Uncertain health and wealth inequality

Thomas Monk*

December 2019

Abstract

Precautionary saving is a key driver of wealth inequality within models of the Bewley-Huggett-Aiyagari canon. However, models with savings rates calibrated solely to idiosyncratic income risk find it difficult to replicate the vast wealth inequality empirically observed in the United States. This paper looks at a potential source of increased precautionary savings - idiosyncratic medical expenses shocks. This paper: i. establishes an identification procedure for medical expenditure shocks across the entire life cycle, ii. finds that idiosyncratic shocks are very highly persistent, iii. establishes the extent to which these shocks contribute to wealth inequality through the effect on savings behaviour.

Risk is a pervasive feature over the life cycle. Understanding the quantity and type of risk agents face is an important means to understanding their savings behaviour and therefore the shape of the wealth distribution that this induces. Agents also face deterministic reasons for saving, pertinently the sudden income drop that is faced upon retirement. This paper presents a new perspective on an important source of risk throughout the entire life cycle, out of pocket medical expenditure risk, and its importance in forming the vast wealth inequality seen in the United States.

The first aim of this paper is to establish a method of estimating a stochastic process of medical expense cost shocks. Firstly, out of pocket costs are decomposed, utilising a similar procedure as performed by Storesletten, Telmer et al. (2004) and Deaton and Paxson (1994), and an informal graphical analysis is undertaken. These results suggest that the shocks are highly persistent over the life cycle, almost reaching the unit root. This implies that the variance of medical cost shocks increases linearly over time, and is highest nearer the end of the life cycle - an important feature in generating

*London School of Economics, t.d.monk@lse.ac.uk. This paper is adapted from my dissertation at University College London. I would like to thank Vincent Sterk for providing valuable support and advice throughout the dissertation process.

substantial precautionary savings. These results are then more formally established within a minimum distance estimation framework, targeting cross sectional moments.

Secondly, the estimated process is placed within an eighty period heterogeneous agent life cycle model, with lifespan risk, in order to measure the consequences to wealth inequality. Agents are hit with idiosyncratic income and health shocks, which are modelled as mandatory shifts in the budget constraint of an agent. This model suggests that medical expenditure shocks are important to understanding savings behaviour, increasing the Gini coefficient over a benchmark model by 20 percentage points to 0.79 and closely matching the empirically observed distribution of wealth.

Medical costs form a large proportion of expenditure in the elderly of the United States. Without labour income, agents need a buffer stock of savings to smooth consumption across potential expenditure shocks. Importantly, even if large costs are never fully realised over the life cycle, simply the knowledge that these costs can occur in the future can influence saving behaviour at every point in the agents' life cycle. As the variance of these shocks is increasing over time, through the highly persistent process, agents will be inclined to retain savings even when old in order to maintain smooth consumption: the precautionary motive is strong. Further, this model also explores wealth inequality from the other perspective: why agents don't save. With medical expenditure throughout the life cycle potentially shifting an agent's budget constraint to the left, agents are never in a position to fully take up the life cycle and precautionary saving they wish to.

Previous work has been performed towards quantifying medical expenditure, however much has focused on the retirement phase of the life cycle, most pertinently by De Nardi, French et al. (2010), French and Jones (2004) and De Nardi, Fella et al. (2016). Less work has been done on the process of medical expenditure over the entire life cycle of an agent. Hubbard, Skinner et al. (1995) look across the life cycle, but medical expenditure is just one part of extensive model of social security, and further use cross sectional data from just one survey year. Feenberg and Skinner (1992) perform a wide analysis, though due to an absence of good data sources, a complex identification procedure assessing households who have deducted medical expenses from their tax returns is used. Kotlikoff (1986) suggests the importance of medical costs in amassing savings, and identifies a need for including medical cost shocks within a full stochastic model of income and lifetime risk.

This paper will use a pool of contemporary panel data from the *MEPS* and *HRS* surveys to explore the medical cost process across the entire life cycle, from the ages of 21 to 100, and will explicitly look at the its connection to wealth inequality within a wider calibrated model of risk, from the perspective of precautionary savings.

The paper is structured as follows. Section 2 provides context to the heterogeneous agent structures used in modelling idiosyncratic risk, and their success in accounting for wealth inequality. Section 3 outlines the data sets used within the estimation process, and how their characteristics provide motivation for the following discussion. Section 4 develops a two stage identification strategy for estimating a process of medical expenditure shocks, which is found to be highly persistent. Section 5 places this process within a structural life cycle model of idiosyncratic income and health cost risk, and finds that empirically grounded expenditure shocks can account for a marked increase in wealth inequality.

1 Motivation

Inequality within advanced economies is a pertinent issue in modern economic and political discourse, with renewed public interest following the publication of Piketty's (2014) *Capital in the Twenty-First Century*. Empirically both income and wealth are highly skewed towards the right of the distribution, concentrated within a small fraction of society - the richest 10% hold over two-thirds of the total wealth of the United States. (Cagetti and De Nardi, 2008). Wealth inequality is not fully described by attaching its mechanism solely to labour income, as placed within the canonical Aiyagari (1994) model. There is a rich literature, summarised in De Nardi (2015), with a wide variety of mechanisms, which attempt to match more closely the empirically observed distribution.

Krusell and Smith (1998) suggest that introducing heterogeneity in patience should allow the wealth distribution to concentrate within an incomplete markets economy. The mechanism of action includes idiosyncratic income risk, but further places a weight of preference upon savings rate: agents who are less asset rich have *chosen* to be poorer.

Castañeda, Díaz-Giménez et al. (2003) develop a highly specified model with a number of components, which aims to match a number of empirical moments of the wealth distribution of the United States. Most pertinently, the model calibrates an earnings process with two interesting features: i) extremely wide variance of income levels, ii) relatively high probability of leaving the highest income level (equivalently, a low proportion of agents residing within the top income levels). De Nardi, Fella et al. (2016) find that there is not a substantial amount of empirical evidence for extremely large earnings risk when excluding an important subclass of economic agent: the entrepreneur. Quadrini (1999) has further established the importance of entrepreneurship in explaining the savings behaviours of the very wealthy. Cagetti and De Nardi (2006) model the behaviour of entrepreneurs explicitly - in a move away from completely precautionary motives.

De Nardi (2004) looks at intergenerational wealth transmission through two mechanisms: accidental bequests (solely b_t) or voluntary bequests which are embedded directly into the utility function (providing a '*warm glow*'). Accidental bequests are simply redistributed wealth leftover from an uncertain life length, whilst voluntary bequests provide motivation for old aged individuals to hold savings until the point of death. Empirically, Gale and Scholz (1994) have found that bequests appear to form up to 60% of wealth accumulation.

The literature on savings mechanisms allows one to go some way to matching the empirically observed wealth distributions. The precautionary savings motive in particular is clearly a powerful incentive for wealth accumulation across the life cycle. The two best performing models surveyed in this paper modified the savings behaviour directly of the very asset rich in order to match the thick right Pareto tail that is seen in the empirical wealth distribution. This paper aims to take a view from the entire life cycle and cross section of agents in order to generate organic precautionary savings, through *idiosyncratic medical cost shocks*.

Since Modigliani's (1980) life cycle model of consumption, it had been understood that a household will amass savings until they retire, and then spend the savings they have for the rest of the period they to remain alive, in a manner which matches their discounting rate and expected life span. However, empirical studies have shown that the elderly often save, at least do not dissave (Danziger, Van Der Gaag et al., 1982), or simply continue to hold assets at a rate higher than suggested by the basic life cycle model (Hurd, 1989) - uncertain medical expenditure can go some way to explaining this (Palumbo, 1999).

Medical expenses, as this paper understands them, effectively shift an agent's budget constraint, as opposed to a limit on earnings or a utility shock. Uncertainty about the magnitude and timing of these shocks encourage individuals to save in anticipation of offsetting potential large future costs. Importantly, these shocks can occur at any point over the life cycle, and empirically appear to reach their peak (in levels) in old age (De Nardi, French et al., 2010).

The analysis that follows is built upon the initial work of Palumbo (1999), Kotlikoff (1986) and Hubbard, Skinner et al. (1995). Kotlikoff (1986) looks at the different financing methods of medical expenditure, and finds that these are likely to substantially drive aggregate savings. Hubbard, Skinner et al. (1995) have looked at the impact of social security on savings, explicitly modelling a lifetime health care risk shock within a life cycle mode. Their simulations show a strong role for precautionary savings due to medical uncertainty. Palumbo (1999) develops a dynamic structural model, and reaches a similar conclusion. Ozkan (2011) and De Nardi, French et al. (2010) have looked at medical expenditure by income quintle over the life cycle, but have not looked at the entire distribution of wealth. Both assert that from retirement, richer agents pay higher medical costs that poorer agents, and estimate very high medical expenditure for the wealthy when old.

This paper will look at a rich set of recent panel data, using a new estimation technique, in order to establish a stochastic process of medical cost shocks. Further, the literature is yet to explicitly study the impact on wealth inequality across the cross section, which this paper aims to analyse within a structural life cycle model.

2 Data

2.1 Overview

This paper utilises data from the *Medical Expenditure Panel Survey* (*MEPS*). MEPS is a US government funded two-year, large scale panel survey of individuals and their families. The data collected includes accurate and specific information on total and out of pocket medical expenditure, health care use, and insurance coverage. Further, a wide range of demographic detail is available, including socio-economic characteristics, labour information and income breakdowns.

MEPS has two components: the Household Component (HC), collecting data directly from individuals and the Insurance Component (IC), which collects data from sampled private and public sector employers on details of their health insurance coverage. This paper will solely use data from the Household Component.

The MEPS Survey begun in 1996, and surveys around 13,000 nationally representative households in an overlapping panel design. Each year, a new panel of households is selected, whilst one of the previous panels drops out. This leads to around 33,000 individuals within the survey in any given year, with an attrition rate of approximately 5%. Figure 1 demonstrates the structure of MEPS graphically.

This paper also uses data from the Health and Retirement Study (HRS) dataset (University of Michigan, 2016). The HRS is a longitudinal panel survey of older American individuals aged 51 or over and their households. Its main goal is to provide high quality panel data that enables research in into retirement, saving, expenditure, health and economic well-being. Specifically the RAND HRS data set is used which pools together multiple panels of the HRS set, merging potentially heterogeneous variable coding into one coherent file (RAND Center for the Study of Aging, 2016).

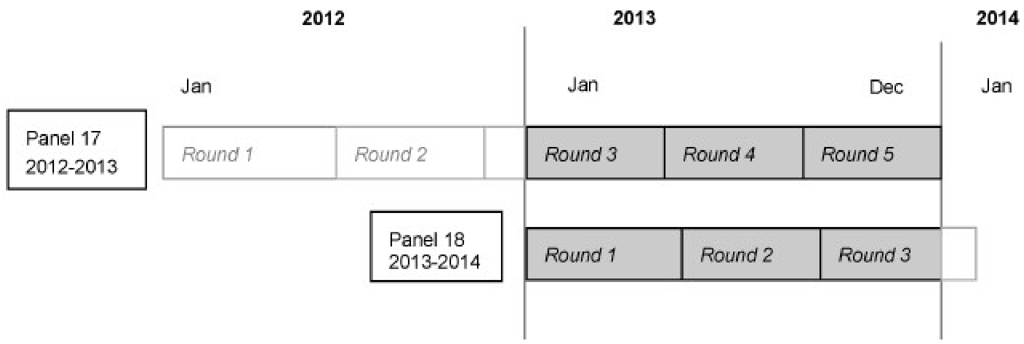

Figure 1: MEPS panel design
(Agency for Healthcare Research and Quality, 2015)

2.2 Sample

This paper utilises a pooled data set of thirteen MEPS years, from 2002 until 2014 inclusive, providing 454,748 data points across 257,942 individuals. The survey provides a wide range of medical expenditure sources, including: Medicaid, Medicare, private insurance, other public insurance (i.e veterans care, TRICARE etc.), drug costs, and out of pocket expenditure. For the analysis that follows, the sample has been selected according to the following restrictions:

- Individuals aged between 21 and 84 inclusive.
  - MEPS top codes individuals to the age of 84. Therefore, mean and other moment information regarding ages at the top code of 85 will be inaccurate.
  - Individuals below the age of 21 have been dropped for two reasons. Firstly, there is high variability in the expenditures of younger agents due to a high percentage of zero codes. Further, in line with the literature, savings behaviour does not commence until the age that the individual joins the labour market.
- The MEPS survey provides information across three waves each year, as outlined in the figure above. The selected sample contains only end of calendar year values for comparison across years.
- Medical expenditure, which is nominal within the datasets, has been inflated to 2014 USD ($). Following Dunn, Grosse et al. (2016), total expenditure has deflated with the GDP deflator, and out of pocket expenditure has been deflated with the Consumer Prices Index.
- Following French and Jones (2004), medical expenditure has been bottom coded to $250, which allows for a log transformation when medical expenditure is zero. This transformation is not completely benign, and may have an economically relevant impact. However, when mean levels are used this transformation is ignored.

The HRS sample has also been cleaned, pooling data from the same years as the MEPS survey, 2002 until 2014 inclusive. This provides 122,564 data points across 27,648 individuals - an average of 4.4 years within each panel for each agent. Similar transformations have been to the HRS data as the MEPS data. This data is not directly used in the estimations, but is used as a robustness check in section 3.2.1 and to provide context in the following section.

2.3 Core facts

To understand the context behind medical cost risk, it is initially useful to look at the cross sectional detail of the sample in the aggregate. To enable this, summary statistics have been constructed, utilising the MEPS survey weights provided to create representative sample moments. Across 5-year cohorts banded by age at the time of the panel survey, table B.1 summarises all of the MEPS sample data points.

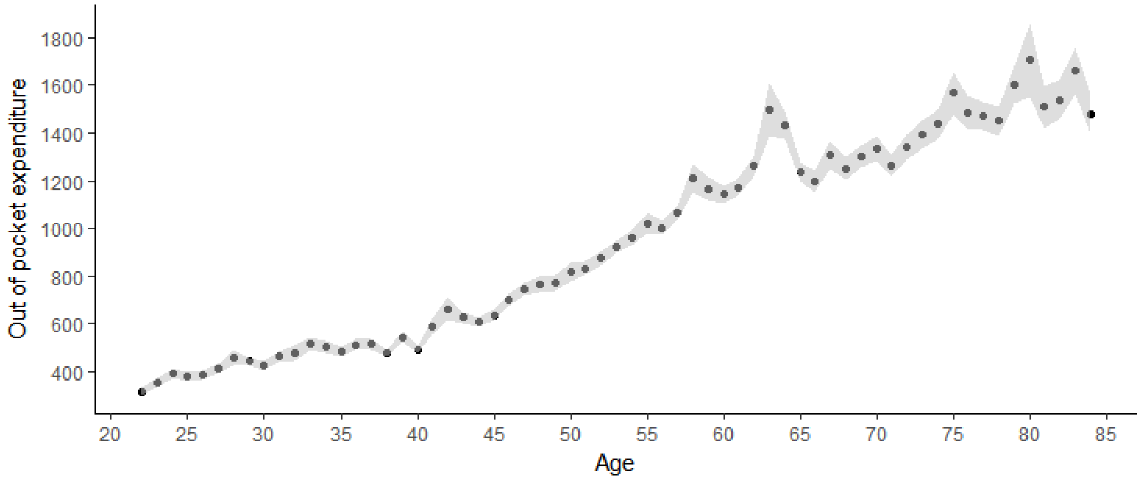

Figure 2: Mean out of pocket medical expenditure, by age

As shown in figure 2, mean out of pocket medical expenditure is clearly strongly increasing over the life cycle. Starting at around \$350 dollars, expenditure rises four fold over a 60 year period to between \$1,400-1,800. Total expenditure is a less important metric, as it will not directly hit the budget constraint, but is a useful indication of the extent of insurance in the health care market and the overall health risk at a given age. Total medical expenditure, figure C.1, rises at a steeper gradient than out of pocket expenditure: it starts around \$1,900 and rises to over \$10,000 at the upper age band.

Importantly, the standard error over time for out of pocket expenditure appears to become proportionally larger over the life cycle. This will be verified later on in this paper, but it looks as if it is empirically true. This does not look to be an effect of the band's population size, as similar populations are within the very lowest and very highest age cohorts.

The figures discussed above suggest that an increasing proportion of expenditure is covered by medical insurance as an agent ages: figure C.2 shows that this is the case, with around 80% of expenses when young, increasing to around 87% when old, paid by an insurance source. This finding is echoed in figure C.5, which shows that the proportion of uninsured agents decreases over the life cycle. In the retirement stage of the life cycle, this is strongly influenced by the uptake of Medicare, which nearly all individuals over the age of 65 are eligible for. Figure C.4 displays this clearly, as around 5% appear publicly insured prior to the age of 65, rising steeply to 13% by the end of the life cycle.

As the MEPS dataset can only provide medical expenditure data up to the age of 85, Figure C.3 shows expenditure from the HRS dataset. As the sample sizes are vastly more limited amongst the very old (93+), there is clearly far more inherent variability in the sample. However, from 85-95, medical expenditures appear to increase at the same pace. De Nardi, French et al. (2016) have taken a comprehensive look at the HRS data set, and have found that total medical expenditure increases fourfold over the retirement phase. They also find that total medical expenditure is highly concentrated towards the end of the life cycle. As Hurd, Rohwedder et al. (2009) notes, mean levels of out of pocket spending are inherently higher HRS than MEPS due to some amount of measurement error at the points where expenditure is extremely concentrated. Agents in the top 5 percent of the distribution account for 35% of all medical spending, and total spending is correlated with income. However, the proportion of out of pocket expenditure is found to be near constant across the income distribution. The impact of end of life spending of the very old on the estimation process is discussed in section 3.2.1 of this paper.

Finally, it is important to assess the potential real impact of medical expenditure on the budget constraint. Figure C.6 shows combined earnings and pension income over the life-cycle, not including capital income. The profile is hump shaped, as studied by the wide literature on the income process, showing clearly the career development and the retirement phases of the life cycle. Figure C.7 displays out of pocket expenditure as a proportion of combined income. Expenditure steeply rises proportionally after the age of 60 - this is the dual effect of losing earnings income and also requiring high medical expenditure as one ages.

Saving for retirement is clearly an extremely important economic mechanism - looking at nominal medical expenditure puts this need in sharp focus - but to now assess the *precautionary* motives, as opposed to the deterministic, the variation of these expenditures is the important moment. The following section turns to a method of identifying this risk.

3 A parametric model of medical cost shocks

Two features of the data above seem imperative: medical expenditure increases sharply with age, but also *variance* of medical expenditure appears to increase with age - i.e there is medical cost 'inequality'. This profile of expenditure lends itself to a strand of identification methods used in the literature on the income process, beginning with Deaton and Paxson (1994) and developed by Storesletten, Telmer et al. (2004).

Intuitively, the life cycle of our medical costs in a developed nation (given the institutional structure and financial system available) can be seen by a system of persistent idiosyncratic shocks - a system determined by our age, our genetics, and some amount of 'luck' over our lifetime. Each individual takes on a fixed effect of inclination towards bad medical outcomes, leading to higher medical costs - these are fixed at birth and stay with the individual throughout their life. Secondly, idiosyncratic shocks to health costs are received at each period of the life cycle, some permanent, and some transitory. The permanent shocks persist at some rate across the rest of the agent's life cycle.

The above describes an error components process model of medical expenditure cost shocks across the life cycle. This section identifies this process through two methods: an informal graphical analysis, which provides valuable insight into the process over the cross section, followed by a formal GMM minimum distance estimation.

3.1 Stage I: Variance decomposition

To understand the risk of medical expenditure, the unexplained component needs to be observed to initiate the analysis. The first stage of the estimation procedure extracts this:

$$m_{it}^h = x_{it}^h \beta + \sum_{t=1}^{T-1} Y_t + \pi_{it}^h \tag{1}$$

m_{it}^h describes the logarithm of total out of pocket medical expenditure, for individual i, of age h, in year t. x_{it}^h is a vector of explanatory demographic variables. Y_t are dummies for each year of the panel. Years run from 2002 until 2014, and therefore t runs from 1 to 13. As discussed in section 2.2, the age of the sample runs from 21 to 80, and therefore h runs from 1 to 60.

This paper will use Ordinary Least Squares to extract the residual, π_{it}^h. There is strong precedent for this in the literature: see, for example French and Jones (2004). Health care costs are skewed to the right of the distribution - following the bottom-coding performed on the variables a log transformation is performed on medical costs to address this. If zero values were instead dropped, as in Hurd (1989), a substantial amount of data would be lost. Manning and Mullahy (2001) suggests that performing OLS upon log transformed variables works well with heavy tailed distributions. However, the coefficients will not be fully reliable as cost measures. This analysis will not directly use the coefficients in the analysis, and one is only interested in the residuals of this regression - however, the coefficients are useful for context. The usual OLS assumptions apply, including exogeniety: $\mathbb{E}(\pi_{it}^h) = 0$.[1] This implies the following:

$$Var(\pi_{it}^h) = \mathbb{E}(\pi_{it}^h)^2 = \sigma_{ht}^2 \tag{2}$$

Therefore, estimated variance equates to the following, where $\hat{m}_{it}^h$ is the OLS projection of the model in equation 1.

$$\hat{\sigma}_{ht}^2 = \mathbb{E}(\hat{\pi}_{it}^h)^2 \tag{3}$$
$$\hat{\pi}_{it}^h = m_{it}^h - \hat{m}_{it}^h$$

Table 1 shows the results of the OLS regression under two specifications, (1) including full demographic details, (2) just including a quadratic function of age as dependent variables. The regression shows what was seen graphically in section 2.1, (log) out of pocket medical expenditure is an increasing function of age. The remaining signs of the dependent variable coefficients appear to affect medical expenditure as expected. The negative coefficient on the male dummy is likely to be longevity bias - women live longer than men in general and are therefore more likely to hit higher medical costs. Unlike the findings of French and Jones (2004) for retired agents, log income negatively affects medical expenditure, as would be expected.

[1] French and Jones (2004) argue, however, that insurance coverage is not not completely exogenous: agents could purchase insurance in response to a medical cost shock, or lose their workplace insurance following catastrophic illness.

	(1)		(2)	
Variable	Coefficient	Robust S.E	Coefficient	Robust S.E
Age	.00442***	(.00058)	.01073***	(.00057)
Age^2	.00014***	(6.20e-06)	.000071***	(6.01e-06)
Male	-.1768***	(.00310)		
Married	.0138***	(.00327)		
Log Income	-.0027***	(.00036)		
Private Ins.	-.2762***	(.00466)		
Public Ins.	-.1914***	(.00382)		
Constant	5.826***	(.01354)	5.438***	(.01217)

Year dummies omitted

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

Table 1: OLS regression for Stage I estimation, with robust standard errors

3.2 Stage IIa: Graphical analysis

The residuals ($\hat{\pi}_{it}^{h}$) from the Stage I estimation are now used to estimate a medical expenditure process. The identification of this process is driven by the difference in cross sectional variance between ages. This section will demonstrate graphically that the cross sectional variance of medical expenses increases over the lifecycle of the agent.

Firstly, it is important to disentangle the *age* effects on medical costs from *cohort* effects. Cohort effects encompass some part of cost variation - the variability of the institutional environment and the effects of public insurance expenditure. The amount of public expenditure on healthcare has a direct effect to the individual costs, but also a potential indirect effect through prices.

Following Deaton and Paxson's (1994) and Storesletten, Telmer et al.'s (2004) technique for decomposing the income process, each individual is assigned a cohort by the year of their birth, denoted c. σ_{it}^{h} denotes the residual of the stage I decomposition of the logarithm of out of pocket medical expenditure, by individual i of age h at time t. This is now decomposed into cohort effects, a_h, and age effects, b_h.

$$\hat{\sigma}_{ih}^{c} = a_c + b_h + e_{ih} \tag{4}$$

As explained above, the analysis that follows will focus on the age effects b_h. The cohort effects do not affect the analysis but clearly play a role in producing variability.

Now an error components model, similar to French and Jones (2004), can be placed upon the residuals. More specifically, following Storesletten, Telmer et al. (2004), the variation in costs can be seen as arising from two sources:

$$\hat{\sigma}^c_{ih} = x_c + u_{ih} \tag{5}$$

where u_{ih} is a specific shock to an individual of age h, while x_c is a shock specfic to the cohort. We can now place a structure on the process that this individual, age dependent shock takes throughout the lifecycle of an agent:

$$u_{ih} = \alpha_i + z_{ih} + \epsilon_{ih} \tag{6}$$

α_i is an individual fixed effect which is realised at the start of the lifecycle and retained throughout it. z_{ih} and ϵ_{ih} are age and individual specific shocks, realized at each point of the life cycle. ϵ_{ih} is a transitory medical cost shock[2], whereas z_{ih} is a permanent shock, which can be further decomposed into an autoregressive component $\rho z_{i,h-1}$ and the innovation η_{ih}:

$$z_{ih} = \rho z_{i,h-1} + \eta_{ih} \tag{7}$$

It is assumed that all shocks within the system are orthogonal and i.i.d., with mean 0.

$$\begin{aligned} \alpha_i &\sim \mathcal{N}(0, \sigma^2_\alpha) \\ \epsilon_{ih} &\sim \mathcal{N}(0, \sigma^2_\epsilon) \\ \eta_{ih} &\sim \mathcal{N}(0, \sigma^2_\eta) \\ z_{i0} &= 0 \end{aligned} \tag{8}$$

Therefore, leading from the assumptions in equation 8, and the error components model in equation 6, at each age h, the age effect from the above decomposition b_h has

[2] ϵ_{ih}, as well as including transitory shocks, embeds *i.i.d* measurement error.

a cross-sectional variance of the following:

$$b_h = Var(u_{ih}) = \sigma_\alpha^2 + \sigma_\eta^2 \sum_{j=0}^{h-1} \rho^{2j} + \sigma_\epsilon^2 \tag{9}$$

Equation 9 is proved in Appendix D, and simply follows from the distributional assumptions placed on error components. Importantly, the summation is not simplified, due to the presumed value of the auto-regressive parameter being close to unity.

The equation forms the heart of the formal identification strategy that follows below, and provide the majority of the moment conditions in the formal estimation. However, it is useful to first look look at a graphical analysis of the variances the decomposition above implies, and how the identification strategy exploits the features of the cross sectional variance of expenditure.

The important parameters, b_h, which are the cross sectional variances of the expenditure shock at age h, can be recovered by simple OLS (without a constant) on the residuals π_{ih}. However, the coefficients must then be scaled to match an unconditional variance within the distribution, so that comparisons can be made cross-sectionally. Following Deaton and Paxson (1994), the coefficients are scaled to b_h^* through the choice of a reference age - this analysis will use 42. b_h^* exactly identifies the cross sectional variance at age h. The following system of linear equations can then be solved to obtain the relevant parameters in the error components system, following Storesletten, Telmer et al. (2004):

$$\begin{aligned} \mathbb{E}[(\pi_{ih}^c)^2|h] - \bar{a} - b_h &= 0, \text{H moments} \\ \mathbb{E}[(\pi_{ih}^c)^2|c] - \bar{b} - a_c &= 0, \text{C moments} \\ b_h + (m - b_{42}) - b_h^* &= 0 \\ \mathbb{E}[(\pi_{ih}^c)^2|h = 42] - m &= 0 \end{aligned} \tag{10}$$

where H is the number of distinct ages present in the sample, and C is the number of distinct ages in the sample. Therefore, $H + C + 2$ equations in total are to be solved to obtain b_h^* - solving the system of equations obtains the results in Figure 3.

Figure 3 shows clearly that the cross sectional variance is increasing over the life

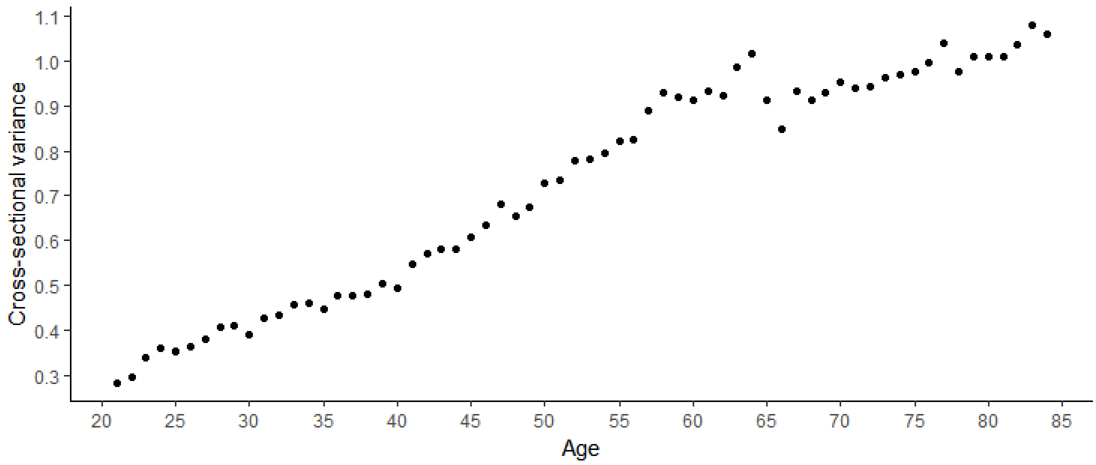

Figure 3: Cross-sectional variance

cycle. One can loosely identify important parameters of the error components model from the properties of the line - Storesletten, Telmer et al. (2004) describe this algorithm as impressionistic. The first derivative roughly identifies the variance of the permanent shocks σ_η^2, whilst the second derivative identifies the persistence parameter ρ. Very roughly, $\sigma_\alpha + \sigma_\eta$ can be identified as being around the intercept of the data. It is clear that the line is roughly straight, implying a second derivative of 1 - medical cost shocks are *highly persistent*. The gradient of the line is approximately 0.01 - $\eta \sim \mathcal{N}(0, 0.01)$. The intercept of the line is roughly 0.35 ($\approx \sqrt{0.59}$).

To more fully identify the parameters, we can place equation 9 into the system.

$$b_h^* = Var(\pi_{it}^h) = \sigma_\alpha^2 + \sigma_\eta^2 \sum_{j=0}^{h-1} \rho^{2j} + \sigma_\epsilon^2, \text{H moments} \tag{11}$$

This now sets $2H + C + 2$ equations in total can now be solved to estimate the four parameters σ_α, σ_ϵ, σ_η and ρ. This initial analysis will exactly identify the system using three different ages and corresponding b_h points. [3] The subsequent section will place these moment conditions inside a minimum distance estimator to ensure full robustness. As $\sigma_\epsilon^2 + \sigma_\alpha^2$ is simply a linear combination, this procedure will not be able to identify

[3]The ages used in the analysis above are 30, 45 and 55 - only three are needed as σ_α and σ_ϵ cannot be seperately identified. The analysis has been tested as robust to using a range of different ages, and is confirmed with the GMM procedure below.

them separately.

Analysis	$\sigma_\alpha + \sigma_\epsilon$	σ_η	ρ
Rough graphical estimates	0.59	0.1	1
System of non-linear equations	0.515	0.128	0.996

Table 2: Stage IIa estimation

This exactly identified system confirms the results of the graphical analysis - medical costs are extremely persistent - almost permanent. The transitory shock is slightly smaller than the fixed effects, but over the life cycle these effects sum to become substantial, through their high persistence.

The analysis can also be tested for robustness across demographic variables. Appendix E plots the cross sectional variances across gender and marital status. Females appear to have a higher cross sectional variance than males up, with the difference declining until the point of Medicare eligibility at age 65 where the two converge. This suggests that women are likely to be less well insured against these shocks than their male counterparts. Across marital status, there appears to be limited difference across the life cycle, and the coefficients confirm the aggregate levels above.

3.2.1 Limitations of the MEPS dataset: a robustness check

Before continuing with the formal analysis, it is worth discussing the limitations of the MEPS dataset which are pertinent to this paper. Firstly, MEPS top-codes all ages to 85. This implies an inaccuracy in the moment information of ages at the top code, and further limits our potential for analysis of individuals aged 84 or younger. Secondly, the dataset is limited to non-institutionalised individuals. De Nardi, French et al. (2016) have shown that a large portion of healthcare costs in the elderly (aged 65 and older) is driven by institutional costs (see figure C.8), and this proportion increases as individuals age. It is clearly important to address these issues to ensure accurate estimation for medical costs moments in the very old aged. This section will outline the methods used in this paper to tackle this.

The Health and Retirement Study (HRS) dataset is a rich and complex longitudinal panel survey of older individuals and households in the United States. Positively, it includes medical expenditure related to long term care. Further there is no top coding in age; with a low attrition rate (Sonnega, Faul et al., 2014) individuals are often surveyed until death.

However, the HRS panel population is only surveyed every two years. Reconciling this two-yearly data with the yearly data produced by MEPS specifically is difficult. MEPS, as discussed, only provides a one-year lag, which HRS excludes completely. Therefore, the moment conditions that are used in the formal analysis in a merged data set would be mutually exclusive across the two data sources. One could simply average across the two years to create two one year figures but this would remove the variance differentials by which the estimation procedure is driven. French and Jones (2004) produced a methodology for producing estimated yearly process parameter values based on two year expenditure data, but producing process estimations for the entire life cycle using simulated data from these processes would place an inherent distribution upon the data and bias further analysis over the entire cross section.

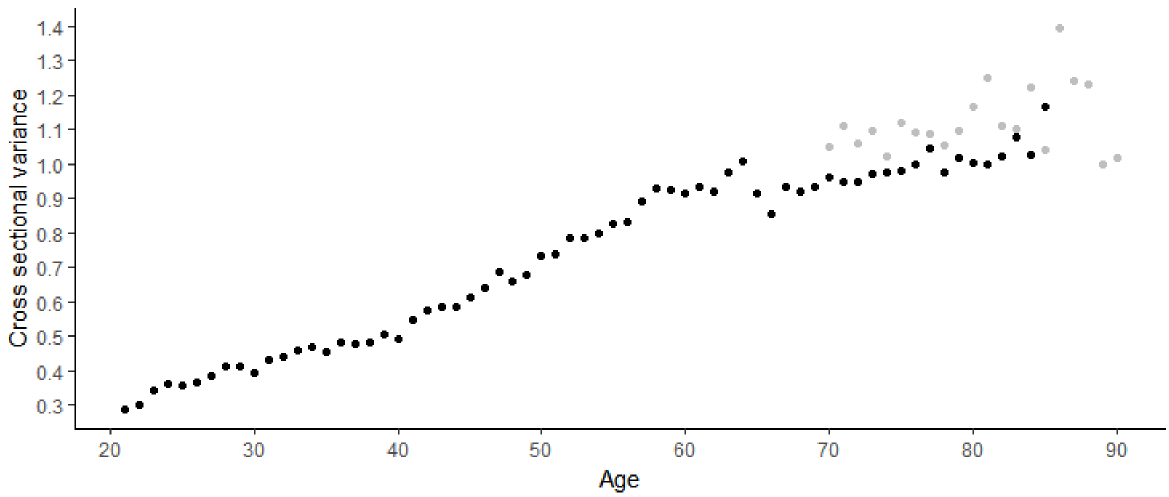

Figure 4: HRS (grey) and MEPS (black) cross sectional variances

The approach in this paper is to combine the two datasets in a limited manner, producing the parametric estimates through the residual variances of the more limited MEPS dataset, but using a merged dataset to estimate overall level information utilised in the model, which is less important for savings reasons due to being deterministic. Cross sectional moments of the two datasets suggest this approach is reasonable.

Figure 4 shows the cross sectional variance of the residuals of a regression similar to the one performed above, MEPS variances in black, whilst the HRS variances are in grey. Note that the variances are not scaled as above, but the relative sizes remain important.

The HRS variances are slightly higher than MEPS set, which is likely to be attributed both to differences in institutional status and also sample size. However, disregarding the final years of life whose variance is completely driven by the extremely small sample size ($N \approx 10$), the variances appear to increase at a similar rate across the entire life cycle. This vindicates the error components model estimated above, and is clearly well matched by a highly persistent process.

3.3 Stage IIb: Formal analysis

The analysis above appears robust to graphical, informal parametric treatments with a mix of identifying equations, across demographic characteristics, and to further comparison to different data sets. The formal analysis within a GMM framework will confirm these findings fully, provide standard errors, and will be able to fully identify the parameters within the error components model.

Utilising a GMM methodology also allows the possibility of widening the information set included in the analysis - to more accurately identify the persistence parameter the estimation will now include a set of autocorrelation moment conditions, given the two years of data available for most individuals within the MEPS panel.

The autocovariance of the residual of medical expenditure between ages h and $h-1$ for individual i can be shown, in a similar method to the proof of equation 9, c.f. Sopchokchai (2015), to equal the following:

$$\mathbb{E}(\pi_{it}^{h}\pi_{i,t-1}^{h-1}) = \sigma_{\alpha}^{2} + \rho \sum_{j=1}^{h-1} \rho^{2(j-1)}\sigma_{\eta}^{2} \tag{12}$$

Equations 9 and 12 underlie the GMM estimation used in this section. The two sets entail 2H - 1 total moment conditions to use within the system.

The MEPS dataset is unbalanced, and therefore to obtain a full estimation this analysis will follow the techniques used for income and consumption process estimation of Blundell, Pistaferri et al. (2008). In summary, minimum distance estimation will occur across a empirical weighted variance-covariance matrix against the variance and autocovariance moment conditions above in equations 9 and 12.

The empirical variance-covariance matrix is calculated in the following manner. For every individual within the sample, i, define d_i as an H x 1 vector, with each element an indicator, $d_{ih} \in \{0, 1\}$ as to whether the agent is present in the sample at age h. Conformably with d_i, define m_i as a H x 1 vector of residuals from the Stage I estimation of individual i at age h.[4]

The VCV matrix (C), of dimensions H x H, is therefore defined as the following (an element wise division):

$$\mathbf{m} \equiv vech\left\{\frac{\sum_{i=1}^{I} m_i m_i'}{\sum_{i=1}^{I} d_i d_i'}\right\} \tag{13}$$

Each element of the vector $\mathbf{m}$ contains the cross sectional permuted co-variances of the medical cost residual, $cov(m_h, m_{h+s})$ at age h and $h + s$.

Estimation can subsequently be performed following the techniques of Chamberlain (1984), a minimum distance estimation (MDE) which minimizes the delta between across the empirical VCV matrix and the modelled covariance matrix. The MDE estimator solves the following problem:

$$\min_{\Theta} \, [\mathbf{m} - f(\theta)]' \, \Omega \, [\mathbf{m} - f(\theta)] \tag{14}$$

where $f(\theta)$ is a vector conforming to the vectorised VCV matrix $\mathbf{m}$, with dimensions H(H+1)/2, 1; Ω is a weighting matrix, and Θ is a set of parameters - in this case our error component parameters σ_α, σ_ϵ, σ_η and ρ.

Regardless of the choice of weighting matrix, the estimator is consistent under loose regularity conditions (for a proof, see (Chamberlain, 1984)). However the estimator may be made asymptotically more efficient by using a suitable weighting matrix. A potential asymptotically efficient weighting matrix would be the inverse of the asymptotic variance V^{-1} where

$$V = \frac{\sum_{i=1}^{I} (\mathbf{m} - \mathbf{m}_i)(\mathbf{m} - \mathbf{m}_i)'}{ss'}$$

and where $s = vech\{\sum_{i=1}^{I} d_i d_i'\}$. However, as discussed in Altonji and Segal (1996) this weighting matrix has substantial bias in small samples due to the correlation in errors between $\mathbf{m}$ and V. A majority of the literature on the estimation of the income

[4]Missing residuals are marked at 0. As individuals are followed for a maximum of two years, the matrix will be extremely sparse, with a minimum of A-2 zero entries.

process use an equally weighted minimum distance estimator, $\Omega = I$, which this paper will follow. Standard errors will be bootstrapped.[5]

3.3.1 Estimation results

Table 3 details the results of the estimation. What is immediately clear is that the medical shocks are certainly near permanent, the permanent innovation is slightly decreased, and the sum of the transitory and fixed effects are slightly increased. Overall, the parameters appear robust to the inclusion of autocovariance moments, and the overidentification. The fixed effects are quickly dominated by the build up of permanent innovations over the life cycle - a potential interpretation of this is once a medical problem which requires out of pocket expenditure presents itself within a life time it is unlikely to completely dissipate. Alternatively, those who have been sick in the past within a life cycle, and require out of pocket spending, are more likely to require it in the future.

Analysis	σ_α	σ_ϵ	σ_η	ρ
Rough graphical estimates	0.59		0.1	1
System of non-linear equations	0.515		0.128	0.996
MDE estimation	0.236	0.406	0.107	0.995
Standard errors	(0.050)	(0.015)	(0.004)	(0.0008)

Table 3: Stage IIb MDE estimation

There have been few other studies on the health cost process. French and Jones (2004) looked at a wide range of possible processes, and fitted an ARMA(1, 1) process to health cost data within the HRS (inherently limited to ages 51+). This process was then decomposed into AR(1) plus white noise, with one year shocks approximated from 2 year HRS data, and subsequently fitted to a model of the distribution of catastrophic health care shocks. The decomposed and fitted processes are respectively French (1) and French (2) in Table 4. Feenberg and Skinner (1992) performed an estimation of the medical expenditure process based on the tax filings of those who have deducted the

[5] 2000 random samples, with replacement, of N individuals will be taken from the overall sample of N, and the estimation procedure above is then performed. Standard errors are taken from cross-sample variation.

Standard errors could also be computed in the following manner: $(\mathbf{G}'\Omega\mathbf{G})^{-1}\mathbf{G}'\Omega\mathbf{V}\Omega\mathbf{G}(\mathbf{G}'\Omega\mathbf{G})^{-1}$, where $\mathbf{G}$ is the gradient matrix of the objective VCV $\partial f(\theta^*)/\partial\theta$, evaluated at the parameter $\theta^* \in \Theta$.

costs, whilst Hubbard, Skinner et al. (1995) utilise data from the 1977 National Health Care Expenditure Survey.

Estimation	σ_ϵ	σ_η [6]	ρ
MDE estimation	0.406	0.107	0.995
French, HRS data (1)	1.019	0.362	0.922
French, HRS data (2)	1.347	0.281	0.925
Feenberg & Skinner	0.316	0.230	0.896
Hubbard et al.	0.469	0.418	0.901

Table 4: Comparison to other estimates

Firstly, the estimate for σ_ϵ, the transitory effect, is far smaller than French's but roughly in line with both Hubbard and Freenberg & Skinner's. Firstly, as none of the models report an individual fixed effect, this is not included in the tables, but may be the source of the disparity. Secondly, French and Jones (2004) discuss potential reasons as to why their estimates are so vastly different to the previous models, and concludes that a possible reason is due to their inclusion of health care costs of those who have died mid sample, and therefore have higher health costs. In general, age seems indicative of where the differences lie - French and Jones solely look at agents over the age of 50, as this is all that HRS captures. This does not capture the behaviour of medical cost shocks at the start of the life cycle, potentially missing prior volatility. Although the process is stationary over the life cycle, not capturing the lower disparities in individual data throughout the bulk of the working life cycle could contribute to these differences.

Turning to the smaller differences in the MDE estimates of σ_ϵ to the Hubbard or Freenberg results, these could simply be attributed to rising general variance of medical expenditure of time. However, the source of the data may again be an issue - the tax returns provided for Freenberg's estimation could simply be less variable than the survey results. As French and Jones (2004) note, only items are costly enough to be itemised are included within the sample. Given that Hubbard et al. based their estimates in part on the Tobit tax model of Freenberg, these results inherent similar properties. Alternatively, measurement error may be higher in the MEPS sample, as this is solely captured by σ_ϵ.

The most important reason for the difference in the processes is the near unit root of

[6] σ_η is obtained from the variance of the autoregressive component of French and Jones's (2004) estimations.

the MDE estimation - directly translating from the evidence seen in the cross sectional variance graphs. The estimate for the innovation to the permanent shock σ_η is clearly the smallest of the set. However, it is important to note that this is coupled with the higher persistence of the AR(1) shock and therefore this parameter has much more of an effect across the life cycle than the other model's estimates. Intuitively - the other models predict more equally distributed variance across the life cycle, whilst this estimation sees variance as a linear function of age - the cross sectional variance of shocks is highest upon reaching old age. French and Jones, in only capturing the expenditure process of the already old, perhaps misses the persistence across the entire life cycle.

In summary, this model has shown that there is robust evidence that medical cost shocks are extremely persistent over the life cycle. Although the fixed effect is larger than the conditional variance of the permanent shock, this gets amplified across the life cycle. Similar to Storesletten, Telmer et al.'s (2004) findings for the income process, the almost unit root of the medical cost process is a direct implication of the linear increase in cross sectional variance over the life cycle.

4 Life cycle model

Having determined an estimated medical expenditure process, this section will now tackle the core question of this paper: what is the effect of the uncertainty of these shocks on the wealth distribution of an economy? This paper presents a stochastic finite horizon life cycle model of this cost risk, similar to Palumbo (1999) and Hubbard, Skinner et al. (1995). Agents will face three sources of uncertainty: individuals face idiosyncratic medical expenditure shocks as estimated by the procedures above, they also face idiosyncratic income shocks and face risk regarding their longevity (i.e. a non-zero risk of dying at every point in the life cycle). This model therefore allows the precautionary savings motives discussed throughout the previous sections to be explored and tested through calibration and parameterisation with contemporary data and estimation techniques.

4.1 Outline

Consider the agent's problem:

$$\max_{c_h, a_{h+1}} \mathbb{E} \sum_{h=1}^{H} \beta^h \phi_h u(c_h) \tag{15}$$

$$\text{s.t. } c_h + a'_{h+1} \leq (1 + r)a_h + y_h - m_h + b_t \tag{16}$$

The agent will choose optimal c_h and a_{h+1}, consumption and asset holding respectively, in order to maximise their expected utility over their life cycle subject to their budget constraint. The risks discussed present themselves through the following parameters:

ϕ_h denotes conditional survival probabilities of reaching age $h + 1$, having attained age h, and act as a discount rate. It is assumed that households do not have a full market available to insure against this risk - for example, an actuarially fair annuities market could mitigate against this risk and therefore it would lose its bite on consumption patterns (Hubbard, 1987; Friedman and Warshawsky, 1990; Hubbard, Skinner et al., 1995).

m_h denotes medical expenditure at age h. This is modelled as an exogenous shock to the budget constraint. In some papers, health costs are modelled as providing utility (or alleviating disutility), but this paper will model them as a negative shock to the budget constraint, following the models of De Nardi, French et al. (2010); Palumbo (1999); Kotlikoff (1988) and Hubbard, Skinner et al. (1995). This paper also abstracts from the potential endogeneity of health costs through (observed) health status. As estimated above, health cost shocks are assumed to follow the following process:

$$\ln m_h = \kappa_h + \alpha + z_h + \epsilon_h \tag{17}$$

$$z_h = \rho z_{h-1} + \eta_h$$

As estimated previously, α is a fixed effect drawn at birth from the distribution $\alpha \sim N(0, \sigma^2_\alpha)$; ϵ_h is a transitory shock, drawn at each age h from the distribution $\epsilon_h \sim N(0, \sigma^2_\epsilon)$ and z_h is a permanent AR(1) shock with innovation $\eta_h \sim N(0, \sigma^2_\eta)$. κ_h determines the deterministic profile of medical expenditure at age h.

In comparison to other models, the health shocks here are directly modelled. Kotlikoff's (1988) model assumes that the shocks are independent over the life cycle, and are drawn from a simplified distribution not from empirical micro data. Palumbo (1999) utilises uncertainty in health status to matched regressed expenditure values to produce cost uncertainty. De Nardi, French et al. (2016) also models endogenous medical costs, in part, and Ozkan (2011) has modeled a form of cost through preventative medical expenditure. This paper directly models a process for medical cost shocks, empirically grounded in the micro data - Hubbard, Skinner et al. (1995) have estimated a similar model, but uses material from Feenberg and Skinner (1992) for their estimates. The model this paper is estimating is fully self-contained.

y_h denotes income in year t. This follows a similar process to medical expenditure, following Storesletten, Telmer et al. (2004); Hubbard, Skinner et al. (1995) etc; with conformable assumptions to the above. After the age of retirement agents receive a pension.

$$\ln y_h = \kappa^y_h + \alpha^y + z^y_h + \epsilon^y_h \tag{18}$$

$$z^y_h = \rho^y z^y_{h-1} + \eta^y_h$$

b_t denotes government transfers. In line with De Nardi, French et al. (2010), Hubbard, Skinner et al. (1995) a consumption floor, $\bar{c}$ will hold:

$$b_t = \max\left[\bar{c} + m_t - (1+r)a_h - y_h, 0\right] \tag{19}$$

Exogenous government transfers will ensure that an agent's consumption doesn't fall below the floor - this is important for savings when negative shocks enter into the budget constraint, and provides a simple model of social security. The floor could be funded with taxes, but this model is abstracted from government resource allocation.

4.2 Calibration

This paper will calibrate the model described above using various sources of microdata, in order to closely match the empirical reality.

Time. One period of the model, h, is one calendar year within the life cycle. Agents start their lives at age 21, and are certainly dead by age 100. Conditional on surviving, they retire at age 65.

Mortality. Survival rates are calibrated to actuarial life tables from the National Center for Health Statistics for the year 2002 - the starting year of the MEPS sample used.

Patience. In combination with the mortality rates, the time preference parameter β discounts the future utility stream. The discount factor is set to 0.96, in line with values used by Huggett (1993), Cagetti and De Nardi (2006) and Hubbard, Skinner et al. (1995).

Utility. The CRRA (isoelastic) utility function is widely used in the literature, c.f. Storesletten, Telmer et al. (2004), Cagetti and De Nardi (2006) etc.

$$u(c) = \frac{c^{(1-\gamma)} - 1}{1-\gamma}$$

The choice of γ is important for three reasons directly related to the savings behaviour of agents. Firstly, it functions as the coefficient of relative risk aversion; secondly, $1/\gamma$ is the intertemporal elasticity of consumption. Most pertinently, $\gamma + 1$ drives the precautionary savings motive, and is a measure of prudence. Through Jensen's inequality, it can be shown that a convex marginal utility function (u''') generates prudence - an additional unit of consumption is more valuable to an agent when consumption is low, as opposed to when consumption is high. Kimball (1990) defines a measure of absolute prudence as $P(w) \equiv -\frac{u'''(c)}{u''(c)}$.

In line with a large section of the literature, this analysis will set γ to 1.5 (see Castañeda, Díaz-Giménez et al. (2003), c.f. Huggett (1993), Mehra and Prescott (1985) for a discussion.

Medical expenditure. Medical expenditure shocks are calibrated to the model estimated in this paper: $\sigma_\alpha = 0.236$, $\sigma_\epsilon = 0.406$, $\sigma_\eta = 0.107$ and $\rho = 0.995$. The deterministic factor, κ_h, is calibrated to the age profile of income levels within the sample set of the MEPS and HRS datasets.

Income. The income process is modelled as an autoregressive income shock with one lag, potential fixed and transitory effects are not implemented due to computational limitations. The persistence and innovations to this process use the estimation performed by Storesletten, Telmer et al. (2004): $\rho^y = 0.977, \eta_h^y = 0.033$ The deterministic part of income at each age h, κ^y, is calibrated to the age profile of income levels (income being defined as total earnings and pensions[7]) within the pooled MEPS dataset used for the estimation above, and the HRS.

4.3 Solution method

This model is unable to solved analytically with closed form solutions. Therefore, this paper will solve the stochastic life cycle model using discretised approximations of the shocks.

There are three state variables: asset choice last period (a), income state last period (s), and medical state last period (e). The AR(1) shocks of the income and medical cost components must be also discretised. The standard in the literature is to utilise Tauchen's (1986) method of transforming an AR(1) process into a discrete set of Markov-chain states. However, this method is not robust to highly persistent processes. Therefore, this paper will use the method of discretisation proposed by Rouwenhorst, established in the macroeconomic literature by Kopecky and Suen (2010). [8]

The following regularity constraints are also imposed: (1) the choice variable must always be at least the value of the consumption floor, $c_t + a_{t+1} \geq \bar{c}$, (2) asset choice must

[7]WAGEP and PENSP in the MEPS dataset.

[8]The processes were discretised into 7 states with mean zero, and 7x7 corresponding Markov-chains.

be non-negative (a strict no borrowing constraint): $a_{t+1} \geq 0$.

Following the solution method of İmrohoroğlu, İmrohoroğlu et al. (2001), discretise the asset space: let $A = \{a_1, a_2, ..., a_m\}$ denote the grid of possible asset holding points. Conformable entities S and E exist for income shocks and medical expenditures respectively. At the beginning of each period h, the agent finds themselves in state $(a, s, e) \in A \times S \times E$. Let $(a_j, s_j, e_j) \in \Omega_j$ define the set of constrained possible states, such that the budget constraint and regularity constraints hold at age j for $j \in \{21, 22, ..., 100\}$. $V_j(a, s, e)$ can now be defined as the solution to the agent's dynamic problem (the Bellman equation).

$$V_j(a, s, e) = \max_{a'} \left\{ u(c) + \beta \phi_{j+1} \mathbb{E} V_{j+1}(a', s', e') \right\} \tag{20}$$

Given that the agent is dead with certainty at age $H + 1$, the value function and the corresponding decision rule can be obtained through solving the problem backwards from the last period of life. Substituting the budget constraint into the Bellman equation allows the agent to only have one choice variable, their asset holdings.

In order to develop aggregate statistics, the economy is Monte Carlo simulated through forward iteration over the life cycle, given the optimal decision rules and value function generated using the procedure above. Life cycle profiles are formed and full simulated panel data of each economic variable is generated. 2,000 individuals are simulated and followed over their life cycle. The initial state, medical and income shocks are drawn from the distributions detailed in section 4.2.[9]

The aim is to observe the economy at a steady state - this is approximated by the distribution of the simulated households (weighted by survival probability) at each age h, through the law of large numbers. This distribution is a snapshot across many different cohorts of agents at the point they reach age h. This snapshot is assumed stationary, as the economy is in a steady state. Aggregate economic variables can then be constructed using the cross-sectional information from the steady state snapshot.[10]

[9]The transitory and fixed effect are not included in the discussion above, for brevity, but must also be discretised. The transitory shocks are transformed into a 7 element vector, with each element i matching the mass of the normal distribution at percentile $i/7$. The agent has a 1/7 chance of choosing any element within that vector at each period. Computing the economy with fixed effects is more computational intensive task due to the fact that unique decision rules and value functions must be calculated for for possible effect. For that reason, it is discretised in a similar fashion, but over three states - a positive, negative or neutral effect (of mean zero). 1/3 of the agents simulated are assigned to each fixed state.

[10]This is performed through weighting the asset densities of each age group by survival probabilities

4.4 Results

The results of the model described suggests that adding uncertain medical expenditure to the life cycle increases wealth inequality within an economy quite significantly. The table below displays the benchmark model, which only includes idiosyncratic income risk, and the full model with modelled medical expenditure shocks.

		Share of total wealth (%)				
Source of distribution	Gini	Top 40%	20%	10%	5%	1%
Data (SCF, 2001)	0.79	94	82	69	57	32
Baseline model	0.60	88	59	34	18	4
With medical expenditure shocks	0.80	99	87	65	40	10

Table 5: Modeled wealth distribution with medical expenditure shocks

Firstly, the Gini coefficient is well matched - this is the overall measure of inequality in the economy. The percentile information shows exactly how this wealth is distributed. The top 20 per cent to 5 per cent are very well matched to the data - a large proportion of wealth sits at the right tail of the distribution. However, the top 1 per cent and the overall top 40 per cent are less well matched. To understand why this model works well is also to understand its limitations - both of which give important insight into the importance and mechanism of health costs in understanding inequality. The intuition as to why this occurs is from two important factors: the precautionary motive is strong, the agent is well inclined to build up savings for future medical costs shocks, but the medical costs themselves bring with them a potential inability to build up even basic life-cycle requirements for retirement saving.

The precautionary motive is evident and well established - the need to save for the potential of high medical cost shocks in the future is built into the value function. The agent is well motivated to build substantial buffer stocks for the two precautionary factors he faces: income shocks and health cost shocks. This can be seen in the Table 6 below. Artificially increasing the variance of the innovation of the medical cost shocks increases the Gini coefficient by 10 percentage points, and increase the density at the right hand tail of the distribution - wealth density in the best off in society appears to be an increasing function of the variance of medical costs shocks.

and a population growth rate, n, calibrated to 1 per cent.

Source of distribution	Gini	Share (%) 10%	5%	1%
$\sigma_\eta = 0.017$	0.8	65	40	10
$\sigma_\eta = 0.5$	0.9	73	52	17

Table 6: Sensitivity test: increased variance of cost innovations

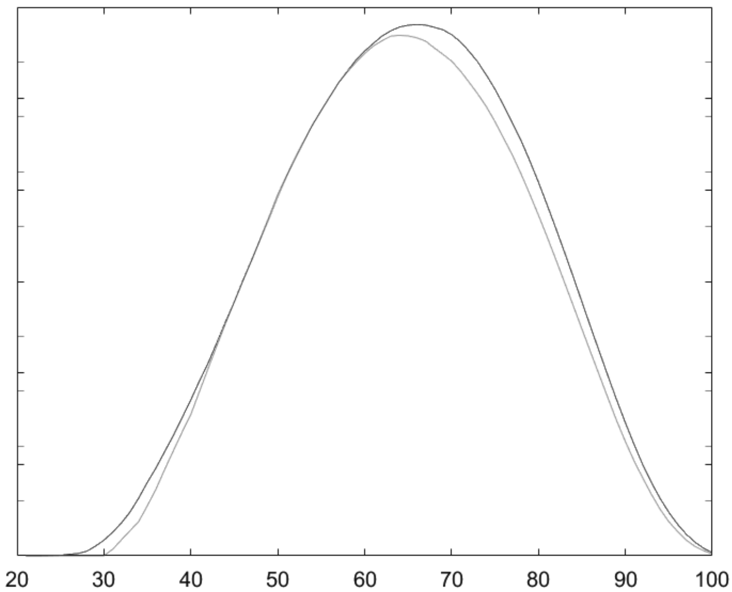

Figure 5: Savings policy functions, light gray represents benchmark model

The asset functions of the agents provide an illuminating picture of the savings implications of medical costs. Figure 5 shows cross-sectional mean assets at each age in the lifecycle for both the benchmark and the cost shock model. Both functions have been scaled to appear on the same y-axis, but peak savings in the cost shock model are around 50-60 per cent of those in the benchmark.[11] The striking feature is the clear shift to the right of the peak. The benchmark model shows Modigliani's (1980) assertion - agents will save until the point of retirement (in this case, age 65), and subsequently spend through their savings. However, the potential of large medical cost into the old age makes it vital for savings to continue into later life, and therefore savings remain until late in the life cycle. Further, savings are deaccumulated slower as the agent reaches the

[11]The difference in granularity is due to the resolution of the discretised grid sizes, as the benchmark model is simpler a more highly resolved grid can be used. However, the model is mathematically equivalent.

	21-25	26-30	31-35	36-40	41-45	46-50	80-100	Peak
21-25	1							
26-30	0.3057	1						
31-35	0.1811	0.8121	1					
36-40	0.1422	0.714	0.9498	1				
41-45	0.1197	0.6844	0.9118	0.9802	1			
46-50	0.1308	0.6821	0.8915	0.9588	0.9879	1		
80-100	0.1258	0.4982	0.6792	0.7561	0.7971	0.8265	1	
Peak	0.1211	0.6356	0.8315	0.9038	0.94	0.9642	0.9081	1

Table 7: Correlation matrix of average asset holdings, by age band

end of their life due to the potential for costs late into their lives.

However, the mean values do not portray the full picture of asset accumulation, and thus the overall wealth distribution. It is also important to consider the effects on the budget constraint of the medical cost risk throughout the life cycle. Wealth inequality is driven by a second factor: the inability to save. Faced with variable cost shocks at the beginning of the life cycle, and coupled with idiosyncratic income risk, savings may not be able to be formed at all whilst an agent is in the working portion of their lives. As Figure F.1 shows - a Lorenz curve of wealth distribution in the modelled and benchmark economies - a large proportion of agents hold no assets at all, a feature seen in the empirical distribution of wealth. The number of agents holding no wealth is markedly increased from the benchmark economy.

This corroborates the findings of Hubbard, Skinner et al. (1994) who asks why *don't* people save. Rather than attributing wealth inequality to irrational myopia (even the chosen myopia of Krusell and Smith (1998)), this model allows for a low level of wealth accumulation can be explained. Table 7 displays the correlation matrix of average asset holdings of the simulated households, averaged over age bins within their life cycles. What is shown clearly is that the amount of assets accumulated in the first few years (up to age 30) of the life cycle is reasonably to highly correlated with both maximum assets accumulated and assets available in the last few years of life. If an agent is unable to save in the early years of their life, due to the negative shocks to the budget constraint that they receive, the agent is unlikely to make up that saving in later life.[12] The amount of low asset holding households is substantial due to the action not just at the point where the shock hits the budget constraint, but the effect this on the decision making process over the whole of the rest of the life cycle.

Although the model does increase the wealth inequality at the top of the distribution, it doesn't fully fit the extreme right of the wealth distribution, and is more effective in changing the savings behaviour throughout the population. To improve the fit in the top 5% to 1% of the economy, this model is well suited to be placed with a fully featured model, alongside bequests and entrepreneurs savings motives in order to generate a tight fit to the empirical data at the very end of the right tail.

5 Conclusion

Through accounting for the profile of health care shocks across the entire life cycle, this paper has presented two important findings. Firstly, a full identification procedure for medical expenditure shocks has been established. The analysis of the cross sectional variance of shocks over the life cycle has shown that it is compelling to model health care shocks as not just a highly persistent process, but a process with a near unit root - medical cost shocks are almost permanent. This is clearly implied by the linear shape of the cross sectional variances across the life cycle, as captured by the graphical analysis. Placing these observations inside a more complete GMM minimum distance estimation system confirms that these findings are robust.

Secondly this paper developed a structural, stochastic heterogeneous agent life cycle model of health care expenditure shocks, including idiosyncratic income and medical cost risk. Accounting for and calibrating the model to the empirically observed process brings the modelled distribution of wealth much closer to the empirical reality, increasing the Gini coefficient by 20 base points compared to a baseline model. This result is driven by a clear precautionary savings motive, and also establishes a mechanism for the potential inability to save due to unexpected shifts in the budget constraint. Wealth formed in the very early stages of the life cycle is highly correlated with peak and final asset holdings.

Some potential limitations have been raised in this paper. Measurement error within

[12]This is also due to the consumption smoothing feature of the utility function.

the MEPS data set could not be insignificant. This has been managed in the late life cycle through the analysis of HRS data but, as discussed, without full information on nursing home expenditure the analysis be under estimating the key parameters. However, as discussed in French and Jones (2004) and Storesletten, Telmer et al. (2004), any other measurement error is captured in the transitory shock, which therefore could be overestimated.

The model itself could be extended in a number of interesting ways. There is clear scope for this model to be analysed within general equilibrium. The life cycle model estimated does not account for the prices of labour or assets, which could produce differing optimal savings functions. Placing medical expenditure shocks within a fully specified general equilibrium model, with overlapping generations, bequests and social security, could produce valuable results, and help quantify the effect of medical expenditure shocks and its interaction with savings at the end of the life cycle.

Further extensions could be made to the mechanism of action of medical cost shocks. Endogoneising the medical costs shocks is a natural step - this could be performed through constructing a state of health status, from which health costs are derived. Alternatively, health costs could impact consumption through the utility function, not just the budget constraint. This would reflect the fact that medical expenditure variability is to some extent driven by consumer choice - low wealth agents could perhaps choose a lower quality or quantity of health care.

This paper sets itself within the vast canon of literature on wealth inequality. Idiosyncratic medical expenditure shocks, and the mechanism through which they act, appear to play a non-trivial role in the development of inequality in the United States and could provide another rich avenue of research into the thick right tail of the empirically observed distribution.

References

Agency for Healthcare Research and Quality (2015). *Medical Expenditure Panel Survey Documentation*.

Aiyagari, S. R. (1994). *Uninsured Idiosyncratic Risk and Aggregate Saving*. *The Quarterly Journal of Economics*, 109(3), 659–684.

Altonji, J. G. and Segal, L. M. (1996). *Small-sample bias in GMM estimation of covariance structures*. *Journal of Business & Economic Statistics*, 14(3), 353–366.

Blundell, R., Pistaferri, L. and Preston, I. (2008). *Consumption inequality and partial insurance*. *The American Economic Review*, 98(5), 1887–1921.

Cagetti, M. and De Nardi, M. (2006). *Entrepreneurship, Frictions, and Wealth*. *Journal of political Economy*, 114(5), 835–870.

Cagetti, M. and De Nardi, M. (2008). *Wealth Inequality: Data and Models*. *Macroeconomic Dynamics*, 12(S2).

Castañeda, A., Díaz-Giménez, J. and Ríos-Rull, J.-V. (2003). *Accounting for the U.S. Earnings and Wealth Inequality*. *Journal of Political Economy*, 111(4), 818–857.

Chamberlain, G. (1984). *Panel Data*. *Handbook of Econometrics*, 2, 1247–1318.

Danziger, S., Van Der Gaag, J., Smolensky, E. and Taussig, M. K. (1982). *The Life-Cycle Hypothesis and the Consumption Behavior of the Elderly*. *Journal of Post Keynesian Economics*, 5(2), 208–227.

De Nardi, M. (2004). *Wealth Inequality and Intergenerational Links*. *The Review of Economic Studies*, 71(3), 743–768.

De Nardi, M. (2015). *Quantitative Models of Wealth Inequality: A Survey*. *National Bureau of Economic Research*.

De Nardi, M., Fella, G. and Pardo, G. P. (2016). *The Implications of Richer Earnings Dynamics for Consumption and Wealth*.

De Nardi, M., French, E. and Jones, J. (2010). *Why Do the Elderly Save? The Role of Medical Expenses*. *Journal of Political Economy*, 118(1), 39–75.

De Nardi, M., French, E., Jones, J. B. and McCauley, J. (2016). *Medical Spending of the US Elderly*. *Fiscal Studies*, 37(3-4), 717–747.

Deaton, A. and Paxson, C. (1994). *Intertemporal Choice and Inequality*. *Journal of Political Economy*, 102(3), 437–467.

Dunn, A., Grosse, S. D. and Zuvekas, S. H. (2016). *Adjusting Health Expenditures for Inflation: A Review of Measures for Health Services Research in the United States*. *Health Services Research*, pp. n/a–n/a.

Federal Reserve (2001). *Survey of Consumer Finances*.

Feenberg, D. and Skinner, J. (1992). *The Risk and Duration of Catastrophic Health Care Expenditures*. *National Bureau of Economic Research*.

French, E. and Jones, J. (2004). *On the distribution and dynamics of health care costs*. *Journal of Applied Econometrics*, 19(6), 705–721.

Friedman, B. M. and Warshawsky, M. J. (1990). *The Cost of Annuities: Implications for Saving Behavior and Bequests*. *The Quarterly Journal of Economics*, 105(1), 135–154.

Gale, W. G. and Scholz, J. K. (1994). *Intergenerational Transfers and the Accumulation of Wealth*. *The Journal of Economic Perspectives*, 8(4), 145–160.

Hubbard, R. G. (1987). *Uncertain lifetimes, pensions, and individual saving*. In *Issues in Pension Economics*, pp. 175–210. University of Chicago Press.

Hubbard, R. G., Skinner, J. and Zeldes, S. P. (1994). *Expanding the Life-Cycle Model: Precautionary Saving and Public Policy*. *The American Economic Review*, 84(2), 174–179.

Hubbard, R. G., Skinner, J. and Zeldes, S. P. (1995). *Precautionary saving and social insurance*. *Journal of Political Economy*, 103(2), 360–399.

Huggett, M. (1993). *The risk-free rate in heterogeneous-agent incomplete-insurance economies*. *Journal of Economic Dynamics and Control*, 17(5), 953–969.

Hurd, M. D. (1989). *Mortality risk and bequests*. *Econometrica: Journal of the Econometric Society*, pp. 779–813.

Hurd, M. D., Rohwedder, S. et al. (2009). *The level and risk of out-of-pocket health care spending*. *Michigan Retirement Research Center Working Paper*, 218.

İmrohoroğlu, A., İmrohoroğlu, S. and Joines, D. H. (2001). *Computing models of social security*. In *Computational Methods for the Study of Dynamic Economies*, edited by Marimon, R. and Scott, A. Oxford University Press.

Kimball, M. S. (1990). *Precautionary saving in the small and in the large*. *Econometrica: Journal of the Econometric Society*, pp. 53–73.

Kopecky, K. A. and Suen, R. M. (2010). *Finite State Markov-Chain Approximations to Highly Persistent Processes*. *Review of Economic Dynamics*, 13(3), 701–714.

Kotlikoff, L. J. (1986). *Health expenditures and precautionary savings*. *National Bureau of Economic Research*.

Kotlikoff, L. J. (1988). *Intergenerational transfers and savings*. *Journal of Economic Perspectives*, 2(2), 41–58.

Krusell, P. and Smith, A. (1998). *Income and Wealth Heterogeneity in the Macroeconomy*. *Journal of Political Economy*, 106(5), 867–896.

Krusell, P. and Smith, A. (2015). *General-equilibrium models of wealth inequality based on uninsurable idiosyncratic risk: Bewley-Huggett-Aiyagari Models*. Central Bank of Chile. Available from http://aida.wss.yale.edu/smith/chile/wealthinequalityeabcn2.pdf.

Manning, W. G. and Mullahy, J. (2001). *Estimating log models: to transform or not to transform?* *Journal of health economics*, 20(4), 461–494.

Mehra, R. and Prescott, E. C. (1985). *The Equity Premium: A Puzzle*. *Journal of Monetary Economics*, 15(2), 145–161.

Modigliani, F. (1980). *Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data*. In *The Collected Papers of Franco Modigliani*, volume 6. MIT Press.

National Center for Health Statistics (2004). *United States Life Tables 2002*, volume 53. US Department of Health, Education, and Welfare, Public Health Service, Health Resources Administration, National Center for Health Statistics.

Ozkan, S. (2011). *Income Inequality and Health Care Expenditures Over the Life Cycle*. *Manuscript, Federal Reserve Board.*

Palumbo, M. G. (1999). *Uncertain medical expenses and precautionary saving near the end of the life cycle*. *The Review of Economic Studies*, 66(2), 395–421.

Piketty, T. (2014). *Capital in the Twenty-First Century*. Cambridge Massachusetts: Harvard University Press.

Quadrini, V. (1999). *The importance of entrepreneurship for wealth concentration and mobility*. *Review of Income and Wealth*, 45(1), 1–19.

Quadrini, V. and Ríos-Rull, J.-V. (1997). *Understanding the U.S. Distribution of Wealth*. *Federal Reserve Bank of Minneapolis. Quarterly Review - Federal Reserve Bank of Minneapolis*, 21(2), 22.

RAND Center for the Study of Aging (2016). *RAND HRS Data Version P. (2002-2014)*. Produced by the RAND Center for the Study of Aging, with funding from the National Institute on Aging and the Social Security Administration).

Sonnega, A., Faul, J. D., Ofstedal, M. B., Langa, K. M., Phillips, J. W. and Weir, D. R. (2014). *Cohort Profile: the Health and Retirement Study (HRS)*. *International Journal of Epidemiology*, 43(2), 576–585.

Sopchokchai, D. (2015). *For Better or For Worse: Income Processes and Intra-household Risk Sharing*. Monograph.

Storesletten, K., Telmer, C. and Yaron, A. (2004). *Consumption and Risk Sharing over the Life Cycle*. *Journal of Monetary Economics*, 51(3), 609–633.

Tauchen, G. (1986). *Finite State Markov-Chain Approximations to Univariate and Vector Autoregressions*. *Economics Letters*, 20(2), 177–181.

University of Michigan (2016). *Health and Retirement Study 2002-2014*. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740).

A Wealth inequality within heterogeneous agent models

Source of distribution	Gini	Share of total wealth (%) Top 40%	20%	10%	5%	1%
Empirical (SCF, 2001)	0.79	94	82	69	57	32
Aiyagari (1994) baseline (σ = 0.2)	0.38	/	41	9.9	/	3.2
Increased variance (σ = 0.4)	0.41	/	45	12	/	4
Krusell and Smith (1998) baseline	0.26	/	35	20	11	3
With preference hetrogeneity	0.66	/	74	61	46	20
Cagetti and De Nardi (2006) baseline	0.6	95	58	/	20	4
With entrepeneurship	0.8	94	83	/	60	31
De Nardi (2004) baseline	0.67	90	69	/	27	7
With voluntary bequests	0.74	95	76	/	37	14

/ denotes unreported wealth percentiles

Table A.1: Wealth inequality within heterogeneous agent models
Data sourced from: Federal Reserve (2001), Quadrini and Ríos-Rull (1997), Krusell and Smith (2015), Cagetti and De Nardi (2006), De Nardi (2015)

	s' = 1	s' = 2	s' = 3	s' = 4
s = 1	96.24	1.14	0.39	0.006
s = 2	3.07	94.33	0.37	0
s = 3	1.5	0.43	95.82	0.02
s = 4	10.66	0.49	6.11	80.51

Table A.2: Markov transition matrix between endowment states
(Castañeda, Díaz-Giménez et al., 2003)

B Summary statistics of MEPS dataset

Age (N)	Out of pocket expenditure	Total expenditure	Percent male	Total income	Public insurance	Private insurance	Uninsured
21-25	349.661	1900.366	0.509342	16459.53	0.02141	0.417406	0.537191
(28,665)	(10.948)	(79.251)	(0.005)	(201.83)	(0.002)	(0.007)	(0.007)
26-30	429.2248	2349.715	0.489836	28832.44	0.034351	0.354251	0.512217
(29,202)	(11.9)	(66.561)	(0.004)	(352.01)	(0.002)	(0.007)	(0.007)
31-35	492.7714	2858.455	0.494012	34741.66	0.042004	0.334838	0.481313
(29,288)	(14.095)	(77.307)	(0.004)	(410.11)	(0.002)	(0.006)	(0.006)
36-40	510.505	2912.28	0.491974	37876.46	0.044413	0.361752	0.460711
(29,271)	(10.65)	(64.685)	(0.004)	(426.55)	(0.002)	(0.006)	(0.005)
41-45	619.6736	3500.145	0.490391	38577.58	0.045748	0.387177	0.442616
(29,255)	(13.856)	(122.818)	(0.004)	(463.91)	(0.002)	(0.006)	(0.006)
46-50	755.1034	4217.53	0.485901	38808.75	0.045236	0.396811	0.436141
(29,354)	(14.645)	(91.765)	(0.004)	(439.96)	(0.002)	(0.006)	(0.005)
51-55	913.0449	5411.704	0.493058	39251.62	0.041353	0.384373	0.444812
(27,386)	(15.543)	(122.667)	(0.004)	(480.46)	(0.002)	(0.006)	(0.005)
56-60	1102.464	6662.364	0.482051	36995.33	0.048673	0.371368	0.42902
(23,400)	(21.098)	(193.433)	(0.005)	(484.24)	(0.003)	(0.006)	(0.006)
61-65	1305.167	7981.847	0.478818	29461.41	0.059298	0.375788	0.40518
(18,615)	(25.443)	(192.713)	(0.005)	(503.31)	(0.003)	(0.007)	(0.007)
66-70	1275.688	8468.134	0.469364	19172.19	0.072924	0.422588	0.36064
(14,399)	(28.538)	(209.088)	(0.005)	(510.12)	(0.004)	(0.008)	(0.007)
71-75	1389.236	9613.77	0.446417	12930.86	0.096062	0.45712	0.318736
(11,139)	(29.465)	(228.086)	(0.006)	(449.87)	(0.004)	(0.009)	(0.008)
76-80	1525.935	10208.1	0.420467	10313.37	0.116393	0.473638	0.303549
(8,690)	(39.429)	(231.308)	(0.007)	(475.23)	(0.005)	(0.009)	(0.009)
81-85	1555.25	10658.37	0.394346	8510.86	0.135455	0.484296	0.285417
(5,091)	(52.396)	(299.383)	(0.009)	(363.27)	(0.008)	(0.012)	(0.012)

N and standard error in parentheses

Table B.1: MEPS summary statistics

C MEPS summary figures

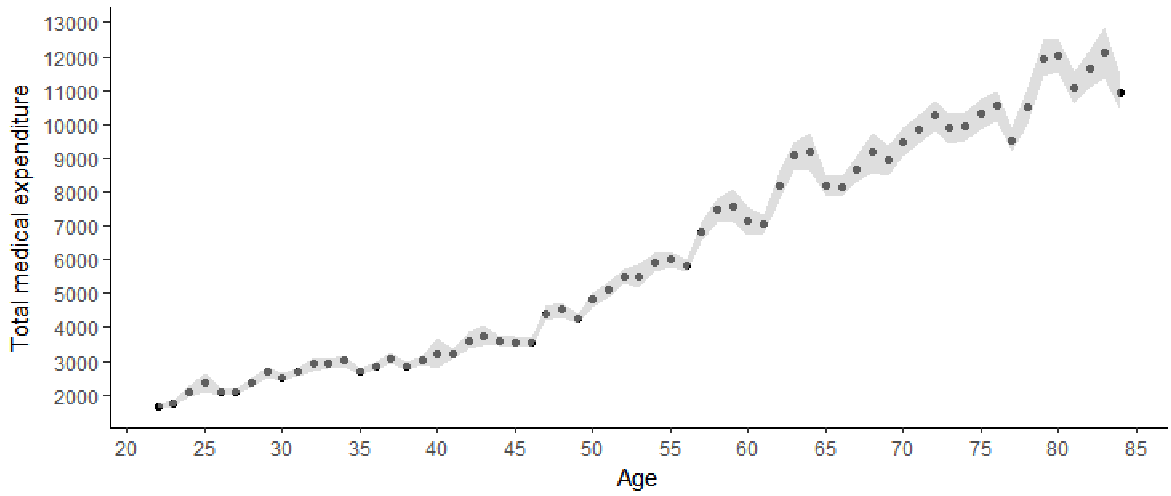

Figure C.1: Mean total medical expenditure, by age

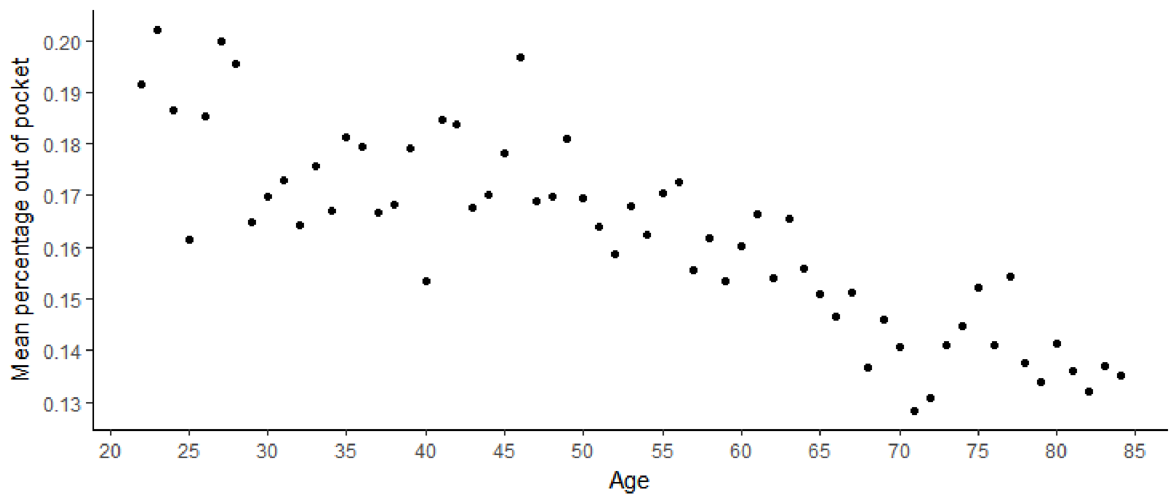

Figure C.2: Percentage out of pocket medical expenditure, by age

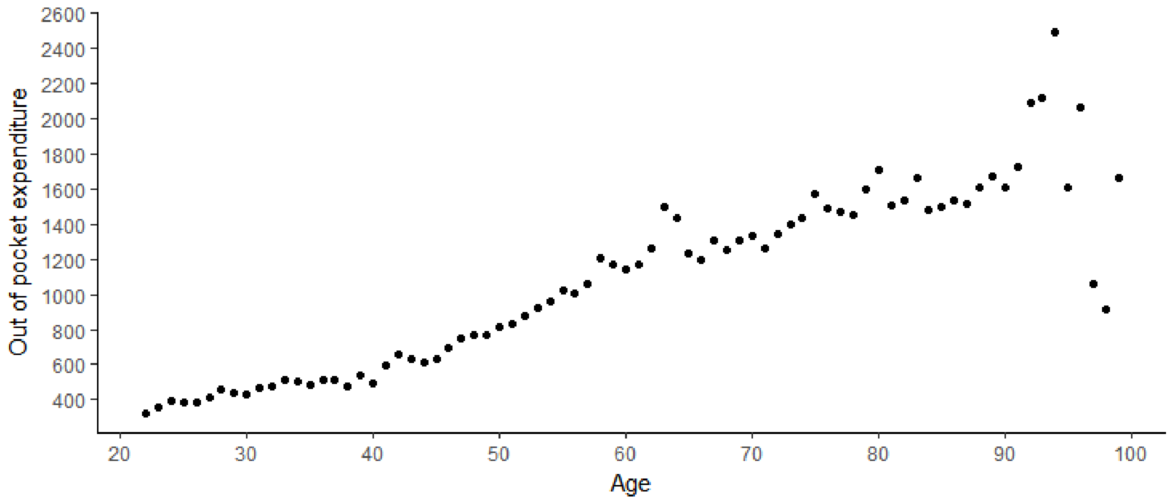

Figure C.3: Out of pocket expenditure, including HRS, by age

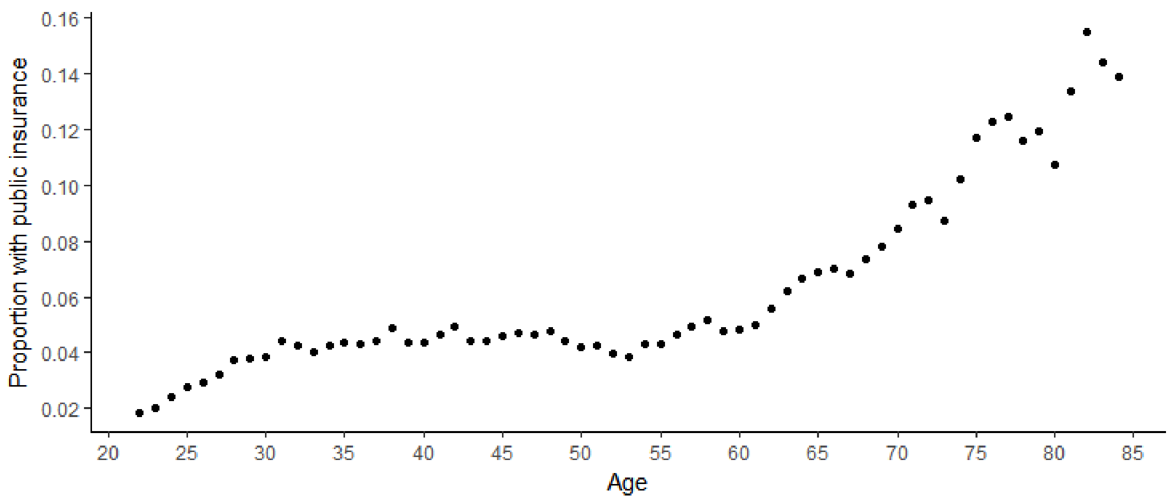

Figure C.4: Proportion with public medical insurance, by age

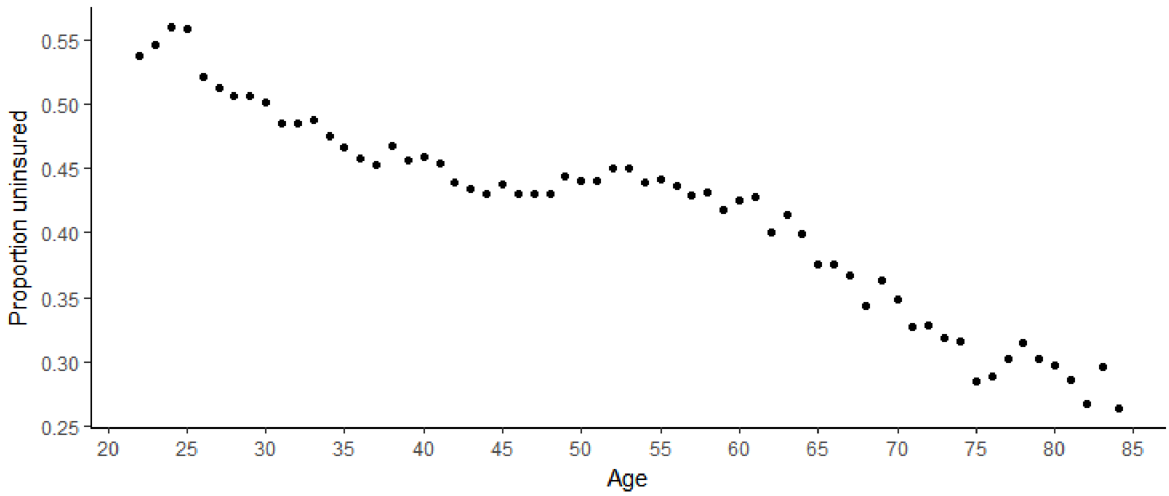

Figure C.5: Proportion with no medical insurance, by age

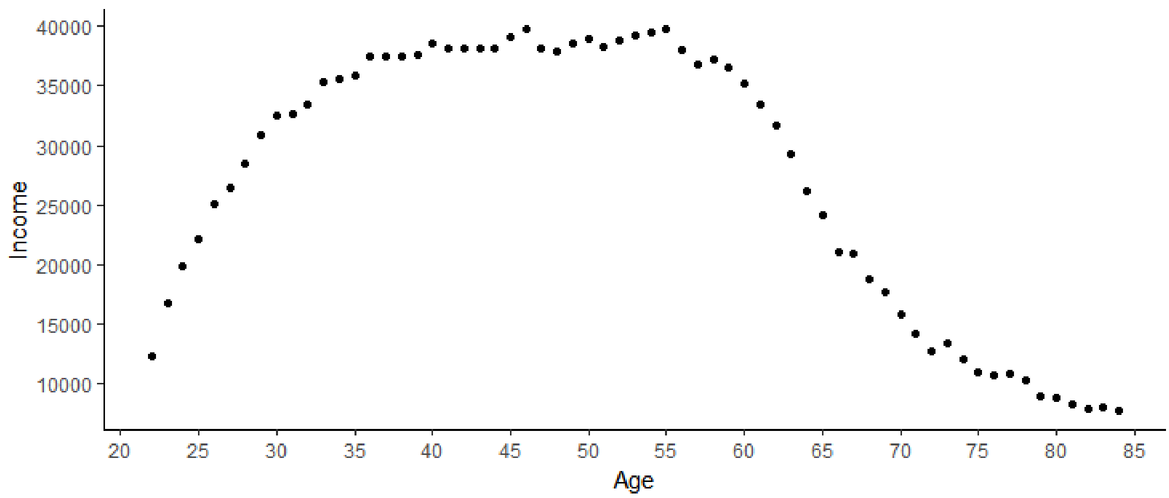

Figure C.6: Earnings and pension income, by age

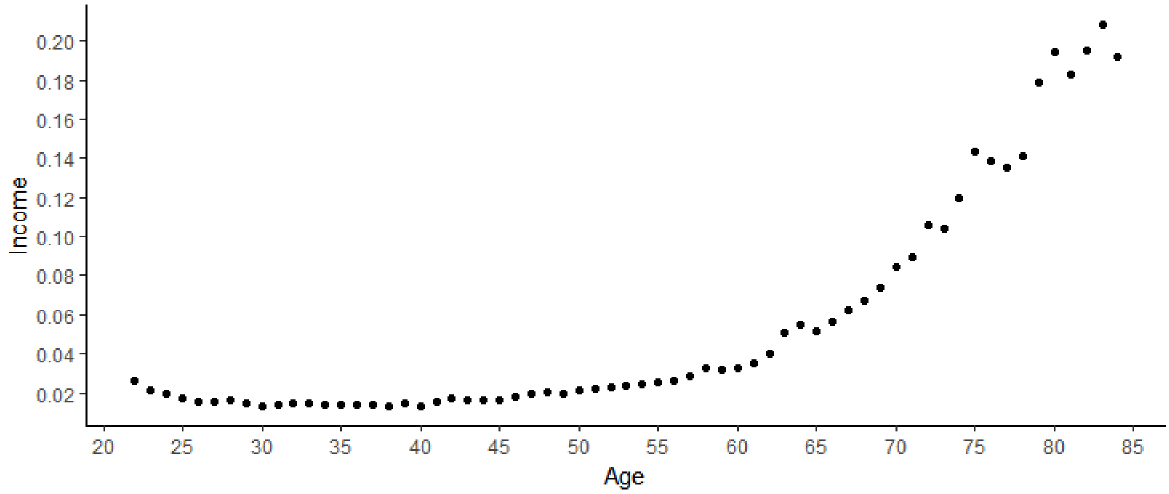

Figure C.7: Proportion of medical expenditure to total combined earnings and pension income

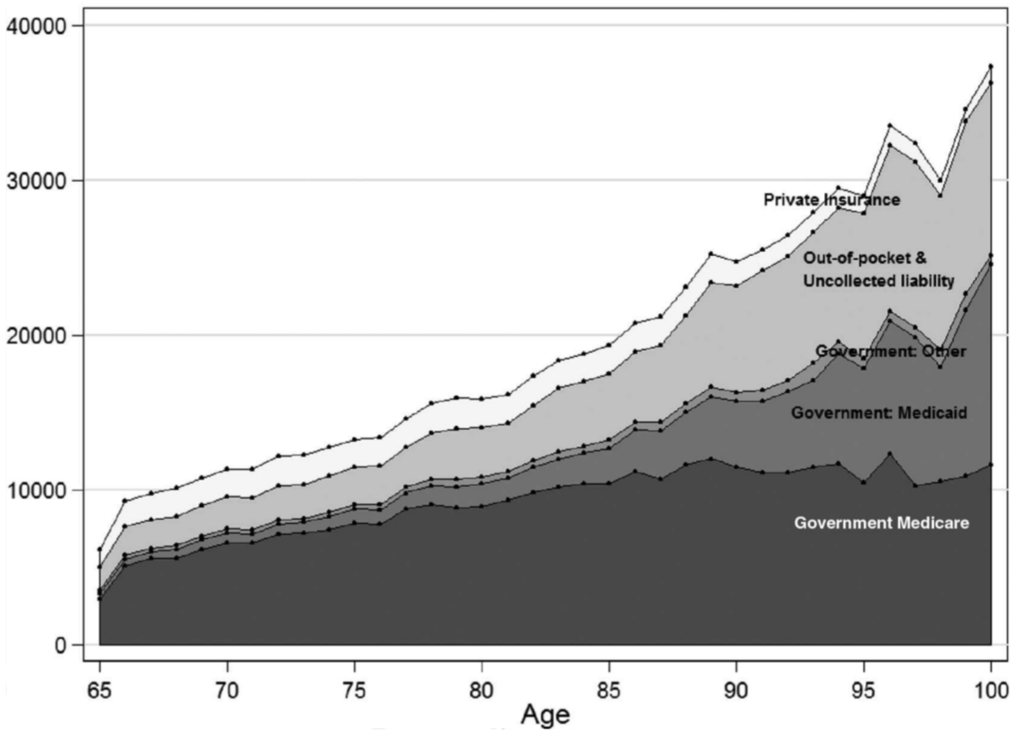

Figure C.8: Proportion of medical expenditure, by payer type (De Nardi, French et al., 2016)

D Identification Stage IIa: moment condition proof

Proof. $Var(u_{ih}) = \sigma_\alpha^2 + \sigma_\eta^2 \sum_{j=0}^{h-1} \rho^{2j} + \sigma_\epsilon^2$

$$
\begin{aligned}
u_{ih} &= \alpha_i + z_{ih} + \epsilon_{ih} \\
Var(u_{ih}) &= Var(\alpha_i) + Var(z_{ih}) + Var(\epsilon_{ih}) \\
&= \sigma_\alpha^2 + \sigma_\epsilon^2 + Var(\rho z_{i,h-1} + \eta_{ih}) && \text{(from i.i.d assumptions)} \\
&= \sigma_\alpha^2 + \sigma_\epsilon^2 + \sigma_\eta^2 + \rho^2\, Var(z_{i,h-1}) \\
&= \sigma_\alpha^2 + \sigma_\epsilon^2 + \sigma_\eta^2 + \rho^2\, Var(\rho z_{i,h-2} + \eta_{i,h-2}) \\
&= \sigma_\alpha^2 + \sigma_\epsilon^2 + \sigma_\eta^2 + \rho^2 \left[\sigma_\eta^2 + \rho^2 Var(z_{i,h-2})\right] \\
&= \sigma_\alpha^2 + \sigma_\epsilon^2 + \sigma_\eta^2 + \rho^2\sigma_\eta^2 + \rho^4\, Var(z_{i,h-2}) \\
&= \ldots && \text{(repeat to h=0)} \\
&= \sigma_\alpha^2 + \sigma_\eta^2 \sum_{j=0}^{h-1} \rho^{2j} + \sigma_\epsilon^2
\end{aligned}
$$

□

E Identification Stage IIa: cross sectional variances

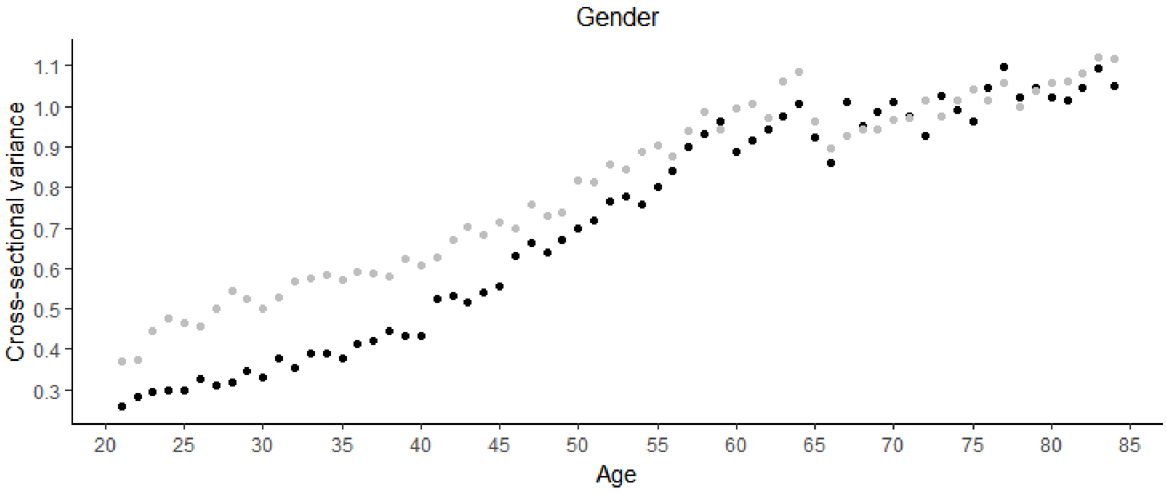

Figure E.1: Cross sectional variance by gender, male (black), female (grey)

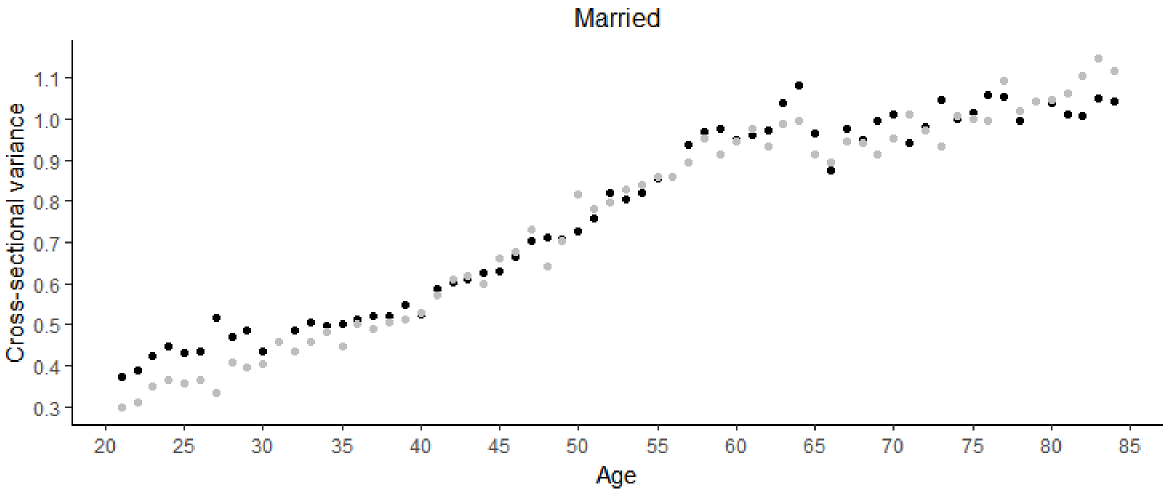

Figure E.2: Cross sectional variance by marital status, married (black), unmarried (grey)

F Life cycle model: Lorenz curve

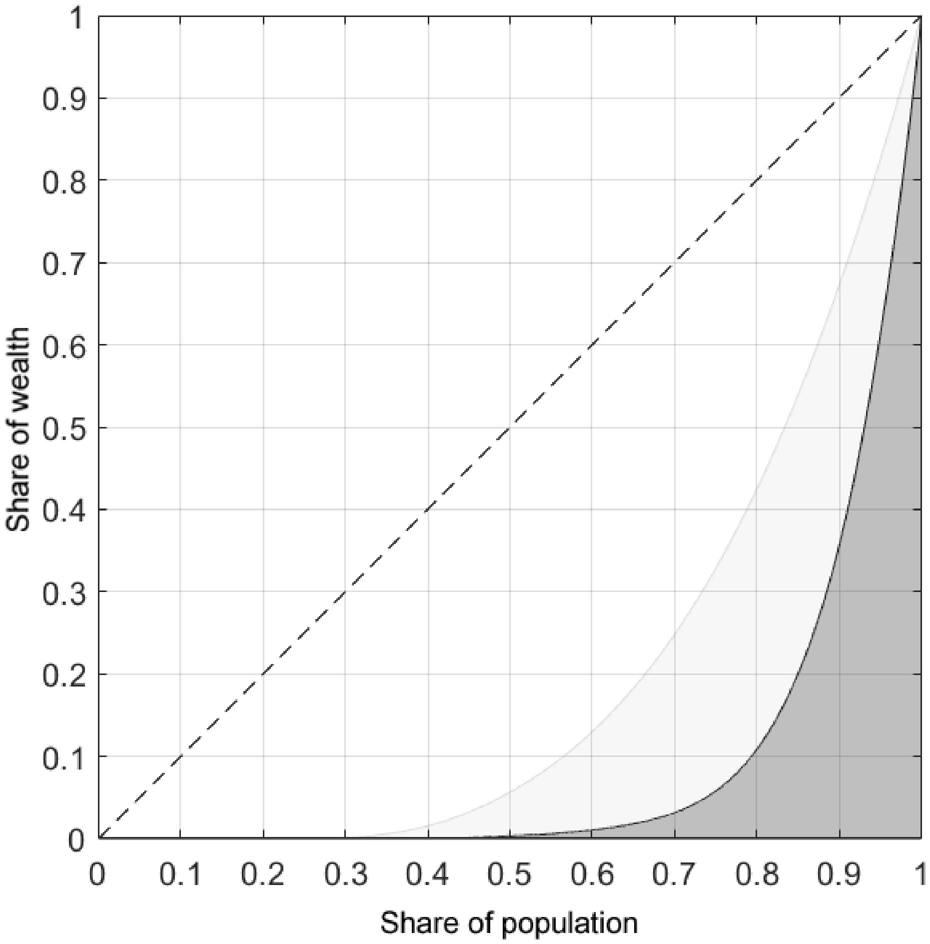

Figure F.1: Lorenz curves of the model and benchmark economies