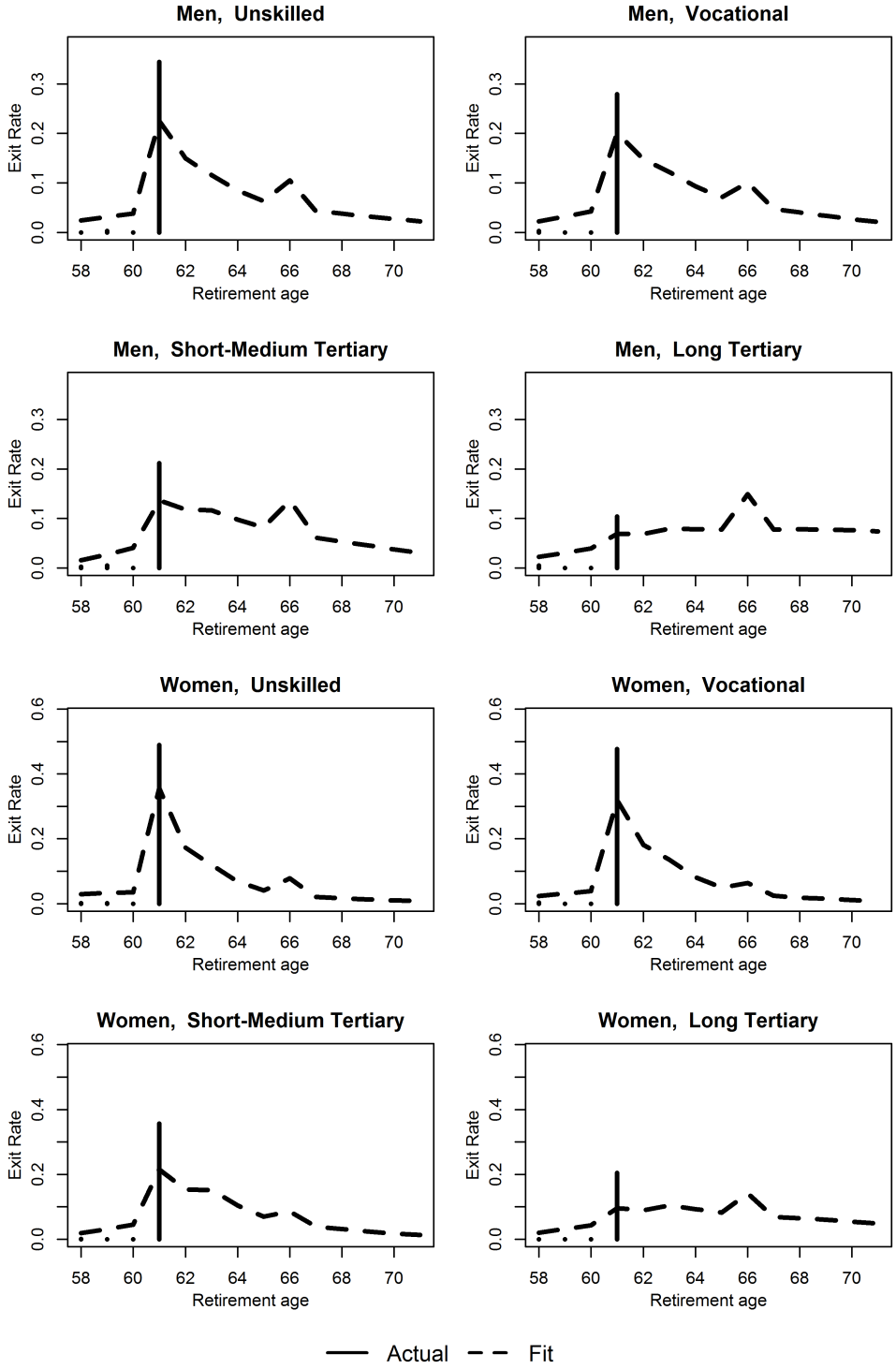
Men	**N**	α_0	α_1	β	σ	ρ	$\mathbf{d^0_{65}}$	$\mathbf{d^1_{65}}$	**k**
Unskilled	22462	0.0055	-2e-05	0.860	1.60e-06	1.96	1.24e-06	1.12e-06	1.27
Vocational	39890	0.0073	-5e-05	0.844	4.34e-06	1.87	2.72e-06	2.35e-06	1.11
Short-Medium	13084	0.0097	-5e-05	0.810	2.21e-05	1.73	1.54e-05	1.09e-05	0.92
Long Tertiary	6528	0.0056	-6e-05	0.834	7.69e-07	2.02	5.18e-07	4.91e-07	1.10
Women	**N**	α_0	α_1	β	σ	ρ	$\mathbf{d^0_{65}}$	$\mathbf{d^1_{65}}$	**k**
Unskilled	25016	0.0045	-5e-05	0.865	6.99e-07	2.00	8.02e-07	7.94e-06	1.31
Vocational	33361	0.0059	-6e-05	0.850	2.69e-06	1.90	2.01e-06	1.94e-06	1.17
Short-Medium	17210	0.091	0	0.812	6.28e-05	1.63	3.80e-05	3.82e-05	0.96
Long Tertiary	3247	0.0080	-1e-05	0.809	1.02e-05	1.80	6.93e-06	6.07e-06	0.96

Figure 25: Predicted vs. Actual Exit Rates - Not ERP eligible (without leisure preference heterogeneity)



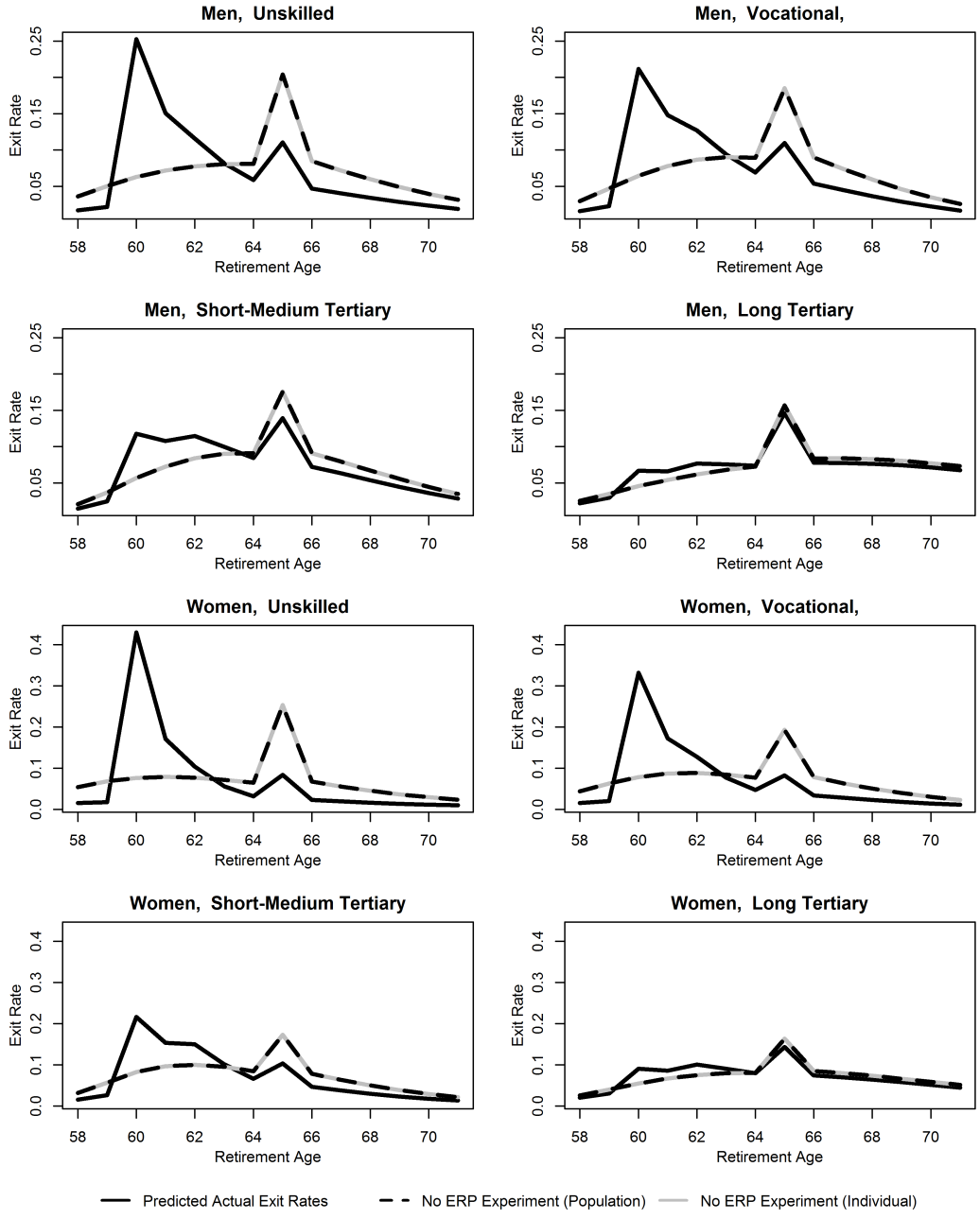
The graphs display the actual and predicted exit rates into retirement for birth cohorts 1942-1944, using the simpler model without leisure preference heterogeneity. The vertical black lines plot the actual retirement age distribution. The dashed black lines plot our model's predicted retirement distribution when we use only the population distribution of k. The solid gray lines show the predicted retirement distribution modeled using the estimated individual-specific k-distributions.

Figure 26: Predicted vs. actual exit rates, external validation test on cohort 1954 (without leisure preference heterogeneity)



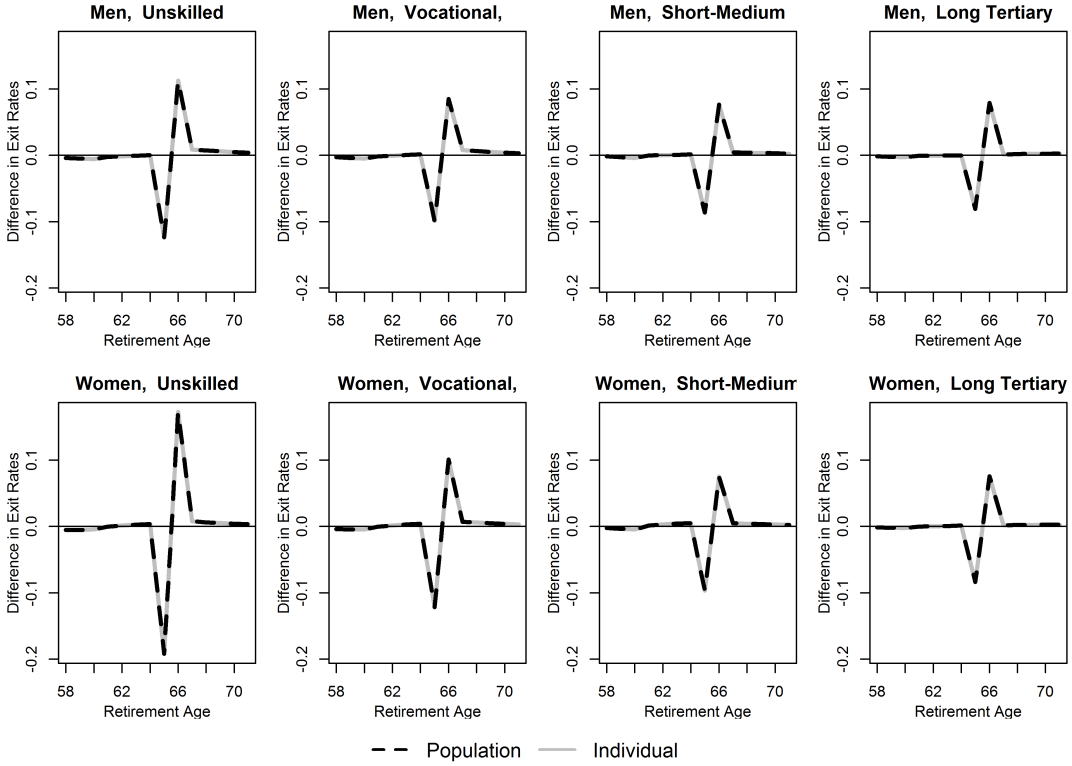
The graphs display the actual and predicted exit rates into retirement for individuals born in the second half of 1954, who participate in the Early Retirement program, using the simple model without leisure preference heterogeneity. The vertical solid black lines plot the actual retirement age distribution. The dashed black lines plot our model's predicted retirement distribution when we use only the population distribution of k. The solid gray lines show the predicted retirement distribution modeled using the estimated individual-specific k-distributions.

Figure 27: Retirement Effect of Baseline Experiment: No ERP scheme (without leisure preference heterogeneity)



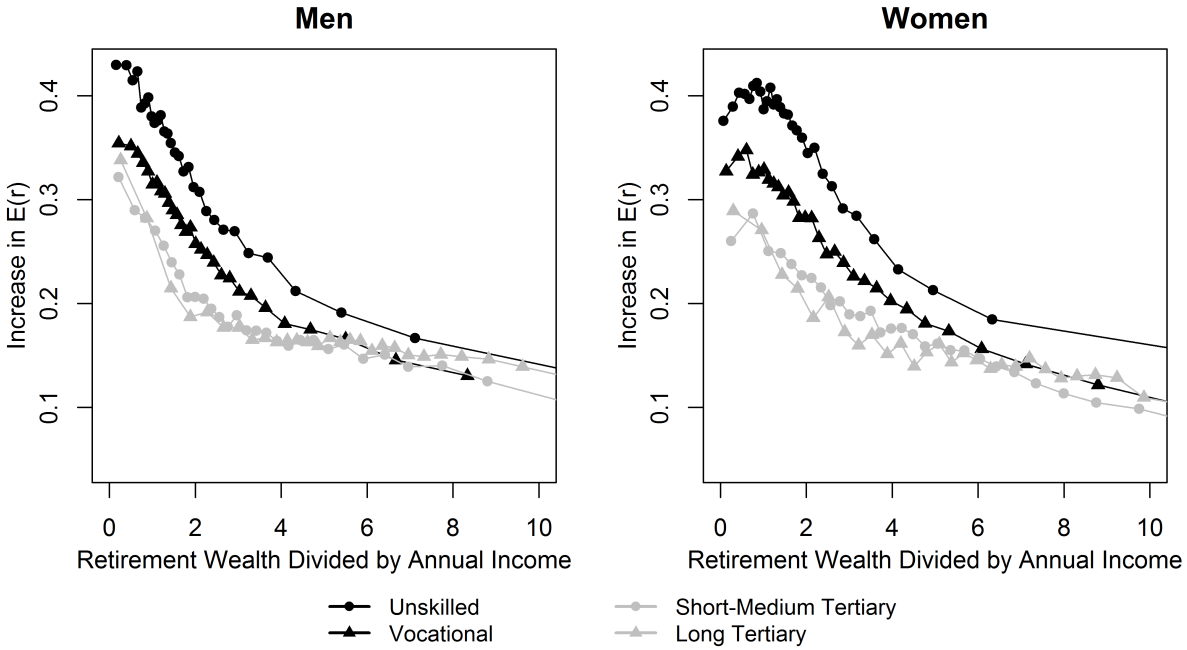
The solid lines depict the model's predicted exit rate distributions for the entire estimation sample consisting of cohorts 1942-1952 using the simple model without leisure preference heterogeneity (these are similar to those plotted in 25, but include more cohorts). The dashed black lines plot the simpler model's predicted retirement distribution in the baseline experiment, where the ERP scheme is abolished. The solid gray lines show the predicted retirement distribution modeled using the estimated individual-specific k-distributions, also using the simpler model without leisure preference heterogeneity.

Figure 28: Retirement Effect of One Year Increase of the NRA (without leisure preference heterogeneity)



The figures plot, for each gender- and education specific group, the differences in predicted exit rates between the baseline experiment where the ERP scheme is abolished and an experiment in which the ERP scheme is abolished and the NRA is increased from 65 to 66, when we use the simple model without leisure preference heterogeneity to predict.

Figure 29: Retirement Effect of One Year Increase of the NRA (without leisure preference heterogeneity)



For each gender- and education specific group, the figure shows how the simple model without leisure preference heterogeneity predicts that an increase of the NRA from age 65 to 66 would have affected the expected retirement age of cohorts 1942-1952, given that the early retirement scheme was abolished. The effect is measured by years, as the change in expected retirement rates.

Chapter 2

The Economic Impact of Healthcare Quality

The Economic Impact of Healthcare Quality

Anne-Line Koch Helsø[1], Nicola Pierri[2], and Adelina Yanyue Wang [*2]

[1]Copenhagen University

[2]Stanford University

Abstract

We study the costs of hospitalizations on patients' earnings and labor supply and evaluate the quality of treatment based on its ability to mitigate the labor market consequences of a given diagnosis. We link the universe of hospital admissions in Denmark to full-population tax data which allows us to measure the labor market consequences of hospital admissions, controlling for diagnosis, co-morbidities, and detailed patient- and local labor market characteristics. We measure the heterogeneity in treatment quality across hospitals and find a 4 percentage points difference in lost earnings between the best and worst hospital, all else equal. Conservative estimates imply that bringing all Danish hospitals to, at least, the median quality, would lead to a saving of DKK 456 million per year in foregone earnings for inpatient admissions only. We also document a large decline in the labor cost of hospitalizations over time, with large variations across diseases. We find that the average post-hospitalization reduction in labor supply has declined by 13.6 percentage points from 1998 to 2012, and a conservative estimate implies that healthcare improvements explain at least 40% of the estimated decline in labor cost.

[*]Supported by a Thomas Parry Research Fellowship award from the Integrated Benefits Institute (IBI). Many results of this paper are still preliminary and should not be used for policy design. We thank Tore Olsen for essential advice about data sources in the early stages of this project. We thank, for their insightful comments, Tim Bresnahan, Mark Duggan, Liran Einav, Matt Gentzkow, Caroline Hoxby, Thomas H. Jørgensen, Grant Miller, Torben Heien Nielsen, Petra Persson, Luigi Pistaferri, Bertel Schjerning, and seminar participants at Stanford University.

1 Introduction

Adverse health shocks are extremely costly, both to the individual and to society at large. The average OECD spending on health as a share of GDP was 8.9% in 2016, and a staggering 17.2% in the United States. Direct medical costs, however, only make up a fraction of the total costs of morbidity, and earnings losses might comprise the majority of total life-cycle costs from bad health experienced by working-age individuals [De Nardi et al., 2017]. Not surprisingly, there is considerable interest in measuring and improving the quality and efficiency of healthcare. However, the extensive literature that measures and quantifies the variation in treatment quality are based on medical outcomes, such as mortality and readmission rates, rather than the expensive labor market costs associated with adverse health shocks.

In this paper, we present the first (to our knowledge) attempt to measure the quality of a hospital treatment by its ability to mitigate the labor market consequences of a given diagnosis. The intuition supporting this procedure is that good healthcare should mitigate the detrimental effects of illnesses and injuries on patients' health and, consequently, on their labor market outcomes. All else equal, the better the treatment, the smaller the drop in patients' labor earnings and participation.

We link the universe of hospital admissions in Denmark with data on the labor market outcomes of the entire working-age population in Denmark from 1995 to 2015. This unique dataset allows us to measure the negative effects of hospitalization on the patient's earnings, which we will use to infer the quality of treatment received by the patient. More specifically, we consider the heterogeneity in treatment quality across two different dimensions: across hospitals and over time.

We find sizable heterogeneity in the labor cost of a hospitalization across different hospitals: our point estimates indicate a 4 percentage point difference in lost earnings between the best and worst hospital, all else equal, when we consider patients with their first non-pregnancy related inpatient admission in four years. Conservative estimates imply that bringing all Danish hospitals to, at least, the median quality, would lead to a saving of 456 million DKR per year in foregone earnings for inpatient admissions only. Rankings based on our measure are positively correlated with rankings based on

traditional quality metrics, with a significant Spearman rank correlation of around 30%.

The validity of our quality measure faces two main concerns related to unobservable patient severity, which have been raised by previous literature (see Geweke et al. [2003], Doyle [2011], and Doyle Jr et al. [2015]). First, one could worry that the best hospitals treat a larger share of the unobservably sickest patients, which would bias their estimated quality downwards (and vice versa for the worst hospitals). Second, better hospitals might be located closer to unobservably "better" (e.g. with higher human capital) individuals, and therefore admit a large share of them. These individuals might be more resilient to health shocks, which would also invalidate our estimation procedure. Although all of our specifications include detailed controls for the main diagnosis, other co-morbidities, and past medical history, which should minimize the magnitude of unobservable patient severity, these are still relevant worries. To deal with both concerns, we adopt two complementary empirical strategies to test whether being admitted to a higher quality hospital (according to our estimates) leads to a positive causal effect on post-admission earnings.

First, we instrument the quality of the hospital where a patient is admitted with the neighborhood (or "Church District") where she lives (similarly to Geweke et al. [2003]). In fact, Danes are more likely to go to the closest hospital (neighborhoods explain more than 70% of the variation in hospital quality). The IV estimates confirm that being admitted to a better hospital increase normalized earnings (in the year of admission and the following 3) as much as predicted by the main OLS model estimates, thus supporting the causal interpretation of the quality measure.

Second, we follow Doyle [2011] and focus on patients admitted for acute non-deferrable conditions outside their home location (that is, in a hospital which is not "common" for patients from the same neighborhood). These patients are more likely to be admitted because of unforeseen health shock while they were outside their daily routine and less likely to have consciously chosen a specific hospital. Studying this sub-sample of admissions, we find that patients coming from the same location are better off being admitted to a higher quality hospital. We also show that home location does not predict post-admissions earnings dynamics when patients are treated "away" from their local hospital, suggesting that unobserved heterogeneity clustered at local level is not a significant concern for our

quality measures.

We also document a significant decline in the labor cost of hospitalizations over time. For different diseases, we compare the change in employment probability and labor earnings of patients after their hospital admissions in different years. We find that there has been a significant decline in both earnings and labor participation drops following a hospital admission during the period 1998-2012. More specifically, we find that the post-hospitalization reduction in labor supply has declined by 8.8% on the extensive (participation), and 4.9% on the intensive margin (earnings conditional on working) on average, as measured by the accumulated change in the three years following the admission. Thus, the total change in estimated labor costs amounts to 13.6 percentage points. This corresponds to an average reduction of extensive margin labor market costs of about 50%, and a 25% reduction of intensive-margin labor cost of a disease. We also find a huge variation in the reduction of labor market cost across different diseases,

Given the assumption that factors not related to the improved treatment quality should depend only on the severity of the disease, and not the nature of the disease itself, a conservative estimate implies that at least 40% of the estimated change are due to improved healthcare treatment over time. Other factors which could explain the decrease in labor cost could be related to e.g. labor market changes aimed at decreasing sick leave.

Our paper contributes to a large literature on healthcare quality measurements, for which practitioners, policymakers, and academics have produced several metrics of health and healthcare quality. Risk-adjusted mortality and readmission rates are the most commonly used measures of hospital quality. For instance, in the US, the Affordable Care Act (through the Hospital Readmissions Reduction Program) has targeted the reduction of the readmission rates following hospital admissions for a few specific conditions[1] as the relevant goal for quality improvement and cost reduction [Gupta, 2017]. Other hospital quality measures rely on patients' self-reported ability to perform activities of daily living (ADL), such as bathing or dressing, or patients' satisfaction with their treatments or their queuing time. More technical measures of treatment quality track specific biological indexes to study the impact of a treatment on the evolution of a specific medical condi-

[1]Heart attack, heart failure, pneumonia, Chronic Obstructive Pulmonary Disorder (COPD), and Hip/Knee replacement.

tion, such as hemoglobin concentration for anemia [Miller et al., 2012] or white blood cell concentration for HIV [Hamilton et al., 2016].

A large health-economics literature has measured hospital quality focusing on emergency room (ER) admissions for heart attacks, heart failures, and other acute and severe conditions. For instance, see Geweke et al. [2003], Doyle Jr et al. [2015], and Hull [2018]. These studies also mostly consider readmission and mortality rates as the outcome measure.

Another strand of the health economics literature considers the economic consequences of severe health shocks. Using U.S. data, [Dobkin et al., 2018] documents that hospital admissions are followed by severe and long-lasting worsening of economic conditions. They estimate an average annual decline in labor market earnings of about 17 percent of pre-admission earnings during the first 3 years after admission. Using Danish register data, [Fadlon and Nielsen, 2015] show that negative health shocks substantially harm individual labor supply and disrupt households' consumption patterns. [Gilleskie, 1998] show that employees' poor health might increase work absence and hinder workers' ability to adequately perform the job tasks, leading to substantial costs for employers and governments. Our work connects these two strands of literature by using the economic consequences following a hospitalization to infer the treatment of the received quality.

Our measure of healthcare quality differs in several respects to the ones employed by the previous literature, which are often based on mortality and readmission rates. Mortality and readmission rates, however, are not only very coarse and extreme outcomes, but they are also mainly informative for very severe and acute conditions. As we consider earnings losses, we're able to compute a quality measure for a much broader range of conditions, including also the less severe diseases. An additional advantage of our measure is that is has a direct economic interpretation in terms of forgone earnings. The economic impact of physical and mental impairments, which is mostly ignored by traditional measures, can be a sizable part of the risk associated with such negative events. With a labor market outcome variable, it is crucial that we also control for local labor market characteristics and differences in caseworker procedures across municipalities. Luckily, our rich set of variables in the Danish register data allows us to do that. The disadvantage of our

measure is that in only focuses on the working-age population, and as such our measure is uninformative of the treatment quality of children and seniors.

Relatedly, our paper also adds to the literature that documents significant variation in healthcare utilization, spending, and outcome across different regions and hospitals. While there has been strong evidence of large geographic variation in healthcare utilization and spending (Finkelstein et al. [2016]), there has been relatively little evidence showing how such heterogeneity translates into heterogeneity in actual health outcomes. Existing literature often find small geographic variation in health outcomes, such as mortality, despite large differences in spending and treatment practices. However, our paper demonstrates that there can be significant differences in healthcare quality even in a relatively homogeneous healthcare system as the Danish.

The paper proceeds as follows: Section 2 describes data sources, institutional setting, and presents some raw data patterns. Section 3 presents our measure of heterogeneity across hospitals. Section 4 investigates whether this heterogeneity can be interpreted as a reliable measure of hospital quality. Section 5 deals with the evolution of the labor market consequences of hospital admissions over time. Section 6 concludes.

2 Data

2.1 Data Sources

The Danish National Hospital Register ("Landspatientregisteret"/LPR) contains nationwide data on all hospital interactions since 1995. Available information includes a 4-digit ICD10 main/action diagnosis code and other related co-diagnoses, patient type (inpatient, outpatient, ER), start- and end-date of treatment and the hospital and hospital department at which they received the treatment. The data also contain information on the conducted treatments, operations and examinations. Using a unique personal identifier, we are able to link the hospital data to other full-population registers. We link data from the Danish National Hospital Register (LPR) to tax- and labor market registers so that we observe a detailed work history and all interactions with the hospital system in Denmark for the full population of Danes aged 25-59 years from 1995 to 2015. Our final data set

includes an extensive list of the labor market and demographic variables, including earnings, transfer income, municipality, marital status, wealth, spousal earnings and wealth, education, occupation and socioeconomic status. As our data includes the full population of working-age Danes, we're also able to construct education-specific local labor market statistics, such as e.g. unemployment, median income etc., of each municipality.

In this study, we will focus exclusively on in-patient hospital admissions. To mitigate the problem of co-morbidities affecting the treatment outcomes, we restrict our sample to include only "index"-admissions. An "index" admission is defined as the first non birth/pregnancy related inpatient hospital admission for at least 3 years. We follow the labor-market outcome of the patients 3 years prior to-, during, and 3 years following their treatment year. Our data contains a total of 1,1 million index hospitalizations. Table 1 shows some descriptive statistics of the earnings and labor force participation before, in, and after the year of an index admission for our main analysis sample. It also shows similar statistics for a sample including all observed inpatient hospital admissions (only including one hospitalization per year per person), also in 1998-2012 for individuals who are 28-57 years old once hospitalized, and for a control group (which we use in the over time analysis) who don't experience any index admission during the observed years.

The table shows that roughly 1/3rd of all observed hospitalizations for the specified ages ad years are index admission observations - counting only one observations per person per year, index admissions make up 50%. Those who experience an index admission are on average slightly younger (0.4 years) compared to those who experience any inpatient admission, and have slightly lower earnings conditional on being in the labor force. However, the participation rate of the index admission sample is significantly higher, with pre-admission participation being 0.915 compared to a pre-hospitalization participation rate of 0.68 for all inpatient admissions. Both groups, however, experience a drop in participation rates of roughly four percentage points in the hospitalization year, and another three percentage points in the following year. Also, mortality rates are slightly lower in the index-admission subsample. The control group includes individuals who are either of very poor health (with too many inpatient hospitalizations to have an index-admission observation), or of very good health (without any non-birth related inpatient admission

during the period). The participation rate of this group is also lower (0.82) compared to the pre-hospitalization participation rate of the index-admission sample (0.915), but earnings conditional on working are similar to that of the index admission patients. As such, we loose some representativeness when we restrict to index-admission observations only.

2.2 Institutional Settings

The Danish health care system is universal and provides free and equal access to health care for all citizens. Health care expenses are primarily tax financed and the vast majority of hospitals are publicly owned. During each year from 1998-2012, private hospitals accounted for less than 2% of total hospital expenditures. Government spending on healthcare increased by 30% from 2000 to 2010, mainly driven by an increase in hospital expenditures. Hospital financing is based on a system of politically fixed budgets that account for heterogeneity in patient mix. Following a large structural reform in 2007, the 16 counties responsible for running the hospitals were merged into 5 regions.

Compared to e.g. the U.S., there are fewer but larger hospitals in Denmark. Since the 1990s, several mergers and hospital closures have resulted in a further concentration of hospitals. Often, the general practitioners act as a gatekeeper to specialized treatment. A person's residence address determines his/hers "default" hospital, to which he/she is referred to by default. However, since 1993, patients were granted the right to freely choose among all somatic hospitals in Denmark, if the requested hospital has the capacity. Since 2002, an extended free choice reform has granted patients the right to choose a state financed treatment at a private hospital if waiting times at the public hospitals are too long (2 months before 2007 and 1 month after). However, approximately 91% of patients in our sample period are admitted to hospitals in the same region where they live.

2.3 Descriptive Patterns

To motivate our analysis, and to give examples of what the data looks like, we show examples of how individuals' earnings evolve before and after a hospital admission for different types of diseases. These illustrative examples show suggestive evidence that

earnings trajectories following hospital admissions vary for different types of diseases, over time and across hospitals.

To do so, we estimate the following linear model

$$y_{i,t} = X_{i,t}\beta + \gamma_t + \sum_{-5 \leq r \leq 5} \delta_r + \epsilon_{it} \tag{1}$$

The dependent variable is a labor market outcome of individual i at year t. Due to the descriptive nature of this exercise, we include only a small set of controls, $X_{i,t}$: age, gender and education, and γ_t denotes calender-year time fixed effects. For individuals who experience an index hospital admission, the subscript r denotes the year relative to the hospital admission year where $r = 0$[2]. We add a 10% random subsample of the population who don't experience an index admission from 1995-2015 as a control group, and normalize $\delta = 0$ for these individuals. As such, the δ coefficients estimate the differences in labor market outcome for individuals hospitalized in year $t - r$ with respect to other (non-hospitalized) individuals with similar characteristics X. For illustrative purposes, we estimate equation (1) for four different diseases. These specific types of diseases are chosen because they are quite common and known to most people, and then they differ a lot in terms of their severity and duration. We consider intervertebral spondylosis (spinal degeneration), acute cerebrovascular disease, breast cancer and fracture of the lower limb[3].

First, we consider both the intensive- and extensive margin labor market consequences of a hospital admission. We compute the intensive margins by using $log(earning)$ as the outcome variable - thereby we only include positive earnings observations. We compute the extensive margins by using a dummy variable for positive earnings as the outcome variable for all observations, with dead individuals included with zero earnings (excluding those who die does not change any of the figures notably).

The δ coefficients of equation 1 are plotted in Figure 1, where the left column of Figures show the intensive margin responses, and the right column show the extensive

[2]For instance, $r = -2$ indicates the individual i experience an index admission in year $t + 2$. Note that in this section only, we restrict to index admissions where there is no other inpatient admission in year t-5 through t, while for all the main analysis the index admissions are defined as no other inpatient admissions in year t-3 through t.

[3]These disease classes are grouped according to the Clinical Classification Software, CCS. In the empirical analysis of hospital heterogeneity, we use a finer diagnosis classification.

margin responses. The plots show that patients' earnings and labor market participation sharply decline around the year of an hospital admission. The patterns for the two margins are quite different: while the drop in the intensive margin tends to recover - especially for the less severe diseases - the extensive margin response continues to decline for all diseases.

In the left column showing the intensive margin decline in labor supply, we see that and the sizes of the drop, as well as the degrees to which earnings rebound, vary a lot by disease. Patients with intervertebral spondylosis (spinal degeneration) experience an intensive-margin decline in earnings of 8% , which is halved to a 4% decline 3 years after the admission, and their labor force participation drops by 12%. Patients who suffer from an acute cerebrovascular disease experience an intensive-margin decrease in earnings of 12% and a drop in labor force participation of almost 20%. Breast cancer patients experience a somewhat smaller drop in labor supply with a 5% decrease in the intensive-margin, and a steady decline in labor force participation mounting in a 14% decrease five years after admission. Following a fracture of the lower limp, patients suffer an 8% drop in intensive-margin earnings, but are almost fully recovered after 3 years, and their labor force participation permanently decrease by 7%.

Has healthcare quality improved since 2000? We use our simple specification to provide evidence of advancement in the effectiveness of medial treatment in preventing patient earnings loss. We re-estimate equation (1) dividing patients according to the year of their hospital admissions: 2000-2002, 2003-2004, 2005-2007 and 2008-2010. To capture both extensive and intensive margin responses, the dependent variable is the log of labor earnings plus a constant (1 DKK), so that we're able to combine the intensive and extensive margin labor market outcomes in one outcome measure. Because of the arbitrary choice of this constant (1 DKK), we only rely on this specific outcome measure for illustrative purposes, and we will refrain from use it in the main analysis (see section 3.1). Results for the four different diseases are presented in Figure 2. We detect a striking heterogeneity over time for the different diseases. For instance, the earnings losses associated with intervertebral spondylosis (spinal degeneration) for patients treated in 2008-2010 is only borderline significant smaller than the loss of patients treated in 2000-2002. On the other

hand, patients admitted for cancer of breast or acute cerebrovascular diseases seem to be much better off in the later period compared to the early 2000s. In section 3, we present additional results about hetereogeneity in quality improvements.

Do healthcare quality vary across different hospitals? Again, we consider equation (1) to provide some suggestive answers to this question - now allowing the δ coefficients to vary according to the index admission's hospital. In Figure 3 we present results related to four randomly selected hospitals. The pre-hospitalization trends in earnings of individuals who suffer from acute cerebrovascular diseases are not significantly different across the different hospitals which they are admitted to. However, the earnings losses following the hospitalization are much more severe for patients of hospital A and D rather than B or C. In general, for all four diseases, patients of hospital A and D seem to suffer the most from their disease. However, these descriptive results do not in detail control for patient characteristics, specific diagnosis types, local labor market characteristics etc. In section 3 we present a more in-depth investigation of heterogeneity of hospital treatment quality.

3 Healthcare Treatment Across Hospitals

A worker's health status alters the physical and mental cost of working (or searching for an occupation), and is consequently a fundamental factor of labor supply and productivity. Therefore, the quality of healthcare received by a hospital patient should have an effect on his/her post-hospitalization labor market outcomes. Intuitively, as long as the cost of working is decreasing in health, and while productivity is increasing, then the economic costs of a hospital admission should be lower for a patient receiving better care.[4]

Following this intuition, we define the quality of hospital h as the difference in earnings realized after a treatment at hospital h and the earnings which would have been realized after at treatment at some reference hospital 0:

$$q_h = E[y_h|r \geq 0] - E[y_0|r \geq 0] \tag{2}$$

where y_h refers to the labor market outcome of interest following an admission to hos-

[4]This assumption would be violated if, for instance, sicker people might decide to supply more hours worked because their value of leisure is lowered by a negative health shock.

pital h. r denotes the time relative to the hospitalization, such that $r \geq 0$ includes the year of the admission and the following ones. We propose a simple empirical framework for the estimation of q_h under different assumptions on the data generating process for the observable and unobservables elements affecting health, hospital choice, and workers' productivity. In particular, our main specification relies on the assumption that, after controlling for a large set of characteristics at the patient-, case-, and local-level, hospital selection is not driven by the residual unobservable heterogeneity (e.g. unobserved severity). Thus, we run different empirical checks to test the robustness of our main findings to possible violations of this identifying assumptions and we show remarkable stability.

3.1 Empirical strategy

To estimate hospital quality, we consider all inpatient "index" hospital admissions in Denmark during the years 1998-2012, see section 2.1 for more details. We restrict to a set of reasonably large hospitals[5], and we exclude patients with "rare" diagnoses (less than 100 observed index admissions in 1998-2012). We include only Danish-born patients aged 28 - 56 at time of hospitalization, in order to focus on individuals at the peak of their labor supply.

The main outcome of interest is the post hospitalization earnings produced by individual i in year t as a share of their mean earnings in the 3 years prior to the index admission (henceforth referred to as "normalized earnings"). The use of normalized earnings allows us to consider both the extensive (labor market participation and survival) and intensive margins of the labor market consequences of hospital admissions [Kleven et al., 2018]. We let h denote the hospital of the index admission, d the main diagnosis, and $0 \leq r \leq 3$ the time relative to the hospital admission. The analysis rely on the following linear model

$$y_{i,t,d,h} = q_h + X_{i,t}\beta + \gamma_t + \phi_d + \sum_{0 \leq r \leq 3} \delta_{r \times dg} + \epsilon_{i,t,d,h} \tag{3}$$

where $X_{i,t}$ represents a vector of control variables, γ_t are calendar year fixed effects, q_h are hospital fixed effects and ϕ_d are main diagnosis (according to 4-digits ICD10 codes)

[5]For consistency purpose, we include hospitals that were established by 2011 and are the largest 35 in Denmark in terms of index inpatient admissions across all samples of analysis.

fixed effects. The parameters $\delta_{r,dg}$ capture disease group specific evolution of earnings around the hospitalization, where r denotes the time relative to hospitalization and dg the diagnosis group.[6]

The vector $X_{i,t}$ contains several covariates that capture both individual heterogeneity and local labor market conditions. The richness and the level of details of $X_{i,t}$ is one of the strengths of our study with respect to previous literature focusing on US hospitals. At the individual level we include several socio-economic pre-hospitalization characteristics: we flexibly control for earnings (percentile of the earnings distribution), earnings growth, decile of own assets, decile of spouse's wealth and spouse's earnings interacted with gender, gender, age, education, marital status, occupation, and socioeconomic status the year before hospitalization. We further control for the number of co-diagnoses related to the index admission and for other interactions with the public healthcare system, namely number of non-pregnancy related outpatient and ER visits and number of pregnancy related inpatient, outpatient, and ER visits the same year and 3 years before. Finally, we include controls for education-specific labor market conditions, namely unemployment rates, non-in-labor-force rates, and earnings levels for individuals with same education level within the municipality, and the share of individuals receiving disability benefits in the municipality.

The quality estimates q_h are obtained by applying OLS to equation (3), which is consistent under the assumption that the error is uncorrelated with the regressors. This assumption is violated if the hospital choices are affected by either unobserved individual characteristics or unobserved severity of health shocks. In section 4 we discuss these potential biases, and we present evidence suggesting these concerns have a limited role for our main results.

3.2 Results

Figure 4 shows the distribution of estimated hospital FEs. Since the dependent variable is normalized earnings, we interpret the quality measure as the percentage change in earnings (compared to mean pre-hospitalization levels) expected to occur over 3 years

[6]We group the 4-digit ICD10 codes into 271 diagnostic groups.

from the hospital admissions.

We find substantial heterogeneity in hospital quality, with a standard deviation of approximately 1%. An inpatient admission in the lowest-ranked hospital is associated with an additional decrease in annual labor earnings of around 4.10 percentage points compared to the highest-ranked hospital and around 1.45 percentage points compared to the median hospital. Patients treated at hospitals with below-median hospital FE's in our analysis sample would have on average earned an additional 0.7% of their pre-hospitalization earnings (around 2,104 DKK) every year after hospitalization if they were treated at the hospital with median estimated quality. It amounts to 236 million DKK (36.8 million USD) in forgone earnings (up to three years after hospitalization) on average for the index admissions in our analysis period each year, and in total 3.53 billion DKK (550 million USD) throughout the analysis period of 1998 - 2012.

To extrapolate to all inpatient admissions (rather than only the index admissions in our sample), we assume that the labor market consequences, as measured by percent of pre-hospitalization earnings, is the same for the entire sample of inpatient admission observations as those estimated for index-admission patients only. We perform a back-of-the-envelop calculation of the savings from increasing hospital quality following the formula:

$$Savings_t = \sum_{h:q_h<q_{med}} \left[(q_{med} - q_h) \times \left(N_{ht} \times \bar{Y}_{ht} \times 3\right)\right]$$

For all the below-the-median hospitals we calculate the increase in quality needed to achieve the median quality $q_{med} - q_h$. N_{ht} is the number of all non-pregnancy related inpatient admissions of patients aged 25-59 in year t for hospital h. $\bar{Y}_{ht}$ is the mean earnings the year before hospitalization for all the patients included in N_{ht} at hospital h in year t. Then, we multiply this quantity for the number of years for which the patients' earnings are affected by health care quality (which is 3 years in our case). Our estimates imply that if all hospitals had at least the median quality, the Danish economy would save on average at least 456 million DKK (71.4 million USD) in foregone earnings each

year, and 6.84 billion DKK (1.07 billion USD) in total in 1998 - 2012.

We also compute a ranking based on our measure of quality. We compare our ranking to an alternative ranking provided by the Danish medical news journal "Dagens Medicin" in 2011. They rank the Danish hospitals based on three outcomes: treatment quality, patient satisfaction and reputation among hospital personnel. Treatment quality is an average of various different specialty-specific indicators, such as provided procedures, risk-adjusted mortality- and readmission rates - each weighted by their importance for each specialty. We find a Spearman rank correlation of ≈ 0.3 (statistically different from 0); our measure, while positively correlated with this alternative measure, also contributes with additional information.

4 Robustness of the Quality Measure

In this section, we investigate whether the heterogeneity of the labor market consequences of hospital admissions across hospitals (documented in section 3) can be interpreted as a reliable measure of differences in treatment quality.

When we measure hospital quality we aim to capture the changes in normalized earning caused by the admission to one institution rather than another (see equation 2). Estimation of equation (3) by OLS provides consistent estimates of such parameters under the assumption that the error term is uncorrelated with the hospital choice. That is, after controlling for observable factors at the patient- and local labor market-level, the remaining unobservable characteristics of admissions must be equally distributed across different hospitals.

The literature on the measurement of hospital performance has highlighted two main concerns for this potential endogeneity problem. The *first concern* is selection on unobservable patient severity. Facing a waiting list trade-off, Danish patients with non-urgent treatment requirements are free to choose any hospital in the country with few exceptions. Therefore, the best hospitals might treat a disproportionate share of the sickest patients, all else equal, which would cause a downward bias of their estimated quality.

The *second concern* is that, since hospitals are not randomly allocated across a coun-

try's territory, some hospitals might be closer to a specific subset of patients, and therefore admit more of them. For instance, some hospitals might be located closer to unobservably "better" (or more resilient) patients, causing an upwards bias of their estimated quality. For instance, Doyle Jr et al. [2015] write that *"patients who live relatively close to "high-tech" hospitals could be different from those who do not, in ways that are difficult to control"*. A concrete example could be differences in other health-care related amenities, such as local GP quality.

In order to reduce the amount of residual heterogeneity between patients as much as possible, we include numerous different controls in all of our specifications, including a detailed set of controls for the main diagnosis, co-morbidities, and past medical history. Nonetheless, patients, nurses, and doctors are likely to have additional information about the patient, which is unobserved to the econometricians, and which affects the hospital choice of the patients.

Both endogeneity concerns are likely to be less harmful in our study compared to previous, mainly US focused, literature. This is primarily because of the the greater level of detail in our data set, which allows us to control for a wide range of patient characteristics, which consequently should limit the amount of unobserved heterogeneity. One could also argue that the concerns are smaller because of the specific institutional settings in Denmark, where fewer and larger hospitals are also more likely to serve a more representative mix of patients, and where the relative travel distance to the closest vs. second closest hospital is generally larger. Also, the Danish population is more homogeneous, with lower income inequality and poverty, lower cultural and ethnic diversity, etc. Nonetheless, both concerns are still extremely relevant for our analysis, and in this section we present several empirical strategies to assess the extent to which these potential biases could affect our estimates. We find evidence that our results are remarkably robust.

4.1 Robustness across different sample definitions

The first robustness exercise is based on re-estimating the main equation in 3 using different patient samples. These different set of patients are be subject to different selection dynamics because of economic and technical reasons. Selection on unobserved hetero-

geneity could be a problem for each one of these specifications. However, the ranking produced with all of different samples are very similar, suggesting that selection issues are not a primary driver of our baseline results.

Admissions at default and common hospitals

To reduce the confounds of patient mix, we first repeat our analysis for a subsample of individuals who are treated at hospitals of the same county/region as their residence (henceforth referred to as "default hospitals"), which is around 91.5% of the full analysis sample. We then further restrict the sample to individuals who are admitted to their local common hospital which is defined as the most common hospital for all individuals in their neighborhood (Church District) with inpatient admissions for the same diagnosis group. These patients are 72.8% of the full sample. The estimated hospital FE's in both subsamples are extremely similar to those estimated in the full sample. The Spearman correlation of rankings based on the hospital FE's are above 0.97 across the full sample, the default hospital sample, and the common hospital sample. Thus the estimated hospital quality is unlikely driven by the treatment of patients opting out of their default hospitals.

Acute admissions

Compared with chronic diseases, acute health conditions may allow less room for patient selection due to the urgent need of treatment. Hence we carry out another set of analysis restricting to acute conditions only and compare the results with those from the full sample. We first restrict to index admissions that are classified as acute in the LPR (health record) data, which is around 62% of the full analysis sample. We then further restrict to diagnoses both labeled as acute in the LPR data and whose weekend admission rates are closest to two-sevenths (similar to Doyle Jr et al. [2015]), which is around 10.7% of the full analysis sample. The Spearman correlation of rankings is 0.97 between those estimated in the full sample and in the first acute sample, and 0.79 between those in the full sample and the more restrictive acute sample. The exercise show that our estimation results are thus much unlikely driven fully by patient selection, since they are robust to restricting to even the 10% most acute conditions.

Figure 5 illustrates the correlation of the quality estimates across all the above-mentioned samples. We interpret the large correlation coefficients between the full sample quality estimates and the different subsamples as evidence for the robustness of our estimated treatment quality.

4.2 IV estimates and acute admissions at non-common hospitals

The causal interpretation of the hospital quality estimates implies that being admitted to hospital h leads to a change in normalized earnings of q_h (with respect to the baseline hospital 0). However, as previously described, we worry that patient selection on unobservables could bias these estimates. To mitigate these concerns, we test whether if our estimates are consistent once measured in a setting there patients are being admitted to different hospitals for "random" reasons.

We can learn about the robustness of our quality measure by testing a) whether an admission to a better hospital actually leads to a smaller drop in earnings, and b) whether the size of the effect is the same as predicted by equation (3) (i.e. exactly q_h).

To do so, we exploit two complementary empirical strategies. Each one deals in turn with one of the two main concerns described above. First, we use patients' home location as an instrument for their hospital choice. This strategy addresses the first concern, which is that the best hospitals treat patients who are unobservably more sick. That is, we relax the assumption that hospitals treat patients with (unobservably) similar characteristics (e.g. as the severity of the health shock). However, this strategy is valid only if the unobserved heterogeneity is homogeneous across neighborhoods within the country and, therefore, patients' location affect the dynamics of post-hospitalization normalized earning only through hospital choice.

Then we also consider the second concern, which could invalidate both the main specification and the IV estimates. That is, certain hospitals might be located closer to unobservably "better" (or more resilient) patients, and therefore admit more of them because of proximity. To address this concern, we focus on a small sub-sample of patients who are admitted to hospitals "far" from home for arguably random reasons. In this exercise, we find no evidence that local heterogeneity is significantly biasing our quality

measures.

Instrumental Variables

Let $\widehat{q_h}$ be the quality of hospital h estimated by applying OLS to equation (3). We consider the linear specification:

$$y_{i,t,d,h} = \alpha \cdot \widehat{q_h} + X_{i,t}\beta + \gamma_t + \phi_d + \sum_{0 \leq r \leq 3} \delta_{r \times dg} + \eta_{i,t,d,h} \tag{4}$$

It is worth comparing equations (3) (our main specification) and (4), which are identical except for q_h. In the former, q_h is a hospital fixed effect parameter. In the latter $\widehat{q_h}$ enters as a variable of observed hospital characteristic, while α is the parameter of interest. The set of controls included in (3) and (4) are the same, see section 3 for details. Estimation of (4) by OLS mechanically leads to $\alpha = 1$.

A valid instrument for hospital choice would allow us to identify the effect of being admitted to a hospital with estimated quality $\widehat{q_h}$. Recovering $\alpha = 0$ would then imply that the measure of quality is uninformative, as it only captures heterogeneity in patient mix between hospitals (e.g. differences in case severity). $\alpha > 0$ would imply that hospitals that are better according to our measure are indeed decreasing the cost of hospital admissions, although less than expected. $\alpha = 1$ is the value consistent with our main equation (3) and would imply that OLS estimate are not significantly biased by selection issues. In Appendix A, we present results from a simple simulated model to further clarify this exercise.

Location is the main driver of hospital choice in Denmark. In fact, more than 70% of patients in our sample are admitted to the "common" hospital, which is the hospital admitting the largest share of patients in a given neighborhood with the same medical condition. Therefore, we follow previous literature which use patient address as an instrument for hospital choice, see e.g. Geweke et al. [2003].

We estimate equation (4) using patient neighborhood (Church District) as an instrument for $\widehat{q_h}$, and we present results in column (1) of Table 2. We find that α is neither statistically nor economically different from 1. That is, being admitted to a better hospital (according to our measure of quality) decrease the economic impact of the hospital

admission exactly as much as predicted by the main specification (3). This suggests that selection issues do not affect the quality estimates described in section 3.

This empirical strategy is valid under the assumption that patients' location affect their post-hospitalization earning dynamics only through hospital choice. Although we include a large set of controls at both local- and the individual level, the residual unobserved heterogeneity might invalidate this IV approach. For instance, unobservably "better" patients might be clustered around better hospitals, or local labor markets might have unobservably different characteristics for different hospitals. Furthermore, other local healthcare characteristics, such as GP quality and rehabilitation procedures, could also influence the outcome of patients at a given hospital. The same holds true for differences in case-worker procedures, e.g. requirements for disability pension eligibility, which might also vary across municipalities and across time.

Acute admissions at non-common hospitals

As discussed above, the presence of local heterogeneity might lead OLS and IV estimate to be inconsistent for the true hospital quality. To understand the extent to which patient location reveals information about their resilience to health shocks, and therefore whether local heterogeneity should be a main concern, we focus on a specific subset of patients which are admitted to hospitals outside "far" from home for arguably random reasons.

This exercise is inspired by the identification strategy in Doyle [2011], who consider the hospital admissions of tourists in Florida who experience an acute health shock during their vacation. For our analysis, we consider the subsample of patients who are treated outside their local common hospital and who are admitted with an acute non-deferrable condition (using the more restrictive weekend share definition). Because of the nature of their diagnosis, they are more likely to be randomly admitted to an "uncommon" hospital, as the unexpected health shock is likely to have occurred while they were away from home, and not because they conscientiously chose a hospital further away. This strategy helps us overcome problems with selection- and differences in other local health care characteristics, but it also forces us to drop the vast majority of observations and only consider a small range of acute conditions and few patients.

We first re-run our main quality estimation specification on this subsample, and compare the OLS quality estimations with those in the full sample and the sample for all acute admissions. Figure 6 illustrate that the OLS estimates of hospital quality once we only consider those admitted in non-common hospitals with acute conditions are indeed similar to those estimated with the full sample (correlation coefficient of 0.55), and very similar to those estimated with the sample including all acute admissions (correlation coefficient of 0.69).

We then consider the following specification:

$$y_{i,t,d,h} = \alpha \cdot \widehat{q_h} + \tau_n + X_{i,t}\beta + \gamma_t + \phi_d + \sum_{0 \leq r \leq 3} \delta_{r \times dg} + \epsilon_{i,t,d,h} \tag{5}$$

where n is the neighborhood where patient i lives. Equation (5) differs from (4) as we also include of a set of neighborhood fixed effects (τ_n). The other control are the same as in section 3. The parameter α reveals how (estimated) hospital quality affects the normalized earnings of patients *conditional* on their home location. We estimate (4) by OLS using the acute admissions at non-common hospitals sample. Results are reported in column (2) of Table 2. We find $\alpha \approx 0.65$: being admitted to a better hospital significantly decrease the earnings cost of the hospitalization. The estimated magnitude is smaller than 1 (what predicted by the main equation 3). However, we cannot we cannot reject the null of $\alpha = 1$ at 10% confidence level ($p - value = 0.13$).

To investigate further whether local heterogeneity is affecting the quality estimate of section 3, we consider the alternative specification:

$$y_{i,t,d,h} = \alpha \cdot \widehat{q_h} + \theta \cdot \widehat{q_{n,dg}} + X_{i,t}\beta + \gamma_t + \phi_d + \sum_{0 \leq r \leq 3} \delta_{r \times dg} + \epsilon_{i,t,d,h} \tag{6}$$

where $\widehat{q_{n,dg}}$ is the estimated quality of the "common" hospital for patients of neighborhood (Church District) n and diagnostic group dg. The intuition behind this approach is that if part of what the quality measure captures is local heterogeneity (e.g. patients' unobserved characteristics or the presence of better GPs) rather than actual hospital quality, then patients coming from "better" neighborhoods should do better, even when they are treated outside the local hospital, all else equal. We focus again on acute admissions

at non-common hospitals. OLS estimates of equation (6) are presented in column (3) of Table 2. The estimated α is still close to 0.65, while θ is not different from zero: it is only the hospital where they are admitted–and *not* where they live–that matters for the economic impact of the hospitalization. Consequently, we find no evidence that the measures of hospital quality are affected by patients heterogeneity clustered at the local level or other local unobserved factors.

This exercise shows that there is no correlation between the estimated quality of a hospital and the earning dynamics of patients coming from the same area, when these patients are admitted to different hospitals for reasons arguably orthogonal to treatment quality itself. If the hospital quality was biased by local unobserved heterogeneity, this correlation should be positive.

Summarizing the results presented in this section, we conclude that the instrumental variable estimates suggest that selection issues are unlikely to significantly affect our quality measures, as long as local unobserved heterogeneity it is not too important. The last specification (equation 6) shows that local heterogeneity does not seem to significantly affect our results, mitigating the main concerns and supporting a quality interpretation of the across-hospital heterogeneity.

5 Healthcare Treatment Over Time

5.1 Empirical Strategy

To estimate the variation in treatment quality over time, we consider all inpatient "index" hospital admissions in Denmark during the years 1998-2012 (see section 2.1), together with a 10% control group of individuals without any index admissions during the same period. For each disease group, we consider the following equation:

$$y_{i,t,yg} = \alpha_i + X_{i,t}\beta + \gamma_t + \sum_{-3 \leq r \leq 3, r \neq -1} \delta_{r \times yg} + \epsilon_{i,r,t} \tag{7}$$

where α denotes individual fixed effects and $X_{i,t}$ is a vector of control variables. These includes age fixed effects, municipality- and education-specific unemployment rates and log median earnings in year t, and the share of individuals on disability benefits in the

municipality in year t. We also include dummies for whether the patient is age 40-50 or above age 50 when hospitalized, a dummy for weather a hospitalized individuals had manual labor in time $r = -1$, fixed effects for the number of out-patient admissions and ER visits during the previous three years, fixed effects for the number of co-diagnoses, and fixed effects for the ICD10 chapter of the primary co-diagnosis related to the index-admission. γ_t denotes calendar-year fixed effects. $\delta_{r \times yg}$ are the parameters of interest, and these denote the "time relative to hospitalization" fixed effects interacted with the year-group of treatment: 1998-2000, 2001-2003, 2004-2006, 2007-2009 and finally 2010-2012. All year groups have relative-time reference group $r = -1$. We identify the change in labor market consequences of a specific disease as the difference in $\delta_{r \times yg}$ for $r > 0$ across different year groups.

As we include a control group, we cannot use normalized earnings as the outcome measure as in Section 3. Instead, we consider two different outcome measures: 1) a dummy indicator for earnings larger than 50,000 DKK, and 2) log earnings, conditional on earnings being larger than 50,000 DKK. While the first measure captures the extensive margin labor participation consequences, the second captures the intensive-margin changes in earnings. Thus, the two measures should be interpreted together. If the extensive-margin drop in labor participation following a hospitalization decreases over time, this could potentially cause an increase in the intensive margin consequences over time, as better treatment could result in more individuals remaining in the labor force, but with lower earnings.

We categorize the different diseases by a clinical classification code, which groups our more than 9000 different 4-digit ICD10 codes into 271 diagnostic groups. Out of these, we consider the 150 groups with most observations, making up 97.5% of the total index admission observations, and each covering more than 850 index-observations.

A measure of the extensive-margin cost of a disease is given by $\delta_{r=3 \times yg}$ from equation 7, where the outcome variable is a dummy for whether earnings exceeds 50.000 DKK. We consider the relative time dummy for the 3rd year following the index admission, as the extensive-margin cost of a disease accumulates/continues to decline in the years following a hospital admission, as shown in Figure 1. As such, we pick the last year available within our 3-year index admission definition to capture the full accumulated effect of the different

diseases.

Our measure of the intensive-margin cost of a disease is given by $\delta_{r=1\times yg}$ as measured in equation 7, where the outcome variable is log earnings, and we only include earnings observations larger than 50,000 DKK. We consider the first year following the index admission, as this is when patients often experience the largest intensive-margin earnings loss, as earnings tend to recover for the less durable severities (as we document in section 2.3, Figure 1).

5.2 Results

Figure 7 shows the estimated $\delta_{r=3\times yg}$ coefficients for all 5 year groups, covering the period 1998-2012. The diagnostic groups are sorted by their estimated severity in year 2010-2012, $\delta_{r=3\times yg=2010-12}$. These coefficients measure the severity of the specific diseases, and the differences across the year groups capture the development of the disease's extensive-margin labor market consequences over time. A general finding across diseases is that there is a decline in the estimated severity over time, as the negative labor participation consequences of the disease decrease for the later year groups. For most diseases, we even observe a steady decrease in the estimated severities, despite the relatively short time intervals of 3 years. It also seems that the decrease accelerated over time.

According to our estimation, acute cerebrovascular disease is the most severe in terms of the labor participation drop following a hospitalization. In 1998-2000, 19% of patients were dropped out of the labor force 3 years after their hospitalization. This number, however, decreased to only 6% in 2010-2012. Spondylosis (spinal degeneration) is also extremely costly for labor participation, where it caused 13% to drop out of the labor force in 1998-2012. This number dropped to 11% in 2010-12. As such, the reduction in labor participation cost is much smaller for Spondylosis compared to acute cerebrovascular disease. Other very severe diseases include poisoning by other medications and drugs (including e.g. aspirin overdose), poisoning by psychotropic agents, alcohol-related mental disorders and multiple sclerosis, causing roughly 10% of patients to drop out of the labor force.

However, since we also keep the individuals who die, giving them zero earnings, some

of the change is driven by a decrease in mortality rates. Figure 11 repeats the same analysis, but exclude dead individuals. Here we find that 14% of the surviving patients dropped out of the labor force 3 years after their hospitalization for acute cerebrovascular disease in 1998-2000, compared to 9% in 2010-2012. These findings suggest that changes in mortality rates drive more than half of the estimated change in labor participation drop for this disease. Mortality explains a much smaller share for the less severe diseases.

Figure 8 shows the estimated intensive-margin severity measures given by $\delta_{r=1\times yg}$ in equation 7 for all 5 year groups, covering the period 1998-2012 and including all of the 150 most prevalent diagnostic groups. These are sorted by the estimated intensive-margin severity in year 2010-2012, $\delta_{r=1\times yg=2010-12}$. Compared to the extensive margin results presented in Figure 7, the intensive-margin costs of the different diseases are generally smaller. Also, while we do observe an overall decrease in disease severity over the years, the decrease is not as steady for the intensive margin as compared to the extensive margin. This, however, is not surprising, as a decrease in the extensive margin labor cost is likely to come at a "cost" with respect to the intensive margin labor cost. It seems reasonable to assume that individuals who would have previously dropped out in the labor force, but now remain, do so with declining earnings.

Figure 10 shows the scatter plot of the estimated intensive- vs. extensive margin severities for the year groups 1998-2000 and 2010-2015, together with the 45 degree line. The plot shows that the order of the estimated severities are somewhat similar for the two margins: if a disease have large extensive-margin costs, it also have large intensive-margin costs. We also see that severities have declined in both dimensions, but where the extensive margins severity tended to dominate in year 1998-2000, it seems that the severities are more equal for the two margins in 2010-2012. As such, the decline in severity over time was largest for the extensive margin. Again, this is not very surprising, and it is in fact noticeably that we observe such large decreases in the intensive-margin costs of the diseases, despite the fact that more individuals also remain in the labor force (large increases in the extensive margin).

For the extensive margin, we use a dummy variable for whether a person have positive earnings in a given year as the outcome variable and measure the change in extensive-

margin consequences of a specific disease as $\delta_{r=1\times yg=2010-2012} - \delta_{r=1\times yg=1998-2000}$. We consider a dead person as a person with zero earnings, so that the extensive margin can be affected by changes in mortality rates. However, results are quantitatively similar if we drop dead people from the sample.

While Figure 7 and Figure 8 plot the severity estimated in only one year ($r = 3$ and $r = 1$ respectively), we know that the labor market outcome consequences of diseases lasts for several years. As such, when we measure the "accumulated" or total costs of a disease, it is reasonable to summarize the costs over several years following the hospitalization. We will compute the total change in the labor costs of a disease as the accumulated labor-costs for each of the 3 years following a hospitalization:

$$\text{Accumulated Change} = \sum_{r=1}^{3} \delta_{r,yg=2010-2012} - \sum_{r=1}^{3} \delta_{r,yg=1998-200} \tag{8}$$

where the δ coefficients are given by equation 7. Figure 9 displays the estimated accumulated changes in severity over time, as measured by both the intensive and extensive margin, for all of the 150 most prevalent diseases. The plot confirms the previous finding that the change in estimated severity declined more on the extensive participation margin compared to the intensive log earnings margin. For the diseases with the most significant changes in severity over time, the decline in the accumulated extensive margin severity varied from roughly 7-27 percentage points, while the decline in accumulated intensive-margin severity were more constant around 4-6 percentage points.

According to our estimates, lung cancer was the disease for which the largest increase happened. During the three years following a hospitalization for lung cancer, the accumulated decline in labor force participation decreased by 27percentage points from 1998-2000 to 2010-2012 on the extensive margin. In general, we see that many diseases related to cancer have especially large declines in their labor cost over time, with secondary malignancies, breast- and colon cancer topping the list. The accumulated labor costs of acute cerebrovascular disease also declined by 20 percentage points on the extensive margin, and 10% on the intensive margin. For acute myocardial infarction (AMI), the change was 15 percentage points on the extensive margin and 8% on the extensive margin. Diseases that have seen only little or no significant decline in labor market costs are benign neoplasm

of uterus, varicose veins of lower extremity and abdominal hernia.

Weighted by the number of patients during all the observed years 1998-2012, we estimate that the average accumulated 3-year decline in participation rates following a hospitalization to have decreased by 8.8percentage points from 1998-2000 to 2010-2012. Similarly for the intensive margin, we estimate that the 3-year accumulated earnings decline following a hospitalization has dropped by 4.9 percentage points. This corresponds to an average decline in the labor force participation costs of a hospitalization of 50%, and a 25% reduction in corresponding intensive-margin earnings costs. The corresponding total change, computed as the sum of the intensive- and extensive margin changes, is 13.6percentage points.

5.3 Possible explanations for the decrease in the labor cost of hospitalizations

Part of the observed changes are probably caused by other factors than improved treatment quality. For instance, changes in the labor market and a general health improvement could also explain the observed variation in estimated severity over time.

A potential driver of the observed reduction in labor market consequences of hospitalization could be the introduction of the flex-job scheme in 1998. The flex job scheme provides wage subsidies for individuals with reduced working capacity, to enable partly disabled individuals to enter the labor force. As the scheme grew in popularity during the 2000's, it could explain part of the decrease in participation rates over time. As we're able to identify everyone who participate in the flex-job program in our data, we're able to run a "counter factual" robustness check to investigate whether the flex-job scheme itself is driving the observed improvement: we assume that all individuals who participated in the flex job scheme would have gone one full disability benefits instead, and have 0 earnings. Thereby we assume the "worst case" scenario of what would have happened, had the flex-job scheme not been introduced. Under this counter factual scenario, however, we do not find very different results. In fact, we estimate slightly larger declines in both the intensive- and extensive labor market responses following a hospitalization. As such, we do not expect the introduction of this particular scheme to drive our results. This

is probably due to the fact that the share of index-admission patients participating in the flex-job scheme is quite low. For index-admission patients hospitalized in 2000, 0.4% participated in the program in 1999 and 1% in 2001. For index-admission patients who were hospitalized in 2011, 3.3% participated in 2010 vs. 3.7% in 2012.

However, there have also been other labor market regulations targeted the sick, and several labor market reforms specifically aimed at decreasing sick leave were implemented during the period which we consider. Improved health and cultural changes (e.g. the common perception of how sick one should be to not work, or the pressure from the employers to return to work faster) could also explain some of the change.

While these factors might vary across disease severity - e.g. reforms targeting long-term sick leave, we expect them to have the same effect across different diagnostic groups, conditional on the disease severity. If we assume that the health care quality of all diseases have improved or remained constant over time, then we can compute a "back-of-the-envelope" lower bound estimate of how much of the improvement is due to increased treatment quality. Let the change in labor costs of a diagnostic group, $\Delta\text{LC}(dg)$ equal the sum of the change in treatment quality of the diagnostic group $\Delta\text{TQ}(dg)$ and other "labor market" factors, which are only specific to the severity of the disease $\text{LM}(\text{severity}(dg))$:

$$\Delta LC(dg) = \Delta TQ(dg) + \Delta LM(\text{severity}(dg)) \tag{9}$$

For each defined "severity" class, we consider the diagnostic group which had the lowest decline in estimated severity over time. Now, as a lower bound (we assume $TQ \geq 0$), we assume that the treatment quality of that given diagnosis did not improve at all ($\Delta\text{TQ}(dg) = 0$). Thus, we attribute the entire observed change in severity over time to other factors, ΔLM. As we assume that the contribution from ΔLM is severity dependent, but don't vary for different diagnostic groups *within* a severity group, we're able to identify $\Delta\text{LM}(\text{severity})$ as:

$$\Delta\text{LM}(\text{severity}) = \min_{dg \in \text{severity}} \Delta\text{LC}(dg) \tag{10}$$

Of course, the coarseness of the split in severities will affects how much of the change in labor cost which we attribute to improvements of treatment quality, with a finer spit leading to smaller changes in treatment quality. We consider the "total" labor cost to

be the sum of the extensive- and intensive margin labor costs, and the severity to be the sum of intensive- and extensive margin labor costs in 1998-2000. According to our estimation, the total 3-year accumulated labor costs declined by 13.6 percentage points over the period. Relying on the back-of-the envelope strategy presented in equations 9 and 10 and assuming a split of severities in 3 equally large groups, we find that 8.8 percentage points come from health care improvements. For 5 groups of severities, this number drops to 7.7 percentage points, and 10 groups gives us 5.3 percentage points. As such, a conservative estimate would imply that improvements in healthcare quality alone have caused the labor costs of a hospitalization to decrease with at least 5.3 percentage points.

6 Conclusion

We introduce a novel measure of the quality of a hospital treatment based on its ability to mitigate the labor market consequences of a given diagnosis. We measure the labor costs of hospitalizations by the drop in patients' earnings and labor supply, and we evaluate the quality of treatment based on its ability to mitigate the labor market consequences of a given diagnosis.

To do this, we link the universe of hospital admissions in Denmark to full-population tax data on the labor market outcomes of the entire working-age population in Denmark from 1995-2015. This allows us to measure the labor market consequences of any hospitalization, controlling for diagnosis, co-morbidities, and a very rich set of patient- and local labor market characteristics. To mitigate the problem of co-morbidities affecting the treatment outcomes, we consider all patients who experience their first non-pregnancy related inpatient admission in four years.

We find significant heterogeneity in the labor cost of a hospitalization across different hospitals, and our point estimates indicate that there is a 4 percentage point difference in lost earnings between the best and worst hospital, all else equal. Conservative estimates imply that bringing all Danish hospitals to, at least, the median quality, would lead to a saving of DKK 456 million per year in foregone earnings for inpatient admissions only. Rankings based on our measure are positively correlated with rankings based on

traditional quality metrics.

As for traditional mortality and readmission based measures, our quality estimates could be biased if unobserved patient characteristics are different across hospitals. Two main concerns that have been mentioned in the literature are 1) patients who are unobservably sicker select into the better hospitals, and thereby generate a downward bias on the estimated quality of the good hospitals, and 2) unobserved characteristics of patients are not evenly distributed across geographic regions and could bias the estimated quality of the hospitals located in unobservably "good"/"bad" neighborhoods. While we expect these concerns to be mitigated by the fact that we control for a broad range of patient characteristics (arguably broader than in previous papers), we perform two complementary empirical strategies to test whether being admitted to a higher quality hospital (according to our estimates) leads to a positive causal effect on post-admission earnings. Both tests suggest that the above mentioned concerns are not biasing our estimated quality measure in any significant way.

We also document a significant decline in the labor cost of hospitalizations over time. For different diseases, we compare the change in employment probability and labor earnings of patients after their hospital admissions in different years. We find that there has been a significant decline in both earnings and labor participation drops following a hospital admission during the period 1998-2012. More specifically, we find that the post-hospitalization reduction in labor supply has declined by 8.8% on the extensive (participation), and 4.9% on the intensive margin (earnings conditional on working) on average, as measured by the accumulated change in the three years following the admission. Thus, the total change in estimated labor costs amounts to 13.6 percentage points. This corresponds to an average reduction of extensive margin labor market costs of about 50%, and a 25% reduction of the intensive-margin labor cost of a disease. We consider the hypothesis that part of this decline is due to changes in labor market institutions (e.g., policies aimed to decrease sick leave) rather than hospital quality itself. A conservative estimate implies that healthcare improvements explain at least 40% of the estimated decline in labor cost.

References

De Nardi, M., Pashchenko, S., and Porapakkarm, P. The lifetime costs of bad health. Technical report, National Bureau of Economic Research, 2017.

Dobkin, C., Finkelstein, A., Kluender, R., and Notowidigdo, M. J. The economic consequences of hospital admissions. *American Economic Review*, 108(2):308–52, 2018.

Doyle, J. J. Returns to local-area health care spending: Evidence from health shocks to patients far from home. *American Economic Journal: Applied Economics*, 3(3):221–43, 2011.

Doyle Jr, J. J., Graves, J. A., Gruber, J., and Kleiner, S. A. Measuring returns to hospital care: Evidence from ambulance referral patterns. *Journal of Political Economy*, 123 (1):170–214, 2015.

Fadlon, I. and Nielsen, T. H. Family labor supply responses to severe health shocks. *NBER Working Paper*, 2015.

Finkelstein, A., Gentzkow, M., and Williams, H. Sources of geographic variation in health care: Evidence from patient migration*. *The Quarterly Journal of Economics*, 131(4): 1681–1726, 2016. doi: 10.1093/qje/qjw023. URL `http://dx.doi.org/10.1093/qje/qjw023`.

Geweke, J., Gowrisankaran, G., and Town, R. J. Bayesian inference for hospital quality in a selection model. *Econometrica*, 71(4):1215–1238, 2003.

Gilleskie, D. B. A dynamic stochastic model of medical care use and work absence. *Econometrica*, pages 1–45, 1998.

Gupta, A. Impacts of performance pay for hospitals: The readmissions reduction program. *WP*, 2017.

Hamilton, B. H., Hincapié, A., Miller, R. A., and Papageorge, N. W. Innovation and diffusion of medical treatment. *Technical report, Working Paper*, 2016.

Hull, P. Estimating hospital quality with quasi-experimental data. *WP*, 2018.

Kleven, H., Landais, C., and Søgaard, J. E. Children and gender inequality: Evidence from denmark. *NBER working paper*, 2018.

Miller, G., Luo, R., Zhang, L., Sylvia, S., Shi, Y., Foo, P., Zhao, Q., Martorell, R., Medina, A., and Rozelle, S. Effectiveness of provider incentives for anaemia reduction in rural china: a cluster randomised trial. *BMJ*, 345:e4809, 2012.

Tables

Table 1: Descriptive Statistics

	All Admissions			Index Admissions			Control
	r=-1	r=0	r=1	r=-1	r=0	r=1	-
Mean earnings	204,742	201,737	196,661	327,639	313,750	307,559	301,963
Mean log earn | earn $>$ 50,000	12.48	12.49	12.53	12.68	12.69	12.70	12.70
Share w/ earn $>$ 50,000	0.678	0.642	0.611	0.915	0.872	0.840	0.82
Share alive	1,000	0,984	0,966	1,000	0,989	0,981	0.97
Mean age	42.83	43.84	44.84	42.42	43.42	44.40	42.38
# Observations	3,104,532	3,104,632	3,099,490	1,098,162	1,113,321	1,111,576	3,312,609

The table shows descriptive statistics for three different samples. r specifies the year of the observation relative to the hospitalization year. We observe a total of 3.1 million non-birth related index admission observations of patients born in Denmark aged 28-56 at the time of hospitalization in year 1998-2012. This is the sample which we refer to as "all admissions". Counting only one admission per person per year, we have approximately 2 million observations. The index admission sample is the sample we use for our main analysis, and includes the subsample of patients who experience the first observed non-pregnancy related inpatient hospital admission for at least 3 years. The control group contains a 10% random sample of everyone who didn't experience an index admission, with similar birth-year and age restrictions as in the index-admission sample.

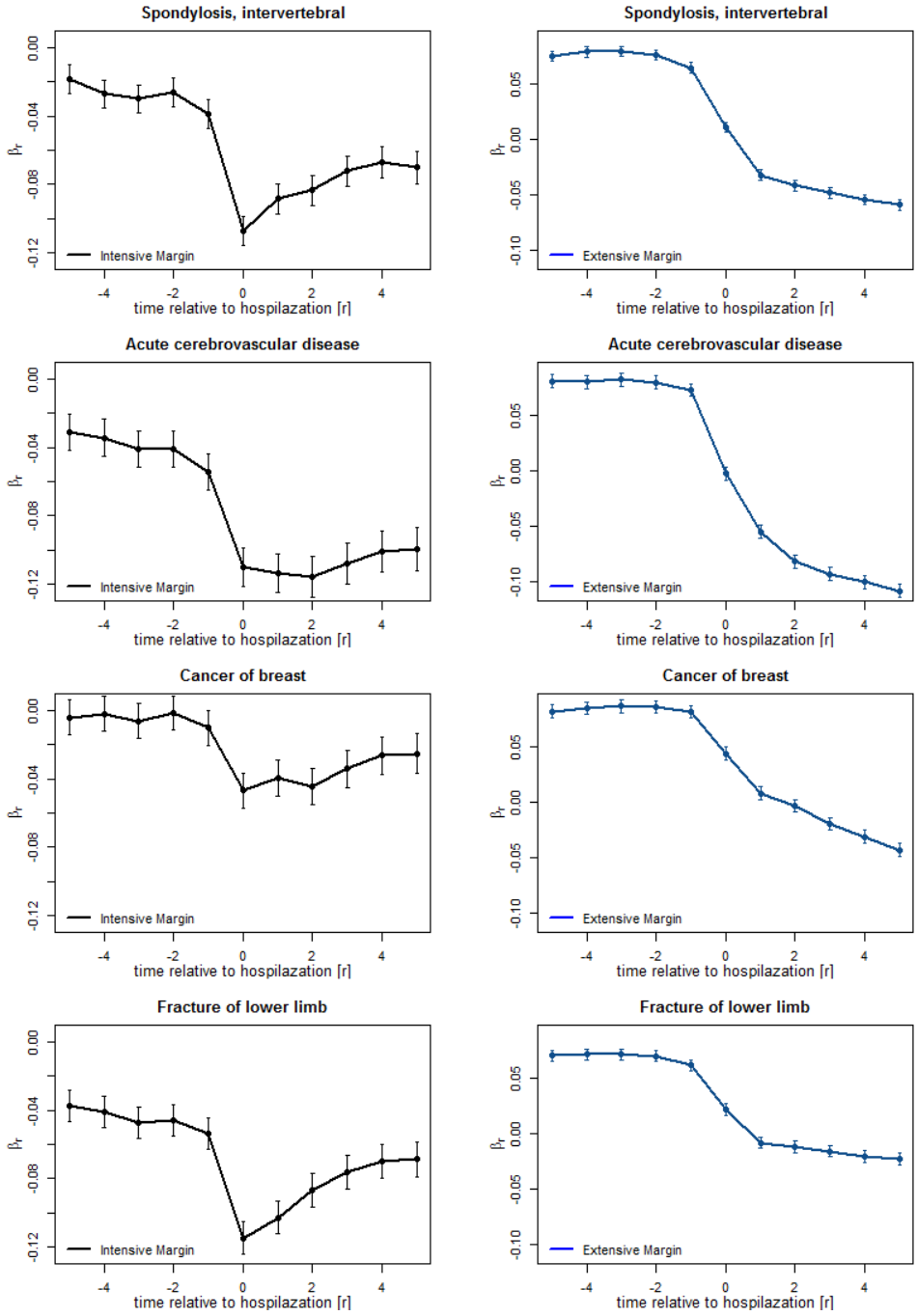
Table 2: Robustness of the hospital quality measure

VARIABLES	Normalized earnings	Normalized earnings	Normalized earnings
Estimator	IV	OLS	OLS
	(1)	(2)	(3)
Hospital Quality	0.989***	0.653***	0.690**
(estimated)	(0.0737)	(0.220)	(0.258)
Quality of "Common"			0.0225
Hospital (estimated)			(0.271)
Church District FEs		×	
Controls	×	×	×
Observations	3,160,852	74,768	64,549
Sample	All	Non-deferrable acute	Non-deferrable acute
	(top 31 hospitals)	not at "Common"	not at "Common"
		hospital	hospital
R-squared	0.143	0.254	0.209

Column (1): results of estimating equation $y_{i,t,d,h} = \alpha \cdot \widehat{q_h} + X_{i,t}\beta + \eta_{i,t,d,h}$, where y are the normalized earnings of patient i in year t. The patient has been admitted to hospital h in a year between t and $t-3$ with a main diagnosis d. Labor earnings are normalized by the average labor earnings during the three years before the hospital admission. $\widehat{q_h}$ is the estimated quality of the hospital h, see equation (3). $\widehat{q_h}$ is treated as an endogenous variable and it is instrumented with a vector of dummy variables indicating the neighborhood (Church District) where patient i lives during the admission year. For sample selection criteria and details about the vector of controls, see sections 3 and 4 and, in particular, equation (4). Column (2): results of estimating equation $y_{i,t,d,h} = \alpha \cdot \widehat{q_h} + \tau_n + X_{i,t}\beta + \epsilon_{i,t,d,h}$. It differs from column (1) because of the inclusion of a full set of fixed effects to control for the neighborhood where patient i lived during the hospitalization year. Sample includes only patients admitted for non-deferrable acute conditions to any hospital different from the most common one among patients coming from the same neighborhood and having similar diagnosis. Estimation performed via OLS. For details, see section 4 and equation (5) in particular. Column (3): results of estimating equation $y_{i,t,d,h} = \alpha \cdot \widehat{q_h} + \theta \cdot \widehat{q_{n,dg}} + X_{i,t}\beta + \epsilon_{i,t,d,h}$. It differs from column (2) because the neighborhood fixed effects are substitute with $\widehat{q_{n,dg}}$, which is the estimated quality of the most common hospital for patients coming from the same neighborhood as i and having similar diagnosis. Observations in column (3) are less than column (2) because the quality is estimated only for the top 31 hospitals. Estimation performed via OLS. For details, see section 4 and equation (6) in particular. Standard errors in parentheses are clustered at the municipality level.*** p<0.01, ** p<0.05, * p<0.1

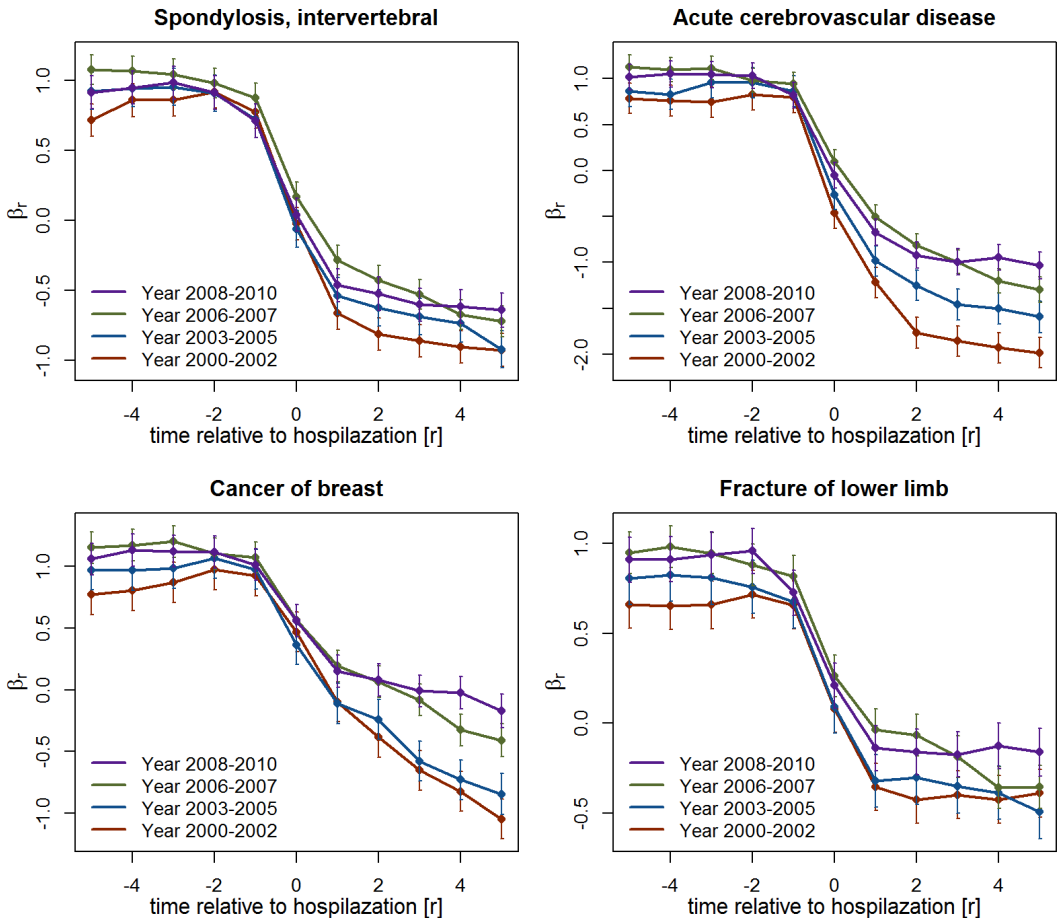
Figures

Figure 1: Labor supply response to an index hospital admission



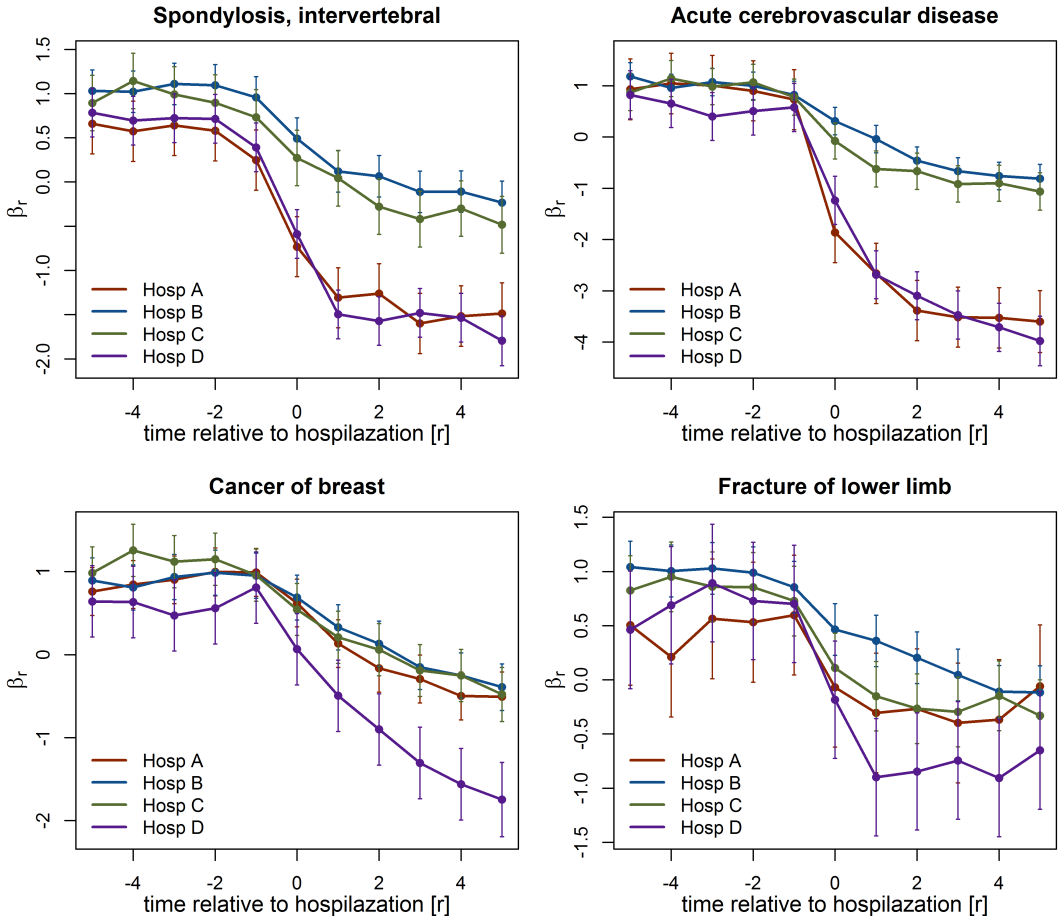
The intensive margin plots show the β_r coefficients from equation 1 with outcome variable $log(earnings)$ (excluding non-positive earnings observations). The intensive margin plot show the β_r coefficients from equation 1 with the dummy outcome variable for positive earnings. The β_r coefficients estimate the differences in labor market outcome for individuals hospitalized in year $t - r$ with respect to other (non-hospitalized) individuals with similar characteristics. Hospital admission year is 0. See section 2.3 for details.

Figure 2: Labor supply response to an index hospital admission: heterogeneity over time



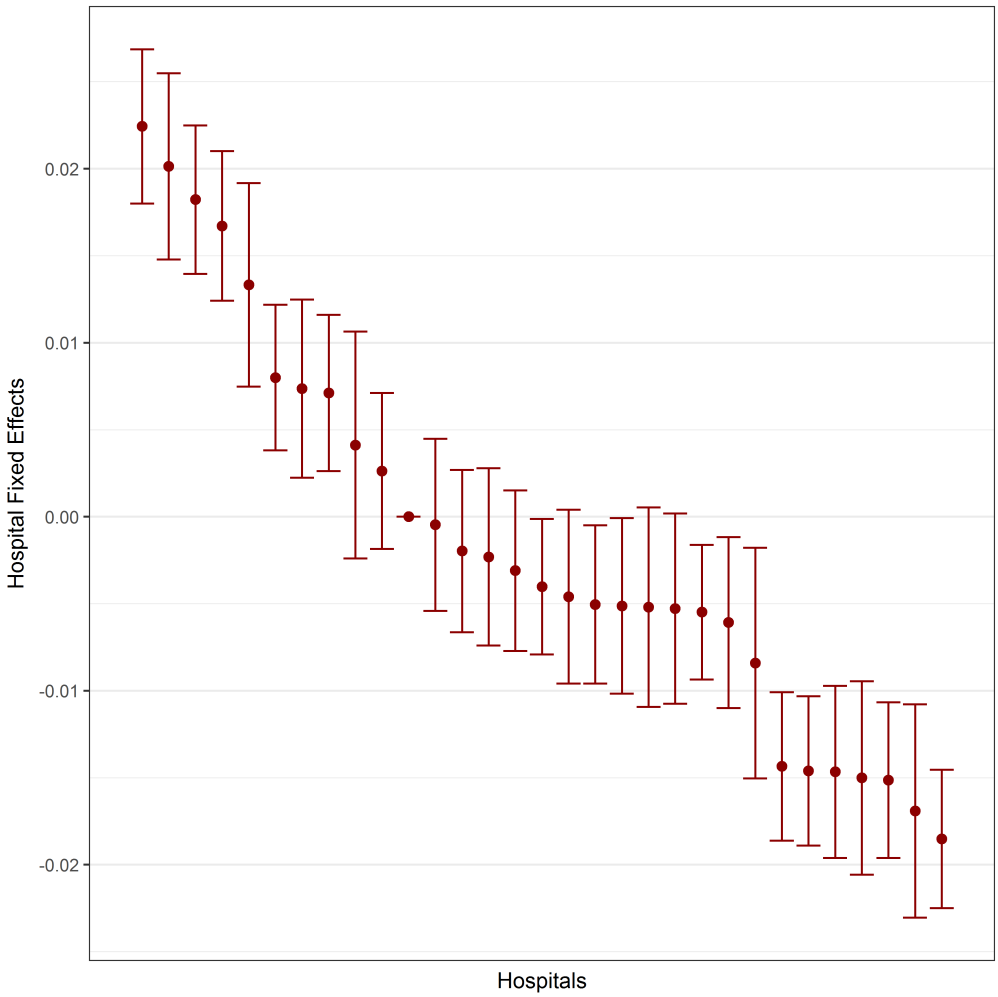
Event Study of log earnings before and after an "index"-hospital admission. Hospitalization year is 0. Each line refers to patients hospitalized in a given period. See section 2.3 for details.

Figure 3: Labor supply response to an index hospital admission: heterogeneity across hospitals



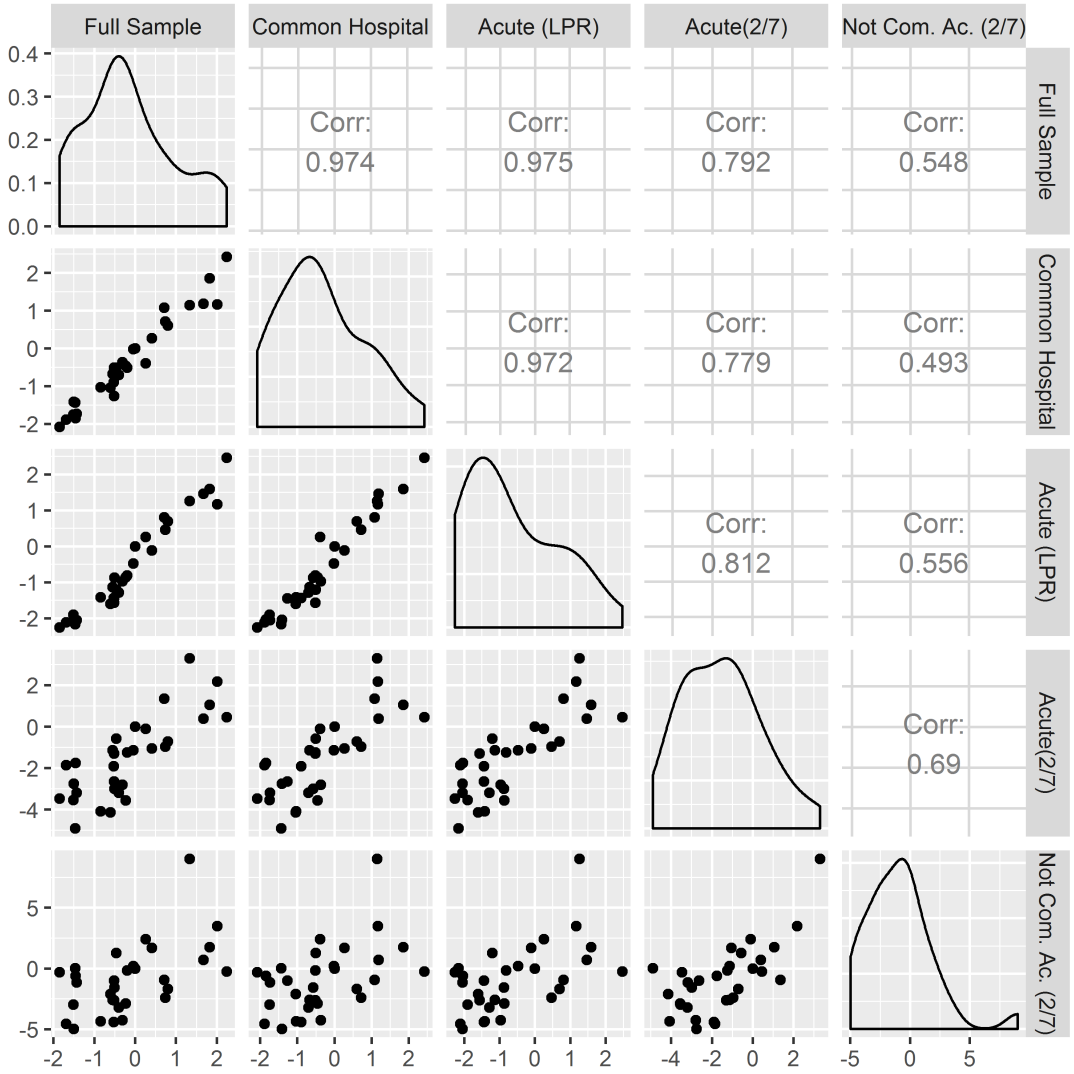
Event Study of log earnings before and after an "index"-hospital admission. Hospitalization year is 0. Each line refers to patients hospitalized in a given hospitals. See section 2.3 for details.

Figure 4: Estimated Hospital FEs on full sample



The hospital fixed effects are estimated using the full sample of 1,175,981 index admission observations. The OLS regression specification is described in Section 3.1

Figure 5: Correlation Matrix of Subsample Robustness Checks



The figure shows the correlation matrix plot of the estimated hospital FEs using different subsamples for robustness check.

Figure 6: Comparison of full sample and acute not common hospital FEs

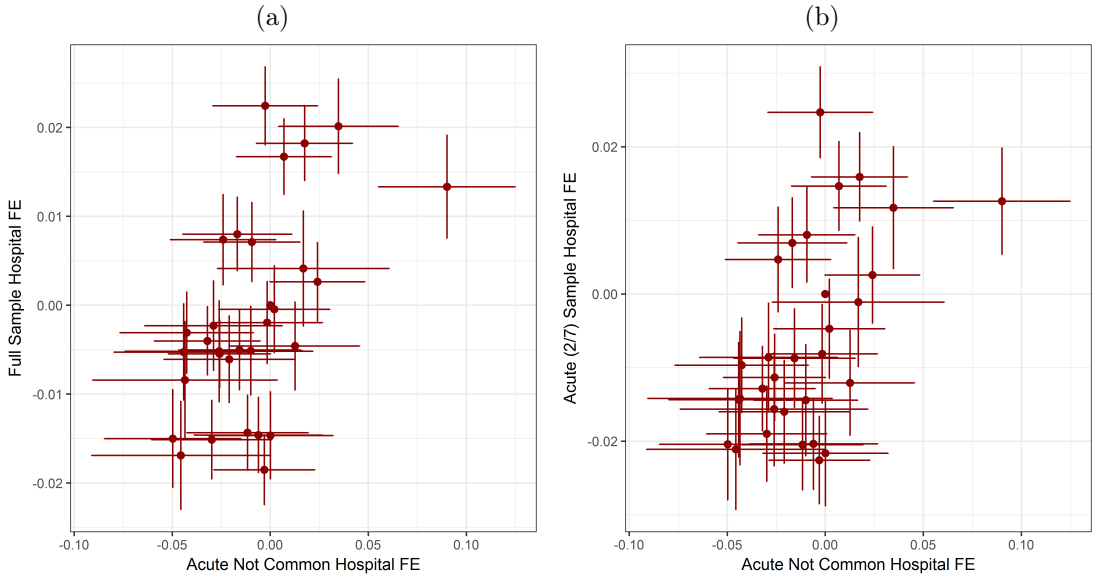
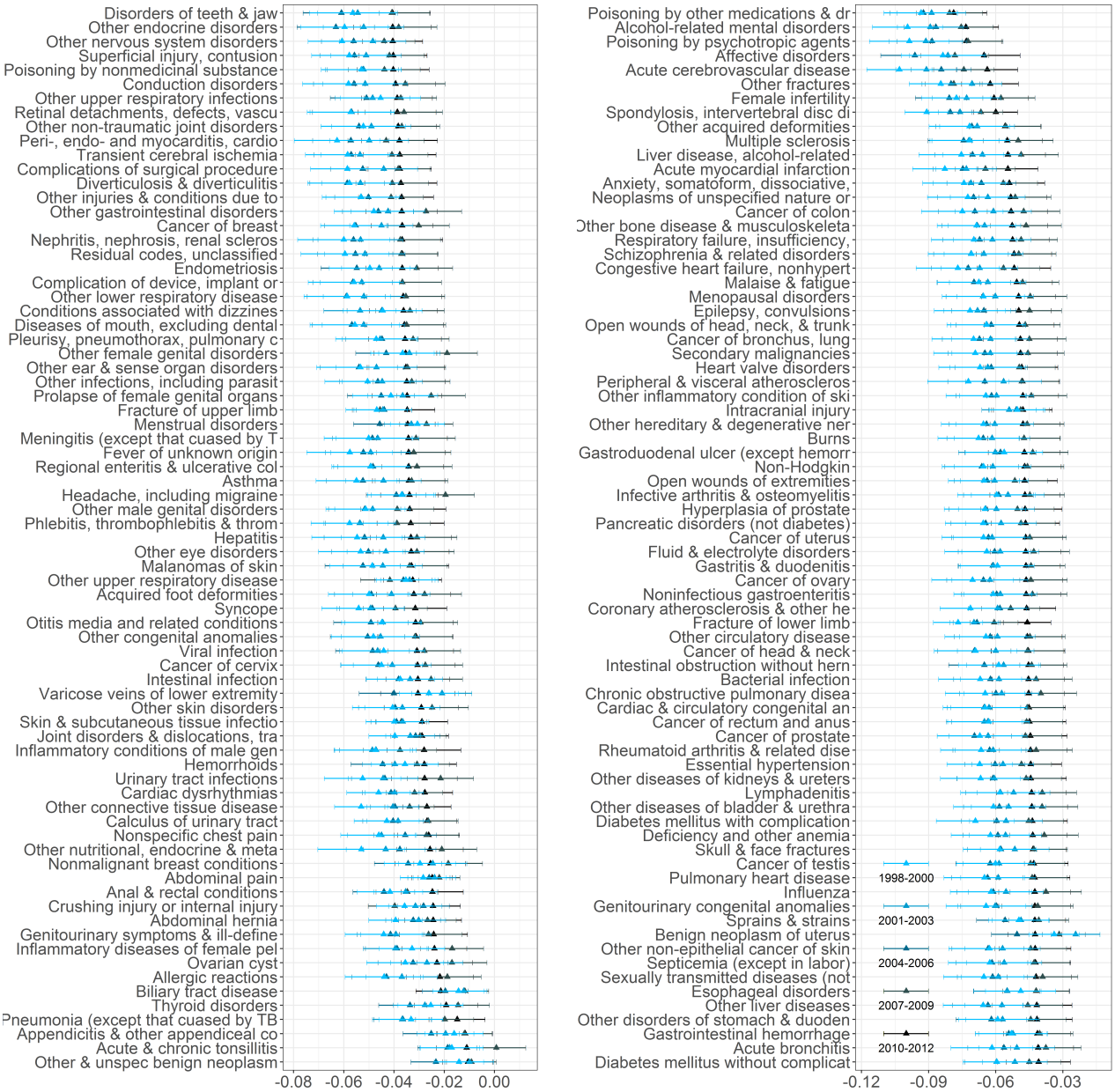


Figure a) shows the scatter plot of the hospital FE estimated on the full sample (vertical axis) and the subsample of acute patients treated outside their local default hospital (horisontal axis), with standard errors.

Figure 7: Estimated labor force participation loss three years after a hospitalization by disease and year of treatment

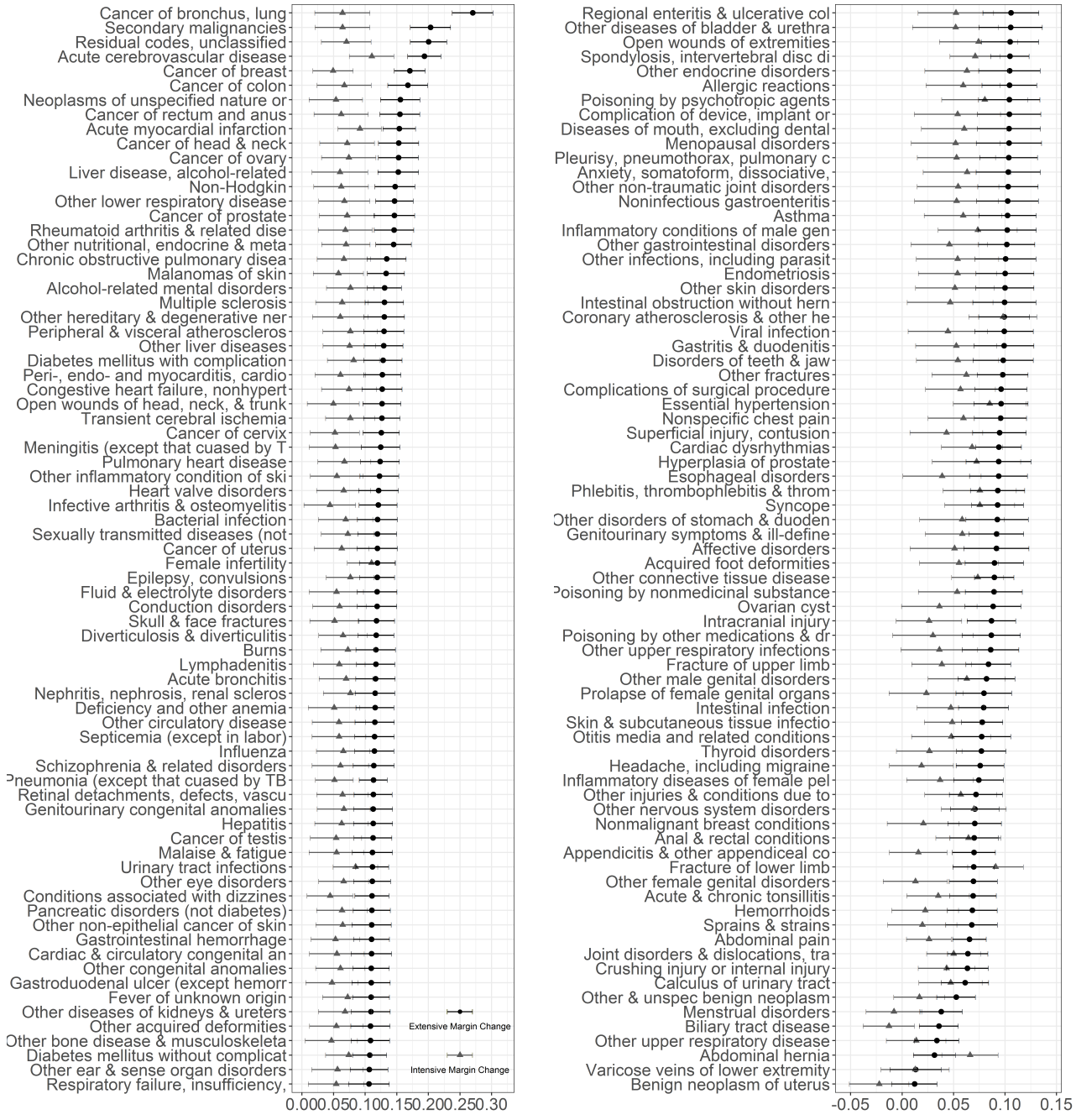
The figure shows, for each of the 150 most frequently observed disease groups, the difference in participation rates 3 years after an index admission. This difference is estimated in equation 7, where the outcome variable is a dummy for whether if earnings in a given year exceeds 50,000 DKK, and equals the estimated parameter of $\delta_{3\times yg}$.

Figure 8: Estimated earnings loss in year r=1 by disease and year of treatment



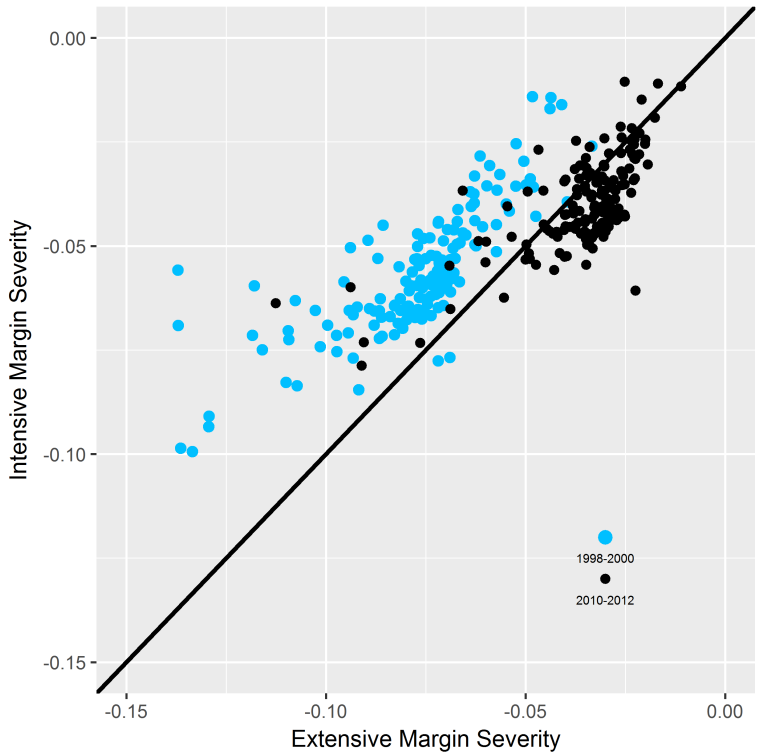
The figure shows, for each of the 150 most frequently observed disease groups, the difference in participation rates 1 year after the index admission. This difference is estimated in equation 7, where the outcome variable is log earnings, conditional on earnings exceeding 50,000 DKK, and equals the estimated parameter of $\delta_{1\times yg}$.

Figure 9: Change in accumulated labor market cost of a hospitalization by disease group, 1998-2001 to 2010-2012, $r \in \{1, 2, 3\}$



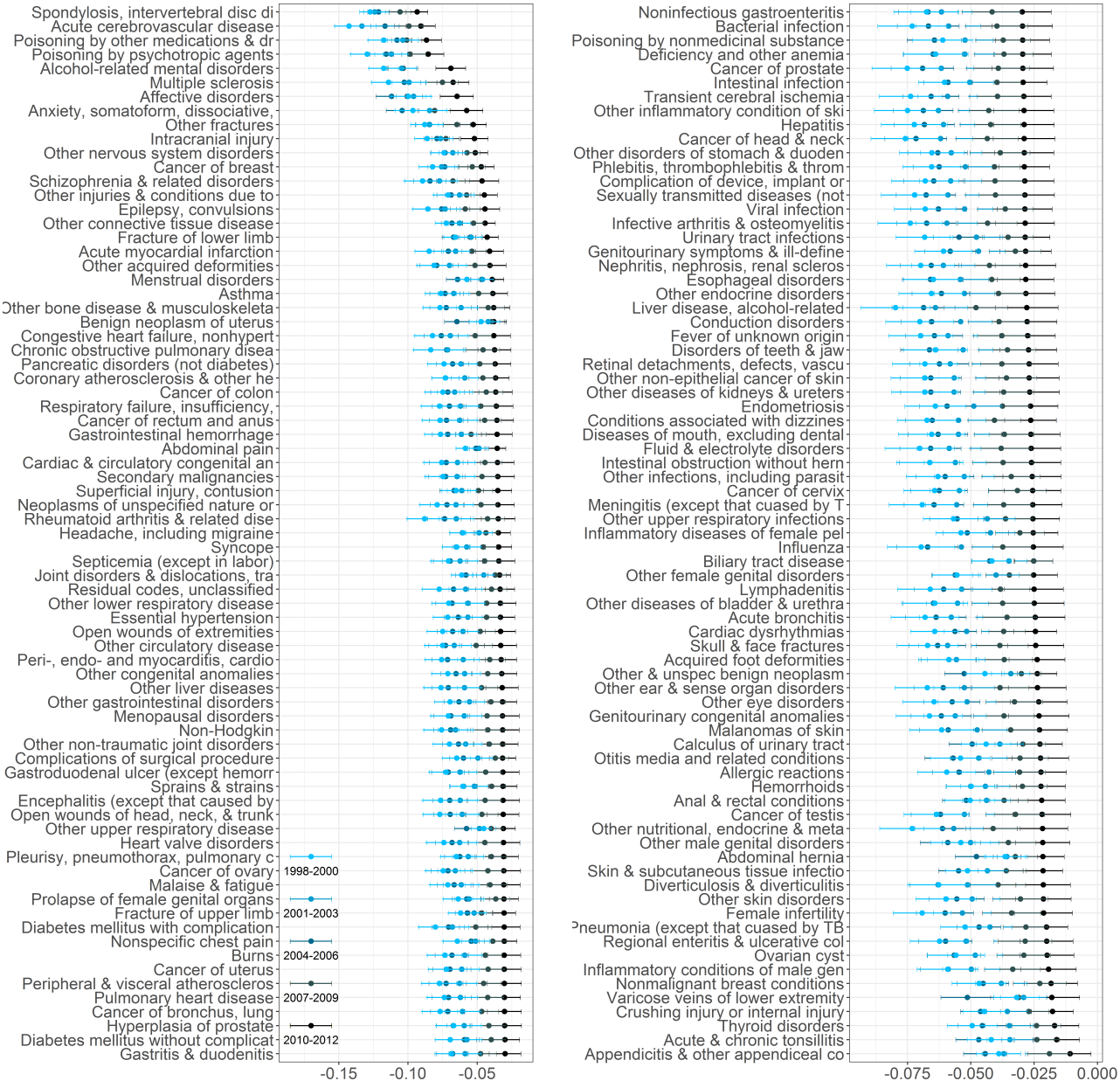
The figure shows the estimated accumulated changes in severity over time, both for the intensive and extensive margin, for all of the 150 most prevalent diseases. We meassure the accumulated change as $\sum_{r=1}^{3} \delta_{r,yg=2010-2012} - \sum_{r=1}^{3} \delta_{r,yg=1998-200}$, where the δ coefficients are estimated in equation 7. The black dots denote the change in β_5 over time with 95% confidence intervals, and the gray triangles and lines denote the difference as measured by the intensive margin (log earnings conditional on working).

Figure 10: Histogram of estimated pathology-specific improvements, 2000-2002 to 2008-2010



The scatter plot plots the estimated severity measured by its intensive-margin labor costs vs. its extensive margin labor cost. The y-axis denotes the intensive cost and is measured in equation refeq: time, where the outcome variable is log earnings, conditional on earnings exceeding 50,000 DKK, and equals the estimated parameter of $\delta_{r=1\times yg=2010-2012}$. The x-axis denotes the extensive cost of the disease and is measured in equation refeq: time, where the outcome variable is a dummy for whether if earnings in a given year exceeds 50,000 DKK, and equals the estimated parameter of $\delta_{r=3\times yg=2010-2012}$

Figure 11: Estimated labor force participation loss in year r=3 by disease and year of treatment, excluding deaths



The figure shows, for each of the 150 most frequently observed disease groups, the difference in participation rates 3 years after an index admission. This difference is estimated in equation 7, where the outcome variable is a dummy for whether if earnings in a given year exceeds 50,000 DKK, and equals the estimated parameter of $\delta_{r=3\times yg}$

A Location IV: Simulation

We simulate a simple model of endogenous hospital selection and heterogeneous hospital quality to clarify the exercise presented in section 4 (equation 4).

We consider a population of N patients, who are equally divided between neighborhoods A and B. There are two hospitals in this county, hospital 1 with quality $q_1 > 0$ and hospital 0 with quality normalized to 0. The two neighborhoods differ only because A is closer to hospital 1, while B is closer to hospital 0. The normalized earnings for patient i admitted to hospital h is:

$$Y_i = c_0 + q_1 \cdot \mathbf{1}_{h=1}(i) + \epsilon_i \tag{11}$$

where $1_{h=1}(i)$ is a dummy variable indicating whether the patient i is admitted to hospital 1 (takes value 0 if admitted to hospital 0), c_0 is a constant, and ϵ_i is a normally distributed random shock.

We assume the hospital selection process takes the form:

$$\mathbf{1}_{h=1}(i) = 1 \iff \Phi\left(\eta_i + c_1 \cdot \mathbf{1}_{n=A}(i) + G \cdot \epsilon_i\right) > \frac{1}{2}$$

where $\Phi(\cdot)$ is the CDF of a standard normal distribution, $\mathbf{1}_{n=A}(i)$ is a dummy variable indicating whether patient i lives in neighborhood A, η_i is a random shock affecting hospital choice (uncorrelated with ϵ_i). $c_1 > 0$ means that patients from neighborhood A are more likely to go to hospital 1.

The parameter G disciplines endogenous hospital selection, that is whether hospital 1 admits "worse" or "better" patients. If $G = 0$ the patients admitted to both hospitals are equal. If $G < 0$ then the best hospital admits sicker patients. If $G > 0$ then it admits more resilient patients.

As in section 3, the researcher estimates q_h by applying OLS to equation (11). Let the resulting estimate be $\widehat{q_h} = \widehat{q_1} \cdot \mathbf{1}_{h=1}(i)$. Then, as in section 4, she uses patient location as an instrument for hospital quality, in order to test the robustness of her OLS measure. That is, she consider the linear model:

$$Y_i = a_0 + \alpha \cdot \widehat{q_h} + \zeta_i$$

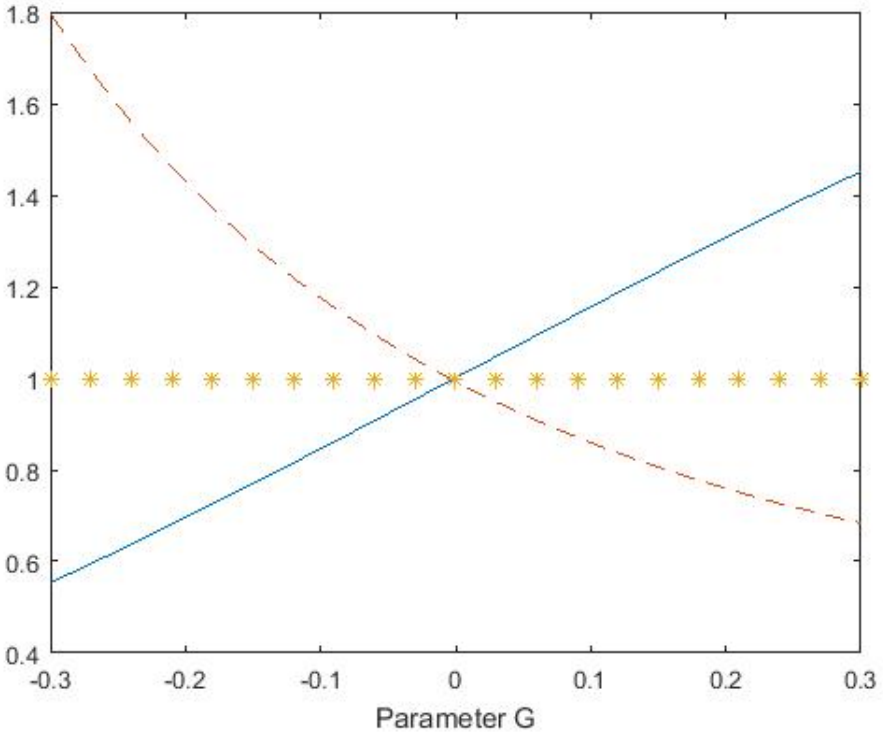
She instruments the variable $\widehat{q_h}$ with the dummy $\mathbf{1}_{n=A}(i)$. Let the resulting estimate for α be α_{IV}, while the OLS estimate is α_{OLS}. How can we use these estimates to get information on the underlying selection process?

We set $q_1 = 1$ and choose arbitrary values for the other parameters.[7] We let the selection parameter G take value on a grid from -0.3 to 0.3. We simulate the model 200 times for each value of G, compute $\widehat{q_1}$, α_{IV}, and α_{OLS} and plot the sample average.

Results are presented in Figure 12. The solid line indicates the average $\widehat{q_1}$. As expected, OLS estimates of quality recover the true parameter only in absence of endogenous selection (i.e. $G = 0$). When $G < 0$, the quality of hospital 1 is underestimated because it treats sicker patients, while when $G > 0$ it is overestimated. The stars indicates α_{OLS}, which is mechanically equal to 1 for all simulations.

[7] Namely, $c_0 = 1$, $c_1 = .5$, $N = 20,000$, ϵ and η distributed as *iid* standard normal.

The dotted line represents the average α_{IV}. As expected, when there is no endogenous selection, and the estimated quality is unbiased for the true quality, the parameter is equal to 1. That is, being admitted to a hospital with quality $\widehat{q_1} = 1$ has actually had an effect of size $\widehat{q_1}$. When $G > 0$ (and therefore $\widehat{q_1} > q_1$), the average α_{IV} is below 1; in fact, being admitted to the good hospital has an effect smaller than what is predicted by the OLS quality estimates. The opposite is true when $G < 0$.



Results of the simulated IV model described in Appendix A. The solid line indicates the average estimated quality $\widehat{q_1}$, the dotted line represents the average α_{IV}, and the stars represent the average α_{OLS}. Averages are taken over 200 simulations of the model for each value of the endogenous selection parameter G (on the x-axis).

Chapter 3

Intergenerational Income Mobility in Denmark and the U.S.

Intergenerational Income Mobility in Denmark and the U.S.

Anne-Line Koch Helsø

Copenhagen University

Abstract

Abstract: I present a novel cross-country comparison of intergenerational income mobility in Denmark and the U.S. Unlike existing studies, I rely on high-quality administrative data for both countries. I find that Denmark is 50-100% more mobile than the U.S. While the difference varies across specifications, it is stable across different income measures. I contrast my findings to the existing literature which finds larger cross-country differences and show that my results are more robust to sample selection and measurement error biases.

Keywords: Comparative analysis of systems; inequality; social mobility

JEL classifications: I28, I32, J62, P57

1 Introduction

Social mobility has become a highly salient issue for policymakers in recent years where researchers have considered cross-national rankings of mobility to understand how different policies associate with differences in equality of opportunity across nations.

Due to recent access to U.S. tax data, I am able to present the first comprehensive cross-country comparison of intergenerational income mobility in Denmark and the U.S. which is based on administrative data for both countries. Specifically, I consider the findings in Chetty et al. [2014] and Mitnik et al. [2015]. I closely follow their applied sample selection criteria, income concepts and family definitions in order to compute directly comparable mobility estimates for Denmark using administrative data.

I find consistent evidence of significantly higher intergenerational income mobility in Denmark compared to the U.S. My findings comprise many different specifications and measures of intergenerational income mobility and suggest that mobility in Denmark is roughly 50-100% higher compared to the U.S.

Cross-country differences vary across specifications and mobility measures: I find the largest cross-country differences in intergenerational income elasticities (IGEs) when children's family income is compared to parents' family income - for this specification, Denmark is 100% more mobile as the U.S. for each gender. For children's individual income, I find the largest cross-country differences for sons ($\approx 50\%$) compared to daughters ($\approx 25\%$), which could be caused by cross-country differences in female labor force participation rather than differences in equality of opportunities. For positional rank mobility estimates, cross-country differences are roughly halved.

Albeit large and significant, the estimated difference in mobility of around 50-100% is significantly smaller compared to previous studies which find that mobility in Denmark is 3 to 5 times higher compared to the U.S.[1] One explanation is that the Danish registry data, starting in 1980, has only recently reached a long enough time span to efficiently

[1]See e.g. Jantti et al. [2006], Hussain et al. [2009], Corak [2013] and Bratberg et al. [2017]

measure the income persistence between two generations[2]. As such, previous mobility measures have overstated the level of mobility in Denmark. With Danish data covering a period of 35 years from 1980-2015, I am able to better overcome the attenuation- and selection biases that have influenced previous studies based on fewer years of data. I find that approximately 24% of fathers' earnings advantage is passed on to their sons, which is about 60-100% higher than previous estimates for Denmark.

While not the main focus of their paper, Landersø and Heckman [2017] measure both parents and children at optimal ages, but still find that Denmark is 3.5 times more mobile than the U.S. when measured by wage earnings. They also find that cross-country differences vary greatly for income measures, e.g. including/excluding capital income, taxes and public transfers. I show that the variation across income mobility measures is driven by the specific measurement choices in Landersø and Heckman [2017], including cross-country differences in income specifications and sample selections. As my cross-country comparisons is based on on highly comparable sample selection criteria, income concepts, family definitions and data qualities for the two countries, my findings of a 50-100% difference is robust to the choice of income measure.

2 Measuring Intergenerational Mobility

The literature studying intergenerational income mobility is long and encompasses a wide variety of approaches seeking to measure the degree to which a child's social and economic opportunities depend upon that of his/her parents. The aim is to estimate the dependence between lifetime income or earnings of parents and offspring, with strong associations implying a low degree of intergenerational mobility and vice versa.

The canonical mobility measure, de facto a measure of *income persistence*, is the elasticity of child income with respect to parent income, commonly termed the *intergen-*

[2]The same problem does not apply to most U.S. studies, as these rely on survey data starting in the late 1960s and 70s.

erational elasticity of income (IGE). A high IGE coefficient is equivalent to a low degree of mobility (high degree of income persistence), and vice versa. The IGE is estimated by regressing log child income ($\ln Y^C$) on log parent income ($\ln Y^P$):

$$\ln Y^C = \alpha + \beta^{IGE} \ln Y^P + \epsilon \tag{1}$$

The slope coefficient β^{IGE} provides a measure of intergenerational persistence, with a higher value of β^{IGE} corresponding to a lower degree of mobility. β^{IGE} can also be expressed in terms of the correlation and relative standard deviation of parents' and children's log income:

$$\beta^{IGE} = Corr(\ln Y^C, \ln Y^P) \frac{SD(\ln Y^C)}{SD(\ln Y^P)} \tag{2}$$

The above equation clearly shows how the IGE is affected by the marginal income distributions of each generation. If inequality is rising over time, such that $SD(Y^C) > SD(Y^P)$, this will result in increasing IGE estimates.

A comprehensive literature has investigated the possible biases that might result in downward biased estimates. The literature agrees that the ideal measure of income should approximate permanent income, averaging yearly income observations over several years [Solon, 1992]. As measurement error in an explanatory variable is known to lead to *attenuation bias*, Solon [2002] and Mazumder [2005] suggests that as many as 9 years (and preferably more) of parental income observations are needed to avoid downward bias due to measurement error of permanent income due to transitory income shocks. Grawe [2006] and Haider and Solon [2006] show that *life cycle bias* (rooted in heterogeneous age-income profiles) is best avoided when income is measured at mid-age (between early thirties and mid forties) for both parents and children.

As zero-income observations are excluded from the analysis, estimating earnings based on too few observations leads to *selection bias*, as individuals with weaker labor force attachments are more likely to be excluded from the analysis, as emphasized by Mitnik

et al. [2015]. As the conventional IGE estimator in fact pertains to the *geometric* mean of children's income rather than the expectation or arithmetic mean, Mitnik and Grusky [2017] propose an alternative estimator: the IGE of expectations, IGE_E (see appendix A.2). The IGE_E correctly computes the impact on *log expected* child income of an increase in log parental income, but more importantly, the IGE_E estimator allows us to include zero-income children in the sample and thereby overcome the sample selection bias embedded in the conventional IGE estimate.

Another popular mobility measure is the rank-rank correlation, which is a measure of *positional* mobility. The rank-rank correlation is equal to the Spearman correlation between parent and child income, but one can also interpret the coefficient as the expected increase in child income rank when parental income rank increases by one. An advantage of the rank-rank correlation is that it allows for the inclusion of zero-income children and parents.

3 Cross-Country Comparisons using Administrative Data from both countries

Previous cross-country comparisons of income mobility in the U.S. and the Scandinavian countries rely on different types and qualities of data. They compare high-quality full population Danish administrative data to few U.S. survey data observations with high attrition and measurement error. Also, they often compare different cohorts and ages of measurement for the two countries, as the U.S. survey data such as the PSID an NLSY start in the late 1960s and 1970s while the Danish register data started in 1980. However, recent studies based on high-quality U.S. administrative data have enabled comprehensive cross-country comparisons of income mobility in Denmark and the U.S. that use administrative records for both countries. In this section, I carefully imitate the sample selection rules, income concepts, parent/child and spouse definitions applied in Chetty et al. [2014] and Mitnik et al. [2015] and compute directly comparable mobility

estimates for Denmark. Income levels are deflated using country-specific CPIs, and I use a PPP adjusted exchange rate of $100 to DKK 776.

3.1 Comparing to Chetty et al. (2014)

Chetty et al. [2014] use as many as 9.9 million parent/child observations and compare children's family income to that of their parents. Parents and children are matched by dependent claiming, and spouses are identified by joint tax filing. Parents' income is measured as mean family income between 1996 and 2000, when their children are between the ages of 15 and 20.[3] Child family income is measured in 2011-2012 when the children are in their early thirties. Their income measure is total pre-tax income and includes labor earnings (payroll and self-employment income), capital income, unemployment insurance, social security- and disability benefits. As such, their measure excludes non-taxable cash transfers, such as TANF (temporary assistance for needy families), SSI (supplemental security income), in-kind benefits such as food stamps, and all refundable tax credits such as the EITC (earned income tax credit).

For Denmark, I define the parents to be the legal parents living with the child in 1996. This definition results in 30.5% single parents compared to 30.6% in Chetty et al. [2014]. The mean age of fathers in 1996 is 43.5 in the U.S. sample and 44.0 in the Danish sample. Mothers' mean ages are 41.1 in both samples. Spouses, both in the parent and child generation, are defined by marriage, resulting in 65.5% of children without spouses in 2012, compared to only 44.3% in Chetty et al. [2014].[4]

A perfect imitation of the total pre-tax income definition used in Chetty et al. [2014] does not exist, due to fundamental differences in benefit schemes and eligibility rules in the two countries. To ensure that my results are robust to the choice of income measure

[3]For parents who are initially married and then divorce, this measure tracks the mean family income of each parent over time. For a parent who is initially single and then marries, the family income measure tracks individual income before, and total family income after, marriage

[4]If cohabiting partners are also considered spouses, I get 30.1% of single children in 2012. Results are robust to using a spousal definition that includes cohabiting partners, see Appendix Table A1

used for Denmark, I consider three different measures of income:

- DK_{chetty}: labor earnings, capital income, unemployment insurance, retirement- and disability benefits
- DK_{trans}: labor earnings, capital income, and *all transfers*
- DK_{no_trans}: labor earnings and capital income

Here, labor earnings include both payroll income and business profits from self employment. While the income definition DK_{chetty} best resembles the one in Chetty et al. [2014], I also consider alternative income measures which include all (DK_{trans}) and no (DK_{no_trans}) transfers as robustness checks. The main difference between the income definitions DK_{chetty} and DK_{trans} is that DK_{trans} includes cash assistance (kontanthjælp), leave benefits (orlovsydelse) and child benefits (børnepenge), while DK_{chetty} does not. Using the DK_{chetty} income definition, the resulting data consist of 151,367 observations with average child- and parental income of above $\$10^5$ and 157,543 observations once zero-income children are included.

In Table 1 I present the estimates comparable to those in Chetty et al. [2014] Table I. I find similar results across all three income specifications for Denmark, and the estimated coefficients for Denmark are consistently lower compared to their U.S equivalents, indicating a higher degree of intergenerational income mobility in Denmark. IGE and IGE_E estimates of parents' and children's family income are approximately twice as large for the U.S., with slightly lower IGE estimates for women compared to men. When restricting the sample to observations between the 10th and 90th percentile of parental income, IGE and IGE_E estimates increase for both countries, due to non-linearities as illustrated in Figure 1.

[5]In the Danish data, interest income is 3rd party reported to the tax authorities by financial institutions, resulting in quite a few interest-only income observations of just a few DKKR for individuals with no earnings. I exclude these observations to match the effective income criteria in Chetty et al. [2014] who do not include interest income of non-filing individuals.

Table 1: Comparing to Chetty et al. [2014]

Sample	**Estimator**	**U.S.**	DK_{chetty}	DK_{trans}	DK_{no_trans}
All	IGE	0.344 (0.000)	0.171 (0.004)	0.154 (0.003)	0.154 (0.004)
Parental income in P10-P90	IGE	0.452 (0.001)	0.237 (0.007)	0.187 (0.005)	0.256 (0.007)
Men	IGE	0.349 (0.001)	0.178 (0.005)	0.184 (0.004)	0.159 (0.005)
Women	IGE	0.342 (0.001)	0.164 (0.005)	0.125 (0.004)	0.150 (0.005)
All (incl. zero-income children)	IGE_E	0.335 (0.008)	0.178 (0.005)	0.166 (0.005)	0.153 (0.004)
Parental income in P10-P90 (incl. zero-income children)	IGE_E	0.414 (0.004)	0.209 (0.009)	0.182 (0.009)	0.210 (0.0082)
All (incl. zero-income children)	rank-rank	0.341 (0.000)	0.205 (0.003)	0.174 (0.003)	0.212 (0.003)
Men (incl. zero-income children)	rank-rank	0.336 (0.000)	0.218 (0.003)	0.199 (0.003)	0.225 (0.003)
Women (incl. zero-income children)	rank-rank	0.346 (0.000)	0.192 (0.004)	0.150 (0.004)	0.199 (0.004)
All (incl. zero-income children), child's individual earnings vs. parents' family income	rank-rank	0.282 (0.000)	0.239 (0.003)	0.234 (0.003)	0.235 (0.003)
Sons' (incl. zero-inc.) individual earnings rank vs. parent family income rank	rank-rank	0.313 (0.000)	0.250 (0.004)	0.246 (0.004)	0.246 (0.004)
Daughters' (incl. zero-inc.) individual earnings rank vs. parent family income rank	rank-rank	0.249 (0.000)	0.224 (0.003)	0.219 (0.003)	0.220 (0.003)

Standard errors are listed in parentheses. U.S. results are from Chetty et al. [2014] Table I, the Danish results are based on my own calculations based on Danish Register data. DK1 family income definition includes labor earnings (including self-employment income), capital income, unemployment insurance, retirement- and disability benefits. DK2 includes labor earnings (including self-employment income), capital income, and *all transfers* , DK3 includes labor earnings (including self-employment income) and capital income

For the IGE estimates[6], one might worry that five years of parental income is not enough to properly deal with attenuation bias and that measuring the child generation in their early 30s is too early to properly deal with life-cycle bias, as pointed out by Mazumder [2015]. These biases, and the fact that Danish IGEs in children's family income are smaller compared to individual income, explain why the Danish estimates in Table 1 are smaller than the estimates presented in the remaining analysis. However, if only both countries' estimates are "equally" biased, this is less important for a cross-country comparison.

A relevant worry in this regard is that that cross-country differences in, for instance, average graduation ages could influence the comparison by leading to a more considerable life-cycle bias for Denmark than the U.S., and hence an overstatement of income mobility in Denmark compared to the U.S. In Appendix Table A1 I show that results are robust when I measure the Danish children three years later, from 2014-15, however with slightly smaller differences between the countries. The comparison in Section 3.2 measures parents for 9 years and children in their late 30s, and this also reveals similar results.

The U.S. rank-rank correlation estimates of family income are 50% larger than the Danish.[7] For both countries, as for the IGE, rank-rank correlations by family income do not vary a lot by gender. When the child's income measure is changed from family income rank to individual earnings rank, U.S. estimates decrease while Danish estimates increase, resulting in U.S. rank-rank correlations being only 24% larger than the Danish. This could indicate that children of high-income parents are more likely to marry in the U.S., but less so in Denmark. When the data are divided by gender, I find that rank-rank correlations in individual earnings are much closer for women compared to men, with a difference of approximately 10% for women and 30% for men[8].

[6] As noted by Chetty et al. [2014] and Nybom and Stuhler [2016], rank-rank estimates are less sensitive to attenuation and life-cycle biases

[7] The Danish rank-rank correlation in family income of 0.205 is slightly higher compared to the Danish estimate of 0.18 already presented in Chetty et al. [2014].

[8] These results are in keeping with Corak et al. [2014] who also find that, regarding positional father/son mobility, Sweden and the U.S. are more similar as compared to IGE mobility estimates. AlsoLandersø and Heckman [2017] find that U.S. rank-rank correlations are 30% larger for the U.S. than Denmark,

Chetty et al. [2014] Figure I.B displays the binned scatter plots of the relationship between child/parent log income levels for each parental income percentile and Chetty et al. [2014] Figure II.A plots the mean child income rank for each parent income percentile. In Figure 1 and Figure 2 I plot these curves together with the Danish equivalents.

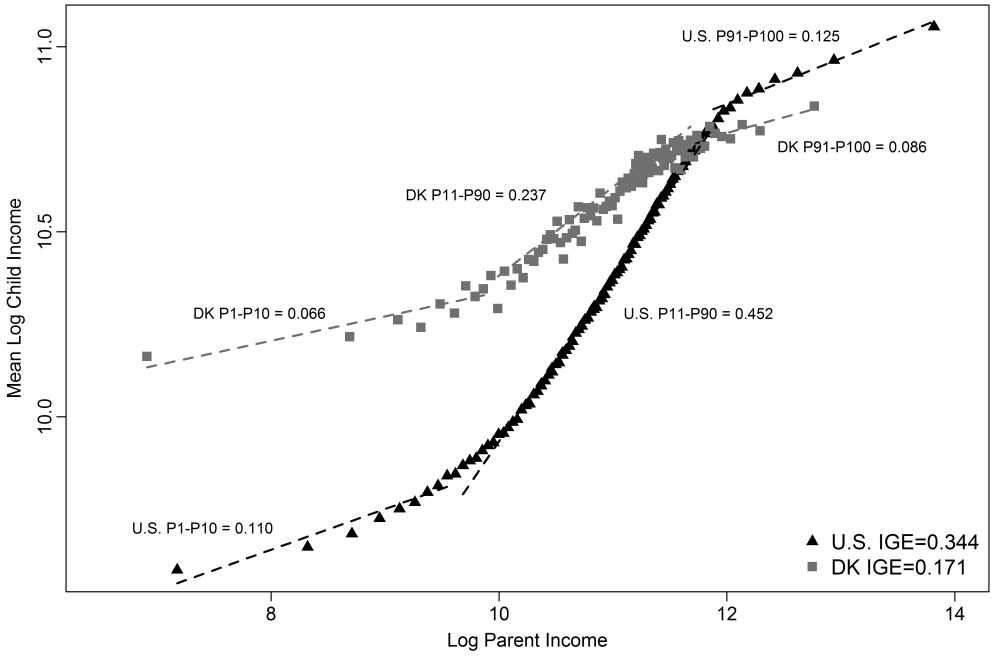
Figure 1 shows that both countries have S-shaped IGE curves, indicating that intergenerational income mobility varies across the parental distribution for both countries, with lowest mobility (steepest slope) for children of middle-income parents[9]. Such strong non-linearities emphasize that comparisons of IGE coefficients alone are problematic, as the underlying assumption of linearity clearly do not hold. Region-specific estimates for the bottom and top percentiles in the parent income distributions suggest that, while Denmark is more mobile than the U.S. across all levels of parental income, the difference is smallest for children of high-income parents.

The rank-rank mobility curves in Figure 2 show the expected child rank for each parental income percentile for the two countries, and these are fairly similar to the cross-country comparison already presented in Chetty et al. [2014]. While the U.S. rank-rank slope is close to being linear, the Danish rank-rank curve has steeper slopes at the tails of the parental income distribution. For the top and bottom parent income deciles, U.S. rank-rank slopes are only slightly larger than the Danish. However, for parental income percentile 11-90, the U.S. rank-rank slope is more than twice as large as the Danish. As such, rank-rank coefficient comparisons are likely to be misleading with respect to transmission mechanisms in the central parts of the parental income distribution and uninformative in the tails.

considering sons' and daughters' pooled individual gross income

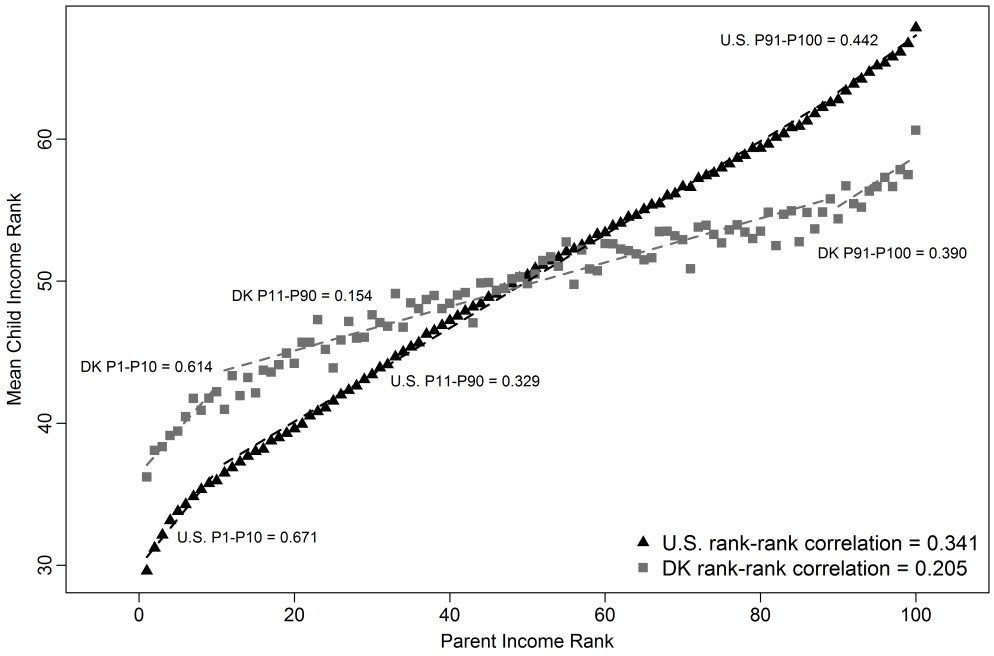
[9]While Bratsberg et al. [2007] also estimate an S-shaped mobility curve for Denmark, they find that the U.S. relationship is close to being linear (but only based on 1999 observations from the NLSY79). Landersø and Heckman [2017] also find S-shaped mobility curves for Denmark

Figure 1: Mobility Curve Comparison



For each parent family income percentile, the mean log of parent family income vs. the mean log of child family income are plotted for the U.S. and DK, respectively. U.S. results are from Chetty et al. [2014] Figure I.B, Danish results are based on my own calculations. Region-specific slopes for P1-P10, P11-P90 and P91-P100 in the parent income distribution are computed using the aggregate bin-scatter values

Figure 2: Rank-rank Curve Comparison



For each parent family income rank, the expected rank of children is plotted. U.S. results are from Chetty et al. [2014] Figure II.A, Danish results are based on my own calculations. Region-specific slopes for P1-P10, P11-P90 and P91-P100 in the parent income distribution are computed using the aggregate bin-scatter values

3.2 Comparing to Mitnik et al. (2015)

Mazumder [2015] and Mitnik et al. [2015] suggest that the IGE estimates presented in Chetty et al. [2014] are downward-biased as children are measured at too early ages (29-32) and parents for too few years (5). To ensure that my results are robust to using later ages, I also compute Danish mobility estimates equivalent to those in Mitnik et al. [2015]. Also, Mitnik et al. [2015] compute IGE_E mobility curves for sons' and daughter's individual earnings.

Mitnik et al. [2015] use the Statistics of Income Mobility Panel (SOI-M) which is based on a stratified random sample of 1987 tax returns. Their data set has a much broader time span compared to Chetty et al. [2014], which enables them to better address life-cycle- and attenuation biases. The trade-off is a smaller sample, counting 12,466 observations. Mitnik et al. [2015] consider children born in 1972-1975. Parents' family income is measured as a 9-year average when children are 15-23 years old[10] and children's mean family income is measured in 2010 (when they are 35-38 years old)[11]. Parents and children are matched by dependent claiming in 1987, and spouses are identified by joint tax filing. Total family income for the U.S. is similar to the specification in Chetty et al. [2014], but further excludes the *non-taxable* portion of pensions, annuities, and social security income. Disposable income is approximated by subtracting out net federal taxes (including refundable tax credits such as the EITC) from total income. State taxes are not subtracted, and some non-taxable transfers (e.g., Temporary Assistance for Needy Families) are not included. Mitnik et al. [2015] include 65% of self-employment income in their wage definition. *For Denmark*, I construct a similar family income measure which includes labor earnings (including 100% of self-employment income), capital income, and unemployment insurance. Disposable (after-tax) income is directly observable in the Danish data. For wages, I also include 65% of self-employment income.

[10] When children are 23 years old, only 2.4% of fathers and 0.7% of mothers are retired in the Danish data, so retirement (parental life-cycle bias) is not expected to influence results.

[11] Even though Mitnik et al. [2015] only include one year of child income, we don't have to worry about selection bias since the IGE_E estimator allows us to include zero-income children.

Family-income mobility estimates from Mitnik et al. [2015] are listed in Table 2.[12] Consistent with the comparison to Chetty et al. [2014], the results suggest that income persistence in family income as measured by the IGE is roughly twice as large in the U.S. compared to Denmark. The higher levels of IGE_E estimates compared to the IGE estimates in Table 1 are probably caused by smaller attenuation (parents are measured in more years), life-cycle (children are measured later), and selection (zero-income children are included) biases.

Table 2: Comparing to Mitnik et al. [2015]: Total and disposable family income

Sample	**Estimator**	U.S.	$DK_{marriage}$	$DK_{cohabitation}$
Total family income of parents and sons	IGE_E	0.48 (0.44-0.52)	0.21 (0.20-0.22)	0.17 (0.17-0.18)
Total family income of parents and daughters	IGE_E	0.46 (0.42-0.50)	0.17 (0.16-0.19)	0.15 (0.15-0.16)
Dispoable family income of parents and sons	IGE_E	0.46 (0.42-0.51)	0.23 (0.22-0.24)	0.20 (0.19-0.21)
Disposable family income of parents and daughters	IGE_E	0.44 (0.40-0.48)	0.18 (0.17-0.19)	0.16 (0.15-0.17)

95% Confidence intervals (bootstrapped) are listed in parentheses. U.S. estimates are from Mitnik et al. [2015] Table 18, Danish results are based on 260,183 register data observations. $DK_{marriage}$ only consider married couples, while $DK_{cohabitation}$ also includes cohabiting spouses

As illustrated in Figure 1, both countries' mobility curves are highly non-linear, and a comparison of IGE coefficients based on the full support of the parental income distribution is therefore problematic. Table 3 show the region-specific IGE_E coefficients for total family income for children born to parents in percentile 1-9, 10-50, 50-90 and 91-100 respectively. Despite a large uncertainty with respect to U.S. estimates, Table 3 confirms the finding of S-shaped mobility curves for both countries, with largest cross-country differences in income mobility for low- to medium income parents.

Mitnik et al. [2015] also consider the elasticity of sons' and daughters' individual earnings with respect to parents' disposable income. These measures are especially interest-

[12] I only report comparisons of constant-elasticity IGE_E estimates - comparisons to their non-parametric IGE_E estimates reveal similar results.

Table 3: Comparing to Mitnik et al. [2015]: Region-specific IGE_E coefficients in total family income

		< P10	**P10-P50**	**P50-P90**	**> P90**
Total family income of parents and sons	U.S.	0.14 (-0.07-0.50)	0.45 (0.33-0.18)	0.69 (0.50-0.90)	0.37 (0.26-0.47)
	DK	0.06 (0.05-0.07)	0.19 (0.18-0.20)	0.42 (0.39-0.45)	0.32 (0.28-0.37)
Total family income of parents and daughters	U.S.	0.22 (0.02-0.67)	0.36 (0.21-0.49)	0.63 (0.47-0.78)	0.42 (0.33-0.52)
	DK	0.05 (0.04-0.06)	0.18 (0.17-0.19)	0.36 (0.33-0.39)	0.26 (0.20-0.33)

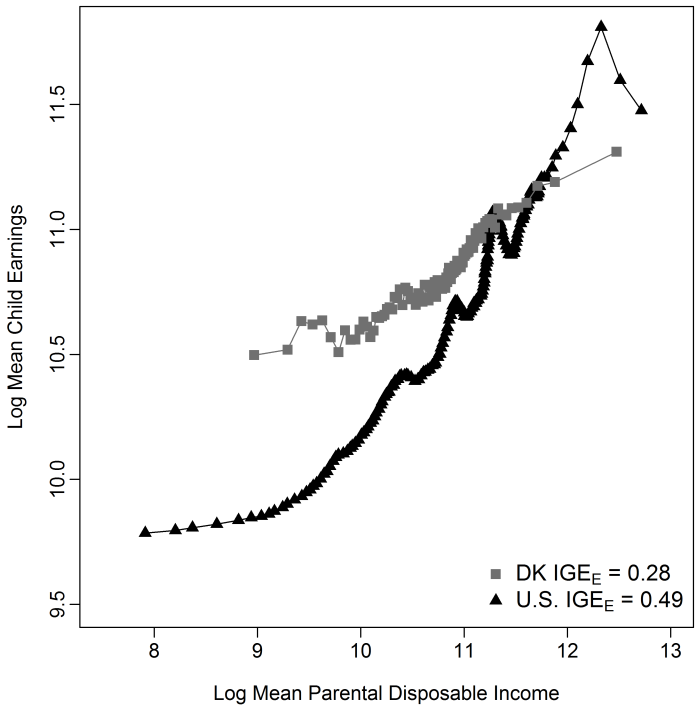
95% Confidence intervals (bootstrapped) are listed in parentheses. U.S. estimates are taken from Mitnik et al. [2015] Table 14. Danish results are based on 260,183 register data observations. Region-specific IGE_E coefficients are computed with a spline model, see Mitnik et al. [2015] p. 24-25.

ing as they compare how parents' disposable income associates with the *individual* labor market outcomes of their children. Figure 3 shows the corresponding mobility curve comparisons of sons' and daughters' individual earnings compared to their parents' disposable family income, respectively.

Both point estimates and mobility curves suggest that Danish men are significantly more mobile compared to men in the U.S. The differences in individual earnings IGEs, however, are much smaller for women, which is consistent with my findings in the Chetty et al. [2014] comparison. Mitnik et al. [2015] does not explain what causes the bumpy patterns of the U.S. mobility curves. The small "hump" in the Danish mobility curves for the children of low-earning parents is caused by a shift in family compositions right around the 7th parental income percentile, shifting from one to two-parent households. The downward trend reflects that children of a single parent with an average household income do better compared to children with two low-income parents. Mitnik et al. [2015] show that the U.S. difference in earnings mobility for men and women is mainly driven by married women tending to withdraw from employment as their spouses' earnings increase. As such, the smaller difference in estimated mobility for women is likely driven by differences in female labor supply in the two countries.

Figure 3: Comparing to Mitnik et al. [2015]: IGE_E mobility curves of Sons' and daughters' earnings vs. parents' disposable family income

(a) Sons



(b) Daughters

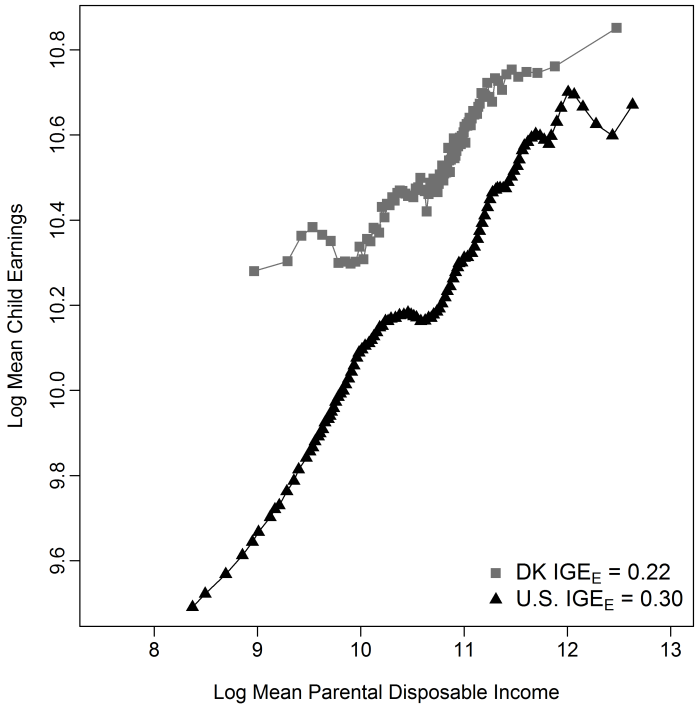


Figure 3(a): For each disposable parent family income percentile, the log of mean parent family income is plotted against the log of the mean of sons' individual earnings. U.S. results are from Mitnik et al. [2015] Figure 8, Danish results are based on my own calculations. **Figure 3(b)**: For each disposable parent family income percentile, the log of mean parent family income is plotted against the log of the mean of daughters' individual earnings. U.S. results are from Mitnik et al. [2015] Figure 8, Danish results are based on my own calculations.

4 Existing Comparisons

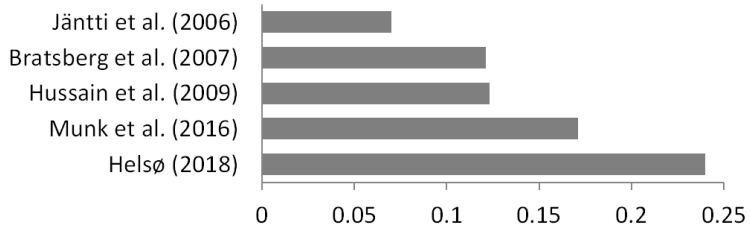
Several studies have compared the level of intergenerational mobility in the U.S. to that in the Scandinavian countries. Most studies have found evidence of much higher mobility in the Scandinavian countries compared to the U.S., finding that the father-son earnings IGE for the U.S. is 300-500% times higher than in Denmark. This is much higher compared to the 50-100% difference found in this paper. In Section 4.1 I show that previous father-son comparisons have overstated mobility levels in Denmark due to measurement- and selection biases. In a recent comparison, Landersø and Heckman [2017] finds that the difference in mobility levels between Denmark and the U.S. vary greatly with the choice of income measure, e.g. if transfers are included or not. In Section 4.2 I show that this finding is specific to the measurement choices in Landersø and Heckman [2017], including cross-country differences in income specifications and sample selections. As my comparisons in Section 3 rely on highly comparable sample selection criteria, income concepts, family definitions and data qualities for the two countries, my findings of a 50-100% difference should remain stable across different income measures.

4.1 Father-Son Earnings Elasticities

Many cross-country comparisons, including the "Great Gatsby" Curve (Krueger [2012] and Corak [2013])[13] consider the father-son earnings IGE. Since 2006, estimates of the father-son earnings IGE for Denmark have increased over time, see Figure 4. The increase is likely driven by the expansion of years of available observations in the Danish register data - starting in 1980 - which has enabled researchers to better deal with attenuation- and life-cycle bias as described in Section 2. To optimally measure the IGE coefficient, more than 9 years of parental income and several years of child income is needed to overcome attenuation and selection biases, and both generations should be measured in their early 30s-late 40s to avoid life-cycle bias.

[13]The "Great Gatsby" show a positive relationship between the Gini coefficients and the father-son earnings elasticities in several countries

Figure 4: Previous father-son earnings IGE estimates for Denmark



Jantti et al. [2006] measure the father-son earnings IGE for Denmark to be as low as 0.07[14]. Based on just one year of fathers' and sons' earnings, this estimate is likely to be downward-biased due to 1) attenuation bias caused by measurement error in parental earnings, and 2) sample selection bias of fathers and sons, as those with weaker labor force attachment are more likely to be observed with zero earnings and thereby excluded. The same can be said for the IGE estimate of 0.121 obtained in Bratsberg et al. [2007], who measure both fathers and sons for two years. Hussain et al. [2009] arrive at an estimate of 0.123 when they measure fathers' income as a five year average in 1984-88, and sons for just one year in 2002 when they are 30-40 year old. Munk et al. [2016] arrive at an IGE estimate of 0.17, measuring both sons and fathers for five years (fathers in 1980-1984 and sons in 2004-2008, sons were between 35-42 years old in 2008).

The previous studies also vary by how they measure earnings. Jantti et al. [2006] include both wages and self-employment income, while Hussain et al. [2009] excludes self-employment income due to its large yearly variations. Bratsberg et al. [2007] include self-employment income for the remaining countries in their analysis, but do not include it for Denmark.[15] The overall consensus in the literature is to measure income as either a proxy for "living standards" or "earning abilities" (or something in between). As self-employment is comprised in both concepts, I would argue that it should be included in the earnings measure as in Mazumder [2015] and Chetty et al. [2014].

[14] Correcting for the attenuation biased caused by the measurement error of parental income, Corak [2006] reaches an estimate of 0.15

[15] Munk et al. [2016] do not in detail describe their earnings definition, but they compare their results to Björklund et al. [2012] who include self-employment income.

With data spanning from 1980-2015, I am able to compute father/son earnings IGEs with measurements of permanent income spanning over a reasonable number of years and ages for both parents and children, avoiding most life-cycle, attenuation, and selection biases that have influenced previous father-son IGE estimates. For birth cohorts 1972-1975 I measure fathers' earnings (including self-employment income) in 1980-1990 and children's earnings (including self-employment income) in 2009-2015.[16] I estimate the father-son earnings IGE to be 0.24 for Denmark.

Using the same birth cohorts and earnings definition (including self-employment income), but changing the ages of measurement to match those in the previous literature, I get a corresponding estimate of 0.10 for Jantti et al. [2006] and Bratsberg et al. [2007], 0.11 for Hussain et al. [2009] and 0.17 for Munk et al. [2016]. The large difference between these estimates and 0.24 once fathers and sons are measured at optimal ages, confirm that the previous findings are in fact downward biased as they are measured at too few and at sub-optimal ages. The preferred estimate of 0.24 drops to 0.09 when I discard self-employment income.

Despite some variation, most recent U.S. estimates of father-son IGEs are generally around 0.5: Corak [2006] reports 41 estimates which range from 0.09 to 0.61 with a preferred estimate of 0.47. Mazumder [2005] estimates a father-son earnings IGE for the U.S. of 0.6, using the 1984 Survey of Income and Program Participation (SIPP) matched to the Social Security Administration's Summary Earnings Records (SER). Mazumder [2015] use the PSID data, averaging over many years for both sons and fathers, and suggests a value of 0.5-0.6. These findings suggest that father-son mobility in Denmark is roughly twice as high as compared to the U.S., which is in keeping with my previous findings.

[16] I discard observations for which sons are observed in less than 3 years, parents in less than 6 years. I further discard observations with average annual earnings below 30 USD

4.2 Robustness to different income measures

In a recent comparison of social mobility in Denmark and the U.S., Landersø and Heckman [2017] (in the following denoted by L&H) find that cross-country differences in income mobility vary greatly with the choice of income measure used. While the primary focus of L&H is on cross-country differences in educational mobility, their income mobility findings suggest that the notion of higher income mobility in Denmark is not a general fact, but that it is contingent on the choice of income measure. Measured by wage earnings, they find that Denmark is 3.5 times more mobile than the U.S.[17] while they find equal levels of mobility once measured by gross income excluding transfers (see Table A2 Sample #1).

Given this finding, one might worry that the cross-country comparisons presented in section 3 does not hold across all income measures. Lacking direct access to U.S. tax data, I cannot compute administrative data comparisons for different income measures. To show that my findings are robust to different income measures, I instead replicate the IGE findings in L&H for Denmark and show that the estimated variation across different income measures is driven by two measurement decisions: first, L&H do not include self-employment income in their wage earnings definition for Denmark, which results in low wage-earning IGEs for Denmark[18]. Second, they compare children's individual income to the sum of their parents' income - this asymmetry in income concepts across generations generates an upward bias on pre-transfer income measures for Denmark, explaining the large variation in before- and after-transfer income IGEs for Denmark. The inclusion of single-parent households for the U.S, combined with very few and high measurement error observations for the U.S., seems to cause the large but insignificant differences in pre- and post transfer measures for the U.S. See Appendix A.3 for a detailed explanation of these robustness checks.

Table 4 illustrates that the variation in IGE estimates across the different income

[17] Given that they measure both parents and children at optimal ages, their estimated wage earnings IGE for Denmark of 0.083 contradicts the findings in Section 4.1

[18] L&H are not the first to have excluded self-employment income when computing Danish IGE estimates, see Section 4.1

measures disappear when I include self-employment income in the Danish wage earnings definition and only consider fathers and their sons. Using the same data sources as L&H, I find that U.S. IGE estimates are roughly 50% larger (but not statistically different) than their Danish counterparts across all income measures, which is in keeping with my previous findings. The difference in IGE estimates in L&H are caused by the above described measurement- and data related issues, which do not apply to the comparison presented in Section 3 where I precisely imitate the U.S. sample- and income concepts for Denmark and rely on high quality administrative data for both countries. As such, I expect my comparison to be robust to different choices of income measures.

Table 4: IGE comparisons for different income measures, L&H and father-son

		# Obs.	**Gross Inc. Excl. Trans.**	**Gross Inc. Incl. Trans.**	**Wage Earnings**	**Wage Earnings and Trans.**
L&H	U.S.	621	0.312 (0.055)	0.446 (0.054)	0.289 (0.044)	0.419 (0.058)
	DK	149,190	0.352 (0.004)	0.271 (0.003)	0.083 (0.003)	0.063 (0.003)
	$\frac{U.S.}{DK}$		0.886	1.646	3.482	6.651
Father-Son	U.S.	318	0.375 (0.069)	0.376 (0.067)	0.312 (0.067)	0.295 (0.063)
	DK	80,636	0.245 (0.006)	0.245 (0.005)	0.210 (0.007)	0.213 (0.005)
	$\frac{U.S.}{DK}$		1.531	1.535	1.486	1.385

Standard errors are listed in parentheses. L&H results are taken directly from Landersø and Heckman [2017] Table 1. The Father-Son comparison is based on the same data as L&H, but only considers fathers and sons. **DK:** Administrative data. Sample includes children born 1973-1975. Parental income is computed as the average of observed income from the child's 7th to 15th year. Child income is observed in 2010-2012, and require both the mother and father (father) and child to be observed during all years. Observations with annual income below $30 are discarded in the father-son comparison (smallest annual income in U.S. data is $285). **U.S.:** PSID data. Sample includes children born 1972-1978. Child income is computed as the average in the years 2007-2013, from when the child is at least 30 years old. Observations are included whenever one year of income is observed for both child and father.

Given the extremely low number of U.S. PSID observations included in Table 4, already suffering from high non-random survey attrition (see Schoeni and Wiemers [2015]) and measurement error[19], one should be cautious to draw any conclusion about the variation

[19]Meyer et al. [2009] finds that, while measurement errors in earnings are approximately zero means in the PSID, transfers are systematically underreported with 30-70%, varying for different transfer types

in U.S. estimates across different income measures. Rather, the IGE estimates presented in Table 4 should serve as a robustness check rather than an attempt to measure the optimal father-son IGEs.[20]

5 Conclusion

I present the first comprehensive cross-country comparison of intergenerational income mobility in Denmark and the U.S., which is based on administrative data for both countries. I consider the findings in Chetty et al. [2014] and Mitnik et al. [2015] and compute directly comparable mobility estimates for Denmark. I find that mobility in Denmark is 50-100% higher than that in the U.S. Cross-country differences are largest once children's family income measures are considered. For individual earnings I find gender variations in cross-country differences, with largest differences for men. Cross-country differences measured by intergenerational income elasticities (IGEs) are roughly twice as large as when measured by rank-rank correlations. I show that the estimated difference of 50-100% is robust to the choice of income measure used, and does not vary when public transfers, taxes or capital income are in-/excluded.

A difference of 50-100% is smaller compared to previous comparisons which have found that Denmark is 3 to 5 times more mobile than the U.S. I present evidence showing that previous estimates have overstated the level of income mobility in Denmark, mainly due to the fact that the Danish register data, starting in 1980, did not cover enough years to sufficiently deal with attenuation-, selection- and life-cycle biases. With data spanning from 1980-2015, I estimate the father-son earnings IGE of Denmark to be 0.24, which is 60-100% higher than previous estimates.

[20]The DK estimate of 0.21 is slightly smaller compared to the previous estimate of 0.24 in Section 4.1 as this is based on fewer years of observations for both parents and children. Mazumder [2015] also use the PSID, but more optimally so as to avoid biases, estimates a father-son earnings IGE of 0.5-0.6, similar for pre- and post transfer earnings

References

Björklund, A., Roine, J., and Waldenström, D. Intergenerational top income mobility in Sweden: Capitalist dynasties in the land of equal opportunity? *Journal of Public Economics*, 96(5-6):474–484, 2012.

Bratberg, E., Davis, J., Mazumder, B., Nybom, M., Schnitzlein, D. D., and Vaage, K. A comparison of intergenerational mobility curves in Germany, Norway, Sweden, and the US. *The Scandinavian Journal of Economics*, 119(1):72–101, 2017.

Bratsberg, B., Røed, K., Raaum, O., Naylor, R., Eriksson, T., et al. Nonlinearities in intergenerational earnings mobility: consequences for cross-country comparisons. *The Economic Journal*, 117(519), 2007.

Chetty, R., Hendren, N., Kline, P., and Saez, E. Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, 129(4):1553–1623, 2014.

Corak, M. Do poor children become poor adults? lessons from a cross-country comparison of generational earnings mobility. In *Dynamics of inequality and poverty*, pages 143–188. Emerald Group Publishing Limited, 2006.

Corak, M. Income inequality, equality of opportunity, and intergenerational mobility. *Journal of Economic Perspectives*, 27(3):79–102, 2013.

Corak, M., Lindquist, M. J., and Mazumder, B. A comparison of upward and downward intergenerational mobility in canada, sweden and the united states. *Labour Economics*, 30:185–200, 2014.

Grawe, N. D. Lifecycle bias in estimates of intergenerational earnings persistence. *Labour economics*, 13(5):551–570, 2006.

Haider, S. and Solon, G. Life-cycle variation in the association between current and lifetime earnings. *The American Economic Review*, 96(4):1308–1320, 2006.

Hussain, M. A., Munk, M. D., Bonke, J., et al. Intergenerational earnings mobilities–how sensitive are they to income measures? *Journal of Income Distribution*, 18(3/4):79–92, 2009.

Jantti, M., Bratsberg, B., Roed, K., Raaum, O., Naylor, R., Osterbacka, E., Bjorklund, A., and Eriksson, T. American exceptionalism in a new light: A comparison of intergenerational earnings mobility in the nordic countries, the united kingdom and the united states. 2006.

Krueger, A. The rise and consequences of inequality in the united states. *Speech at the Center for American Progress, Washington D.C. on January 12, 2012*, 2012.

Landersø, R. and Heckman, J. J. The scandinavian fantasy: The sources of intergenerational mobility in Denmark and the US. *The Scandinavian Journal of Economics*, 119 (1):178–230, 2017.

Mazumder, B. Fortunate sons: New estimates of intergenerational mobility in the united states using social security earnings data. *Review of Economics and Statistics*, 87(2): 235–255, 2005.

Mazumder, B. Estimating the intergenerational elasticity and rank association in the us: Overcoming the current limitations of tax data. 2015.

Meyer, B. D., Mok, W. K., and Sullivan, J. X. The under-reporting of transfers in household surveys: its nature and consequences. Technical report, National Bureau of Economic Research, 2009.

Mitnik, P. and Grusky, D. The Intergenerational Elasticity of What? The Case for Redefining the Workhorse Measure of Economic Mobility. *Stanford Center on Poverty and Inequality Working Paper*, 2017.

Mitnik, P., Bryant, V., Weber, M., and Grusky, D. B. New estimates of intergenerational mobility using administrative data. *Statistics of Income Division working paper, Internal Revenue Service (ht tps://www. irs. gov/pub/irssoi/15rpintergenmobility. pdf)*, 2015.

Munk, M. D., Bonke, J., and Hussain, M. A. Intergenerational top income persistence: Denmark half the size of sweden. *Economics Letters*, 140:31–33, 2016.

Nybom, M. and Stuhler, J. Biases in standard measures of intergenerational income dependence. *Journal of Human Resources*, pages 0715–7290R, 2016.

Schoeni, R. F. and Wiemers, E. E. The implications of selective attrition for estimates of intergenerational elasticity of family income. *Journal of economic inequality*, 13(3): 351, 2015.

Solon, G. Intergenerational income mobility in the united states. *The American Economic Review*, pages 393–408, 1992.

Solon, G. Cross-country differences in intergenerational earnings mobility. *The Journal of Economic Perspectives*, 16(3):59–66, 2002.

A Appendix

A.1 Robustness checks of comparison to Chetty et al. [2014]

Table A1 shows two robustness checks of Table 1. First, the spousal definition is changed from marriage only to also include cohabiting spouses. Second, the Danish child generation is measured three years later, in 2014-2015, to account for later college graduation ages in Denmark compared to the U.S.

Table A1: Robustness Checks of Chetty et al. [2014] comparison

Sample	**Estimator**	**U.S.**	(1) Cohabiting Spouses	(2) Older DK Children
All	IGE	0.344 (0.0004)	0.171 (0.0030)	0.209 (0.0038)
Parental income in P10-P90	IGE	0.452 (0.0007)	0.206 (0.0050)	0.270 (0.0065)
Men	IGE	0.349 (0.0006)	0.178 (0.0043)	0.216 (0.0055)
Women	IGE	0.342 (0.0005)	0.166 (0.0041)	0.202 (0.0053)
All (incl. zero-income children)	IGE_E	0.335 (0.008)	0.157 (0.0028)	0.206 (0.0028)
Parental income in P10-P90 (incl. zero-income children)	IGE_E	0.414 (0.004)	0.157 (0.0033)	0.236 (0.0043)
All (incl. zero-income children)	rank-rank	0.341 (0.0003)	0.238 (0.0025)	0.241 (0.0024)
Men (incl. zero-income children)	rank-rank	0.336 (0.0004)	0.243 (0.0034)	0.251 (0.0034)
Women (incl. zero-income children)	rank-rank	0.346 (0.0004)	0.238 (0.0036)	0.231 (0.0036)
All (incl. zero-income children), child's individual earnings vs. parents' family income	rank-rank	0.241 (0.0003)	0.225 (0.0025)	0.254 (0.0025)
Men (incl. zero-income children), child individual earnings rank vs. parent family income rank	rank-rank	0.313 (0.0003)	0.244 (0.0037)	0.267 (0.0037)
Women (incl. zero-income children), child individual earnings rank vs. parent family income rank	rank-rank	0.249 (0.0003)	0.228 (0.0033)	0.239 (0.0032)

U.S. results are from Chetty et al. [2014] Table I, DK results are own calculations based on Danish Register data, using the income definition DK_{Chetty}. (1) The sample with cohabiting partners included in the spousal definition consists of 151.623 children excluding zero-income children and 156,222 including zero-income children. (2) Children are measured three years later, in 2014-2015. The sample consists of 146.961 children excluding zero-income children and 156,422 including zero-income children.

A.2 The IGE of expectations, IGE_E

The conventionally estimated OLS coefficient of β^{IGE} corresponds to the following estimand:

$$\beta^{IGE} = \frac{dE(\ln Y^C | Y^P = y^P)}{d \ln y^P} \tag{3}$$

Mitnik et al. [2015] note that the conventional IGE estimate has been misinterpreted: scholars have falsely assumed they were estimating the expectation of children's earnings or income ($\ln E(Y^C|Y^P = y^P)$) when in fact their estimate pertains to the *geometric* mean of children's income ($E(\ln Y^C|Y^P = y^P)$). They propose an alternative approach to estimate the IGE which switches the order of the log and the expectation, and thereby correctly match the popular, but wrong, interpretation of the intergenerational elasticity of earnings:

$$\beta^{IGE_E} = \frac{d \log E(Y^C | Y^P = y^P)}{d \log y^P} \tag{4}$$

As such, the IGE of expectations (IGE_E) measures the impact of an increase in log parent income on the log of expected child income (which is identical to the traditional interpretation of the conventional IGE) and can be estimated using the semiparametric Poisson pseudo maximum likelihood estimator (PPML). More importantly, the IGE_E estimate allows U.S. to include zero-income children and thereby overcome much of the sample selection bias that affects the conventional IGE estimate.

While I estimate the IGE_e coefficient with the semiparametric PPML estimator, another way to compute the IGE_e non-parametrically for large sample is by the 2-step estimator proposed by Chetty et al. [2014]. Here, parent income is binned in percentiles, and for each bin $log(E(Y^C|Y^P = y^P)$ is computed and regressed on $log(y^P)$ with OLS. The conventional IGE estimate can also be computed with a similar 2-step estimator, but for each bin of parental income, $E(log(Y^C|Y^P = y^P)$ is computed and regressed on $log(y^P)$ with OLS.

A.3 Robustness to different income measures - a replication exercise

Landersø and Heckman [2017] (L&H) find that cross-country differences in mobility vary greatly by the choice of income measure. In the following, I conduct several robustness checks of the Danish IGE estimates presented in L&H Section 2. The overall objective is to show that my finding of a 50-100% difference in mobility in Denmark and the U.S. is robust to using different income measures.

In Section A.3.1 I show that L&H's very low wage earnings IGE estimates for Denmark are caused by the fact that self-employment income is not included in the Danish wage-earnings definition. In Section A.3.2 I show that the asymmetric income definition of parents and children results in an upward bias for Danish pre-transfer income IGEs. In Section A.3.3 I discuss L&H's concern that standard deviations are driving the estimated IGE differences, how this relates to the findings in Section A.3.1 and A.3.2 which also apply to the NL-IGE results in L&H.

The overall conclusion of this replication exercise is that the estimated variation in cross-country differences presented in L&H is a result their measurement choices, including differences in income specifications and sample selections for the two countries. As my comparisons rely on highly comparable sample selection criteria, income concepts, family definitions and data qualities for the two countries, I do not expect my finding of a 50-100% difference to vary across different income measures.

Table A2 Sample # 1 displays the Danish IGE estimates listed in Landersø and Heckman [2017] Table 1, and Sample #2 shows my replication estimates which closely resembles the results in L&H. The remaining samples in Table A2 show the various robustness checks presented in Section A.3.1 and A.3.2. The sample includes children born 1973-1975. Parental income is computed as the average of observed income from the child's 7th to 15th year, and child income is observed in 2010-2012. The child and both parents are required to be observed during all years. Income levels are deflated using country-specific

CPIs, and I use a PPP adjusted exchange rate of $100 to 776 DKK.

A.3.1 Self-employment income:

In contrast to the U.S. wage earnings definition which includes the labor part of business income,[21] L&H do not include business income from self-employment in their Danish wage earnings definition, and as such they compare different income concepts for the two countries[22].

Once 70% of self-employment income is added (assuming the remaining 30% as capital income), the Danish wage earnings IGEs excluding (including) transfers *more than triples* from 0.084 (0.068) to 0.316 (0.235) (see Table A2, sample # 3), and even more so when 100% of self-employment income is included. [23].

Going forward, I include 100% of self-employment income as in Chetty et al. [2014]. Setting self-employment income to zero results in highly misleading approximations of earning abilities/resources in those cases where earnings *partly* consist of self-employment income. Families where one parent is self-employed (e.g. a farmer, general practitioner, accountant, etc.) and the other a wage-earner, misleadingly appears as low-income parents, resulting in lower IGE estimates.[24].

A.3.2 Close-to-zero earnings observations:

L&H find that Danish IGE measures decrease by approximately 23% once transfers are included in the income measure. However, a comparison of IGE estimates across different income measures does not capture any behavioral responses to policy rules, see appendix

[21]Labor and asset part of business and farm income is split approx. 50/50 in the PSID data. As the self-employment income variable for Denmark (NETOVSKUD) does not include capital income (KAPITVIRK), I choose to include the entire amount. For 50% of NETOVSKUD, wage earnings IGE estimates are 0.293 excluding transfers and 0.203 including transfers

[22]L&H are not the first to have excluded self-employment income when computing Danish IGE estimates, see Section 4.1.

[23]When self-employment income is included, the sample size increase from 149,752 to 157,323 as fewer observations are discarded due to zero earnings. The small drop of 18 observations going from 70 to 100% of self-employment is caused by individuals with negative self-employment income

[24]Other studies have also excluded self-employment income, e.g. Hussain et al. [2009]

Table A2: Replication results and robustness checks of Landersø and Heckman [2017] - Estimated IGEs for Denmark

Sample #	**Data Specification**	**Income Criteria**	**# Obs**	**Gross Income Excl. Trans.**	**Gross Income Incl. Trans.**	**Wage Earnings**	**Wage Earnings and Trans.**	**Net-of-tax Income**
1	Results from L&H	>$0	149,190	0.352 (0.004)	0.271 (0.003)	0.083 (0.003)	0.063 (0.003)	0.221 (0.003)
2	Replication Sample	>$0	149,752	0.351 (0.004)	0.275 (0.003)	0.084 (0.003)	0.068 (0.002)	0.231 (0.003)
3	Including 70% self-employment income in "wage earnings"	> $0	157,323	0.351 (0.007)	0.270 (0.004)	0.316 (0.007)	0.235 (0.004)	0.232 (0.003)
4	Including 100% self-employment income in "wage earnings"	> $0	157,305	0.351 (0.007)	0.270 (0.004)	0.327 (0.007)	0.252 (0.004)	0.232 (0.004)
5	Same as #4	>$1000	154,908	0.297 (0.0043)	0.266 (0.004)	0.281 (0.004)	0.249 (0.004)	0.228 (0.003)
6	Same as #4	>$5000	151,212	0.269 (0.004)	0.263 (0.004)	0.253 (0.004)	0.247 (0.004)	0.226 (0.003)
7	Same as #4, Father-Son	>$0	80,636	0.260 (0.007)	0.247 (0.005)	0.224 (0.007)	0.216 (0.005)	0.221 (0.004)
8	Same as #4, Father-Son	>$30	80,395	0.245 (0.006)	0.245 (0.005)	0.211 (0.007)	0.214 (0.005)	0.223 (0.004)
9	Same as #4, IGE_E	$\geq$ 0 DKKR	165,973	0.377 (0.013)	0.382 (0.009)	0.347 (0.007)	0.343 (0.009)	0.334 (0.013)

Sample #1: Results from Landersø and Heckman [2017] Table 1. **Sample #2:** Replication sample: Children born 1973-1975, parental income measured when child is 7-15 years old, requiring both parents to be observed during all 9 years. Parental income is the sum of legal mother's and legal father's income. Child income is average of observed income in the years 2010-2012, requiring the child to be observed during all three years. Gross Income = PERINDKIALT, Wage earnings = LOENMV (salary), Transfers=OVERFORSINDK (see L&H appendix A23) **Sample #3**: Same as sample #2, but with wage earnings = LOENMV+NETOVSKUD (salary + profits from own business). Sample size has increased as individuals with 0 salary, but positive profits from own business are now included. **Sample #4:** Same as sample #2, but only including fathers' income **Sample #5:** Same as sample #2, but including zero-income children and estimated with the IGE of expectations estimator, IGE_E.

A.4. A possible explanation of why L&H find different IGE estimates for before- and after-transfer income is mechanical rather than behavioral: they compare the *sum* of the father's and mother's income to their children's *individual* income. This asymmetry across generations results in significantly different earnings distributions for parents and children, with more children having close to zero earnings. Due to its functional form, the log transformation magnifies the distance to these close-to-zero income observations, making the IGE estimator sensitive to low-income observations, as noted in Chetty et al. [2014]. As such, the bias mainly applies to pre-tax income, and the overweight of close-to-zero earnings observations in the child generation mechanically drives the estimated differences in before- and after-transfer IGE estimates.[25]

A simple, some might say crude, way of reducing the influence of values close to zero when taking logs is to exclude observations with values close to zero. When I discard income observations of less than \$1000 in either generation, the difference between before- and after transfer IGE decreases from 23% to 11% (Table A2 sample # 3), and similarly by 2% (insignificant) for a cutoff of \$5000, see Table A2 sample # 4. Another way to optimize symmetry across generations is to consider only fathers and sons. While the father-son specification itself is not more robust towards close-to-zero income observations, it's advantage is a more even distribution of the close-to-zero observations across the two generations.[26] Table A2 sample # 7 displays the father-son IGEs for which before- and after transfer IGEs are not significantly different.[27] A third solution is to compute the *IGE of expectations* (IGE_E) as proposed in Mitnik et al. [2015], see Appendix A.2. This measure is less sensitive towards close-to-zero income children and allows for the inclusion of children with zero income. Table A2 Sample # 9 show the corresponding

[25]While L&H Appendix A5 conduct a sensitivity analysis on the effects of alternative treatments of zero earnings, they impute a \$1000 income for the missing and zero-earning observations alone and do not account for the many *close*-to-zero earning children causing the bias.

[26]This is the most frequently used measure for cross-national comparisons, as it is unaffected by cross-national differences in family compositions and female labor force participation.

[27]Discarding 0.3% of the observations with income below \$30, before- and after-transfer estimates are virtually identical, see able A2 sample # 8

IGE_E estimates, which are similar for before- and after-transfer coefficients.[28]

A.3.3 The Role of Inequality and Nonlinear IGEs

Recall that the IGE equals the correlation of parents' and children's log income times the ration of standard deviations ($\frac{SD(\ln Y^C)}{SD(\ln Y^P)}$) as denoted in Equation 2 Section 2. L&H note that "it is the ratio of standard deviations that drives the Danish IGE to levels above the US," and they raise the consern that "it is not meaningful to compare IGE estimates, when arbitrary large or small levels of inequality drive the estimates". What L&H interpret as changes in inequality across generations and income measures, are, to some extent, driven by measurement differences across generations as discussed in Section A.3.2, and I show that the standard error ratios in a father-son comparison no longer drives the estimates.

Table A3 show the correlation- and standard deviation decomposition of the IGE estimates in L&H together with the father-son specification which includes self-employment income for both countries. The table shows that the variation in standard deviation ratios for different income measures is smaller for the father-son specification. The table shows that when L&H include transfers in the income definition, they find that the ratio in standard deviations increase in the U.S., while it decreases for Denmark. For the father-son robustness check, I find that the standard error ratios are decreasing for both countries once transfers are included. As L&H do not describe their sample selection criteria for the U.S. in detail, I cannot with certainty explain why they find larger post-transfer IGEs for the U.S. However, a potential explanation is the measurement errors associated with transfer income income in the PSID data, combined with a U.S. parental definition which include more close-to-zero income observations in the parental generation compared to the child generation. This could be because single parent households are included for the U.S., contrary to the Danish sample selection rules where both parents are required to be observed in the data during all years.

[28] IGE_E coefficients are larger as zero-income children are included, overcoming the downward selection bias inherited in the conventional IGE estimator, where all zero-income observations are discarded.

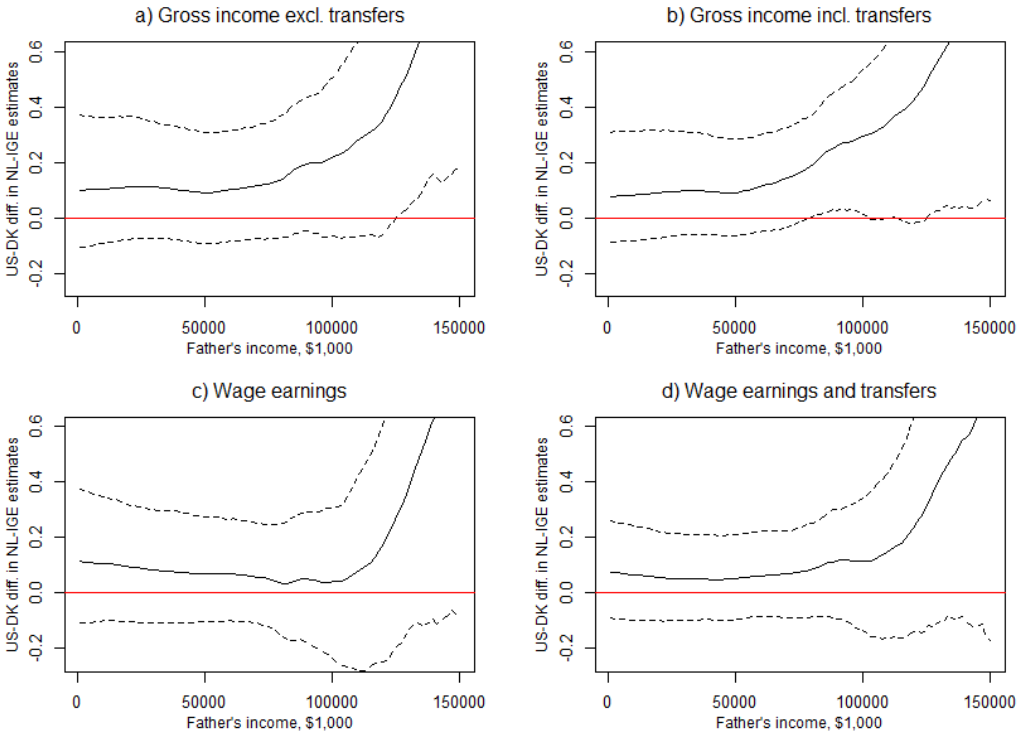
Table A3: Father-Son IGE comparisons for different income measures - including correlation coefficients and standard deviations decomposition

			Gross Inc. Excl. Trans.	**Gross Inc. Incl. Trans.**	**Wage Earnings**	**Wage Earnings and Trans.**
L&H	U.S.	IGE (std.err)	0.312 (0.055)	0.446 (0.054)	0.289 (0.044)	0.419 (0.058)
		Decomposition	$0.268 \cdot \frac{0.977}{0.840}$	$0.318 \cdot \frac{0.906}{0.645}$	$0.256 \cdot \frac{0.970}{0.860}$	$0.280 \cdot \frac{0.923}{0.615}$
	DK	IGE (std.err)	0.352 (0.004)	0.271 (0.003)	0.083 (0.003)	0.063 (0.003)
		Decomposition	$0.201 \cdot \frac{0.860}{0.491}$	$0.214 \cdot \frac{0.375}{0.308}$	$0.081 \cdot \frac{1.004}{0.989}$	$0.075 \cdot \frac{0.561}{0.669}$
Father-Son	U.S.	IGE (std.err)	0.375 (0.069)	0.376 (0.067)	0.312 (0.067)	0.295 (0.063)
		Decomposition	$0.294 \cdot \frac{0.928}{0.727}$	$0.305 \cdot \frac{0.818}{0.665}$	$0.254 \cdot \frac{0.872}{0.710}$	$0.256 \cdot \frac{0.753}{0.653}$
	DK	IGE (std.err)	0.245 (0.006)	0.245 (0.005)	0.210 (0.007)	0.213 (0.005)
		Decomposition	$0.183 \cdot \frac{0.832}{0.621}$	$0.209 \cdot \frac{0.501}{0.428}$	$0.160 \cdot \frac{0.850}{0.648}$	$0.181 \cdot \frac{0.501}{0.425}$

Standard errors are listed in parentheses. The IGE decomposition is given by $Corr(\ln Y^C, \ln Y^P)\frac{SD(\ln Y^C)}{SD(\ln Y^P)}$. L&H results are taken directly from Landersø and Heckman [2017] Table 1. The Father-Son comparison is based on the same data as L&H, but only considers fathers and sons. **DK:** 80,636 administrative data observations. Sample includes children born 1973-1975. Parental income is computed as the average of observed income from the child's 7th to 15th year. Child income is observed in 2010-2012, and require both the mother and father (father) and child to be observed during all years. Observations with annual income below $30 are discarded in the father-son comparison (smallest annual income in U.S. data is $285). **U.S.:** 318 PSID data observations. Sample includes children born 1972-1978. Child income is computed as the average in the years 2007-2013, from when the child is at least 30 years old. Observations are included whenever one year of income is observed for both child and father.

L&H also consider nonlinearities in the IGEs (NL-IGE) in order to analyze how the parent/child association in income changes *across* different levels of parental income. They compute the NL-IGEs as local linear regressions where IGEs are computed locally for different levels of parental income, see Landersø and Heckman [2017] Section 2.4 for further details. While the NL-IGE estimator is more robust to changes in the standard deviation of parental income across income measures, it does not account for the shift in the *ratio* of standard deviations of child an parental income. As such, we would not expect the NL-IGE estimator alone to deal with the bias of many close-to-zero income observations in the child generation. When I only consider fathers and sons, I find that the NL-IGE curves for both countries, across all income measures, are similar in shapes and levels, however imprecisely estimated for the U.S., see Figure A1.

Figure A1: Father-Son NL-IGE Differences. U.S.-DK



The figures plot the estimated NL-IGE differences, US-DK, for different levels of the fathers' income. The samples are identical to the father-son specifications in Table 4 and Table A3. Non-linear IGE estimates computed as described in Landersø and Heckman [2017] with bandwidth=$80.000 / 620.800 DKK (2015 prices). 95% Confidence bands are bootstrapped (50 simulations for Denmark and 1000 simulations for the U.S.) and plotted with dotted lines. Country-specific NL-IGEs are not shown here, but they also have similar shapes across income measures.

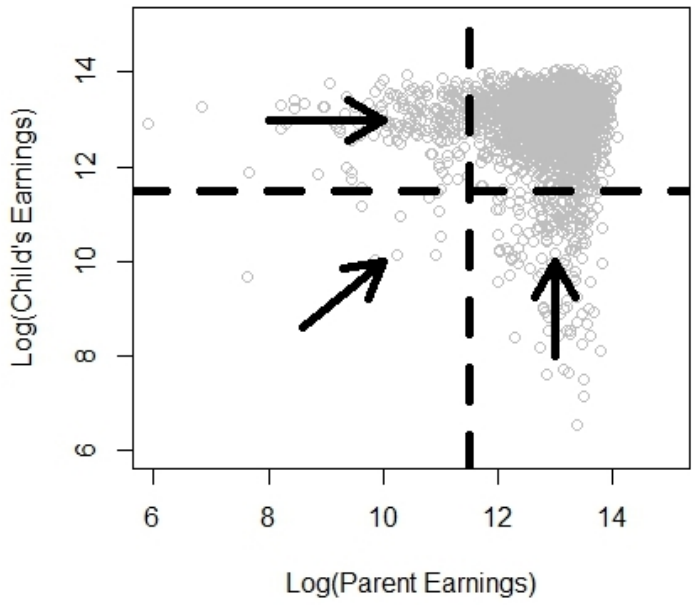
A.4 How transfers/taxes affect the IGE estimate - a visualization

For Denmark, the scatter plot of log(child earnings) and log(parent earnings) has a triangular shape resembling the scatter plot of simulated observations in Figure A2. The triangular shape is a result of right-skewed log-earnings distributions. The data simulated in Figure A2 assume no correlation between parent and child income. As this correlation goes towards 1, the triangle collapses into a straight line.

Assume that transfers are means tested such that one's transfer entitlements depend on your earnings alone. The level of earnings for which individuals are no longer entitled to any transfers divides the parent/children log-earnings sphere into four quadrants. Observations in the upper right quadrant are not entitled to receive any transfers, why these

observations are unchanged in an after-transfers comparison. The bottom left quadrant contains parent/child pairs where only the child is entitled to receive transfers, why these observations will move horizontally towards the right when transfers are added. Likewise are only parents entitled to transfers in the bottom right quadrant, why these observations will move upwards. Both children and parents in the bottom-left quadrant are entitled to receive transfers, why these observations will move northeast in the illustrated diagram in Figure A2.

Figure A2: Illustration of how adding transfers affect the IGE coefficient



Recall that the IGE coefficient is simply the slope of the linear regression of log child income on parent log income. If parents face a transfer system which is more generous compared to that of their children, the horizontal left shift in the upper left square will be bigger than the vertical upward shift in the bottom right square, resulting in an increase in the IGE when adding transfers.

If parents and children have identical earnings distributions and they face identical transfer rules, then the two income definitions result in the same IGE estimates, as the shift is symmetrical and therefore does not affect the linear regression slope. However,

if there are more low earning children than parents - i.e. more observations in the lower right quadrant compared to the upper left quadrant, then the IGE estimate will decrease when we include transfers - this is the case in the Danish IGE sample in Landersø and Heckman [2017]. Vice versa, a larger number of observations in the top left quadrant (more low-earning parents) will result in a higher IGE estimate when accounting for transfers. If we assume that the transfer system and earnings distributions of parents in 1980s in Denmark is basically identical to the transfer system and earnings distribution for their children in 2010's (which is a reasonable assumption), then we would expect similar IGE estimates for pre- and post transfer income. The fact that they are not equal could be caused by the fact that our income concepts for parents and children differ, mechanically causing an overweight of low-earning parents/children. One should therefore be extra careful to ensure comparable measurement of parents' and children's earnings in a before- and after-transfer comparison, as differences in e.g. life cycle timing, business cycles etc. are likely to distort the share of observations in the lower-right and upper-left squares relative to the permanent income.