HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

University of Helsinki
Faculty of Social Sciences
Master's thesis

Adverse Selection and Moral Hazard in Insurance Markets

Evidence on Entrepreneurs in Finland

March 3, 2023

Ella Mattinen

Faculty: Faculty of Social Sciences
Degree Programme: Master's Programme in Economics
Study Track: Research Track
Author: Ella Ingeborg Mattinen
Title: Adverse Selection and Moral Hazard in Insurance Markets: Evidence on Entrepreneurs in Finland
Level: Master's thesis
Month and Year: March 2023
Page count: 73 + 16
Keywords: Social insurance, Moral hazard, Adverse selection, Entrepreneurship, Pension, Sickness allowance, Parental allowance
Supervisor or supervisors: Jani Luoto
Where deposited: E-thesis
Other information:

Abstract

This thesis focuses on the investigation of adverse selection and moral hazard in the Finnish entrepreneurial insurance system. The thesis builds upon previous research on moral hazard and adverse selection in insurance markets, but focuses specifically on the social insurance market for entrepreneurs within sickness and parental allowances. This topic is of interest as mandating social insurance for entrepreneurs is vital, but often lacking due to the flexible nature of entrepreneurial activity. As a result, the social insurance systems may be subject to adverse selection and moral hazard. Rich panel data on insurance contributions of entrepreneurs in Finland allows me to measure the extent of asymmetric information both overall and dynamically. To test the former, I use a positive correlation test. In particular, the test looks at the probability of sickness or having a child in relation to insurance contributions. The results indicate a slight positive correlation between sickness risk and insurance contributions as well as the probability of having a child and insurance contributions. A more significant result is found for the risk of having children in a dynamic sense, showing a strong indication of a rise in insurance contributions around the time of receiving parental allowance. However, the results are more ambiguous in the case of sick pay. These results are robust to several controls as well as two separate identification strategies. Due to the endogeneity of illness or choosing to have children, causal conclusions cannot be drawn from the results.

Acknowledgements

This thesis was written as part of a research project on social insurance to entrepreneurs at the VATT institute for economic research. I am thankful to my colleagues such as Tuomas Matikka, for their help, support, and ideas. I am especially thankful for the comments and support provided by Jarkko Harju throughout this thesis. I would also like to thank Aleksi for proof-reading and pointing out mistakes. I am also thankful for Ossi for helping me through some of the thesis struggles I found myself in. Furthermore, I would like to thank my current employer, Tampere University, for giving me the opportunity to work on my thesis during my working hours. Lastly, I want to express my gratitude to my thesis supervisor Jani Luoto.

Contents

1 Introduction

Entrepreneurship plays a vital role in economic growth. As stated by Audretsch, "entrepreneurship has come to be perceived as an engine of economic and social development throughout the world" (Acs and Audretsch, 2003, p.3). In Finland, there are about 370 000 businesses employing over 1.5 million people (Statistics Finland, 2020). These businesses span many fields as entrepreneurship consists of a diverse range of entrepreneurs. However, due to the inherent uncertainty and unpredictability of entrepreneurial ventures, entrepreneurs often face greater risks compared to wage earners. Furthermore, the insurance systems available to entrepreneurs are generally less established than those available to wage earners. This is due in part to the risky nature of entrepreneurship itself, as well as the fact that entrepreneurs operate in unique and rapidly changing industries, which can make it difficult for insurance companies to develop effective risk management strategies. This is especially the case with social insurance.

Mandating social insurance for entrepreneurs is vital; it can be a first-order welfare improvement as social insurance provides them with financial protection from risks such as old-age, sickness as well as business risks such as bankruptcy. Unlike wage earners, entrepreneurs often experience significant fluctuations in income due to the risky and flexible nature of their work. It is therefore important to allow entrepreneurs to adjust their social insurance coverage to reflect their changing circumstances. However, flexibility in insurance coverage can lead to adverse selection and moral hazard. Adverse selection can occur when higher-risk individuals opt for more extensive insurance plans, while moral hazard can result in increased risk-taking by individuals who have broader insurance coverage.

Moral hazard (MH) and adverse selection (AS) can exist contemporaneously, apart or not exist at all in markets. Insurance markets are one, where adverse selection and moral hazard may play a large role, however, in some cases their presence may not have substantial effects. Moral hazard is considered to be a situation where the action/effort of an agent is only observed by the agent themselves, but not by the principal. Meanwhile, adverse selection means there is asymmetry in information of the agent's risk type. Aforementioned, adverse selection in insurance markets arises when insurees have hidden information about their level of risk. Those with higher levels of risk have incentive to choose more comprehensive insurance plans. Moral hazard arises when the risk of bad events changes as a result of having a different level of insurance. Adverse selection and moral hazard cause inefficiencies in insurance markets that are difficult to control for. Adverse selection may lead to under insurance, but in the presence of moral hazard, a universal insurance mandate is deficient.

Einav and Finkelstein (2011) explain that in insurance markets, adverse selection causes those with the highest willingness to pay to have the highest expected cost for insurance companies. This inefficiency is detrimental, as it causes a dichotomy between the efficient allocation and the equilibrium allocation of insurance. In theory, if policies were put in place which would mandate that everyone purchase insurance, the market could reach an efficient outcome. However, in practice, this may not lead to efficient outcomes due to differences in risk profiles as well as preferences. For example, due to insurance loading, it would be socially optimal to leave individuals whose willingness to pay for insurance is less than their expected cost uninsured [1]. (Einav and Finkelstein, 2011)

As insurance markets are prone to the risk of adverse selection and moral hazard, it leads to difficulties in implementing policies that would yield optimal results. Empirically, through a positive correlation test, we can examine whether adverse selection or moral hazard is present in the market. However, separating these two is rather difficult, as both lead to a positive correlation between insurance levels and insurance claims. In other words, if consumers who have higher insurance coverage make more insurance claims than those with lower coverage, it is difficult to determine whether the cause is moral hazard or adverse selection when using observational data. This would only be possible if there is exogenous variation in the incentives for these factors.

In this thesis, the aim is to investigate **the extent of adverse selection and moral hazard in the Finnish entrepreneurial insurance system**. This thesis builds onto empirical research investigating moral hazard and adverse selection in insurance markets with focus on the social insurance market for entrepreneurs.

[1] Insurance loading refers to a situation where an additional amount is added to the premium for individuals who are higher risk. Insurance companies may require an additional cost to cover the increased potential losses due to such individuals. This is called loading.

Specifically, the main contribution of this thesis is to study moral hazard and adverse selection in the domain of entrepreneurs, which has not been researched heretofore. In addition to investigating the existence of these phenomenons in the market, the aim is to investigate whether insurance contributions anticipate realised risks. Due to asymmetric information about insurees' risks, they can increase their insurance contributions in order to benefit from larger social security payments such as sick pay (SP) or parental allowance (PA). The data used has only been previously used in Benzarti et al. (2020), thus I aim to provide a clear contribution to this field of research.

I intend to investigate the existence of adverse selection and moral hazard in the market using a positive correlation test. Namely, I investigate whether the probability of sickness or having a child increases with insurance contributions. The results indicate that there is slight positive correlation between sickness risk and insurance contributions. A similar result is found for the risk of having children. However, in this case, it is only after controlling for characteristics of individuals. The baseline correlation is found to be negative, implying slight advantageous selection [2].

In addition, when and if insurance contributions are increased as a result of anticipating sickness or children in the future is investigated. This is done by following Kuziemko and Werker (2006) and further, by following Hendren (2017). The former gives us an estimate of whether insurance contributions follow a specific trend prior to illness or children only using a sample of individuals who are affected by one or the other. The latter approach allows me to compare these changes in insurance contribution levels between a treatment and a control. In aggregate, I find that there is strong indication of dynamic adverse selection surrounding the time of receiving parental allowance as both models show an increasing ex ante trend in insurance contributions for the "treated" (individuals who have a child). The results are ambiguous for sick pay as the models do not show cognate results.

Due to the endogeneity of illness or choosing to have children, we cannot draw causal relationships from these results. However, adverse selection comes from

[2] Advantageous selection arises as individuals who are low risk, but have high risk aversion have larger insurance contributions compared to individuals who are high risk and have lower risk aversion.

asymmetries in information about individuals and thus can be interpreted as an inherently endogenous question. Despite the difficulty and complexity of the research questions, these three different approaches allow me to evaluate the associations between risks and actions and therefore provide robust, correlational results. To further validate the findings, more research on both moral hazard and adverse selection in social insurance markets is required.

The thesis is structured as follows: I discuss relevant literature in section 2. Section 3 goes over the institutional background, explaining the Finnish entrepreneurial and social insurance systems with comparison to other countries. In section 4, I go over the data, it's characteristics and the empirical strategy I will use to analyse the presence of asymmetric information in the insurance market. Specifically, section 4.2.1 goes over the positive correlation test, which is followed by section 4.2.2 on the anticipation of absence. This is succeeded with the results and analysis which are presented in section 5. Lastly, section 6 concludes.

2 Literature Review

2.1 Moral Hazard and Adverse Selection

The amount of recent work has burgeoned in the field of moral hazard and adverse selection, especially within insurance markets. Although there has been substantial improvements in understanding and studying the subject, much of it is still left beneath the surface. There are mainly two strands of research stemming from insurance markets: normative theoretical literature and positive empirical research. The former largely focuses on analysing welfare implications of insurance schemes, while the latter addresses the impacts of private or social insurance programs. This literature has particularly focused on ways in which the impacts of asymmetric information can be ameliorated in insurance markets. (Chetty and Finkelstein, 2013)

There exists a compendium of theoretical literature on the impacts of asymmetric information in insurance markets, which can be ascribed to Akerlof (1978) and Rothschild and Stiglitz (1978), who greatly motivated research on this asymmetry. Both papers provide models in which it becomes evident that asymmetric information may lead to under insurance, thus necessitating welfare improvement through government intervention. This theoretical literature continued to surge; in particular the works of Einav et al. (2010) and Einav and Finkelstein (2011), where a simplified model of selection was introduced. This provided motivation for the lagged empirical work on selection and moral hazard. Einav et al. (2010) stipulates that adverse selection is characterised by a downward sloping marginal cost curve, which can be empirically estimated through demand for insurance and the average cost curve. The slope of the curve entails the type of selection: when downward, those with highest willingness to pay are more costly for the insurance company to cover. In this case, the marginal cost curve is decreasing in quantity and increasing in price. On the contrary, when the slope of the cost curve is positive, it entails advantageous selection: those who are most risk averse and least costly, have highest willingness to pay. In their paper, they provide tools which enable one to calculate the welfare cost of inefficient pricing in a market with adverse selection through standard consumer and producer theory. Their ideas are further discussed by Einav and Finkelstein (2011). Namely, by utilising the exogeneity of contracts being offered by insurance companies and assuming the prices of these contracts are endogenously determined, one can make use of the distortions in prices which arise from asymmetric information.

There are, however, only few studies empirically estimating the effects of moral hazard and adverse selection on markets where the welfare consequences of these are quantified. Exceptions, however, include Einav et al. (2010), Hackmann et al. (2015) as well as Seibold et al. (2022). Most past literature has focused on estimating the presence and existence of these phenomenons, which is also the focus of this thesis. For instance, Landais et al. (2021) investigate adverse selection and moral hazard within unemployment insurance markets in Sweden, where unlike in most countries, choice is available for unemployment insurance. A basic mandate is available to everyone, but individuals are given a choice to opt for larger unemployment coverage. This system is much like that of Finland. Landais et al. (2021) find that those opting in for supplemental unemployment insurance coverage are almost twice as likely to become unemployed. The dichotomy between the likelihood of becoming unemployed between those with higher coverage compared to those with lower coverage was found to be a result of a combination of adverse selection and moral hazard. The adverse selection, being a result of private information, leads to variation in willingness to pay for insurance. Moral hazard, on the other hand, can be seen from the responses of consumers to extra coverage received. However, due to the difficulty of controlling for adverse selection and moral hazard and due to the high moral hazard costs and low willingness to pay of low risk individuals, they conclude that a universal mandate is suboptimal. This is generally the consensus because only in the presence of adverse selection alone, should you mandate insurance for everyone.

Kolsrud et al. (2018) investigate the optimal timing of unemployment benefits, providing a compendium of policy implications. In the paper they analyse the relationship between value of insurance and moral hazard cost of unemployment benefits and how these change over the unemployment spell. Unlike Landais et al. (2021), the focus in this paper is on behavioural responses to differences in policies. For instance, they find that unemployment duration changes significantly as a result of benefit levels and the time at which these benefits are paid. While Kolsrud et al. (2018) focus on moral hazard, Hendren (2017) take on a different approach and focus on whether individuals have prior knowledge about job loss and use this private information to select into unemployment insurance contracts. He finds that knowledge about job loss leads to decreases in consumption and increases in spousal labour supply. This suggests the existence of frictions in the unemployment insurance market, which are caused by private information. The methods used in Hendren (2017) are further discussed in subsection 4.2, as they have motivated an empirical strategy used in this thesis.

In discussion of policy implications to fight against a severely adversely selected insurance market, inefficiencies arising from externalities are one of the major impediments of such markets. Often, individuals face prices that do not reflect their willingness to pay, thus causing an inefficiency in the market. Hendren et al. (2021) build onto the work of Landais et al. (2021) in discussion of subsidising more comprehensive plans to ensure their prices reflect optimal incentives of individuals and encourage those with higher willingness to pay, to opt in for more comprehensive insurance schemes. Similarly to Landais et al. (2021), Hendren et al. (2021) use the Scandinavian unemployment insurance market to study choice in unemployment insurance. Due to the heterogeneity in preferences for insurance, conditional on risk, they suggest a Pigouvian approach to counterbalance the adverse selection issue. This entails that a Pigovian subsidy could improve welfare of those on the margin of buying additional coverage. This could improve welfare for those who wish to have more comprehensive coverage, but have lower willingness to pay. By allowing for choice in the market, even in the presence of severe adverse selection, individuals are more likely to choose optimal coverage which maximises social welfare.

Recent work discussing the social determinants of choice quality are discussed by Handel et al. (2020). Like in the paper of Hendren et al. (2021), emphasis is placed on the idea that value of choice in insurance markets is high in the presence of individuals with heterogeneous willingness to pay or utility obtained from insurance. Hendren et al. (2021), however dismiss discussion on the fact that choice increases social welfare only if individuals are able to efficiently choose the option that maximises their utility. Handel et al. (2020) discusses the importance of this and how choice-based policies may increase inequality and in turn, be detrimental to social welfare. They find that approximately 60% of consumers would be better off choosing a more comprehensive insurance contract. This result was calculated using predicted health risks of the individuals and their realised contract choices. Lower deductible plans were found in individuals with lower education and individuals in the lower income quartile were found slightly less likely to choose high deductible plans compared to individuals in high income quartiles. They also found that social and informational networks have significant effects on choice in health insurance plans, which exacerbates inequality. Thus, with contrast to previous research discussed, they find that offering choice over insurance reduces the value of offering high deductible plans due to choice frictions.

Choice frictions may arise, for example, through a dichotomy between actual risk of individuals and their perceived risks. One may falsely perceive their risk as lower or higher than their actual risk, which could reduce the utility one can receive from their insurance plan. Much of previous work has followed the assumptions of Akerlof (1978) and Rothschild and Stiglitz (1978) where individuals are characterised only by their risk types, but recent work has refuted this assumption as individuals may also differ in their risk preferences. For instance, Spinnewijn (2017) analyses the association between risk perceptions and the willingness to pay for insurance relative to the willingness to exert risk reducing effort. He discusses that insurance markets with choice, but lack of adverse selection, may suffer from choice frictions resulting from differences in risk perceptions and risk preferences. That is, those who exhibit less risk averse behaviour, are likely to take less precaution and thus get less insurance coverage. In his model, there are two types of optimisms that can occur: baseline optimism and control optimism. The former being those who are optimistic about their baseline risk level and the latter being those who are more optimistic about the outcome, so change their effort to avoid risk accordingly. He shows that these choice frictions can lead to the separation of the true value of insurance and the value of insurance, which arises from the individual's demand. Thus, as emphasized by Spinnewijn (2017) as well as Landais et al. (2021), one must take precautions in interpreting value of insurance due to the possible presence of such caveats.

This has also been researched empirically by Ericson et al. (2021), who place emphasis on the idea that individuals' risk profiles are comprised of both their risk type and their risk preference. This leads to the inability to empirically distinguish whether an individual has low degree of risk aversion or whether they have overly optimistic beliefs about their risk level. By using data on preferences based on choice and exploiting the variation in comprehensiveness and types of insurance plans, they find evidence which suggests that consumers have distorted perceptions about their risk exposure.

Much of the findings can be extrapolated to the discussion of entrepreneurs. If they are offered choice in insurance, like in any other market, there is a possibility for adverse selection and moral hazard. However, as entrepreneurs are a selected sample, some of the choice frictions in markets may differ from those of the general population. Public policies may also effect entrepreneurs differently, so one must take precautions when discussing policy implications to ameliorate asymmetric information in the insurance market for entrepreneurs.

2.2 Entrepreneurship

Literature has proliferated much less in insurance markets for entrepreneurs compared to other markets. Most of the research has been motivated by the salient differences in insurance schemes provided for entrepreneurs and wage earners. Wagener (2000) takes on a theoretical approach and suggests that in lieu of a dichotomy between wage related pension schemes between workers and entrepreneurs, they should be offered the same level of insurance. The discrimination between entrepreneurs and wage earners in the pension design system largely affects the decisions of individuals in employment choices as these pension systems often provide individuals with basic social protection.

Empirical research has placed much of their focus on studying how the generosity of benefits affect whether one chooses to be self employed or be in the standard form of employment. Xu (2022) as well as Røed and Skogstrøm (2014) study this in the context on unemployment insurance. Xu (2022) finds that higher unemployment benefits decreases the probability that someone chooses self-employment and lengthens the transition time from unemployment into employment. Similarly, Røed and Skogstrøm (2014) find that transitions from unemployment to self-employment are largely affected by unemployment insurance. Namely, the probability of becoming self-employed surges at the point of unemployment insurance exhaustion, which describes the situation after which one cannot accrue any more unemployment benefits as the claim has been paid out. This finding is also corroborated by Kolsrud et al. (2018).

The effect of generosity on employment choices is also studied in the context of health insurance. Jackson (2010) suggests that when faced with health insurance mandates, potential entrepreneurs are more likely to stop seeking an entrepreneurial status and find implications that current entrepreneurs are likely to seek actions that would minimise such mandate costs. Heim and Lurie (2010), on the other hand, study the effects of a policy, which enabled higher deductibility in health insurance premiums. They find that a higher deductible led to an increase in self-employment and a decrease in exit from self-employment. Unlike, Jackson (2010), Bailey (2017) finds that a dependent coverage mandate increased self-employment among disabled young adults significantly, but found evidence of no changes in self-employment among the general population of youths.

Furthermore, Perry and Rosen (2001) study actual insurance purchases made by entrepreneurs and the extent to which they utilise health care compared to wage-earners. They describe their findings as an aberration, as despite having much lower insurance purchases, the utilisation of health care services did not differ significantly between the two groups. Despite entrepreneurs having much lower coverage in insurance plans, they utilised health care services similarly to wage earners. Thus, they provided evidence that the general public policy concern over health care for entrepreneurs is slightly erroneous.

Boeri et al. (2020) take on a different approach where they describe the dichotomy in demand for social security between solo self-employed and self-employed with workers. Using OECD data, they find that solo self-employment is burgeoning, with a decreasing number of entrepreneurs employing workers. As solo self-employed individuals tend to earn less on average, they are also more liquidity constrained. As a result, they are more vulnerable to idiosyncratic shocks. Through the use of survey data, Boeri et al. (2020) find that solo self-employed individuals have higher willingness to pay for social protection when compared to wage earners. They also discuss the difficulty of contemporaneously addressing moral hazard and adverse selection in insurance markets for entrepreneurs while allowing them to alter their working status and incomes flexibly.

Entrepreneurship has also become a topic of interests to researchers in Finland. Namely, Hyrkkänen (2009) describes how entrepreneurs under insure themselves. By using survey data, she also describes the differences in insurance contributions based on characteristics and finds that women tend to report their income more truthfully compared to men, thus on average pay higher insurance contributions relative to their incomes. Differences between insurance payments seem to also be driven by industries. Furthermore, she discusses how under insurance can lead to great welfare losses due to the lack of social security and the risks that come with self-employment.

Benzarti et al. (2020), on the other hand, analyse the effects of a reform in 2011 which relaxed the social insurance mandate for entrepreneurs and led to a significant decrease in social security contributions of entrepreneurs in Finland. They find that the surplus of money from the reduction in social contributions were channelled into business activity. This impact was heterogeneous, however, as younger firms were more likely to use the surplus for increasing business activity while older firms were more likely to purchase stocks with the objective of bettering their fiscal position.

As seen from previous research on entrepreneurship and social insurance, there seems to exist a caveat. No prior research has paid particular reference to the presence of asymmetric information in these markets. The aim of this thesis is to narrow the gap between asymmetric information in insurance markets and insurance markets for entrepreneurs.

2.3 Pensions and saving

A large amount of research has focused on pension schemes and the role of pensions in saving. Central questions in this field are for instance whether mandatory pensions crowd-out personal saving and what the impacts of participant-controlled plans are. This subject in particularly well communicated in Bernheim (2002). With some reference to adverse selection, he questions whether providing mandatory pension schemes would be effective in ameliorating the caveats arising from asymmetric information in private markets. However, the problem of this possibly crowding out personal savings becomes particularly salient. He discusses two separate cases: participants who have no choice over their level of contribution and participants who do. For the former, the general consensus has been that there are hardly any effects of pensions on savings (Gordon and Blinder, 1980; Diamond and Hausman, 1984). However, multiple papers such as Cagan (1965) and Katona (1965) have corroborated the finding that those, who have no control over contributions, may display crowding in effects on savings. Contrarily, not many papers find significant crowding out effects (Munnell, 1976). For the latter, with particular reference to 401(k)s, there seem to be heterogeneous effects depending on the rate of substitution. [3] When participants have control over contributions, policies which stimulate 401(k)s may result in either redirecting money from other saving into pension savings or increase savings in general. Most often this is when the rate of substitution is low or high respectively. Whether either type of pension scheme accumulates wealth or not is still left uncertain due to limitations on data and methodology.

In his model, Bernheim (2002), assumes individuals are not liquidity constrained and they borrow and lend money at the same rate. However, this may not have pertinence in the real world. Liquidity constraints, uncertainty as well as other choice frictions may play a large role in the reaction to policies affecting pension savings. Chetty et al. (2014) also elucidates these by describing a model with passive and active savers. Passive savers are described as individuals whose automatic pension contributions are unaffected by policies or subsidies, whereas active savers who maximise the utility of savings. Through this model, they test how two different policies effect savings. They find that a price subsidy has no impact on passive savers, but it has a positive impact on active savers with regards to pension savings accounts. Moreover, they find that automatic contributions through salary has no effect on savings for active savers, as they can shift savings from one savings account to another, while passive savers were found to have ambiguous results. As 85% of the Danish population are considered as passive savers, it is unlikely that subsidies would increase savings. However, they conclude that

[3] 401(k)s are prominently used in the United States. They are employer-sponsored retirement savings plans which offer tax benefits and aim to help individuals plan their savings for future retirement.

automatic contribution policies may lead to higher rates of saving.

As Finnish entrepreneurs have choice over pension contributions and wage earners do not, the effects of mandates which aim to increase savings may be widely heterogeneous. This may also be the case within these groups due to the difference in nature of participants. Entrepreneurs are unlikely to be representative of the full working population, thus may have a larger share of active or passive savers.

3 Institutional Background

Today, there are almost 370 000 enterprises in Finland, which employ about 40% of the working population (Statistics Finland, 2020). Finland has experienced steady growth in creations of new enterprises since 2001 until the financial and economical crisis of 2008, which had significant, negative and long lasting impacts on the economy. Similarly to most countries, the creation of new enterprises declined significantly as a result of the crisis. This had a negative impact on the GDP as enterprise births are one of the largest sources of employment. (OECD, 2018) Enterprise births may not result in such large surges in employment in Finland, as nearly 90% of enterprises are small, having less than 5 workers. The European Employment Observatory also indicates that self-employment may not contribute to job creation in Finland to a large extent as the willingness to expand their entrepreneurial operations are not high. Hence being one of the reasons why Finland chooses to focus on policies encouraging enterprise expansion and compared to other countries such as France or Germany, focus less on policies encouraging the unemployed to return to the labour market through self-employment. (Hawley et al., 2010)

Social security plays a major role when deciding to become self-employed. Often, the self-employed are considered to be more at risk, as they are not necessarily entitled to the same social protection as wage earners. For instance, in Germany, self-employed individuals can voluntarily choose to take part in health and unemployment insurance schemes. In other European countries, self-employed persons may receive lower insurance benefits or face higher costs. They are especially at a disadvantage regarding pension, illness, disability and paid parental leave. Due to the moral hazard costs of self-employment, some countries, such as Turkey, have decided to not provide the self-employed with any social protection at all. On the contrary, countries such as Denmark, provide the same social security to entrepreneurs as wage earners. This can be complicated, as social insurance systems funded by government-mandated contributions are most often designed for individuals with stable wages obtained from standard forms of employment. Perhaps being one of the reasons why Finland, on the other hand, has an insurance market where entrepreneurs are given choice. This is done similarly in the Czech

Republic and Hungary, where the self-employed are able to exploit choice and in turn manipulate their level of insurance contributions. This, unfortunately, enables self-employed individuals to pay lower contributions, which leads to lower levels of benefits or adverse selection and moral hazard. (Hawley et al., 2010)

In Finland, differentiation is made between wage and non-wage earners in the social insurance program. The social insurance mandate is compulsory for both wage and non-wage workers in Finland. Wage earners are automatically a part of TyEL. Entrepreneurs, on the contrary, must acquire YEL insurance if the following applies to them: they are living in Finland, have had a continual entrepreneurial status for at least 4 months, be between ages of 18 to 68, not be a part of any other private pension system and earn a yearly income of at least 8 261,71 euros (Pohjola Vakuutus, 2022). If these do not apply to them, they must acquire TyEL insurance. Both of these are pension contributions which provide the basis of statutory social security and are determined by their social insurance contributions.

YEL and TyEL are the basis from which pension and social security benefits are determined. Each individual is entitled to allowances such as sickness allowance and parental allowance, vocational rehabilitation benefits, unemployment security as well as pensions such as old age pension or disability pension. The level of these benefits is directly influenced by reported income (YEL) or salary (TyEL). For YEL, the level of the contributions is reflected by their reported income which is set by the entrepreneurs independently. The reported income should represent the yearly wage that someone in their position would obtain for their work effort, given they would be a wage earner. In other words, it represents the monetary value of their work input. The benefits are calculated according to the verified annual earnings; for individuals covered by TyEL, annual earnings are verified by the tax authorities at the end of each year, thus the annual earnings used to calculate benefits is that of two years prior. For individuals covered by YEL, this is not the case. Their reported income is verified by the insurance companies in the same year and their benefits are calculated using the verified income of the previous year.

Sick Pay

Annual Earnings (€)	Replacement Rate (RR)
1 325 - 26 898	70%
26 899 - 50 606	35%
50 607 -	25%

Parental Allowance (Male)

Annual Earnings (€)	1st 30 Days RR	RR
1 325 - 32 892	75%	70%
32 893 - 50 606	75%	40%
50 607 -	32.5%	25%

Parental Allowance (Female)

Annual Earnings (€)	1st 56 Days RR	Next 30 Days RR	RR
1 325 - 32 892	90%	75%	70%
32 893 - 50 606	90%	75%	40%
50 607 -	32.5%	32.5%	25%

Table 1: Benefit system

Notes: The limits for replacement rates are denoted by annual earnings from 2012. The discontinuity points may change slightly on a yearly basis, but remained fairly similar throughout 2001-2015. Note also that there is a set minimum benefit an individual may receive. In 2012 this minimum benefit was €22.96.

Sick pay (and parental allowance) are calculated in the same way for individuals covered by YEL and those covered by TyEL. Until 2015, for annual earnings up to 26 898 euros (32 892 euros for parental allowance) the replacement rate was 70%, after which the rate decreased to 35% (40% for PA) until annual earnings of 50 606 euros. For earnings above 50 606 euros, the replacement rate was 25%.

The relationship between annual earnings and replacement rate for sickness and parental allowance can be seen in table 1 and graphically in figure 1. Additionally as of 2007, mothers are entitled to maternal allowance with a replacement rate of 90% for the first 56 days for annual earnings below 50 607 euros, after which the replacement rate falls to 32.5%. Both mothers and fathers are also entitled to parental allowance with a replacement rate of 75% for the first 30 days with the same annual earnings boundaries. Regarding sickness allowance, it is also important to note the difference in personal responsibility duration between YEL and TyEL participants: if an individual contributes through YEL, they can receive sickness allowance one day after falling ill, however those contributing through TyEL have personal responsibility duration of 9 days, meaning one can only receive sickness allowance after nine consecutive days of illness.

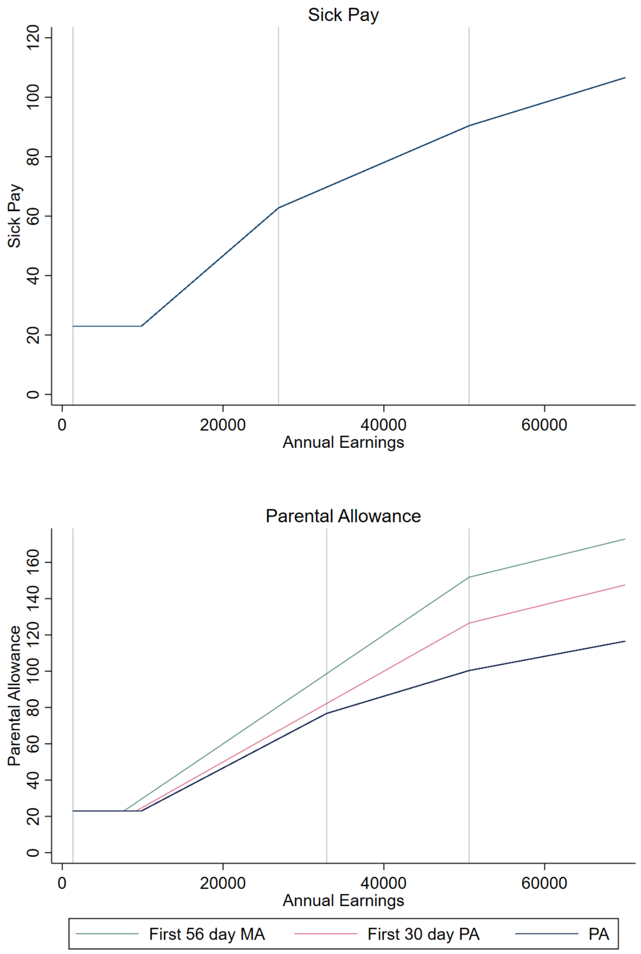

Figure 1: Replacement Rates

Notes: Panel a (above) shows the relationship between prior annual earnings and daily sick pay. The vertical lines represent the discontinuity points at minimum earnings for eligibility and lower as well as upper kink points for the replacement rates. Panel b (below) shows the same relationship for parental allowance, with three different lines. The upper most line shows the relationship for the first 56 days of maternity allowance, the middle line shows the relationship for the first 30 days of parental allowance and the lowest line shows the relationship prior to 2007 or post 2007 for ex post special days.

Due to the flexibility of the system, entrepreneurs often under-insure themselves and the possibility of adverse selection and moral hazard is high. As a result, I expect to find positive correlation between sickness and insurance contributions. For parental allowance, I expect to find less positive correlation as it is likely that having children itself increases contribution levels due to higher risk aversion. Thus, it is unlikely that those who have longer parental leaves have higher contributions. In lieu, I expect to find evidence of dynamic adverse selection for parental allowance as having children is often planned, thus allowing for individuals to increase insurance contributions prior to taking out parental allowance. Although, I expect it to be less the case with sick pay, as it can be difficult to anticipate sickness due to its unexpectedness in most cases. Following Benzarti et al. (2020), I will henceforth refer to those whose pension contributions are paid through YEL as Y owners and those whose pension contributions are paid through TyEL as T owners throughout this thesis.

4 Empirical Context

4.1 Data

I use data from two sources: (1) panel data from two large Finnish pension companies and (2) combined individual level data on income and tax returns. The former contains the social insurance contributions of entrepreneurs and the latter covers characteristics of the individuals and their businesses. Both of these data sets were obtained from Statistics Finland. Thus, they include unique identifiers, which can be used to link them together. Altogether the sample consists of 2,669,000 observations for 543,600 individuals for which 49% of observations are those of T owners and 51% are those of Y owners. This panel data set is unbalanced, as individuals have different number of years available for their insurance contributions. Some individuals may also have gaps between the years in the data.

Insurance Contribution Data. — Data on contribution levels for entrepreneurs was obtained from two large Finnish pension companies, which account for approximately 70% of all entrepreneurs in Finland spanning the years 2001-2014 for both of the companies. This data contains the insurance contribution amount as well as the confirmed reported income for each individual on a yearly basis. Unlike entrepreneurs, wage-earners have no discretion over their social contribution payments as they are calculated directly from their taxable income. These were thus calculated using the income and tax return data based on the basic TyEL contribution for each year.

Income Data. — Individual level income and tax return data contain rich demographic information, including age, annual earnings, gender, education, martial status, number of children as well as received and paid transfers for all individuals residing in Finland. The data covers everyone in Finland from 1987 until 2020, of which years from 2001 to 2014 are used. This data was linked to business data containing tax returns of entrepreneurs at individual and firm level from 2005 until 2015. Namely, this data consisted of sole proprietor's, partnerships as well as main shareholders of businesses. In addition, the data was combined with more comprehensive entrepreneur business data containing all shareholders of businesses, including shareholders who own less than 30% of the shares of the firm.

In this paper, the focus is on social contributions and social insurance benefits received by Y and T owners. Specifically, the focus is on accrued sick pay and parental allowance. The data allows for further analysis on unemployment insurance benefits as well as pensions, but I choose to analyse only sick pay and parental allowance, as the benefits from such are calculated in similar ways. Each individual is entitled to either of these benefits, whereas for supplemental unemployment insurance coverage, this is not the case. One must have insurance contributions exceeding a specified amount to be entitled to supplemental unemployment insurance, for which the sample is very small. For pensions, on the other hand, I am unable to analyse the years after individuals receive a pension, as those receiving one often opt out of the YEL insurance scheme as they no longer wish to partake in entrepreneurial activity. Those who do continue might be a very selected sample.

Observations		236,445				
		Y owners (55.7%)			T owners (44.3%)	
		YEL			TyEL	
	Ext.	Mean	sd	Ext.	Mean	sd
Female	34.0%			32.9%		
Shareholders	31.9%			100%		
Sole-proprietor	50.9%					
Partnership	17.1%					
Age		46.47	10.47		44.24	11.73
Income		39065.97	76564.66		53310.08	132120.30
Insurance Contribution		3669.67	3110.91		7879.07	8479.78
Sick Days (proxy)	7.1% (5.3%)	49.44	60.62	2.1%	58.63	68.99
Parental leave (proxy)	2.0%	65.05	63.37	3.2%	78.89	70.69
Pension	7.2%	11527.30	12503.61	8.9%	21877.26	24609.47

Table 2: Sample statistics (2010)

Notes: Note here that "Ext.", meaning extensive, shows the percentage of individuals in the data which fall into the specified category. For example the extensive margin for sick days means that only 7.1% of individuals in the data in year 2010 have taken sick days. The mean is the average sickness days, conditional on being sick. Note also the 5.3% for sick days shown in brackets refers to the percentage of Y owners who are on sick leave for longer than 9 days. This allows for a better comparison between Y and T owners.

Table 2 presents the summary statistics for the sample for the year 2010. [4] For the baseline sample, there are 236,445 observations, of which 56% are Y owners and the rest are T owners. Each group contains a similar share of women, while the mean age for Y owners is slightly higher compared to T owners. As sole proprietors and partnerships always contribute through YEL, the sample for T owners is made up of solely shareholders whose ownership is less than 50% in the year 2010. In the table, we see both income and insurance contribution to be much higher for T owners. When only shareholders are compared between the two groups, mean incomes are indistinguishable, while insurance contributions remain distinctly dissimilar. The most notable distinction between Y and T owners is the differences in share of individuals receiving sickness allowance. One reason could be the systematic difference for when one is entitled to sickness allowance, but even after obviating the difference, the share of individuals receiving sickness allowance is much higher for Y owners. Despite this, the average number of days on sick leave is higher for T owners. A possible explanation for this distinction could be the difference in mean ages or due to the systematic difference; as sickness allowance is more quickly obtainable, Y owners could be more likely to get it and stay on sick leave for more than 9 days.

4.1.1 Data Limitations

Aforementioned, the data on contributions comes from two insurance companies, which cover 70% of entrepreneurs in Finland. As the data does not cover all entrepreneurs, one may question the validity of the results due to the possible presence of a selection bias. Using summary statistics provided by Statistics Finland (2020) on the full set of entrepreneurs in Finland, the characteristics of the sample can be compared to the statistics on all entrepreneurs. I find no differences between the size of firms or types of industries present, thus obviating the feasible selection bias issue.

A limitation in the data also arises from the absence of number of sick days or days on parental leave. The data provides the benefits paid to each individual in a year, but not the days nor the daily benefit amount one would receive during

[4] Only one year is included for simplicity, but summary statistics for other years have infinitesimal differences to the year 2010.

the days of absence. To circumvent this limitation, I estimate the number of days for each individual, however, this may lead to measurement errors. To ensure the estimates are somewhat accurate, I compare the means and distributions of the days to other studies using data where the number of days is available. I find that the distributions and means do not deviate from other studies to a large extent. While some differences may be explained by measurement errors, especially with outliers, some can be explained by differences in samples. Entrepreneurs are a specific group of individuals who are unlikely to be representative of the full population.

Furthermore, each individual is not observed for the same number of years, thus the panel data is unbalanced. This is unlikely to affect the analysis. As observations are ”missing” at random, a bias arising from the unbalanced nature of the data is implausible. In the panel data, observations are made at one year intervals, so only the number of days of absence due to ill health or having children during the full year can be estimated. I cannot therefore distinguish whether individuals are for example sick for x number of consecutive days, or whether absences are spread out throughout the year. It can be difficult to draw conclusions about whether individuals with more severe illnesses (many consecutive sick days) or individuals with multiple less severe illnesses (spread out sick days) could be driving the results. I can only make conclusions about the differences or the absence of differences between individuals with higher risk (many sick days) and lower risk (less sick days). The case is akin with parental allowance. The birth dates of children are not available in the data, so if a parent gets two children in the same year, I cannot distinguish how much of the parental allowance is for the first child and how much for the next. This may cause a measurement error in the number of days estimated.

4.1.2 Data Characteristics

The differences in social security between Y and T owners in Finland is a heavily discussed topic. It is a well known fact that Y owners have lower insurance contributions than T owners. The implications from this fact are discussed by Hyrkkänen (2009). No clear comparison between these two groups has been shown using other than survey data, which claims that firstly, Y owners under-insure themselves and secondly, that women tend to pay higher contributions than men conditional on income.

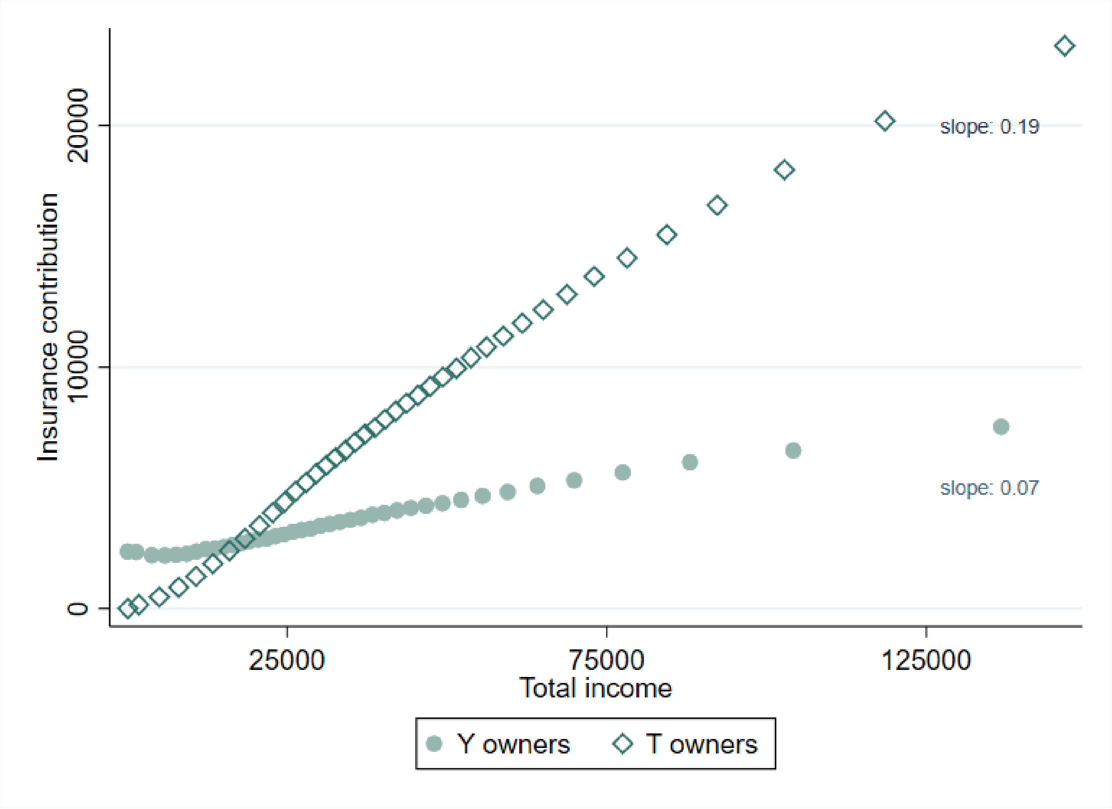

Figure 2

Notes: Each point contains 2.5% of observations with respect to their group. Total income represents the taxable income of individuals. The slope coefficients for both lines are shown below the lines. We can note that the slope of the line for T owners is almost twice that of Y owners.

To further elucidate the first claim, we can use the data to see how insurance contributions change with income for both Y and T owners. Figure 2 presents the contributions of Y and T owners with respect to income. We see a major dichotomy between these two groups; Y owners contribute less to the social insurance system and the relationship between income and insurance contributions for Y owners is largely different from that of T owners. While it is evident that there is a rather linear upward sloping trend for T owners, the relationship is much flatter for Y owners, however, both contribution levels are increasing with income. We can see that up to total annual income of about 10,000, Y owners have higher contributions compared to T owners. This could be explained by the floor set on Y owners for the minimum reported income, from which the contribution level is taken. From 10,000 onward, for the same level of income, insurance contributions for entrepreneurs who pay through TyEL are much higher than those for entrepreneurs paying through YEL. As Y owners are given choice, they seem to be minimising the costs from insurance contributions and while doing so, possibly putting themselves at higher risk if they fall ill or become unemployed.

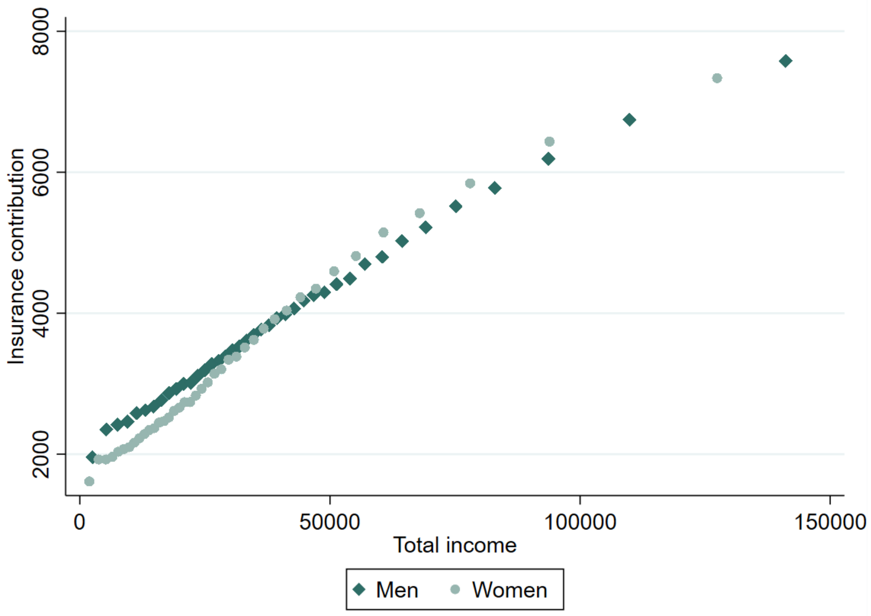

Figure 3

Notes: The figure shows the public pension insurance contributions 2005-2015 for men and women with respect to total annual income. Each point contains 2.5% of the observations within each group and total income is measured as taxable income.

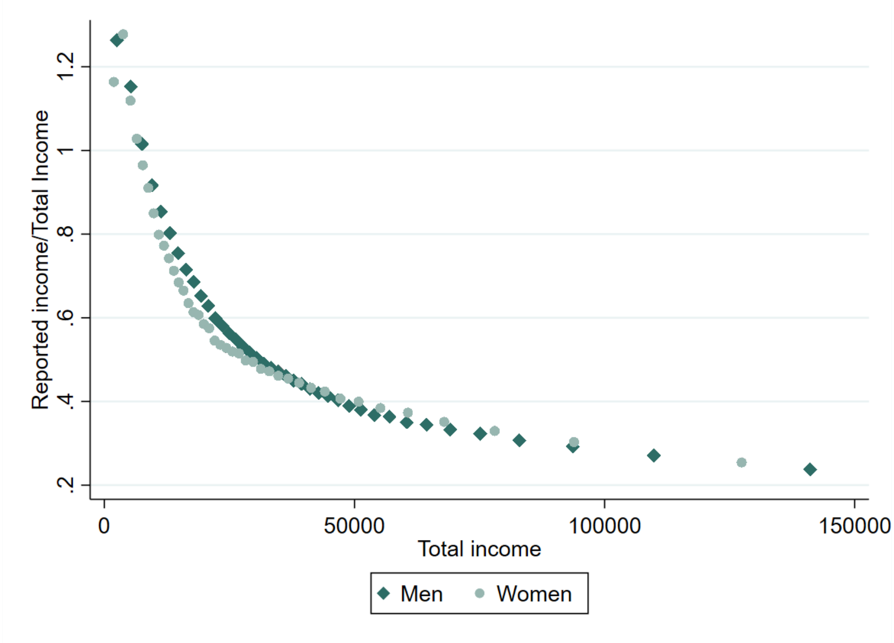

Figure 4

Notes: The figure shows the relationship between reported income as a share of total income and actual income for men and women. Each point contains 2.5% of the observations within that group. Total income is measured as taxable income.

The data also allows us to construe the second claim. Figure 3 shows the insurance contributions (y axis) associated with the total income (x axis) separately for men and women. The two lines do not differ to a large extent. One can note that insurance contributions may be slightly higher for men in the lower income quartiles compared to women, but the opposite is seen in higher income quartiles. Hyrkkänen (2009) specifically points out from survey data, that women's reported income is closer to their real income, thus implying that women report their income more honestly compared to men. Whether one can make conclusions about gender differences in deceptive behaviour, however, is unlikely. Although, for example, Lohse and Qari (2021) study this and find there to be an indistinguishable difference in behaviour between the two sexes when audits reporting their income were computerised.

Reported income and actual income is graphed using the data in Figure 4. The relationship between these is the same regardless of sex. The difference is not driven by sex itself, but rather income, as those in lower income quartiles report their income more truthfully than those in higher income quartiles. On average, women earn less than men (Blau and Kahn, 2017), and this also becomes evident from the data used here. Thus, as the distribution into income quartiles differs between men and women, on average women's reported income is closer to their actual income when compared to the sample of men. The claim is therefore not erroneous, but could be misleading. From Figure 4, we can also note that individuals with very low earnings pay contributions that are on average over 1.5 times their income. One viable reason for this could be that there is a set minimum reported income Y owners must contribute. Individuals in lower income quartiles have also been found to be more risk averse and have higher risks of falling ill, which could partly explain the decreasing ratio between reported income and total income.

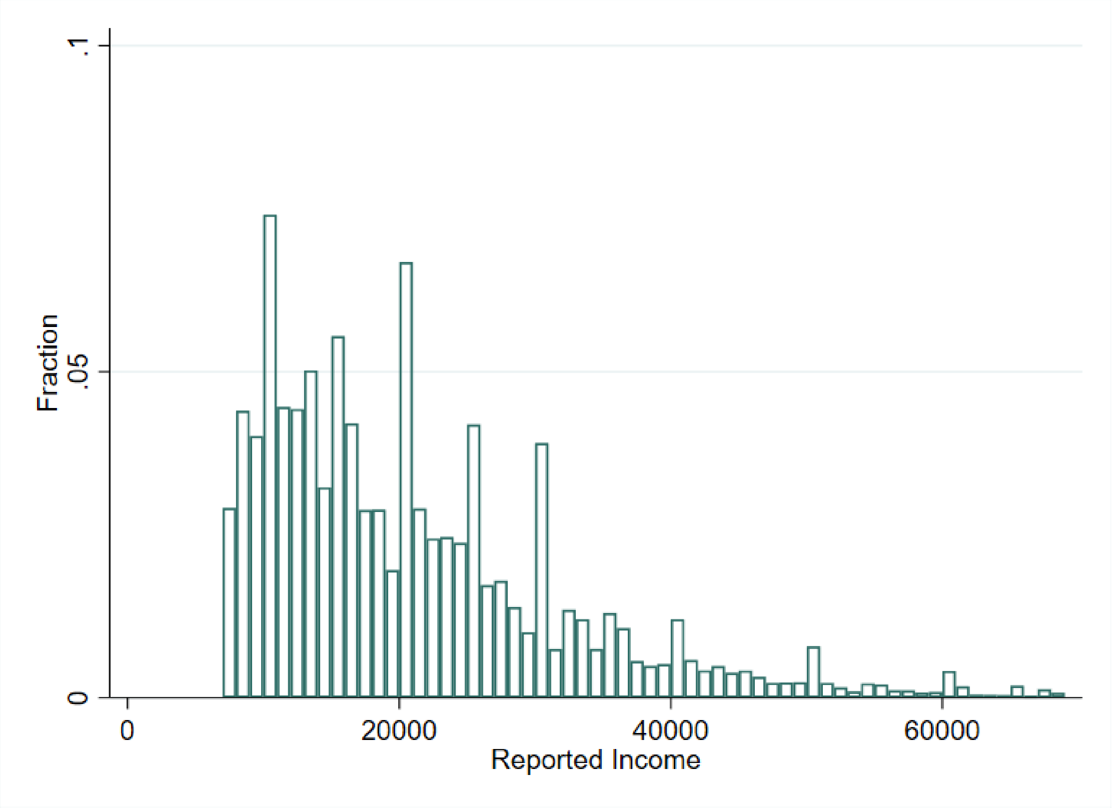

Figure 5

Notes: The figure shows the distribution of reported income for the year 2010. The distributions are similar for each year with bunching at round values.

Interestingly, figure 4 shows that reported income falls down to 50% of total income for individuals earning above the median wage. Even though reported income should be representative of the monetary value of their work input, individuals tend to undervalue their work efforts. This dichotomy between reported income and actual income may also be a result of asymmetric information. That is, individuals may find it difficult to give clear monetary value for their work efforts as the nature of their work is not like that of a typical wage earner. Figure 5 constructs the distribution of reported incomes for each individual. A large share of individuals report their income in the lowest possible income quartile, while most others bunch at round numbers. Rounded values of income are unlikely to be accurate, as most earnings do not fall within round values. Rounding, however, is a very common occurrence with reported incomes; rounding behaviour has been found in, for instance Niskanen and Keloharju (2000) and Schweitzer and Severance-Lossin (1996).

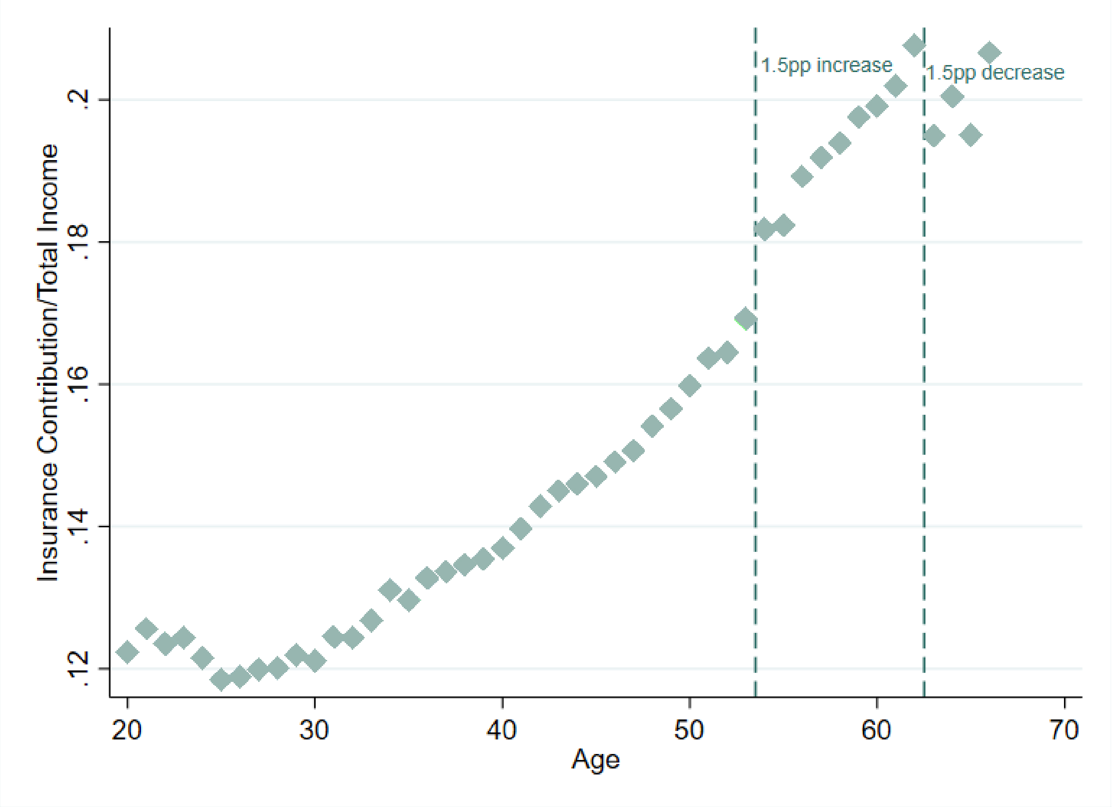

Figure 6

Notes: The figure displays public pension insurance contributions by age. The vertical lines represent percentage point changes in contribution between ages 53 and 62.

Age is another confounding factor that may have large effects on insurance contribution. For instance, Chen et al. (2001) find evidence of large differences in life insurance purchases made by different age cohorts. Individuals face different risks at different ages, thus they are likely to insure themselves differently throughout the years.

Ageing is characterised by gradual accumulation of impairments in bodily functions as well as increased risk to disease and health shocks (WHO, 2015). Due to such increase in health risks, one could expect the willingness to pay for insurance to increase with age. Figure 6 describes the relationship between insurance contributions and age. There is a clear increasing relationship between insurance contributions and age conditional on income. The relationship is smooth, with some discontinuities arising from institutional changes in contribution share; between the ages of 53 and 62, the share of contribution is slightly larger (1.5 percentage points higher). As there is a clear jump in insurance contributions, we can also conclude here that individuals are not actively changing their contribution levels to avoid increased contribution shares. This increasing relationship between age and insurance contributions is also in line with the life-cycle hypothesis as the desire for liquidity is stronger at the earlier stages of the life-cycle while saving in terms of pension seems less relevant. [5]

4.2 Empirical Strategy

This section documents two main empirical strategies. First, I aim to estimate whether adverse selection or moral hazard exists in the social insurance market for entrepreneurs using a simple correlation test. Secondly, I investigate the dynamic relationship between insurance contributions and time of absence through estimating the changes in insurance contributions surrounding the time of absence. For the latter, the aim is to investigate the extent to which individuals anticipate absence due to illness or children and the extent to which individuals act on knowledge of future absence.

4.2.1 Positive Correlation Test

There are several challenges in estimating the extent and effects of moral hazard and adverse selection in insurance markets. This is due to the difficulty of drawing

[5] More data characteristics are available in the appendix A.1. Namely, the contrasting distributions between Y and T owners for level of contributions are given in figure 13. Additionally, figures 14 and 15 show contribution levels by organisational form and by number of workers respectively. Lastly, distributions for number of sick days and length of parental leaves can be found in figure 16.

causal conclusions, which is partly a consequence of adversities arising from distinguishing AS and MH in an empirical context. Both result in a positive correlation in insurance claims and insurance contribution levels. That is, if insurance contributions are higher for those with higher risk levels, it is difficult to know whether high contribution amounts result in higher ex post risk levels (MH) or whether high contribution amounts are a result of ex ante asymmetric information about risk levels (AS). In this thesis, the aim is to evaluate the extent of adverse selection and moral hazard in the market through descriptive analysis and only make suggestive conclusions about whether this is moral hazard or adverse selection. I use a positive correlation test to analyse how insurance contributions correlate with risk levels, which in this case are measured through parental allowance and sick pay.

Einav and Finkelstein (2011) provide a theoretical guide to using a positive correlation test empirically. They suggest a graphical framework which allows one to compare the expected cost of those who are insured more to those who are insured less. Using a price-quantity space, they suggest that if the average cost curve of those with more insurance is consistently above that of those with less insurance, this is suggestive of adverse selection or moral hazard in the market. This test can be implemented by using proxies for expected costs. In Einav et al. (2010), they also show that if the marginal cost curve for insurance is downward sloping, this is indicative of adverse selection as those with higher willingness to pay are also more costly. This is because as price falls, the average cost of contracts decreases; the marginal individuals who choose a contract with higher coverage have lower expected cost than individuals just below the margin. They provide a theoretical framework for quantifying the welfare effects of adverse selection; by knowing the marginal cost curve, average cost curve as well as the demand curve for insurance, we can estimate the welfare losses arising from adverse selection graphically.[6]

In their setting, the average expected cost curve (AC) is computed using the average incremental cost for each individual that chooses a more comprehensive

[6] If the intersection between the marginal cost curve and the demand curve, which gives us the efficient allocation, differs from the intersection between the average cost curve and demand curve, which is the competitive equilibrium allocation, the dead weight loss arising from this difference shows the estimate for the welfare cost due to adverse selection. (Einav et al., 2010)

contract at a given relative price. The incremental cost indicates the difference in cost to the insurer as a result of a difference in having a more comprehensive contract compared to a less comprehensive contract. The AC curve then estimates how the average incremental cost varies relative to price variation in the higher level contract.

With the theoretical framework provided by Einav and Finkelstein (2011) in mind, positive correlation tests have previously been used in Landais et al. (2021) as well as Seibold et al. (2022). Seibold et al. (2022) use a positive correlation test through regression analysis, correlating post-reform private disability insurance take-up with disability risk determined by occupation. In their setting, a reform enables them to measure differences in private insurance take up with respect to different bins of unpriced risk. Using these bins, they are able to calculate the probability of private insurance take-up for each. They find no positive correlation, indicating no adverse selection in the market. Surprisingly, they find a modest negative relationship, suggesting slight advantageous selection. This indicates that high risk individuals are less risk averse while low risk individuals are highly risk averse, resulting in a negative correlation between risk and coverage.

Landais et al. (2021) on the other hand, use the positive correlation test to test for adverse selection or moral hazard in the unemployment insurance market. They correlate the total number of days spent in unemployment in period $t+1$ with the insurance choice made in period t. Here, the insurance choice made in period t gives the required variation in price, where realised risk is a measure of the cost for the insurer. They find substantial positive correlation between unemployment insurance coverage and realised risk.

Following their ideas, the aim of this thesis is to see whether there is a positive correlation between insurance claims for sick pay or parental allowance and insurance contributions. I first estimate the number of days an individual has been ill or on parental leave. If they have not received either of these benefits, the number of days spent out of work will be zero. The number of days out of work will represent the cost of each individual; the more days one takes out of work, the more benefits they accrue. The days are calculated through estimating the amount of benefit an individual would receive each day according to their reported income. This is calculated using the rules mentioned in section 3. This amount is then divided by their accrued benefit amount, which then gives us an estimate of the days one has been ill or the amount of days one has been on parental leave during the year. I calculate the share of days one has been out of work throughout the full panel. This is done in order to account for the unbalanced data; some individuals are present throughout the full panel, 2001-2014, while others are not. Working with means allows me to handle the data in cross-sectional form.

I estimate the correlation between the share of sick days or the share of days on parental leave and mean insurance contribution amount. I estimate the correlation using OLS, thus assuming that the correlation is linear. That is, I estimate the following equation

$$\pi_i = \alpha_0 + \alpha_1 \log(c_i) + \beta X + \varepsilon_i \tag{1}$$

where π_i is a measure of risk (share of days out of work) for an individual i, c_i is the level of contribution, X is a set of controls such as annual income, age, sex and occupation and lastly ε_i is the zero mean error term which may exhibit heteroskedasticity. Both the conditional and unconditional correlations will be measured. In other words, the correlation will be measured with and without controls.

I will then show this relationship graphically. The mean insurance contributions will give us the variation in the so called price of contracts, as individuals are given a choice in the level of insurance through contribution amount. This variation is determined endogenously, thus differs from the settings of other studies. For the graph, I create insurance contribution bins which each contain 5% of the sample. For each bin, I calculate the average days of sickness or parental leave. This will be a measure of probability of sickness or having children, where the outcome will be the average share of days each individual has spent out of work due to either sickness or parenting. If there is a clear correlation between risk and contribution level, whether it be negative or positive, asymmetric information is conducive; individuals react to having choice in the social insurance market where differences in reactions are driven by risk.

As discussed by Einav and Finkelstein (2011), there are several caveats one must consider with the use of positive correlation tests. For instance, aforementioned, using a positive correlation test does not allow one to disentangle adverse selection and moral hazard as both result in positive correlation of insurance claims and coverage. Due to this, one is not able to make any policy implications from findings using this method, as policies addressing either type of asymmetric information are very different. This caveat is one, which I do not aim to overcome. The aim is to only provide descriptive analysis on the insurance market for entrepreneurs and to provide evidence of whether there is a presence of asymmetric information or not. [7] Another caveat which must be considered while using a positive correlation test is the difficulty of conditioning on covariates. These determine whether a positive correlation arises due to self-selection into different insurance contracts or through being offered different contracts as a result of buyer characteristics. I claim this is not an issue in this analysis as each individual must pay social insurance contributions, the level of which is a choice independent of confounding factors other than liquidity constraints. Thus, supply does not play a role and each individual chooses the level at which they wish to contribute to social insurance and this social insurance provides them with the same benefits at similar replacement rates. Thus, differences in insurance claims only arise from dissimilarities in individual characteristics. Most of these characteristic differences between individuals are observable, thus allowing me to compare groups which differ in risk. There is, however, a possibility that some correlation is driven by unobservable characteristics. Lastly, the expected costs of individuals are not always straightforward, thus one must use proxies to evaluate them. Good proxies may not always be available through data, making the estimation of expected costs difficult. In this analysis, I use realised costs, giving me a direct measure of the cost of each individual, through insurance claims, thus allowing me to analyse the theoretical object of expected cost relatively well.

4.2.2 Anticipation of Absence

A positive correlation test measures the general presence of moral hazard or adverse selection, however, adverse selection or moral hazard may also be construed as dynamic. That is, insurance contributions could be increased as a result of

[7] For those interested, this will be discussed in the forthcoming *Social Insurance to Entrepreneurs* by Benzarti et al. on Finnish entrepreneurs

anticipating sickness or children in the future. As both of these are determined using the reported income of the previous year, Y owners can increase their insurance contributions in the year prior to illness or children in order to receive larger benefits. Thus, as there is incentive to increase contributions in such a way, it makes adverse selection conducive. To investigate this, I first use a method following Kuziemko and Werker (2006) which allows me to investigate whether individuals depart from their usual contribution trends during or before the time of absence. To further elucidate this, I then use a method following Hendren (2017). This method uses the same setting, but differs in identification. Instead of only investigating the insurance contributions of a selected sample, I compare them to those of a control group. Using these methods, I aim to find evidence of whether there is ex ante increases in insurance contributions in the years prior to receiving a benefit.

Kuziemko and Werker (2006) investigate whether a country's U.S. aid and U.N. aid increase as a result of election to and exit from the U.N. security council. They use their model to investigate how aid receipts evolve around the time of the election. Namely, if the aid increased significantly in the year prior to election, this would undermine their hypothesis that being elected into the council are driving the results. Following their method, but altering it to fit the context of this thesis, I regress the following equation:

$$\begin{aligned}\log(IC_{it}) =& \alpha + \beta_1 \cdot t_{-2} + \beta_2 \cdot t_{-1,i} + \beta_3 \cdot t_{0,i} \\ & + \beta_4 \cdot t_{1,i} + \beta_5 \cdot t_{2,i} + \gamma_t + \lambda_i + e_{it}\end{aligned} \tag{2}$$

where IC refers to the insurance contribution of each individual i in year t, t_0 is the first year of receiving the benefit, t_{-x} is x years prior to receiving the benefit and t_x is x years after receiving the benefit. These years may differ between individuals, but are set, so that t_0 is the year of illness for each individual. Thus, I normalise the timeline for everyone such that they follow the same pattern. γ_t accounts for the year fixed effects, while λ_i represents the individual fixed effects. Lastly, e_{it} is the zero mean idiosyncratic error term which may exhibit autocorrelation and heteroskedasticity. It measures disturbances that change across t as well as i.

The baseline insurance contribution, which the other years will be compared to on individual level, is t_{-3}, three years prior to receiving either sick pay or parental allowance. Thus, with respect to t_{-3}, we can see how insurance contributions evolve for individuals over time. I use a two-way fixed effects model, where I control for both individual fixed effects as well as period fixed effects.

This specification allows me to address the concern that unobserved individual specific trends or yearly trends are driving the positive association between insurance contributions and receiving benefits. Despite this, the specification does not allow me to draw causal conclusions, but rather strong correlational results. This is due to the absence of a clear control group to which we can compare outcomes to and the endogenous nature of illness or having a child. Each individual in this regression is "treated" in year, t_0, where the timing of t_0 may differ across individuals and changes in contributions are compared to contributions they have made in t_{-3}. "Treatment" here refers to an individual becoming ill or having children.

To further elucidate whether the results are suggestive of knowledge of future absence affecting insurance contribution levels, I turn to the method motivated by Hendren (2017). Hendren (2017) estimates the anticipatory effects of unemployment by measuring changes in consumption with respect to the times surrounding the unemployment. Following him, I estimate the regression:

$$c_{i,t} = \alpha_k + \Delta_k^{FD} B_{i,t-k} + \beta X_{i,t} + u_{i,t} \tag{3}$$

where $c_{i,t} = \log(IC_{t,i}) - \log(IC_{t_{-3},i})$, the change in insurance contribution for individual i with respect to insurance contributions in t_{-3}, $B_{i,t-k}$ is an indicator for whether an individual has received benefits in year t_0 due to sickness or having a child and $X_{i,t}$ is a set of controls. This difference in received benefits will be measured for a range of leads and lags, k, similarly to the previous model; there will be two lags and two leads surrounding the year of absence, t_0 and they will be compared to a third lag, t_{-3}. Δ_k^{FD} is a coefficient which measures the average difference in insurance contribution change between t_0 and t_{-3} for those who receive benefits in year t_0 and those who have not received benefits. To control for trends within years or per individual, the set of controls include yearly as well as individual fixed effects. Lastly, $u_{i,t}$ is the zero mean idiosyncratic error term that may exhibit autocorrelation and heteroskedasticity.

This specification allows me to investigate how individuals who are ill/have a child change their contributions around the time of receiving benefits with respect to individuals who do not receive benefits. All observations will be included in the same model such that even if the years of absence are different, the model will be scaled so that the year of absence is t_0 for everyone. For the control, I use individuals who are not ill and who do not have children respectively for each outcome. The control group is set to include individuals with similar characteristics as the treatment. As we do not have a clear yearly counterfactual, a placebo absence year will be generated randomly for each individual in the control group. This placebo will act as the control's t_0.

This method is subject to selection. Comparing a group of individuals who are ill or get children will likely differ from a group who is not ill or does not get children. Ideally one could use a group of individuals who are on the margin of becoming ill or having children to obtain better estimates, but this information is not available through the data. Thus, again, using these methods, I only aim to find descriptive results of anticipatory behaviour.

For both methods I assume observations to be independent and identically distributed. "Treatment" (getting ill or having children) is absorbed in the second model by construction as there is an indicator for whether one is treated or not. This implies that once individuals are treated, they stay treated. Even though we assume treatment to be absorbed, it is also important to note that receiving benefits is transient, as benefits are only received for an impermanent amount of time. Furthermore, by setting each model in such a way where individuals are "treated" in the same period, t, I am able to hold exposure to treatment constant.

It is also important to note that there is a possibility of serial correlation in the independent variables in both methods which in turn leads to serial correlation in the error terms. This is often the case in fixed effects models which use panel data. To allow the outcome to be dependent across time due to serial correlation one can use a unit clustered variance-covariance structure. Thus, to account for this the error terms are clustered on individual level.

5 Results

In this section I provide descriptive evidence of adverse selection and moral hazard in the insurance market for entrepreneurs in Finland by studying the correlation between insurance contributions and risk. Risk is measured as days of absence due to sickness or having children. Further, I study the dynamic relationship between insurance contributions and the time of absence due to sickness or having children.

5.1 Positive Correlation

In Sickness (and in Health). — The estimation results for equation 1 for sickness risk are shown in table 3 and similarly for parental risk in table 4.

Dependent Variable: Sickness risk	(1)	(2)	(3)	(4)	(6)
Log Insurance Contribution	0.00153	0.00197	0.00072	0.00072	0.00087
	(0.00010)	(0.00012)	(0.00012)	(0.00012)	(0.00013)
Log Annual Income		-0.00105	-0.00090	-0.00088	-0.00078
		(0.00008)	(0.00008)	(0.00008)	(0.00008)
Age			0.00024	0.00024	0.00024
			(0.00001)	(0.00001)	(0.00001)
Female				0.00026	0.00130
				(0.00016)	(0.00019)
Occupation FE	No	No	No	No	Yes

Table 3: Estimation results for sickness risk

Notes: The table presents the estimation results for equation 1 using sickness risk as the dependent variable. Each column uses the full sample of individuals. The table presents the positive correlation estimates for a number of different controls. The estimation results are obtained using OLS with heteroskedastcity-robust standard errors shown in parenthesis.

In column (1) of table 3, I use OLS to estimate the correlation in equation 1. This is done using raw estimates, without controls. The standard errors shown in the table in parenthesis are heteroskedasticity-robust. I obtain a statistically significant slope coefficient of 0.0015.

Aforementioned, the coefficient, α_1, in equation 1 represents a test for the existence of selection through correlation. A positive coefficient is indicative of adverse selection, while a negative coefficient represents advantageous selection. That is, in the presence of adverse selection, the cost of individuals who contribute through YEL is higher the larger the contributions. The curve is therefore upward-sloping due to adverse selection.[8] The point estimate from the specification suggests that a percentage increase in insurance contributions is associated with a 0.0015 percentage point increase in the probability of illness when not conditioning for controls.

Column (2) of table 3 shows that the addition of annual income as a control does not change the results significantly. Income itself is negatively correlated with sickness risk. This is inline with previous work studying the relationships between income, health as well as age. For instance, Deaton and Paxson (1998) find that individuals in higher income cohorts have better health in general, but differences in health are less well-predicted at older ages as the health of individuals deteriorates regardless of income. These results are also supported by Currie et al. (2007), where they find that income plays a positive role in a child's health, and further by Cutler et al. (2008) where they find that wealth plays a considerable role in health.

Adding age as a control in column (3) attenuates the coefficient significantly. Thus, the results confirm that age is a large factor in ones sickness risk; as individuals age, their sickness risk increases, which in turn could explain some of the rise in insurance contributions seen in figure 6. This result is also supported by previous research. [9]

The remaining columns (4) and (5) show the coefficients when a female dummy and occupation fixed effects are added respectively. The coefficient for insurance

[8] Note that this is different to the average cost curve being downward-sloping due to adverse selection. This is because here we are using a cost-coverage space rather than a price-quantity space.

[9] See for instance Deaton and Paxson (1998) or for more recent work by Van Kippersluis et al. (2009).

contributions remains statistically significant and it's value increases slightly after the addition of a female dummy and again after the addition of occupation fixed effects. Despite there being a significant slope coefficient, the relationship between sickness risk and insurance contributions remains small. Namely, after the addition of controls, the slope coefficient is 0.00087, meaning a percentage increase in insurance contributions is associated with a 0.00087 percentage point increase in sickness risk.

These results indicate that the market does suffer from selection through sickness risk, but not to a large extent. Most other studies have found small correlations which are suggestive of selection, thus this result is not surprising. For instance, Perry and Rosen (2001) find that self employed individuals do not utilise health care services less than wage earners despite having a lower share of individuals covered by health insurance. This finding can be extrapolated here as the results could indicate that health insurance, which in this case is in the form of level of contributions does not play a substantial role in whether self employed persons claim sickness allowance or not. Seibold et al. (2022), on the other, hand find slight advantageous selection, if anything, in private disability insurance. However, this result is not highly comparable to results found in this paper as disability risk may not correlate with sickness risk. Health insurance papers such as Einav et al. (2010) also find significant adverse selection in health insurance. Contrary to Einav et al. (2010) paper, the adverse selection in health risks is much lower in these results.

Furthermore, Böckerman et al. (2018) found that there is significant elasticity in the duration of sickness absence with respect to replacement rate in Finland. Thus, they found there to be a behavioural response to differences in replacement rates due to moral hazard. This is in line with the findings in this paper as those with higher contribution payments, which in turn lead to higher benefits, have a higher sickness risk. However, it cannot be determined whether the the duration of absences are longer in the individuals with higher contributions or if they are ill more often. From our findings, it is also difficult to say whether higher contributions lead to larger share of absences or whether sickness risk leads to higher contributions, while in Böckerman et al. (2018) it is evident that differences in absenteeism develop through hidden action.

Insuring the Baby. — Now, we turn to estimating the slope for the outcome variable of parental risk. Similarly to sickness risk, column (1) of table 4 presents the raw estimates, without controls. The standard errors shown in the table in parenthesis are heteroskedasticity-robust. I obtain a statistically significant, negative slope coefficient of 0.00189. This would be indicative of advantageous selection. The result stays fairly unchanged after adding a control for income in column (2). However, once age is controlled for, the coefficient on insurance contributions switches sign to positive, albeit becomes very small. Again, we see that age is driving a lot of the change in becoming a parent. This is not surprising, as most individuals have children at earlier stages of life. In columns (4) and (5), the slope coefficient further increases and becomes significantly larger. Adding the female dummy in column (4) leads to a twofold increase in the coefficient for log of insurance contributions. This suggests that gender drives a lot of the adverse selection. Column (5) indicates that the addition of occupation fixed effects has little effect on the coefficient for log insurance contribution when compared with column (4). The results of the table are indicative of slight adverse selection or moral hazard, but only once one is able to control for characteristics of the individuals.

Dependent Variable: Risk of baby	(1)	(2)	(3)	(4)	(6)
Log Insurance Contribution	-0.00189	-0.00209	0.00036	0.00074	0.00083
	(0.00009)	(0.00010)	(0.00010)	(0.00011)	(0.00011)
Log Annual Income		0.00030	0.00002	0.00054	0.00065
		(0.00006)	(0.00006)	(0.00006)	(0.00006)
Age			-0.00046	-0.00050	-0.00049
			(0.00001)	(0.00001)	(0.00001)
Female				0.00696	0.00620
				(0.00017)	(0.00019)
Occupation FE	No	No	No	No	Yes

Table 4: Estimates for parental risk

Notes: The table presents the estimation results for equation 1 using risk of having a baby as the dependent variable. Each column uses the full sample of individuals. The table presents the positive correlation estimates for a number of different controls. The estimation results are obtained using OLS with heteroskedastcity-robust standard errors shown in parenthesis.

In particular, a percentage increase in insurance contributions is associated with an increase in the probability of having a child by 0.00083 percentage points with controls, but a decrease in the probability of getting a child by 0.00189 percentage points without.

By themselves, these estimates for both sickness and parenting can only provide evidence of the existence of adverse selection or moral hazard, but whether it is one or the other, or both is unclear. I presume that the likelihood of the positive correlation being a result of adverse selection rather than moral hazard is high. This is due to the fact that these individuals already aim to decrease the amount of costs endured as seen in figure 2. Thus, they are likely to have higher payments only if they expect to use the benefits gained from the social insurance system.

The extent of selection is small, but not infinitesimal. Due to the scant amount of literature describing possible moral hazard or adverse selection in parental allowance systems, the results are not comparable to other findings. Previous studies have mostly focused on the effects of parental leave on employment and wages. Others, such as Han et al. (2009), focus on the effects of changes in the duration of paid, parental leave. They find that expansions to parental leaves are associated with a increase in duration of absence by both mothers and fathers, however the magnitudes of which are heterogeneous across different groups. Some of the results from my findings could be explained by moral hazard with regards to previous studies. When individuals are given a chance to have longer paid leave, they take it. Similarly, if one gets higher benefits, they can possibly afford to stay at home for longer. However, as Y owners already contribute much less, higher contribution levels are likely to be a result of asymmetric information rather than hidden action.

Graphical Presentation. — To further elucidate the correlation between risk and insurance contributions, I now turn to the graphical representations of equation 1. Figure 7 depicts the estimation results in binned scatter plots, such that each point represents the average sickness risk for a mean insurance contribution covering 5% of the sample. Panel (a) shows the unconditional correlation of sickness risk and insurance contributions. This corresponds to estimating equation 1 without controlling for mean annual income, age, sex and occupation. This is concomitant with column (1) of table 3. There is a significant positive relationship between sickness risk and insurance contributions, with a slope coefficient of 0.00153. This is shown in the graph next to the fitted line with the standard error in parenthesis. Next, in panel (b) of figure 7, the same relationship is depicted, but conditional on controls. The slope coefficient becomes flatter with a slope nearly half the magnitude. The correlation here can be associated with the results from column (5) in table 3.

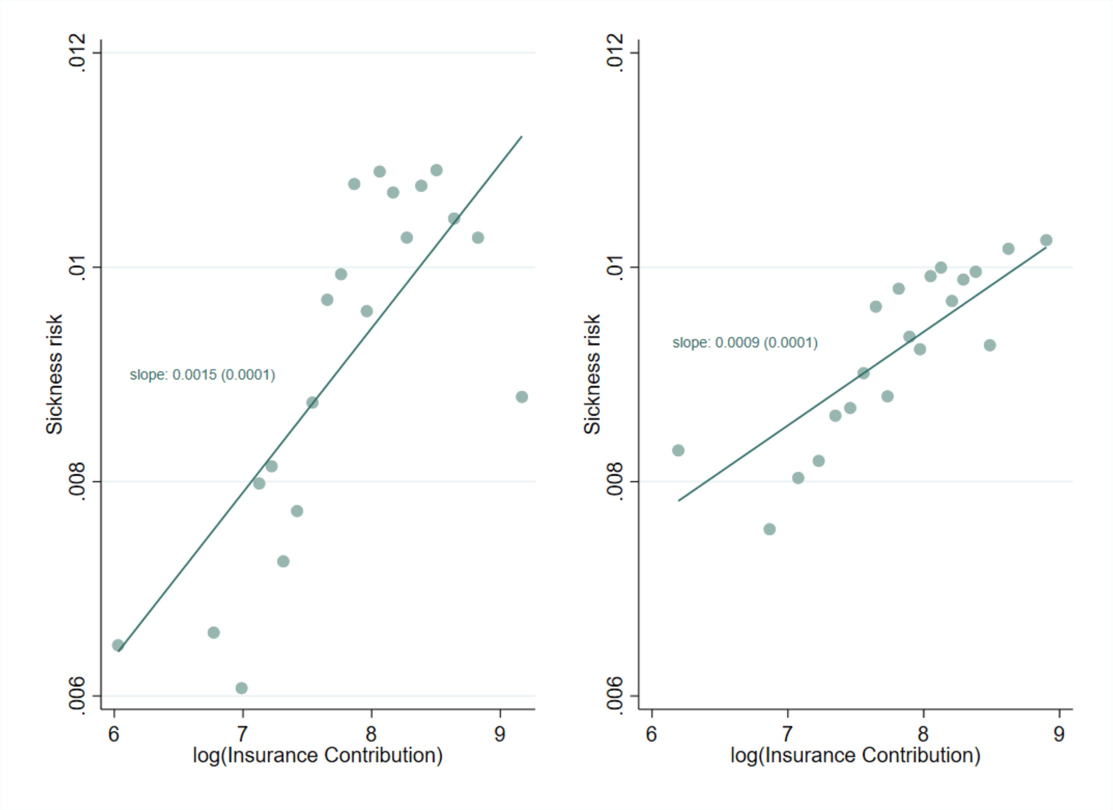

Figure 7: Panel (a) left, Panel (b) right

Notes: The figure shows binned scatterplots depicting the correlation between insurance contributions and sickness risk. Each point contains 5% of the sample. Panel (a) shows the unconditional correlation corresponding to equation 1 with sickness risk as output. Panel (b) shows the correlation, controlling for income, age, sex as well as occupation. Slopes are shown next to lines with robust standard errors in parenthesis.

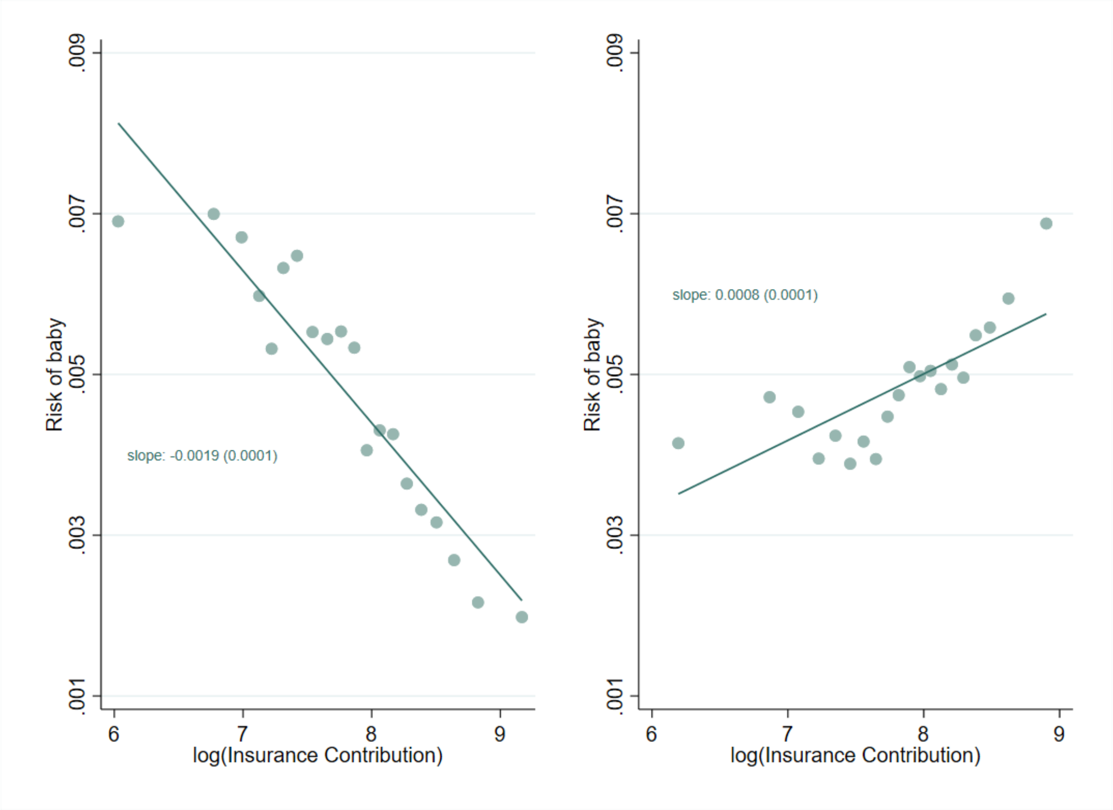

Figure 8: Panel (a) left, Panel (b) right

Notes: The figure shows binned scatterplots depicting the correlation between insurance contributions and parental risk. Each point contains 5% of the sample. Panel (a) shows the unconditional correlation corresponding to equation 1 with parental risk (risk of having a baby) as output. Panel (b) shows the correlation, controlling for income, age, sex as well as occupation. Slopes are shown next to lines with robust standard errors in parenthesis.

Figure 8 shows the graphical positive correlation for parental risk. Panel (a) shows the unconditional relationship of the results for α_1 in binned scatter plots. There is a significant negative relationship, indicating slight advantageous selection with slope of -0.00189. As previously, panel (b) shows the same relationship with controls for mean annual income, age, sex and occupation. Contrary to panel (a), panel (b) shows a significant, positive slope of 0.00083. This is now indicative of adverse selection. the extent of which is almost identical to sickness allowance.

Thus, deviating from expectations, the positive correlation tests seem to validate the claim that there is slight adverse selection in both sickness risk and parental risk; individuals with higher insurance contributions have larger sickness risk than those with lower insurance contributions. Additionally, individuals with higher insurance contributions are more likely to have children and go on longer parental absence when conditioning on characteristics of the individuals.

This result may be surprising as one could expect individuals to have more incentive to contribute if they have knowledge of risks such as sickness or wanting children. In aggregate, many papers have found that those with more insurance are indeed not higher risk. However, there are potential explanations for why the magnitudes of selection in markets has been found to be so small. Firstly, Finkelstein and McGarry (2006) suggest that unobserved preference heterogeneity, especially within risk types, can offset the positive correlation between risk and insurance claims. In other words, if risk preferences are negatively correlated with risk types, this can attenuate the positive correlations arising from adverse selection. Furthermore, Cutler et al. (2008) suggest that preferences and ability may also account for differences in the way people perceive and react to risk, which in turn, is directly linked to insurance decisions.

Secondly, behavioural frictions resulting from misconceptions of risk can preclude the detection of selection in a market. In particular, individuals may be overly optimistic about their level of risk (Spinnewijn, 2017). Lastly, Finkelstein and Poterba (2004) emphasize that absence of selection for one case does not obviate the presence of selection on other cases. Thus, there may be small selection overall within cohorts of different contribution amounts, but this is does not imply that adverse selection in the market is nonexistent, but rather that behavioural responses to choice are heterogeneous among individuals; some individuals may be driving a lot of the adverse selection while others are attenuating the positive correlation.

5.2 Anticipation of Absence

In the following section I present the results for dynamic adverse selection and moral hazard using two methods. First, the results from equation 2, followed by results from regressing equation 3.

5.2.1 Model 1

Insurance Contribution Response to Sickness. — Table 5 shows the estimation results for equation 2 using the logarithm of insurance contributions as the dependent variable and leads and lags surrounding the time of illness, t. The standard errors are robust and clustered on individual level. In the specification in column (1), log insurance contributions is regressed only on the lead and lag year dummies without the use of controls. Standard errors are shown in parenthesis. The coefficients for each lead and lag are statistically significant, with an increasing trend. This trend is disrupted by the year of illness, where a drastic drop in insurance contributions occurs. Column (2) adds individual and year fixed effects; the coefficients change slightly as a result. The coefficients indicate that there are slight increases in insurance contributions in both the first and the second year prior to illness. Namely, there is a significant increase of 0.067 log points in insurance contributions from year $t - 3$ to $t - 1$. The results also reveal that during the year of illness, insurance contributions are associated with a 0.041 log point drop. However, these changes level off in the two years following sickness, shown in rows 4 and 5 of column (2); neither of these are significantly different from zero.

Dependent Variable: log IC	(1)	(2)
t-2	0.01989	0.02801
	(0.00577)	(0.00412)
t-1	0.02289	0.06659
	(0.00550)	(0.00410)
t	-0.08977	-0.04099
	(0.00532)	(0.00418)
t+1	0.02273	0.00289
	(0.00558)	(0.00460)
t+2	0.08845	-0.00628
	(0.00589)	(0.00504)
ID, Year FE	No	Yes

Table 5

Notes: The table presents the estimation results from equation 2 using log of insurance contributions as the dependent variable. Time, t represents the year an individual gets ill. Column (1) depicts the baseline results and column (2) adds individual and year fixed effects. The standard errors are clustered on individual level and are shown in parenthesis. For robustness, more controls are added in the appendix A.2. The appendix also shows results for a sample of individuals who are ill only in year t, thus excludes those who are ill long term.

Aforementioned, during the year of illness, there is a 0.041 log point drop in contributions. This is likely a result of individuals not being able to participate in labour activity due to long term illness. Some individuals stop paying insurance contributions during the year completely, which causes a further strain on the coefficient. Insurance contributions return to baseline in years t+1 and t+2, which could suggest that the rises in years prior to illness result from adverse selection. However, depending on the severity of illness, some individuals could still experience sickness in the following years, indicating that they are still unable to to contribute on the levels that they normally would. As we are only comparing the levels of contributions to three years prior, it is difficult to know whether adverse selection is driving the results, or whether sickness is only driving the trend seen in the latter years. [10]

The pattern of insurance contribution levels over time is indicative of dynamic adverse selection; the most notable increase in insurance contributions comes during the year prior to receiving sick pay, $t-1$. As this is the year from which benefits are based off from, it could indicate that individuals anticipate illness. If we do believe the rises in insurance contributions prior to illness result from illness itself, we assume that individuals are able to predict sickness and evaluate one's own sickness risk. If individuals react to such perceptions through increasing their insurance contributions in order to accrue larger benefits, this increases incentives to go on sickness absence. Thus, the pre-trend rise in insurance contributions could be a result of both adverse selection and moral hazard. Due to the lack of previous research on anticipation of illness, it is difficult to compare the results to other findings, thus possibly threatening the validity of the results. Sickness can be difficult to predict, and often individuals may have misconceptions about their own sickness risks.

I report some heterogeneity for the sample in reactions to receiving sick pay in year t in the appendix A.2. For example, figure 17 presents the estimation results for equation 2 for men and women separately. It becomes evident that both men and women increase insurance contributions prior to illness, but this increase is larger for men in comparison to women. Similarly, figure 18 shows that single individuals raise their insurance contributions slightly more ex ante and ex post sickness in comparison to individuals who are in a relationship. Lastly, no notable differences are seen between individuals with different levels of education. This can be seen from figure 19.

[10] The results are robust to the addition of controls such as relationship status and education. The estimation results for equation 2 while controlling for both of these can be seen in the A.2 in table 11. Education separates individuals into two groups: those with a high school degree and those without a high school degree. Similarly, relationship status is divides individuals into two groups: those who are single and those who are not.

Anticipation of Illness or Correlated Income Shocks. — One could also question whether the changes in contribution levels are a result of changes in income during the period surrounding illness. Replicating the model, but using total annual income as the dependent variable, I obtain the following results shown in table 6.

	(1) log IC	(2) log income
t-2	0.02801	-0.00313
	(0.00412)	(0.00439)
t-1	0.06659	-0.01626
	(0.00410)	(0.00437)
t	-0.04099	0.02436
	(0.00418)	(0.00444)
t+1	0.00289	-0.01435
	(0.00460)	(0.00490)
t+2	-0.00628	-0.00152
	(0.00592)	(0.00537)
ID, Year FE	Yes	Yes

Table 6

Notes: The table presents the estimation results from equation 2 using log of insurance contributions as the dependent variable in column (1) and then log of income as the dependent variable in column (2). Time, *t* represents the year an individual gets ill. Standard errors are clustered on individual level and shown in parenthesis.

Column (1) of table 6 presents the results for the outcome variable of log insurance contributions. These are the same results that are found in column (2) of table 5. Column (2), on the other hand, presents the results for the dependent variable, log income. Comparing the columns, we can detect no similar pattern. With little significance in leads and lags, with the exception of the actual year of illness, we can conclude that the changes in insurance contributions are not a result of changes in income. Interestingly, income increases in the year of illness. Some of the rise can be explained by the benefits that are accrued during that year. While rather ambiguous, it could also be a result of individuals being able to shift income between their firm and themselves more easily.

Figure 9 provides a graphical representation of the results from columns (1) and (2) of table 6.

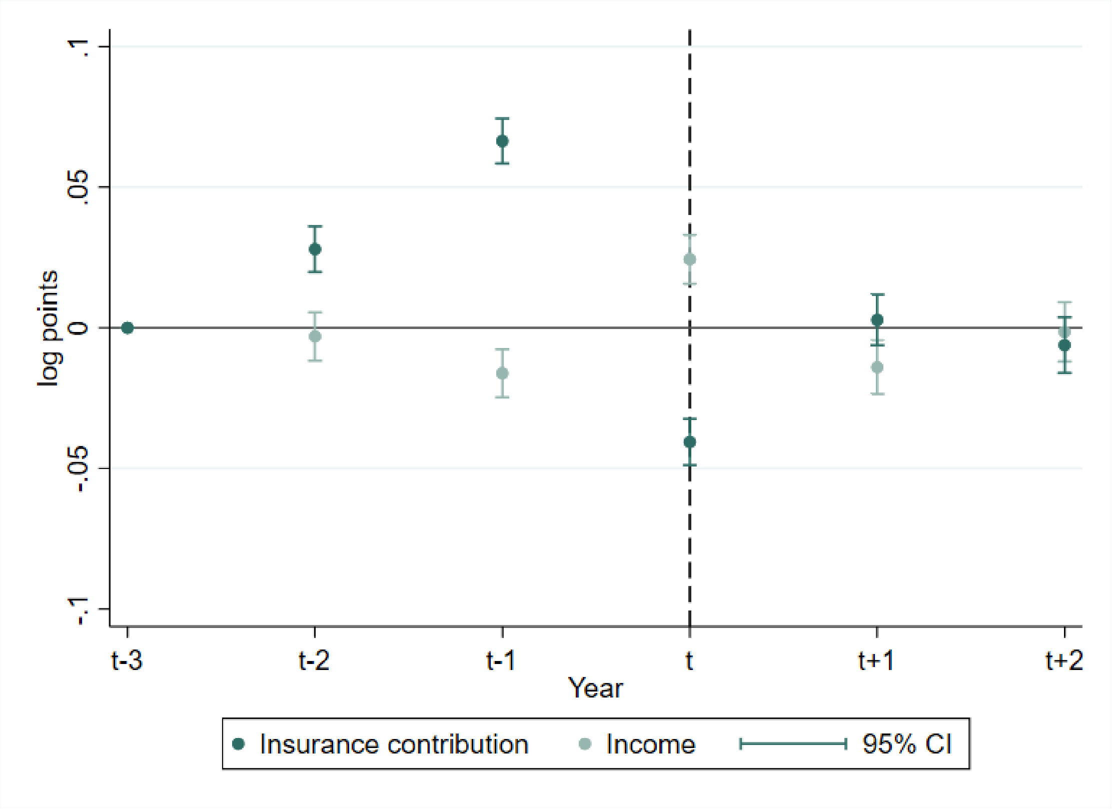

Figure 9

Notes: The figure presents the estimation results from equation 2 using log of insurance contributions as well as log of income as the dependent variables with 95% CI obtained from robust standard errors clustered on individual level. These point estimates correspond to the results presented in table 6. Time, t represents the year an individual gets ill. The sample mean for insurance contributions in year $t-3$ is €3,478 and for income €32,291.

The figure is constructed such that the dashed vertical line demarcates the year of illness, t. The left side of the dashed line shows the results for the lags of t, while the right side presents the coefficients for the leads of t. The horizontal line demarcates the level of contributions in the reference year $t-3$. The most notable increase in contributions is seen on the left side of the dashed vertical line. Namely, there is an increase in year $t-2$ and again in $t-1$. On the contrary, such a trend cannot be seen for income. Instead, income decreases in the year prior to illness and increase in the year of. On the right side of the vertical dashed line, we can see that income as well as insurance contributions hover near the horizontal, zero line, indicating that both fall back to baseline levels. To elucidate the absolute changes in insurance contributions, one can calculate these using the mean insurance contribution for the reference year; this is 3,478 euros. Similarly for income, the sample mean is 32,291 euros in $t-3$.

Insurance Contribution Response to Having a Child. — I now turn to the specification which considers benefits received due to having a child. Table 7 shows the estimation results for equation 2 using the logarithm of insurance contributions as the dependent variable and leads and lags surrounding the time of having a child, t. The standard errors are clustered on individual level and shown in parenthesis.

Dependent Variable: log IC	(1)	(2)
t-2	0.01767	0.03823
	(0.01284)	(0.00996)
t-1	0.04070	0.10233
	(0.01196)	(0.00998)
t	-0.16511	-0.09940
	(0.01134)	(0.01042)
t+1	0.13323	0.08372
	(0.01177)	(0.01173)
t+2	0.23729	0.07981
	(0.01218)	(0.01309)
ID, Year FE	No	Yes

Table 7

Notes: The table presents the estimation results from equation 2 using log of insurance contributions as the dependent variable. Time, t represents the year an individual has a baby. Column (1) depicts the baseline results and column (2) adds individual and year fixed effects. The standard errors are clustered on individual level and shown in parenthesis. For robustness, more controls are added in the appendix A.2.

Column (1) of the table presents the results using only leads and lags and no controls. Similarly to the specification in table 5, there is a steady increase in insurance contributions, with an exception in year t, the year of having a child. This decrease in t is large, but becomes slightly less sizeable after adding individual and yearly fixed effects. Column (2) shows the results for the same specification, but with the addition of controls for individuals and years using fixed effects. The leads become larger and significant, while the lags become less sizeable, whilst remaining significant. Aforementioned, there is still a significantly large decrease in insurance contributions during the year of having a child. Namely there is a decrease of 0.0994 log points from year $t-3$ to t. This is likely due to individuals being absent and thus paying very small, if any, contributions during the first year of having a child. Comparing the results to those obtained for sick pay, the coefficients for both ex-ante and ex-post periods are much larger. Already in year $t-2$ we see a 0.0382 log point difference followed by 0.1023 in year $t-1$. Substantively, we also notice that insurance contribution levels do not revert back to those of year $t-3$, but stay significantly larger. The most probable reason is due to having a child, which can result in an increase in risk aversion and, in turn, an increase in willingness to pay for insurance (Görlitz and Tamm, 2020; Kettlewell, 2019).[11]

Interestingly, Görlitz and Tamm (2020) find that risk aversion increases in both men and women already two years prior to having a first child. This could explain some of the rise we see in the results for both $t-1$ and $t-2$. The results seem to validate the idea that asymmetric information is present both overall as well as dynamically. Even though the increases in insurance contributions could be a result of changes in risk preferences rather than from wanting to accrue higher benefit levels, the source of increase is asymmetric information. However, due to the lack of previous research on adverse selection in parental allowance, comparison of results is difficult.

Heterogeneity for the sample in reactions to receiving parental allowance in year t can be found in the appendix A.2. In particular, figure 20 presents the estimation results for equation 2 for men and women separately. Both men and women increase insurance contributions prior to having a baby, but this increase is larger for men. Surprisingly, men are driving the drop in insurance contributions during the year of having a child, while women are driving the increases in insurance contributions ex post. Figure 21 shows that the changes in insurance contributions are homogeneous for individuals who are single or in a relationship. Lastly, individuals with a high school degree seem to be driving more of the changes in insurance contributions. This can be seen from figure 22.

[11] The results are robust to the addition of controls such as relationship status and education. The estimation results for equation 2 while controlling for both of these can be seen in the Appendix in table 13.

Anticipation of a Baby or Correlated Income Shocks. — While being less likely in the case of having a baby, one could claim that increases in insurance contributions are a result of income only, thus I provide a similar table for parental allowance as for sick pay. The results from equation 2 with log income as the outcome are shown in table 8. Similarly, the standard errors are clustered on individual level and are shown in parenthesis in the table. I find no statistically significant changes in income during the time surrounding the birth of a child. There is, however, a slight increase in income during the year of having a child. This could result from both accrued benefits and the act of shifting money between the firm and the individual.

	(1) log IC	(2) log income
t-2	0.03823	0.01299
	(0.00996)	(0.00890)
t-1	0.10233	-0.00647
	(0.00998)	(0.00891)
t	-0.09940	0.03576
	(0.01042)	(0.00928)
t+1	0.08372	-0.00072
	(0.01173)	(0.01044)
t+2	0.07981	0.00136
	(0.01309)	(0.01166)
ID, Year FE	Yes	Yes

Table 8

Notes: The table presents the estimation results from equation 2 using log of insurance contributions as the dependent variable in column (1) and then log of income as the dependent variable in column (2). Time, t represents the year an individual has a baby. The standard errors are clustered on individual level and are shown in parenthesis.

The results from table 8 are illustrated graphically in figure 10.

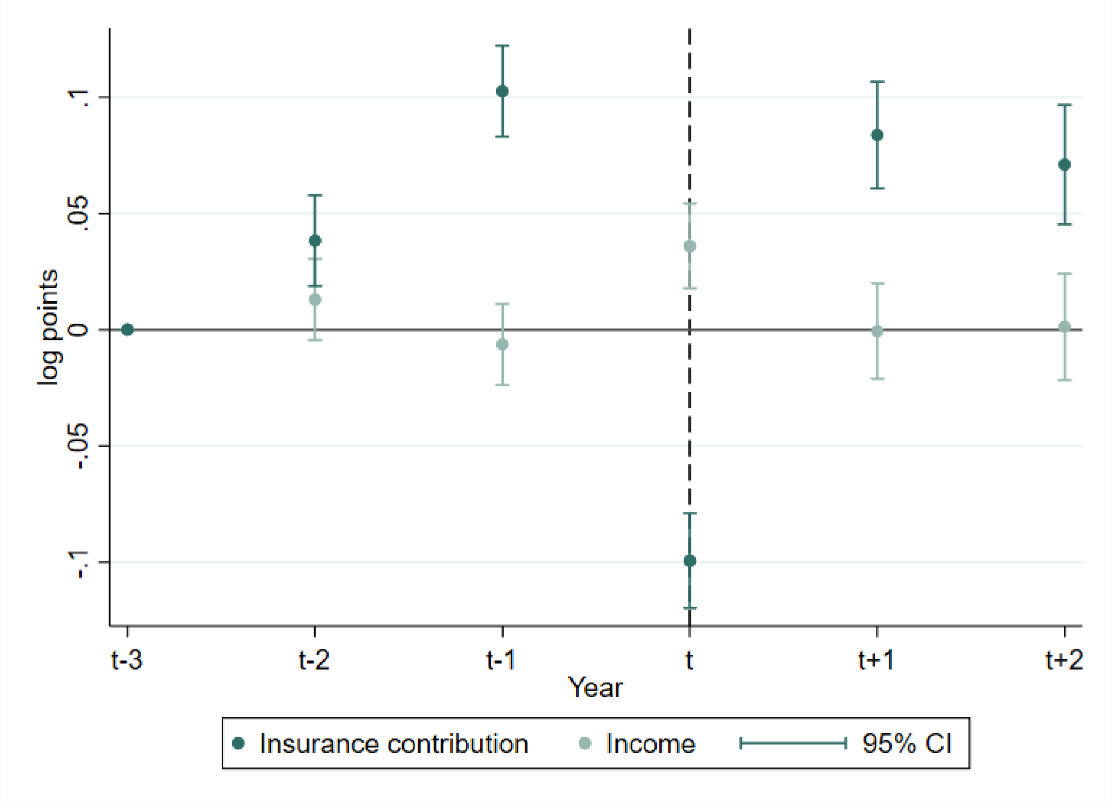

Figure 10

Notes: The figure presents the estimation results from equation 2 using log of insurance contributions as well as log of income as the dependent variables with 95% CI obtained from robust standard errors clustered on individual level. These point estimates correspond to the results presented in table 8. Time, t represents the year an individual has a baby. The sample mean for insurance contributions in year $t-3$ is €2,572 and for income it is €29,601.

Figure 10 is constructed identically to figure 9. The left side of the vertical dashed line, demarcating the year of having a baby, shows large increases in insurance contributions for the lags. Similar increases are not visible for income. While there are increases in insurance contributions for the leads as well, the trend does not seem to be increasing. Again, income hovers around the horizontal, zero line indicating that income does not change from baseline level. The figure indicates strong adverse selection. To elucidate the absolute changes in insurance contributions, one can calculate these using the mean insurance contribution for the reference year; this is 2,572 euros. Similarly for income, the sample mean is 29,601 euros in $t-3$.

Due to the problem of endogeneity of illness or choosing to have children, we cannot draw causal conclusions from these results. However, the nature of adverse selection itself comes from asymmetries in information about individuals, thus being an inherently endogenous question. We can thus shed further light on dynamic adverse selection from a method in which one can compare the outcomes of a possibly adversely selected group to a group which is not (at least in the same way).

5.2.2 Model 2

Response to Sickness. — We now turn to estimating equation 3. Using the sample of individuals who are ill in year t_0 and those to whom t_0 is randomised, table 9 presents the estimates for equation 3 with robust standard errors clustered on individual level shown in parenthesis. The coefficients are difference in differences estimates. In other words, they represent how the treated vary their insurance contributions in comparison to the control.

Dep var: c_{it}	(1)	(2)	(3)
t-2	-0.01048	-0.01255	-0.02144
	(0.00917)	(0.00917)	(0.00690)
t-1	-0.01168	-0.00689	-0.01860
	(0.00915)	(0.00921)	(0.00694)
t	-0.16608	-0.13961	-0.14408
	(0.00909)	(0.00922)	(0.00695)
t+1	-0.12606	-0.07726	-0.10756
	(0.00935)	(0.00944)	(0.00718)
t+2	-0.13695	-0.07043	-0.09676
	(0.00956)	(0.00964)	(0.00735)
Year FE	No	Yes	Yes
Age FE	No	Yes	No
ID FE	No	No	Yes

Table 9

Notes: The table presents the estimation results for Δ_k^{FD} from equation 3. Time, t represents the year an individual gets sick. Column (1) depicts the baseline results, column (2) adds age and year fixed effects, column (3) replaces age fixed effects with individual fixed effects. The standard errors are clustered on individual level and shown in parenthesis. For robustness, more controls are added in the appendix A.3. In the appendix results are also shown for a sample of individuals who are ill only in year t.

Column (1) of table 9 illustrates the baseline difference in difference coefficients with standard errors clustered on individual level shown in parenthesis. There seems to be no evidence of ex ante increases in insurance contributions for the ill individuals when compared to individuals who are not ill. Rather, coefficients for two years prior and one year prior show statistically insignificant coefficients. In the third row, the baseline specification yields a 16.6 percent drop in insurance contributions during the year of illness. This decrease in insurance contributions is upheld throughout the two years following illness. Column (2) shows that controlling for yearly effects as well as age fixed effects delivers similar, but smaller coefficients. Column (3) controls for individual fixed effects rather than age fixed effects. This induces coefficients which are similar in magnitude and statistically indistinguishable from baseline estimates. This suggests that there are no ex ante increases in insurance contributions for individuals prior to illness. [12]

These results are presented in figure 11

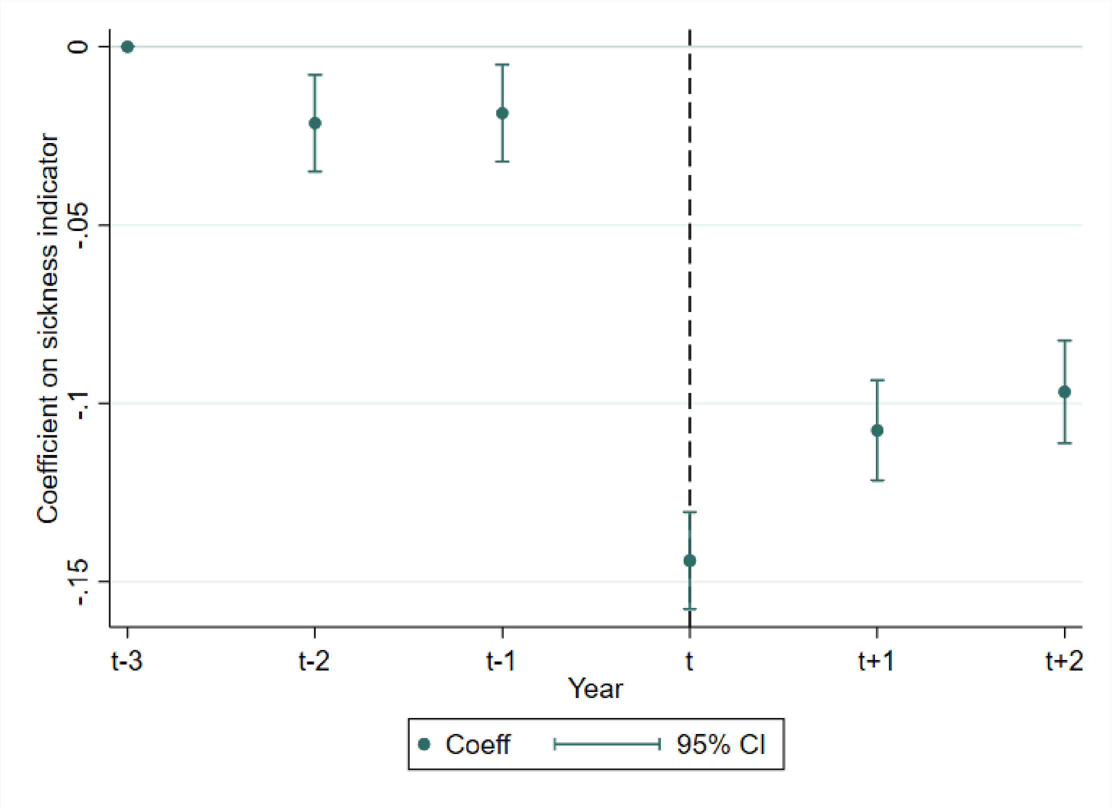

Figure 11

Notes: The figure presents the estimation results for coefficients of Δ_k^{FD} with 95% CI, from equation 3. The point estimates correspond to those of column (3) of table 9. The CI are derived

[12] These results are robust to the addition of controls such as relationship status and education. This is seen in table 14.

from standard errors that are clustered on individual level. These point estimates correspond to the results presented in table 9. Time, t represents the year an individual gets sick.

Divergent Models. — Prima facie, these results seem to contradict those of Model 1. Model 1 suggests that ill individuals increase insurance contributions just before getting sick, while this method suggests that compared to a control, there are no increases in insurance contributions, but rather small decreases. The behaviour of individuals prior to sickness is therefore ambiguous. Each individual, the ones in the treatment as well as those in the control, increase their insurance contributions every year from $t - 3$ to $t - 1$. The extent of these increases seem to be larger for those in the control group. However, from this model, we cannot assume the behaviour of ill individuals would be similar to behaviour of those who do not fall ill during any period. In order to be able to draw such conclusions, I would have to know which individuals are on the verge of getting ill and use these individuals as a control. Obtaining this control group from the data available is, however, impossible.

What both of these methods do suggest, is that the behaviour of the treatment group changes between the two time periods. Prior to illness, insurance contributions are rising, but after illness, they stay relatively unchanged. This could suggest that sickness itself has relatively large effects on individuals insurance contributions in general, but there is no clear evidence that individuals actively change their behaviour as they anticipate sickness. There is, however, clear differences between individuals who are ill and those who are not, suggesting that there is some form of adverse selection. Whether these individuals are aware of their sickness risks themselves, is difficult to say.

There is a lack of previous research on anticipatory behavioural responses to illness. However, some research has focused on the effect of health shocks on risk preferences. The results suggest that post sickness, insurance contributions do not rise, which could indicate that individuals who have been ill, do not change their preferences regarding the level of insurance. This is supported by Kettlewell (2019) and Chuang and Schechter (2015). Decker and Schmitz (2016) find the opposite, suggesting that health shocks increase an individual's risk aversion, with a lasting effect of up to four years. In the results presented here, the lack of increase in insurance contributions after a health shock could be suggestive of health shocks not resulting in changes in risk preferences. As there is lack of risk aversion post sickness, it could also explain the lack of increases in insurance contributions prior to illness.

Other research has focused on adverse selection in disability insurance. For instance, aforementioned, Seibold et al. (2022) found there to be no selection in the private disability insurance market. While the focus in the paper was not on whether individuals anticipate disability, the lack of adverse selection would suggest that the majority of individuals who do obtain disability insurance do not have higher claims even though they may expect disability. Chandra and Samwick (2009) discuss the effects of disability risk on pre-retirement savings and find that disability risk does not increase savings relative to the expected losses in income due to disability. This suggests that disability risk has little effects on savings in general. This result can be extrapolated into our results as it is suggestive of disability or sever illness not affecting savings, which in turn, are reflected by insurance contributions.

Response to Having a Child. — We now turn to the estimation of equation 3 using leads and lags surrounding the time of having a baby. Table 10 illustrates the estimates for Δ_k^{FD} for different leads and lags, k with clustered standard errors shown in parenthesis. These represent the difference of changes in contributions with respect to year, $t-3$, between treatment (have a baby) and control (have a placebo baby).

Dep var: c_{it}	(1)	(2)	(3)
t-2	0.06173	0.05119	0.04826
	(0.01305)	(0.01312)	(0.00973)
t-1	0.07970	0.06814	0.07102
	(0.01301)	(0.01322)	(0.00981)
t	-0.19874	-0.19131	-0.17502
	(0.01279)	(0.01323)	(0.00981)
t+1	-0.05944	-0.03049	-0.04172
	(0.01358)	(0.01404)	(0.01054)
t+2	-0.05507	-0.01015	-0.01429
	(0.01398)	(0.01452)	(0.01088)
Year FE	No	Yes	Yes
Age FE	No	Yes	No
ID FE	No	No	Yes

Table 10

Notes: The table presents the estimation results for Δ_k^{FD} from equation 3. Time, t represents the year an individual has a baby. Column (1) depicts the baseline results, column (2) adds age and year fixed effects, column (3) replaces age fixed effects with individual fixed effects. The standrad errors are clustered on individual level and shown in parenthesis. For robustness, more controls are added in the appendix A.3.

Similarly to table 9, column (1) of table 10 presents the baseline estimates. The table shows standard errors in parenthesis which are clustered on individual level. There is a 6.17 percent increase in year $t-2$ followed by a 7.97 percent increase in year $t-1$. Both estimates are statistically significant, implying that individuals who have children, actively increase their insurance contributions much more than individuals who do not have a baby in year t. In the year of the baby, insurance contributions are lowered by 19.87 percent. The drop is considerably larger here than the drop during sickness. This is most likely due to individuals taking longer absences from work as a result of having a child. The two years following the baby show slightly lower changes in insurance contributions. Column (2) adds year and age fixed effects. These make the estimates slightly smaller, but similar in magnitude. Larger changes in magnitudes are seen in the two years post baby; the coefficient for $t+2$ becomes insignificant, suggesting that the treatment group change their insurance contributions indifferently from the control group. Column (3) replaces age fixed effects with individual fixed effects. The results do not change significantly, but the standard errors become slightly smaller. The results suggest that there are ex ante increases in insurance contributions before having a baby. Namely, there is a 4.8 percent increase in year $t-2$ and a 7.1 percent increase in year $t-1$. There is a large decrease in the insurance contributions during the year of the baby, which is followed by similar changes in contributions for the post baby years for the two groups. [13]

The results are also presented graphically in figure 12.

[13] These results are robust to the addition of controls such as relationship status and education. This is seen in table 16.

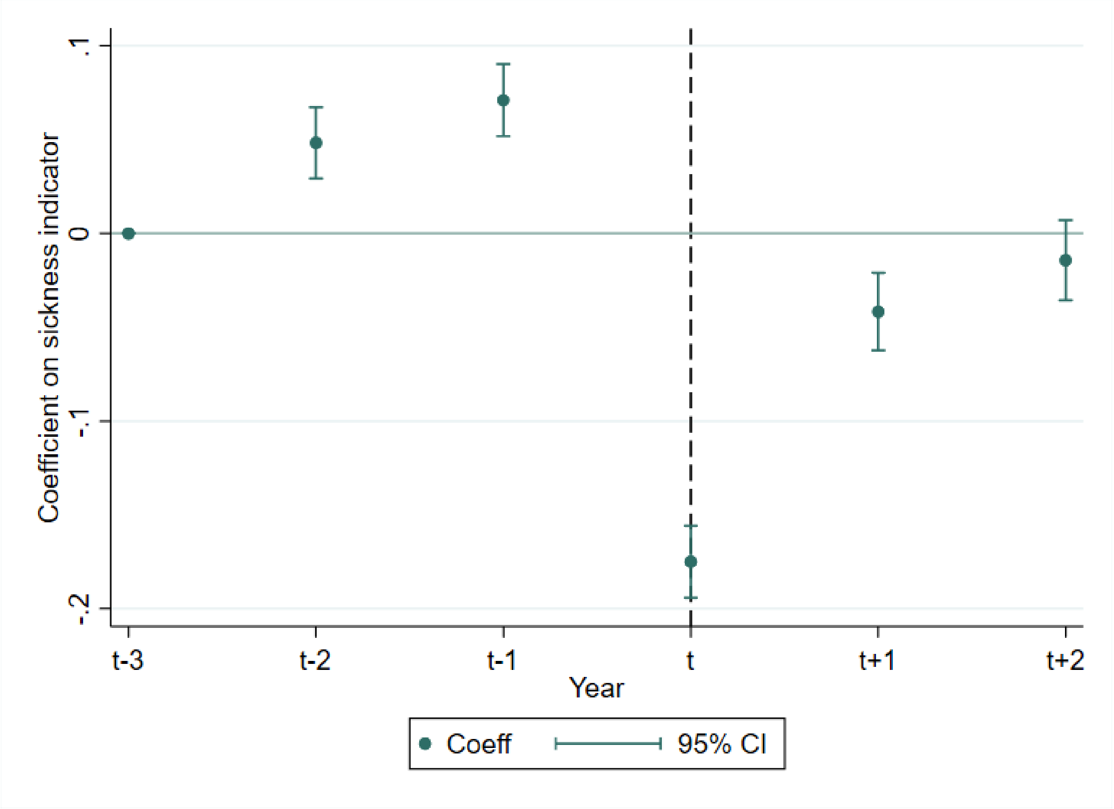

Figure 12

Notes: The figure presents the estimation results for coefficients of Δ_k^{FD} with 95% CI, from equation 3. The CI are derived from standard errors that are clustered on individual level. These point estimates correspond to the results presented in table 10 column (3). Time, t represents the year an individual has a baby.

Divergent Models.– Unlike in the case of sickness, the results from model 2 are cognate with those of model 1 as they seem to follow the same pattern. Thus, there seems to be a consensus of ex ante increases in insurance contributions due to the anticipation of a child using both methods. This is highly suggestive of adverse selection being present in the market dynamically. However, while using the second method, we are still faced with a caveat of having a control group which is not perfectly comparable to the treatment group. Individuals who wish to have children may, for instance, have different risk preferences. Albeit, this does not obviate the existence of selection in the market. Individuals with similar characteristics behave differently during the time surrounding the birth of a baby due to asymmetric information.

Perhaps the most striking difference between the results of the two methods are the magnitudes of the coefficients, especially in the year of a child and the years after. It is evident from using both methods that insurance contributions are increased yearly regardless of whether individuals are in the treatment group or the control group. During the year of having a child, most individuals do not work, thus pay much lower contributions. Thus, as others keep increasing contribution levels, the treatment group decreases them, causing a large negative coefficient in year t. While the treatment group increases payments again in years $t + 1$ and $t + 2$, the magnitudes of these increases are lower for the treatment compared to the control. The models are therefore essentially saying the same thing, but this is expressed either as a single difference or a difference in differences.

6 Conclusion

In this paper, I documented the existence of private information in the public pension insurance market for entrepreneurs. The adverse selection arising from the asymmetric information was demonstrated through a general form of correlation between insurance contributions and risk of absence due to sickness or having children and through dynamic selection, where I tested whether individuals are able to anticipate either getting sick or having children. The empirical work focused on risks measured as length of absences due to sickness or children as well as asymmetric information arising from knowledge of future sickness or future children. This paper argues that adverse selection and/or moral hazard exists in the market, the extent of which is rather small, but apparent.

I began by using data on the insurance contributions and yearly accrued sick pay and parental allowance for 70% of entrepreneurs in Finland throughout 2001 until 2015. I then showed that private information about sickness did not result in ex ante increases in insurance contributions. The findings were shown using two models, one of which showed the changes in contributions levels for a sample of individuals who were sick. The second showed the changes in contribution levels for the same group, but in comparison to individuals who had not fallen ill. The results were slightly ambiguous and thus do not provide enough robust evidence of individuals having knowledge of future sickness and reacting to this through an increase in insurance contributions. However, I found that while sickness may be difficult to anticipate, generally individuals with higher insurance contribution levels have a larger number of days of absence due to sickness. Due to the nature of such an event, the results cannot establish whether this positive correlation follows from moral hazard or adverse selection.

This paper also provided evidence that adverse selection is apparent in parental allowances. I found that individuals who have a child, anticipate such an event and thus increase insurance contributions already two years prior to having a child. I reconciled these findings by presenting two methods, which showed that the results are robust to two different identification strategies. The positive correlation found between insurance contributions and length of absence was argued to be a result of adverse selection rather than moral hazard. This is because an increase in insurance contributions is unlikely going to affect the decision to have children. While there exists little prior evidence to support the claims, previous findings have suggested that having children increases risk aversion, which could explain some of the increases in insurance contributions seen in the results.

The findings highlight that the market suffers from adverse selection to some extent. However, I was not able to estimate the causal effects of the relationship or the impacts of risk preferences or risk types. In particular, differences in risk preferences and risk types can change the analysis of the findings and thus make it difficult to asses policy implications regarding the market. It is worth noting that although selection arising from risk preferences and selection resulting from risk types can counterpoise each other in the insurance market for health, they may reinforce each other in other insurance markets.

Finally, my results suggest a strong correlation which implies that adverse selection is highly present dynamically in parental allowances, but also to an extent generally in both sickness and parenting. An interesting direction for future work would be to look at the causal relationships between insurance contributions and insurance claims and draw conclusions about whether the positive correlation is a result of adverse selection or moral hazard. One could also then provide useful policy implications in order to counterbalance the adversities arising from asymmetric information.

References

Acs, Z. J. and Audretsch, D. B. (2003). Introduction to the handbook of entrepreneurship research. *Handbook of entrepreneurship research: An interdisciplinary survey and introduction*, pages 3–20.

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics*, pages 235–251. Elsevier.

Bailey, J. (2017). Health insurance and the supply of entrepreneurs: New evidence from the affordable care act. *Small Business Economics*, 49(3):627–646.

Benzarti, Y., Harju, J., and Matikka, T. (2020). Does mandating social insurance affect entrepreneurial activity? *American Economic Review: Insights*, 2(2):255–68.

Bernheim, B. D. (2002). Taxation and saving. In *Handbook of public economics*, volume 3, pages 1173–1249. Elsevier.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.

Böckerman, P., Kanninen, O., and Suoniemi, I. (2018). A kink that makes you sick: the effect of sick pay on absence. *Journal of Applied Econometrics*, 33(4):568–579.

Boeri, T., Giupponi, G., Krueger, A. B., and Machin, S. (2020). Solo self-employment and alternative work arrangements: A cross-country perspective on the changing composition of jobs. *Journal of Economic Perspectives*, 34(1):170–95.

Cagan, P. (1965). Possible effects of pension plans on aggregate personal saving. In *The Effect of Pension Plans on Aggregate Saving: Evidence from a Sample Survey*, pages 1–7. NBER.

Chandra, A. and Samwick, A. A. (2009). Disability risk and the value of disability insurance. In *Health at Older Ages: The Causes and Consequences of Declining Disability among the Elderly*, pages 295–336. University of Chicago Press.

Chen, R., Wong, K. A., and Lee, H. C. (2001). Age, period, and cohort effects on life insurance purchases in the us. *Journal of Risk and Insurance*, pages 303–327.

Chetty, R. and Finkelstein, A. (2013). Social insurance: Connecting theory to data. In *Handbook of public economics*, volume 5, pages 111–193. Elsevier.

Chetty, R., Friedman, J. N., Leth-Petersen, S., Nielsen, T. H., and Olsen, T. (2014). Active vs. passive decisions and crowd-out in retirement savings accounts: Evidence from denmark. *The Quarterly Journal of Economics*, 129(3):1141–1219.

Chuang, Y. and Schechter, L. (2015). Stability of experimental and survey measures of risk, time, and social preferences: A review and some new results. *Journal of development economics*, 117:151–170.

Currie, A., Shields, M. A., and Price, S. W. (2007). The child health/family income gradient: Evidence from england. *Journal of health economics*, 26(2):213–232.

Cutler, D. M., Lleras-Muney, A., and Vogl, T. (2008). Socioeconomic status and health: dimensions and mechanisms.

Deaton, A. S. and Paxson, C. H. (1998). Aging and inequality in income and health. *The American Economic Review*, 88(2):248–253.

Decker, S. and Schmitz, H. (2016). Health shocks and risk aversion. *Journal of health economics*, 50:156–170.

Diamond, P. A. and Hausman, J. A. (1984). Individual retirement and savings behavior. *Journal of Public Economics*, 23(1-2):81–114.

Einav, L. and Finkelstein, A. (2011). Selection in insurance markets: Theory and empirics in pictures. *Journal of Economic Perspectives*, 25(1):115–38.

Einav, L., Finkelstein, A., and Cullen, M. R. (2010). Estimating welfare in insurance markets using variation in prices. *The quarterly journal of economics*, 125(3):877–921.

Ericson, K. M., Kircher, P., Spinnewijn, J., and Starc, A. (2021). Inferring risk perceptions and preferences using choice from insurance menus: theory and evidence. *The Economic Journal*, 131(634):713–744.

Finkelstein, A. and McGarry, K. (2006). Multiple dimensions of private information: evidence from the long-term care insurance market. *American Economic Review*, 96(4):938–958.

Finkelstein, A. and Poterba, J. (2004). Adverse selection in insurance markets: Policyholder evidence from the uk annuity market. *Journal of political economy*, 112(1):183–208.

Gordon, R. H. and Blinder, A. S. (1980). Market wages, reservation wages, and retirement decisions. *Journal of public Economics*, 14(2):277–308.

Görlitz, K. and Tamm, M. (2020). Parenthood, risk attitudes and risky behavior. *Journal of Economic Psychology*, 79:102189.

Hackmann, M. B., Kolstad, J. T., and Kowalski, A. E. (2015). Adverse selection and an individual mandate: When theory meets practice. *American Economic Review*, 105(3):1030–66.

Han, W.-J., Ruhm, C., and Waldfogel, J. (2009). Parental leave policies and parents' employment and leave-taking. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 28(1):29–54.

Handel, B. R., Kolstad, J. T., Minten, T., and Spinnewijn, J. (2020). The social determinants of choice quality: evidence from health insurance in the netherlands. Technical report, National Bureau of Economic Research.

Hawley, J., Manoudi, A., Rasell, J., and Scott, D. (2010). European employment observatory review: Self-employment in europe 2010.

Heim, B. T. and Lurie, I. Z. (2010). The effect of self-employed health insurance subsidies on self-employment. *Journal of Public Economics*, 94(11-12):995–1007.

Hendren, N. (2017). Knowledge of future job loss and implications for unemployment insurance. *American Economic Review*, 107(7):1778–1823.

Hendren, N., Landais, C., and Spinnewijn, J. (2021). Choice in insurance markets: A pigouvian approach to social insurance design. *Annual Review of Economics*, 13:457–486.

Hyrkkänen, R. (2009). Onko yrittäjien eläkevakuuttaminen kohdallaan?: Yel-työtulon tasotarkastelua eri näkökulmista.

Jackson, S. (2010). Mulling over massachusetts: Health insurance mandates and entrepreneurs. *Entrepreneurship Theory and Practice*, 34(5):909–932.

Katona, G. (1965). *Private pensions and individual saving.* Number 40. Survey Research Center, Institute for Social Research, University of Michigan.

Kettlewell, N. (2019). Risk preference dynamics around life events. *Journal of Economic Behavior & Organization*, 162:66–84.

Kolsrud, J., Landais, C., Nilsson, P., and Spinnewijn, J. (2018). The optimal timing of unemployment benefits: Theory and evidence from sweden. *American Economic Review*, 108(4-5):985–1033.

Kuziemko, I. and Werker, E. (2006). How much is a seat on the security council worth? foreign aid and bribery at the united nations. *Journal of political economy*, 114(5):905–930.

Landais, C., Nekoei, A., Nilsson, P., Seim, D., and Spinnewijn, J. (2021). Risk-based selection in unemployment insurance: Evidence and implications. *American Economic Review*, 111(4):1315–55.

Lohse, T. and Qari, S. (2021). Gender differences in face-to-face deceptive behavior. *Journal of Economic Behavior & Organization*, 187:1–15.

Munnell, A. H. (1976). Private pensions and savings: new evidence. *Journal of political economy*, 84(5):1013–1032.

Niskanen, J. and Keloharju, M. (2000). Earnings cosmetics in a tax-driven accounting environment: evidence from finnish public firms. *European Accounting Review*, 9(3):443–452.

OECD (2018). Entrepreneurship at a glance. 2018 highlights.

Perry, C. W. and Rosen, H. S. (2001). Insurance and the utilization of medical services among the self-employed.

Pohjola Vakuutus (2022). Yrittäjän lakisääteinen eläkevakuutus eli yel. `https://www.op.fi/yritykset/vakuutukset/henkilovakuutukset/yel-vakuutus#`. Accessed: 25/07/2022.

Røed, K. and Skogstrøm, J. F. (2014). Unemployment insurance and entrepreneurship. *Labour*, 28(4):430–448.

Rothschild, M. and Stiglitz, J. (1978). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. In *Uncertainty in economics*, pages 257–280. Elsevier.

Schweitzer, M. and Severance-Lossin, E. (1996). Rounding in earnings data.

Seibold, A., Seitz, S., and Siegloch, S. (2022). Privatizing disability insurance. *ZEW-Centre for European Economic Research Discussion Paper*, (22-010).

Spinnewijn, J. (2017). Heterogeneity, demand for insurance, and adverse selection. *American Economic Journal: Economic Policy*, 9(1):308–43.

Statistics Finland (2020). Yritykset. `https://www.tilastokeskus.fi/tup/suoluk/suoluk_yritykset.html`. Accessed: 20/08/2022.

Van Kippersluis, H., Van Ourti, T., O'Donnell, O., and Van Doorslaer, E. (2009). Health and income across the life cycle and generations in europe. *Journal of health economics*, 28(4):818–830.

Wagener, A. (2000). Entrepreneurship and social security. *FinanzArchiv/Public Finance Analysis*, pages 284–315.

WHO (2015). *World report on ageing and health.* World Health Organization.

Xu, W. (2022). Social insurance and entrepreneurship: The effect of unemployment benefits on new-business formation. *Strategic Entrepreneurship Journal*, 16(3):522–551.

A Appendix

A.1 Data Characteristics

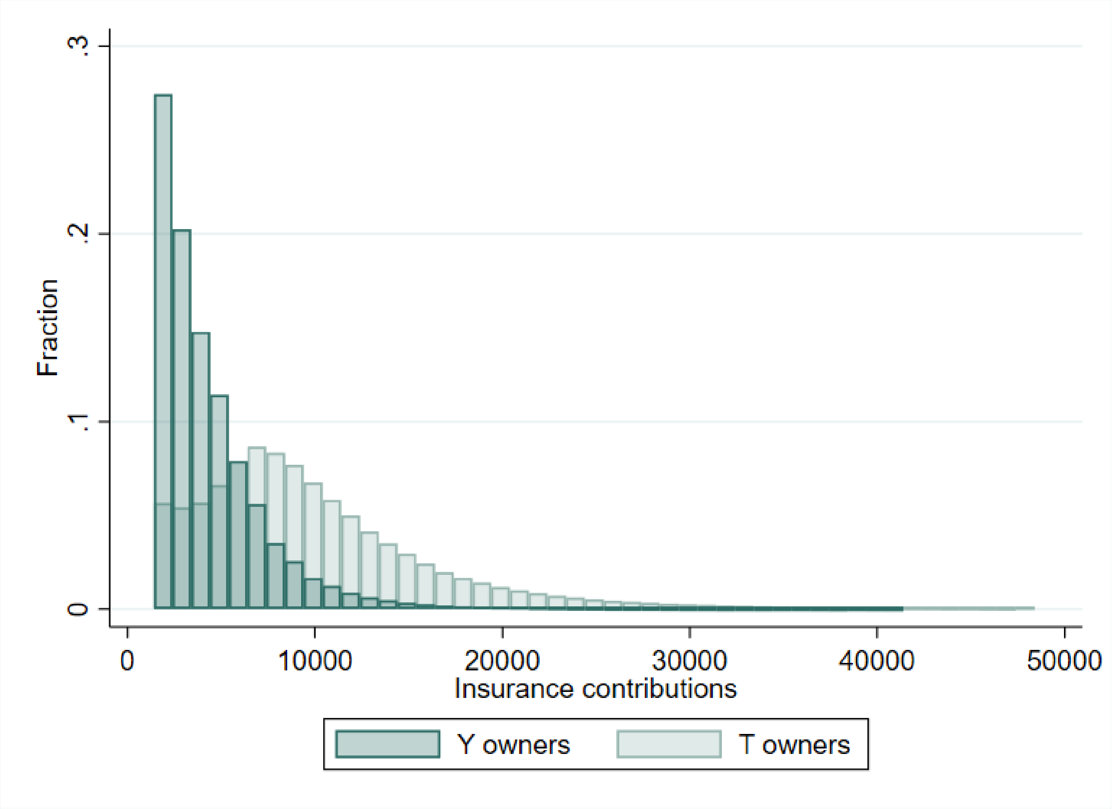

Figure 13

Notes: The figure presents the distributions of insurance contributions for Y and T owners.

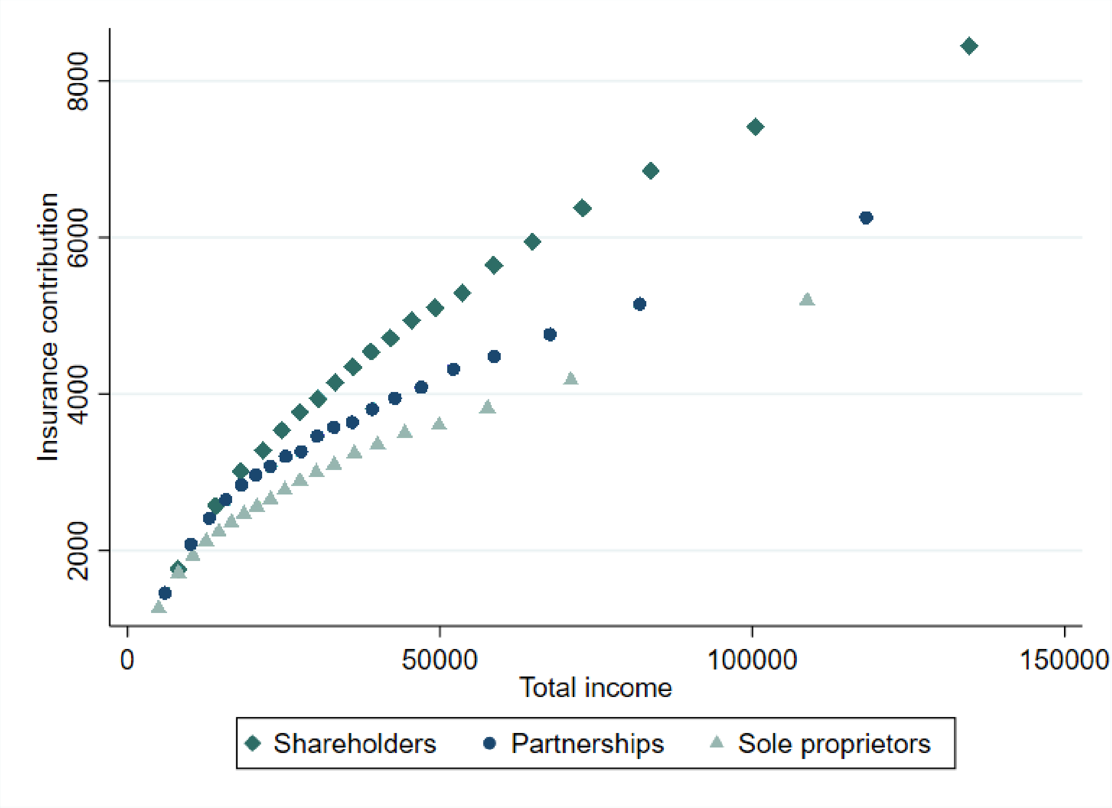

Figure 14

Notes: The figure presents the relationship between public pension insurance contributions and income for the three different organisational forms: shareholders, partnerships and sole proprietors. Each point consists of 5% of observations with respect to their group. Perry and Rosen (2001) investigated the differences between the likelihood of having insurance coverage with respect to organisational form. They found that corporations (shareholders) were most likely to have insurance coverage while partnerships and shareholders were just as likely, but less. This is in line with our findings.

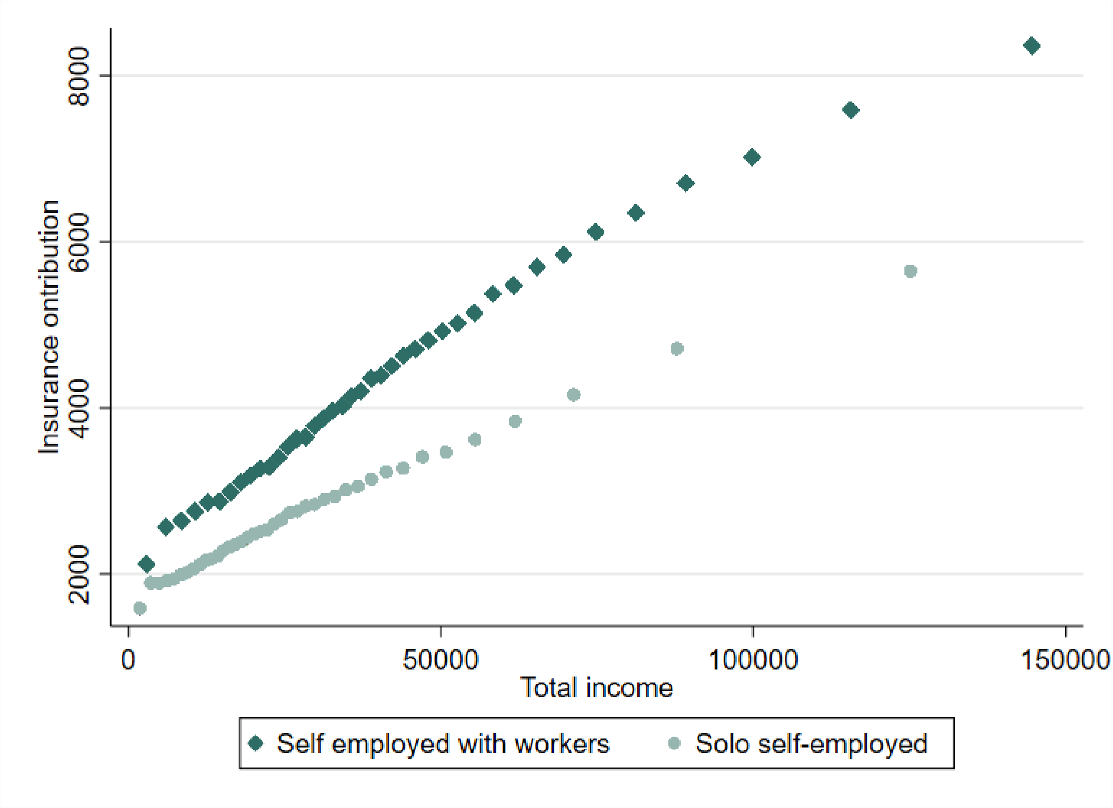

Figure 15

Notes: The figure presents the relationship between public pension insurance contributions and income for individuals who are solo self-employed as well as for those who employ workers. Each point consists of 5% of observations with respect to their group. Boeri et al. (2020) Found that solo self-employed individuals have higher willingness to pay for insurance compared to people in the normal form of employment. The figure shows that while we are not comparing solo self-employed individuals to wage employees, we see a stark difference between individuals who have workers and those who do not. This could be suggestive that solo self employed individuals do not have high willingness to pay for insurance. However, the difference may also arise from factors which affect the minimum contribution amounts. For instance, individuals who have employees must report their income to be at least as high as their highest paid employee.

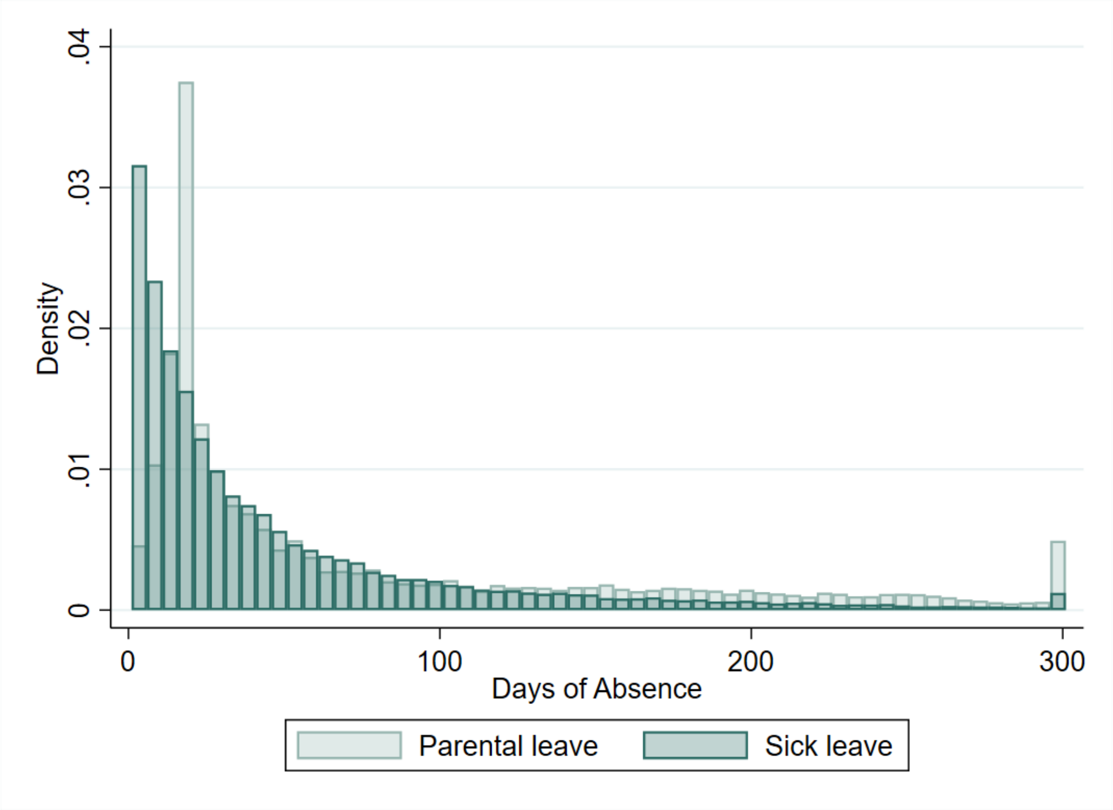

Figure 16

Notes: The figure presents the distribution of absences taken by the sample due to sickness and due to having a child. These are estimated days. The distribution is cutoff at 300 days, such that those taking absences lasting over 300 days are censored at 300.

A.2 Model 1

Dependent Variable: log IC	(1)	(2)	(3)	(4)
t-2	0.01989	0.02801	0.02786	0.02786
	(0.00577)	(0.00412)	(0.00412)	(0.00412)
t-1	0.02289	0.06659	0.06620	0.06620
	(0.00550)	(0.00410)	(0.00410)	(0.00410)
t	-0.08977	-0.04099	-0.04076	-0.04075
	(0.00532)	(0.00418)	(0.00418)	(0.00418)
t+1	0.02273	0.00289	0.00268	0.00270
	(0.00558)	(0.00460)	(0.00460)	(0.00460)
t+2	0.08845	-0.00628	-0.00632	-0.00629
	(0.00589)	(0.00504)	(0.00504)	(0.00504)
ID, Year FE	No	Yes	Yes	Yes
Relationship Status	No	No	Yes	Yes
Education	No	No	No	Yes

Table 11

Notes: The table presents the estimation results from equation 2 using log of insurance contributions as the dependent variable. Time, t represents the year an individual gets ill. Column (1) depicts the baseline results and column (2) adds individual and year fixed effects. Columns (3) and (4) add controls for relationship status and education respectively. Relationship is measured as a binary variable where an individual is either in a relationship or not. Similarly, education is measured as a binary variable for whether an individual has a high school degree or not. Both are controlled for using fixed effects.

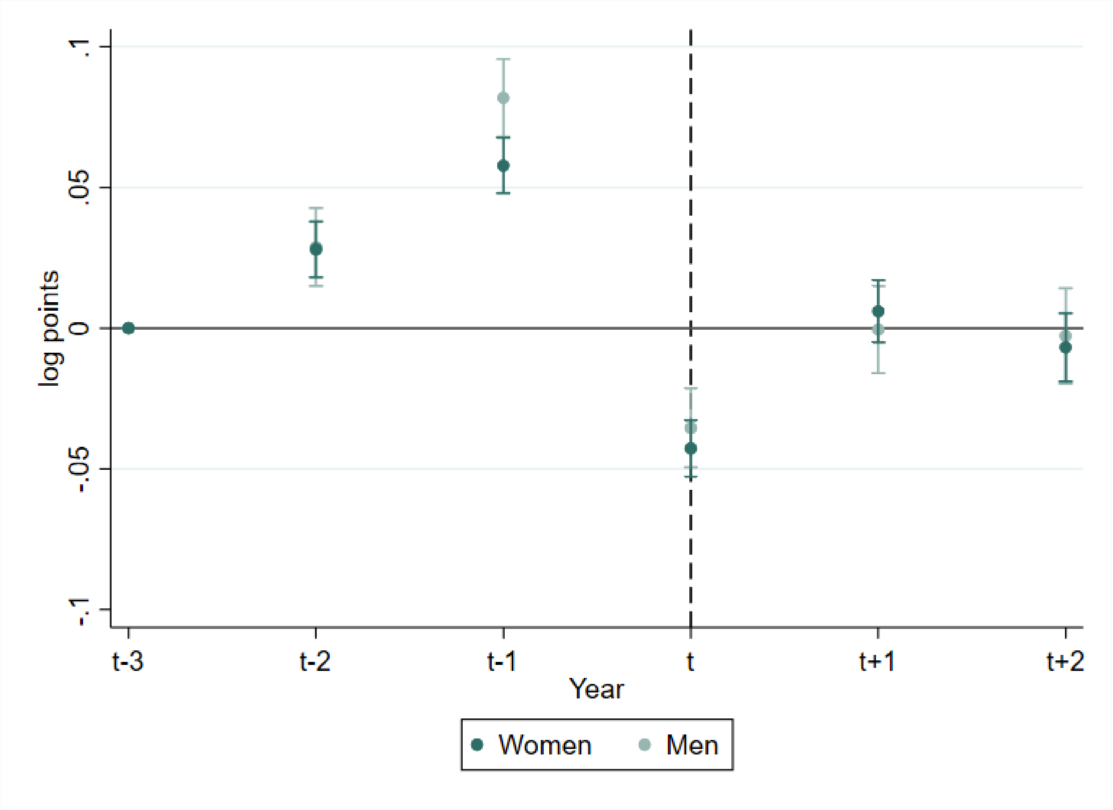

Figure 17

Notes: The figure presents the estimation results with 95% CI from equation 2 using log of insurance contributions as the dependent variable. The CI are derived from standard errors that are clustered on individual level. Time, *t* represents the year an individual gets ill. The figure shows the heterogeneity in responses to sickness allowance between men and women.

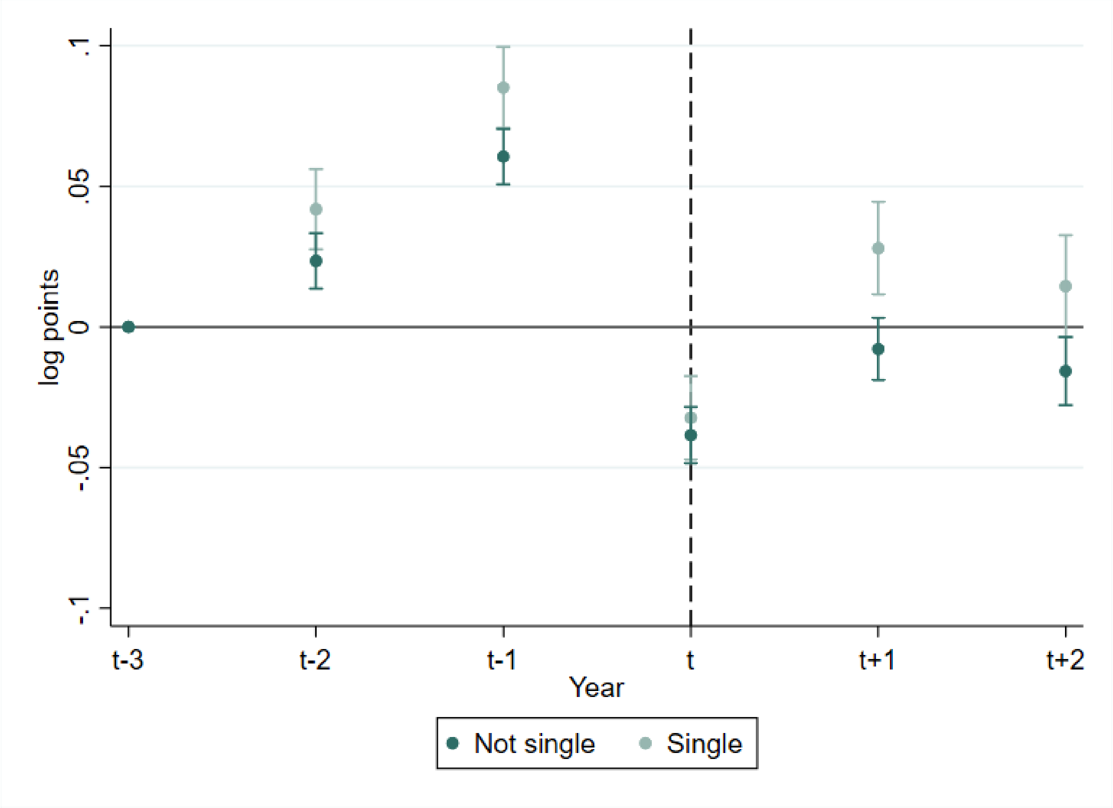

Figure 18

Notes: The figure presents the estimation results with 95% CI from equation 2 using log of insurance contributions as the dependent variable. The CI are derived from standard errors that are clustered on individual level. Time, t represents the year an individual gets ill. The figure shows the heterogeneity in responses to sickness allowance between those who are in a relationship and those who are not.

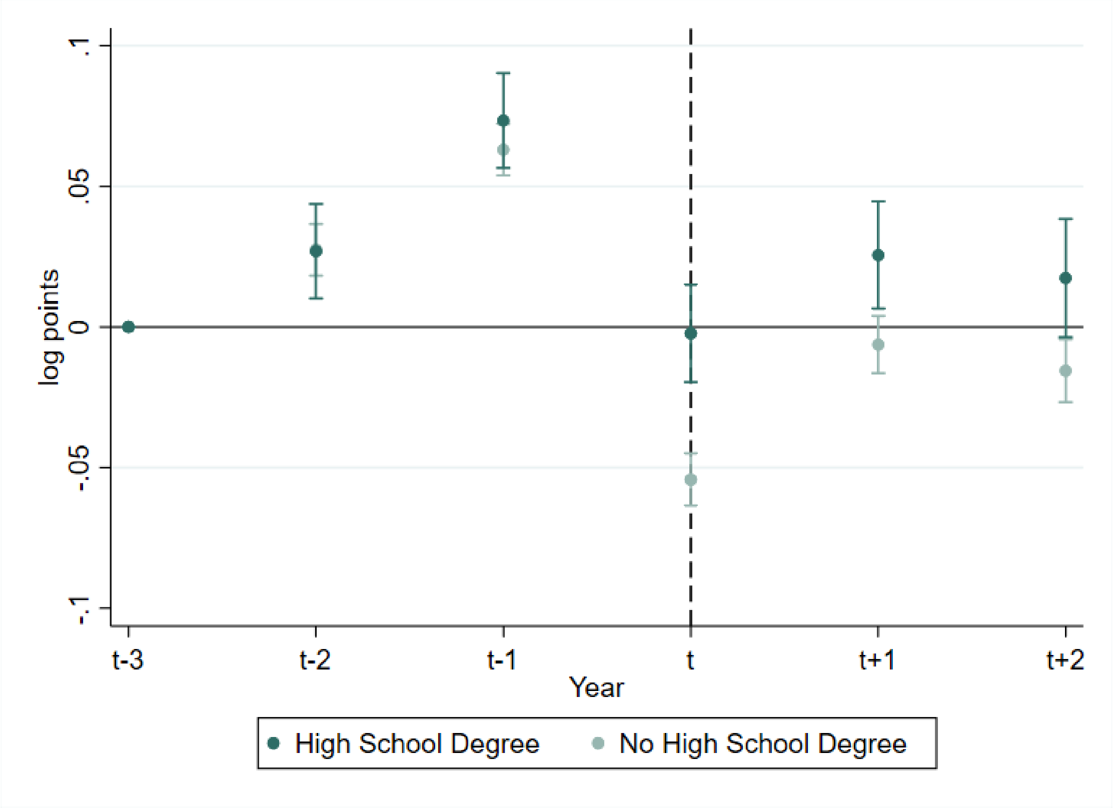

Figure 19

Notes: The figure presents the estimation results with 95% CI from equation 2 using log of insurance contributions as the dependent variable. Time, t represents the year an individual gets ill. The CI are derived from standard errors that are clustered on individual level. The figure shows the heterogeneity in responses to sickness allowance between individuals with a high school degree and those without.

	(1) log IC	(2) log IC (restricted sample)
t-2	0.02801	0.00581
	(0.00412)	(0.01929)
t-1	0.06659	0.02577
	(0.00410)	(0.03560)
t	-0.04099	0.01303
	(0.00418)	(0.05255)
t+1	0.00289	0.08587
	(0.00460)	(0.06970)
t+2	-0.00628	0.08713
	(0.00592)	(0.08690)
ID, Year FE	Yes	Yes

Table 12

Notes: The table presents the estimation results from equation 2 using log of insurance contributions as the dependent variable. Standard errors are clustered on individual level and shown in parenthesis. Column (1) uses the full sample of individuals who are ill in year t. Column (2), however, restricts the sample to only include individuals who are not ill any other year than t. Thus, we exclude individuals who are ill more often and have a longer spell of absence due to illness.

Dependent Variable: log IC	(1)	(2)	(3)	(4)
t-2	0.01767	0.03823	0.03690	0.03685
	(0.01284)	(0.00996)	(0.00998)	(0.0098)
t-1	0.04070	0.10233	0.09925	0.09919
	(0.01196)	(0.00998)	(0.01005)	(0.01005)
t	-0.16511	-0.09940	-0.10391	-0.10402
	(0.01134)	(0.01042)	(0.01053)	(0.01053)
t+1	0.13323	0.08372	0.07917	0.07904
	(0.01177)	(0.01173)	(0.01183)	(0.01183)
t+2	0.23729	0.07981	0.06677	0.06664
	(0.01218)	(0.01309)	(0.01317)	(0.01317)
ID, Year FE	No	Yes	Yes	Yes
Relationship Status	No	No	Yes	Yes
Education	No	No	No	Yes

Table 13

Notes: The table presents the estimation results from equation 2 using log of insurance contributions as the dependent variable. Time, t represents the year an individual has a baby. The standard errors in the table are clustered on individual level and shown in parenthesis. Column (1) depicts the baseline results and column (2) adds individual and year fixed effects. Columns (3) and (4) add controls for relationship status and education respectively. Relationship is measured as a binary variable where an individual is either in a relationship or not. Education is measured as a binary variable for whether an individual has a high school degree or not.

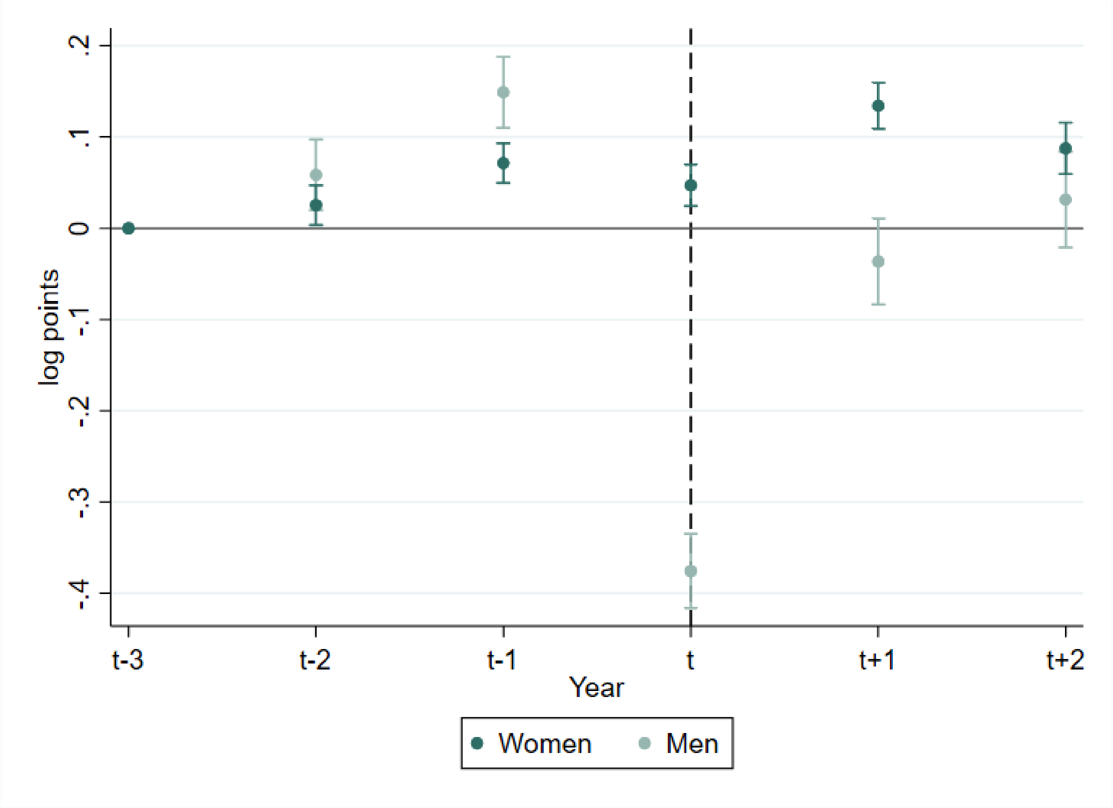

Figure 20

Notes: The figure presents the estimation results with 95% CI from equation 2 using log of insurance contributions as the dependent variable. The CI are derived from standard errors that are clustered on individual level. Time, t represents the year an individual has a baby. The figure shows the heterogeneity in responses to sickness allowance between men and women.

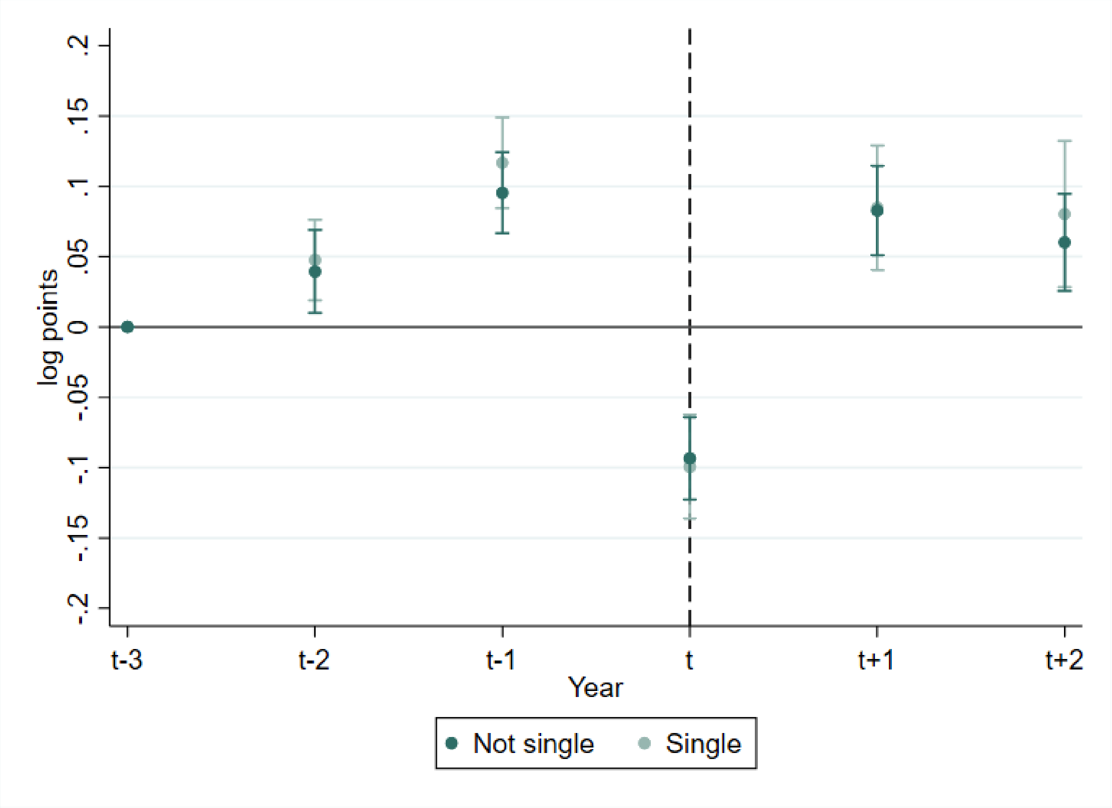

Figure 21

Notes: The figure presents the estimation results with 95% CI from equation 2 using log of insurance contributions as the dependent variable. Time, *t* represents the year an individual has a baby. The CI are derived from standard errors that are clustered on individual level. The figure shows the heterogeneity in responses to sickness allowance between those who are in a relationship and those who are not.

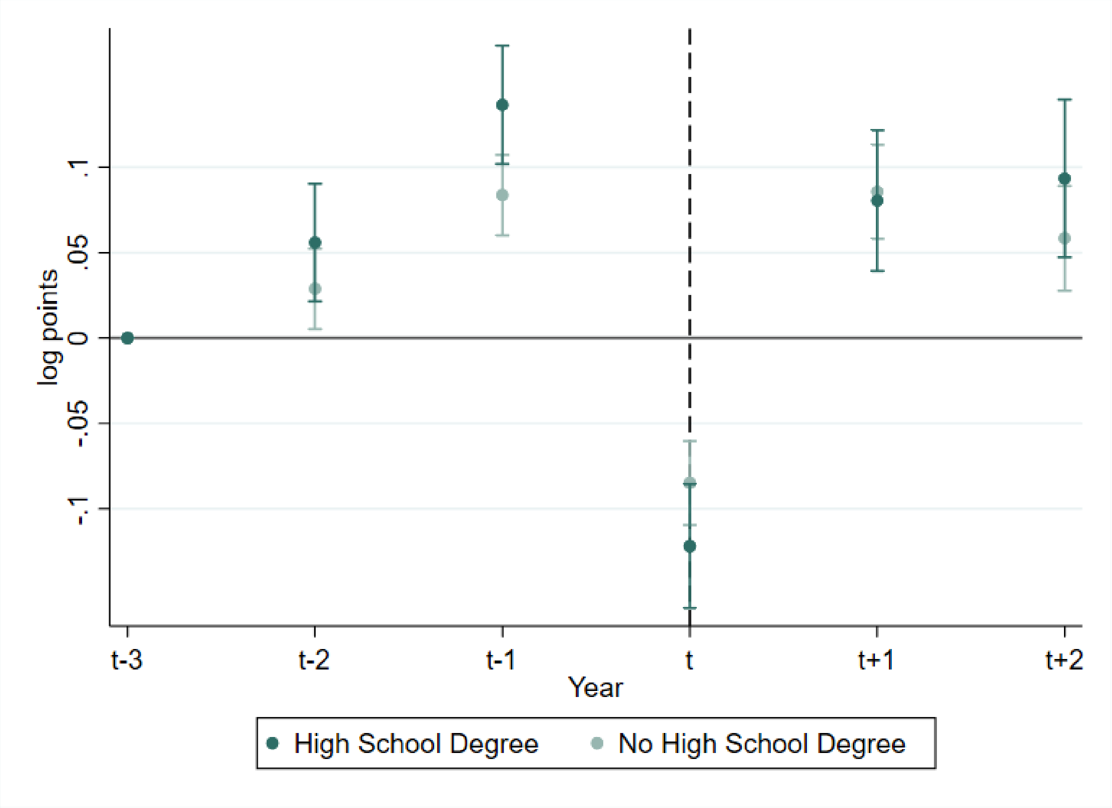

Figure 22

Notes: The figure presents the estimation results with 95% CI from equation 2 using log of insurance contributions as the dependent variable. Time, t represents the year an individual has a baby. The CI are derived from standard errors that are clustered on individual level. The figure shows the heterogeneity in responses to sickness allowance between individuals with a high school degree and those without.

A.3 Model 2

Dependent Variable: c_{it}	(1)	(2)	(3)	(4)
t-2	-0.01048	-0.02144	-0.02145	-0.02145
	(0.00917)	(0.00690)	(0.00690)	(0.00690)
t-1	-0.01168	-0.01860	-0.01872	-0.01872
	(0.00915)	(0.00694)	(0.00694)	(0.00694)
t	-0.16608	-0.14408	-0.14410	-0.14410
	(0.00909)	(0.00695)	(0.00695)	(0.00695)
t+1	-0.12606	-0.10756	-0.10763	-0.10762
	(0.00935)	(0.00718)	(0.00718)	(0.00718)
t+2	-0.13695	-0.09676	-0.09689	-0.09688
	(0.00956)	(0.00735)	(0.00734)	(0.00734)
ID, Year FE	No	Yes	Yes	Yes
Relationship Status	No	No	Yes	Yes
Education	No	No	No	Yes

Table 14

Notes: The table presents the estimation results for Δ_k^{FD} from equation 3. Time, t represents the year an individual gets sick. Standard errors are clustered on individual level and shown in parenthesis. Column (1) depicts the baseline results, column (2) adds individual fixed effects. Columns (3) and (4) add controls for relationship status and education.

	(1) c_{it}	(2) c_{it} (restricted sample)
t-2	-0.02144	0.00547
	(0.00690)	(0.00711)
t-1	-0.01860	0.01313
	(0.00694)	(0.00716)
t	-0.14408	-0.13091
	(0.00695)	(0.00719)
t+1	-0.10756	-0.09868
	(0.00718)	(0.00755)
t+2	-0.09676	-0.09578
	(0.00735)	(0.00785)
ID, Year FE	Yes	Yes

Table 15

Notes: The table presents the estimation results for Δ_k^{FD} from equation 3. Standard errors are clustered on individual level and shown in parenthesis. Column (1) uses the full sample of individuals who are ill in year t. Column (2), however, restricts the sample to only include individuals who are not ill any other year than t. Thus, we exclude individuals who are ill more often and have a longer spell of absence due to illness.

Dependent Variable: c_{it}	(1)	(2)	(3)	(4)
t-2	0.06173	0.04826	0.04644	0.04655
	(0.01305)	(0.00973)	(0.00974)	(0.00974)
t-1	0.0797008	0.0710248	0.06687	0.06703
	(0.01301)	(0.0098109)	(0.00986)	(0.00986)
t	-0.19874	-0.17502	-0.18083	-0.18062
	(0.01279)	(0.00981)	(0.00993)	(0.00993)
t+1	-0.05944	-0.04172	-0.04837	-0.04816
	(0.01358)	(0.01054)	(0.01069)	(0.01069)
t+2	-0.05507	-0.01429	-0.02142	-0.02124
	(0.01398)	(0.01088)	(0.01104)	(0.01105)
ID, Year FE	No	Yes	Yes	Yes
Relationship Status	No	No	Yes	Yes
Education	No	No	No	Yes

Table 16

Notes: The table presents the estimation results for Δ_k^{FD} from equation 3. Time, t represents the year an individual has a baby. Standard errors are clustered on individual level and shown in parenthesis. Column (1) depicts the baseline results, column (2) adds individual fixed effects. Columns (3) and (4) add controls for relationship status and education respectively.