Health status over the life cycle*

Raun van Ooijen [a,c] Rob Alessie [a,c] and Marike Knoef [b,c]

[a] *University of Groningen, the Netherlands*
[b] *Leiden University, the Netherlands*
[c] *Netspar, the Netherlands*

October 29, 2015

Keywords: self-reported health; administrative data, health dynamics; health index; socio-economic status

Abstract

We construct a health measurement model which combines panel data on self-reported health with a rich set of health measures from administrative medical records. Our estimated health model allows us to predict health status for the population at large. We account both for unobserved heterogeneity and for the persistence in unobserved health shocks. To account for inconsistent reporting in self-reported health we propose a model using two measures of self-reported health: the level of self reported health and the self-reported health change. We show that using both measures substantially increases the estimated persistence in health status. We use predicted health status to study the evolution of health as individuals age. Moreover, we analyze how health interacts with economic variables and education. We find a strong gradient in education; the age at which health starts to decline at a greater rate differs by education and gender.

Keywords: self-reported health; administrative data, health dynamics; health index; socio-economic status
JEL classification: C42, I12

*Raun van Ooijen (corresponding author). Adress: Department of Economics, Econometrics, and Finance, P.O. Box 800, 9700 AV Groningen, The Netherlands. Tel.: +31 50 363 8534; fax: +31 50 363 7337; email: R.van.Ooijen@rug.nl. Rob Alessie. Adress: Department of Economics, Econometrics, and Finance, P.O. Box 800, 9700 AV Groningen, The Netherlands. Marike Knoef. Adress: Department of Economics, 2311 ES Leiden, The Netherlands. We are grateful to Richard Blundell, Mariacristina De Nardi, Eric French, Mathilde Godard, Adriaan Kalwij, Arie Kapteyn, Silvia Garcia Mandico, Tom Van Ourti, Paul Rodríguez-Lesmes, and participants at the IFS seminar, the Tilburg Labor and Health seminar, the Netspar pension workshop and the European workshop on econometrics and health economics in Paris, for useful comments and suggestions.

1 Introduction

Health is a very important asset: if you are healthy you are able to work or produce goods at home; health also plays an important role in financial planning and in consumption and saving decisions. Most people probably have a good sense what the word 'health' means; it is, however, difficult to capture health in a single measure. To measure 'true' health we preferably like to combine different sources of information, or dimensions of health, which could explain a person's latent health status. Many studies in this field use self-reported health status (SRH) as a summary measure of health. The SRH variable is typically based on a question in which people are asked to rate their own health on a five-point ordinal scale ranging from 'excellent health' to 'poor health'.

The SRH measure has proven to be a very useful measure of health but is sometimes criticized because it involves some biases. For example Crossley and Kennedy (2002) show that many persons tend to change their self-reported health status within the same survey. Inconsistent reporting over time makes it difficult to analyze the evolution of health of a person over time as it reduces the persistence in health. Psychological factors, such as a person's mood might also influence the reported health status between waves. In addition to reporting bias, two persons with the same underlying health problems might report a different SRH, due to a difference in health perception; e.g. Lindeboom and van Doorslaer (2004). Differences in health perception among age groups result in a biased life cycle health profile. In addition to measurement error and heterogeneity in health perception, SRH seems also sensitive to justification bias as was first mentioned by Bound et al. (1999); people outside of the labor market tend to justify their labor market status by reporting worse health in a survey.

Another disadvantage, at least from our perspective, is that SRH is only available for a relatively small sample of the population. As administrative data from medical registers become more widely available, there is a growing interest to use these data. Important benefits of large administrative data are the possibility to focus on very specific groups in the population and the absence of survey attrition, which is often related to a deterioration in health (Jones et al., 2006).

In this paper, we estimate a health measurement model where we link survey data on SRH to a rich set of objective health conditions from medical records. Relying on objective health measures, instead of self-reported health conditions in survey data, has several advantages.

Objective health conditions do not suffer from justification bias and are less prone to reporting bias (Baker et al., 2004). Even though the questions to measure health outcomes in the Health and Retirement Study (HRS) and other similar surveys such as SHARE, PSID and ELSA are very specific there is ample evidence that self-reported health outcomes in survey data are sensitive to reporting bias. Johnston et al. (2009) show that the large majority of individuals in the UK who are diagnosed with hypertension, which is a very common disease, do not report this in a health questionnaire, though the question is very clear.[1] Also measures like ADL, which are often used as 'objective' measures in health measurement models, may be prone to reporting bias. Shulman et al. (2006), for example, find discrepancies between patients subjective reporting of ADL and IADL and their objective ratings.

On the other hand, SRH might contain information on 'true' health not being captured by the objective health conditions. It is important to take these unobserved individual differences in health into account, otherwise the persistence in health status might be underestimated. We account for unobserved individual differences in health as well as the persistence in unobserved health shocks by exploiting the panel dimension of our data. One problem is that inconsistent reporting over time in SRH (i.e. measurement error) will reduce the estimated persistence in unobserved health. The Longitudinal Internet Study in the Social Sciences (LISS) panel, administered by CentERdata in the Netherlands, allows us to examine the existence of possible inconsistent reporting patterns. In addition to the standard SRH question, respondents are asked to report the change in their health compared to last year. For some respondents we notice inconsistent reporting over time by comparing the SRH measure and the self-reported change in health measure. For instance, some respondents state that their health did not deteriorate over the last year while they report being in worse health compared to last year, or vice versa. These inconsistencies are also found by Benitez-Silva and Ni (2008) and Erdogan-Ciftci et al. (2010) in the Health and Retirement Study (HRS) and Dutch GLOBE data, respectively. They cannot be explained by phenomena such as learning about health status over time, a change in social perception about certain health conditions, or medical innovations, since this would probably result in an up- or downward trend in SRH. We use both measures of SRH in the health measurement model to account for these inconsistencies. We show that this substantially increases the estimated persistence in health.

[1] The wording of such a question in a survey is typically as follows: "Has a doctor ever told you that you have high blood pressure or hypertension?"

Several methods have been put forward to construct a health index using survey data on SRH and on a vector of variables $\boldsymbol{x}$, which measure objective health conditions. A useful overview of the different approaches proposed in the literature is provided by Kapteyn and Meijer (2013), Lindeboom and van Doorslaer (2004) and Cutler et al. (1997). Our method is closely related to the method proposed by Jürges (2007) (who uses a similar approach that were used in Bound, 1991) which boils down to estimating an extended ordered probit model with SRH as dependent variable and objective health measures as explanatory variables. We will extend the approach of Jürges in three directions. First, we exploit the fact that we have panel data on SRH at our disposal. This allows us to take into account unobserved heterogeneity and the persistence in unobserved health shocks. Second, we are able to enrich the LISS survey data with a large set of health indicators stemming from administrative sources. This set of health indicators is very similar to the variables used in the latent health index developed by Poterba et al. (2010) who use measures collected in the HRS. Their latent health index is widely used in recent research, and well able to explain saving and retirement behavior as shown by Kapteyn and Meijer (2013). Related studies to our work by Lange and McKee (2012) and Heiss et al. (2014) also emphasize the importance of using multiple 'objective' measures of health to construct a single index and to account for unobserved heterogeneity in health. Third, we account for inconsistent reporting in SRH over time which, not taking into account, reduces the estimated persistence in health. A good understanding of the persistence in health status is crucial for explaining saving behavior and designing health and long-term care insurance, among other things; see e.g. De Nardi et al. (2010).

The estimated health model allows us to predict health status for the population at large. We account for the stochastic properties of unobserved differences in health. The advantage of using administrative data is that we can focus on specific subgroups that are usually small in surveys–such as the oldest-old—and that we overcome health related attrition in survey data. We use the predicted objective health status for the large administrative data to study the evolution of health as individuals age. Moreover, we analyze how health interacts with socio-economic status (SES). The objective health measure allows us to validate studies on the relationship between SES and health who use subjective health as a dependent variable.[2]

Many studies, such as Case and Deaton (2005) for the United States show, that women report

[2]Another advantage of our health index model is that we can easily measure inequality in health, which is less easy with SRH (as discussed in Van Doorslaer and Jones, 2003).

a lower SRH than men and that the health status of men deteriorates at a faster rate than women. A recent study by Ross et al. (2012) for the United States reports that the relationship between education and health is stronger for women than for men. One explanation for this gender gap is a difference in health perception rather than a difference in the prevalence of chronic disease between men and women. Another explanation could be that women are more inclined to mention health problems then men in a survey. Our health measurement model deals with both measurement problems. According to Van Kippersluis et al. (2010) health declines at a faster rate for low educated than for high educated individual. Contoyannis et al. (2004) analyze the dynamics of health status among British men using SRH, they do not find clear differences in the health persistence by education and income.

Our main results are as follows: First, using either the level of SRH or both the SRH level and change shows significant differences in the estimated persistence in health. For both measures of SRH, objective medical conditions, such as having diabetes, affect SRH in a similar way; using the SRH change as a complement to the level of SRH, however, substantially increases the persistence in health. Second, we find that people of low socio-economic status are more likely to stay in poor health. The age at which health starts to decline at a greater rate arrives earlier for males and persons with a lower level of education. Finally, we show that women on average are in worse health than men due to a higher prevalence of chronic diseases which have a relative detrimental effect on health. Women's health seems to benefit more from having higher education than men. We also provide evidence that income and wealth are protective of health over and above education. These results on the SES and health gradient seem to be in line with the above-mentioned studies using SRH.

The outline of the paper is as follows. Section 2 explains the health measurement model. Section 3 extensively discusses the survey data set and administrative data. Section 4 discusses the estimation results of the health measurement model and the ability to explain the empirical patterns in for example SRH. Section 5 presents descriptive statistics on the persistence in health and the evolution of health over the life cycle. The final section concludes.

2 A longitudinal health measurement model

2.1 Different approaches to model health

Several methods have been put forward to construct a health index using survey data on SRH and on a vector of variables $\boldsymbol{x}$, which measure health conditions or difficulties with activities of daily living.

Our method is closely related to the method proposed by Jürges (2007) who estimates an extended ordered probit model with SRH as dependent variable and objective health measures $\boldsymbol{x}$ as explanatory variables. The extended ordered probit model assumes the existence of a single latent health index y^* which is equal to $\boldsymbol{x}'\boldsymbol{\beta}$. The value of the health index is predicted as follows: $\hat{y}^* = \boldsymbol{x}'\hat{\boldsymbol{\beta}}$. Notice that self-reported health is only used in the construction of the index but not in the prediction of the index. Poterba et al. (2010) use a somewhat different approach: in their model latent health status not only directly influences self-reported health but also all other health measurements. This results in a factor analysis model from which they derive the first principal component as their health index. In the empirical analysis Jürges uses the first wave of the SHARE survey, which includes information of 22,000 individuals aged 50 and above from ten European countries. Like many others the author stresses that the SRH measure is not comparable across countries because of differences in reporting style. For this reason he does not use the standard ordered probit model in his analysis but an extended version of it which allows the threshold parameters to be different across countries, which is known as a 'cut-point shift'. Lindeboom and van Doorslaer (2004) point out that heterogeneity in reporting behavior could not only lead to a cut-point shift but also to a so-called 'index shift' in the β parameters of the latent health index. They propose some likelihood ratio tests to check whether cut-point shifting and index shifting are relevant phenomena. Jürges deliberately does not allow for a country-specific $\boldsymbol{\beta}$ vector so that he does not need to choose a 'reference country' in the cross-country comparison of general health.[3]

Lindeboom and van Doorslaer (2004) show that reporting not only varies across countries. They present evidence for both 'cut-point shifts' and 'index shifts' across age groups and gender: females and older persons are more likely to understate their health status compared to males

[3]To be more precise, Jürges divides the estimated coefficients $\hat{\boldsymbol{\beta}}$ by the difference between the highest and lowest predicted health level to come up with a predicted health index which takes a value between zero and one. These scaled coefficients are referred to as 'implicit disability weights'. Poterba et al. (2010) ranks the predicted health status in percentile scores.

and younger individuals.[4] A possible explanation is that persons compare their health status relative to another person of the same age and gender—in some surveys respondents are actually asked to report their health status relative to another person of the same age. This implies that there is a flattened out age profile in SRH. Since we are interested in modelling 'true' health status over the life cycle it is important to take this difference in reporting style into consideration. Lindeboom and van Doorslaer (2004) find no clear evidence that reporting differs by socio-economic status; which is also shown by McFadden et al. (2009).

2.2 Constructing a health index

In this section we introduce our longitudinal ordered response model in more formal terms. This model should enable us to predict a single latent true health index for individual i in period t by means of a set of health indicators $\boldsymbol{x}_{it}$ from administrative sources. The ordered response model assumes a linear relationship between a latent health index y_{it}^* and $\boldsymbol{x}_{it}$:

$$y_{it}^* = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},\ t = 1, \ldots, T, \tag{1}$$

where ε_{it} represent unobserved factors influencing SRH which are not captured by the explanatory variables, such as lifestyle. We assume that this error term is standard normal distributed conditional upon $\boldsymbol{x}_{it}$: $\varepsilon_{it}|\boldsymbol{x}_{it} \sim N(0,1)$. Notice that the $\boldsymbol{\beta}$ parameter vector may vary across demographic groups. However, we do not allow for index shifting to keep the model parsimonious.

Since we are interested in the evolution of health over the life cycle, it is important to model the persistence in the random effect ε_{it}. Persistence in health can be the result of (1) persistence in observed medical conditions, (2) serial correlation in the error term—for example the experience of recurring health problems after the diagnosis of a chronic diseases such as diabetes—or (3) because of unobserved heterogeneity, for example if unhealthy lifestyle increases the probability of experiencing health problems and this is not captured by the observed variables. In light of these considerations, we assume that the error term ε_{it} can be decomposed into a random individual effect c_i and an idiosyncratic error term u_{it} which represent unobserved

[4]The authors use cross-sectional data from Canada. It is well-known that in a cross-section study we cannot distinguish age effects from cohort effects. In other words the age effect could also be interpreted as a generation effect. In our study we use panel data to allow the threshold parameters to be cohort specific and not age specific.

health shocks:

$$\begin{aligned} \varepsilon_{it} &= c_i + u_{it} \\ c_i &\sim NID(0, \sigma_c^2) \\ u_{it} &\sim N(0, \sigma_u^2) \\ \text{cov}(c_i, u_{it}) &= 0,\ t = 1, \ldots, T. \end{aligned}$$

Given these assumptions, $\sigma_u^2 = 1 - \sigma_c^2$ because $\text{var}(\varepsilon_{it}) = 1$. As we said above, the unobserved health shocks u_{it} are likely to be rather persistent. We therefore model u_{it} by means of an AR(1) process:

$$\begin{aligned} u_{it} &= \gamma u_{it-1} + \zeta_{it} \\ \zeta_{it} &\sim NID(0, \sigma_\zeta^2). \end{aligned}$$

Since $\text{var}(\varepsilon_{it}) = 1$, it holds that $\sigma_\zeta^2 = \sigma_u^2 \cdot (1 - \gamma^2) = (1 - \sigma_c^2) \cdot (1 - \gamma^2)$.

As we said before SRH is measured on a 5-point scale. We assume the following relationship between SRH_{it} and the latent health index y_{it}^*:

$$SRH_{it} = L \text{ if } \lambda_{L-1}^g < y_{it}^* \leqslant \lambda_L^g,\ L = 1, \cdots, 5;\ g = 1, \ldots, G, \tag{2}$$

where $\boldsymbol{\lambda}^g = (\lambda_1^g, \lambda_2^g, \lambda_3^g, \lambda_4^g)'$ are the threshold parameters for demographic group g ($\lambda_0^g = -\infty$ and $\lambda_5^g = \infty$). We allow the thresholds to differ by demographic group g to account for reporting heterogeneity in health ('cut-point' shifting). Based on earlier empirical work by Lindeboom and van Doorslaer, 2004 we distinguish on the basis of the variables 'gender' and 'year-of-birth cohort'.[5]

We further assume that the thresholds as well as the β parameters are constant over time. The β parameters may change over time if medical innovations reduces the impact of certain medical conditions on SRH. In the empirical analysis we will formally test this assumption. We also test for 'index-shifting', the AR(1) structure of the error term and whether the autocorrelation of the error term and the random effect differs between gender and year of birth cohort.

[5]We create four different demographic groups ($G = 4$): males born before 1945, females born before 1945, males born after 1944, and females born after 1944.

2.3 Estimation of the health index model

In this subsection we explain how we estimate the 'structural' parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_c^2, \gamma)'$. For the explanation of the estimation procedure it is relevant to know that our survey data consists of four waves (see the data description in the next section). Estimation is done in several steps. First, we estimate for each demographic group g the multivariate ordered probit model where the dependent variables are SRH in waves 1 till 4.[6] Obviously, the vectors of threshold parameters $\boldsymbol{\lambda}_t^g,\ t = 1, \ldots, 4$ are also wave specific. The multivariate ordered probit model assumes the following relationships between the latent health indices and the explanatory variables:

$$
\begin{aligned}
y_{i1}^* &= \boldsymbol{x}_{i1}'\boldsymbol{\beta}_1^g + \varepsilon_{i1} \\
y_{i2}^* &= \boldsymbol{x}_{i2}'\boldsymbol{\beta}_2^g + \varepsilon_{i2} \\
y_{i3}^* &= \boldsymbol{x}_{i3}'\boldsymbol{\beta}_3^g + \varepsilon_{i3} \\
y_{i4}^* &= \boldsymbol{x}_{i4}'\boldsymbol{\beta}_4^g + \varepsilon_{i4},
\end{aligned}
$$

where the vector $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})'$ is normally distributed conditional upon $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}, \boldsymbol{x}_{i4})'$:

$$
\boldsymbol{\varepsilon}_i | \boldsymbol{x}_i \sim NID \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{21}^g & \rho_{31}^g & \rho_{41}^g \\ \rho_{21}^g & 1 & \rho_{32}^g & \rho_{42}^g \\ \rho_{31}^g & \rho_{32}^g & 1 & \rho_{43}^g \\ \rho_{31}^g & \rho_{32}^g & \rho_{43}^g & 1 \end{pmatrix} \right).
$$

The first step of the estimation procedure yields for each demographic group g an estimate of the following vector of 'auxiliary' parameters $\boldsymbol{\xi}^g = (\boldsymbol{\eta}_1^{g'}, \ldots, \boldsymbol{\eta}_4^{g'}, \boldsymbol{\rho}^{g'})'$ where $\boldsymbol{\eta}_t^g = (\boldsymbol{\beta}_t^{g'}, \boldsymbol{\lambda}_t^{g'})',\ t = 1, \ldots, 4$ and $\boldsymbol{\rho}^g = (\rho_{21}^g, \rho_{31}^g, \rho_{41}^g, \rho_{32}^g, \rho_{42}^g, \rho_{43}^g)'$. In the second step we apply a minimum distance estimation procedure in which we impose the restriction that the parameters of the index functions and the threshold parameters are not wave specific, i.e. $\boldsymbol{\eta}_1^g = \ldots = \boldsymbol{\eta}_4^g = \boldsymbol{\eta}^g = (\boldsymbol{\beta}^{g'}, \boldsymbol{\lambda}^{g'})'$.[7] This second step yields consistent estimates of the vector: $\boldsymbol{\theta}^{*g} = (\boldsymbol{\beta}^{g'}, \boldsymbol{\lambda}^{g'}, \boldsymbol{\rho}^{g'})'$. In the third step we follow Jürges (2007) and apply a minimum distance step in which we impose the restriction that there is no 'index shifting'. In other words, we assume that $\boldsymbol{\beta}^g = \boldsymbol{\beta}$ and $\boldsymbol{\rho}^g = \boldsymbol{\rho}$. In the fourth

[6] We use the Stata module CMP developed by Roodman (2011) for the estimation of the multivariate ordered probit model.

[7] See e.g. Cameron and Trivedi (2005) for more information about the minimum distance estimation procedure.

and final step of the estimation procedure we impose the restriction that the error term ε_{it} can be decomposed into a random individual effect c_i and an AR(1) distributed idiosyncratic term u_{it}. These restrictions imply the following relation between the 'auxiliary' parameter vector $\boldsymbol{\rho}$ and the 'deep' parameters σ_c^2 and γ:

$$\rho_{21} = \rho_{32} = \rho_{43} = (1-\gamma)\sigma_c^2 + \gamma \tag{4a}$$

$$\rho_{31} = \rho_{42} = (1-\gamma^2)\sigma_c^2 + \gamma^2 \tag{4b}$$

$$\rho_{41} = (1-\gamma^3)\sigma_c^2 + \gamma^3. \tag{4c}$$

2.4 Prediction of the health index in a large administrative data set

Next, we use the estimated parameters γ, σ_c^2 and $\boldsymbol{\beta}$ to construct a health index for a large random sample of the Dutch population. The health index is a linear prediction of true health status

$$\hat{y}_{it}^* = \boldsymbol{x}_{it}'\hat{\boldsymbol{\beta}} + \tilde{c}_i + \tilde{u}_{it}. \tag{5}$$

where $\boldsymbol{x}_{it}'\hat{\boldsymbol{\beta}}$ is the estimated conditional expectation of true health status given observed health indicators. As mentioned before, SRH is only used in the construction of the index but not in the prediction of the index. To correct for the stochastic properties of the error term we add a simulated composite residual $\tilde{c}_i + \tilde{u}_{it}$ to $\boldsymbol{x}_{it}'\hat{\boldsymbol{\beta}}$, similar to stochastic regression imputation to restore lost variability in the data (Little et al., 2002).

For the simulation of the composite error term we first assign each person an individual random effect $\tilde{c}_i$ by drawing it from a normal distribution with zero mean and variance $\hat{\sigma}_c^2$. Next, we impute a value of the idiosyncratic error term in the first period $\tilde{u}_{i1}$ by performing a random draw from a normal distribution with zero mean and variance $1-\hat{\sigma}_c^2$. Finally we draw $\tilde{\zeta}_{it},\ t = 2, \cdots, T$ from a zero mean normal distribution with variance $(1-\hat{\gamma}^2)(1-\hat{\sigma}_c^2)$ to simulate $\tilde{u}_{it}$ for subsequent periods exploiting that u_{it} follows an AR(1) process.

3 Data

We distinguish between two samples. The first sample contains survey data on SRH from the Longitudinal Internet Study in the Social Sciences (LISS) between 2007 and 2010. We link this

'LISS sample' to administrative medical data. In addition, we use a large sample of 200,000 individuals on 1 January 2006 from the Dutch municipal population register for which we predict the health index. This sample is linked to administrative medical data and administrative data which contain socio-economic and demographic measures.

3.1 LISS

Survey data are taken from the LISS panel, gathered by CentERdata. This panel is recruited through address-based sampling (no self-selection). Households without a computer and/or internet connection receive an internet connection and computer for free. Residents of institutions and nursing homes are excluded from the survey. This roughly nationally representative household panel (Van der Laan, 2009) receives online questionnaires each month, on different topics. When respondents complete a questionnaire they receive a monthly incentive. A variety of data is available from studies conducted in the LISS panel.[8] In this paper we use the yearly survey on health.

In the LISS panel we select all respondent of the yearly survey on health from 2007 to 2010. This data set consists of 24,486 individual-year observations. To link administrative data to the panel members, an opt-out consent method was used. In September 2011 all panel members received an email asking whether they objected against matching their survey responses with administrative resources (Das and Couper, 2014). A small minority objects against linkage. Unfortunately, not all of our respondents were still participating in September 2011. For these people we have no consent and because of ethical considerations we could not link their survey answers to the administrative data. Administrative records can be retrieved for 17,114 of the 24,486 observations (70%). Nonetheless, in 2007 SRH among those who can be linked to administrative data is not significantly different from the SRH among those who cannot be linked. This suggests that the loss of observations due to linking with administrative data does not yield an endogenous sample selection.[9] When we pool the data for all years (2007-2010), SRH is significantly different at the 5 percent level for those who can and cannot be linked to administrative records.[10] The differences, however, are not substantial as can be seen in table 1. We select all individuals of age 15 and older. The resulting sample consists of 16,720 observations.

[8] For more information, see http://www.lissdata.nl/lissdata/.

[9] The p-value corresponding to the null hypothesis of no difference is equal to 0.733

[10] The p-value corresponding to the null hypothesis of no difference is equal to 0.011

To measure SRH, we use the following two questions in the LISS questionnaire.

How would you describe your health in general? With response options: Poor, Moderate, Good, Very Good and Excellent.

Can you indicate whether your health is poorer or better, compared to last year? With response options: considerably poorer, somewhat poorer, the same, somewhat better and considerably better.

The first question gives us what we call the 'SRH level'. Figure 1 shows the development of SRH over the life cycle by educational attainment. Case and Deaton (2005) report very similar patterns for the United States.

By using the second question, we account for inconsistencies in individual response pattern of health status over time. The 'SRH change' provides additional information about health. For example, when persons are in the highest or lowest health state, a further improvement or reduction can not be observed anymore in the first question but can be observed with the second question. Retrospective questions on the other hand, may suffer from recall problems.[11]

3.2 Administrative sample

From the Municipal Population Register (in Dutch: Gemeentelijke Basisregistratie) we draw a random sample of 200.000 Dutch residents for whom educational attainment is available. This register contains demographic information on age, gender, marital status, among others.

For the sample to be representative, we sample with a higher probability from the older age groups. That is because for middle aged and older individuals, educational attainment is not available in educational registers. For a large part of the older population—approximately 10 percent of the population—educational attainment is registered in the Labor Force Survey (LFS). The LFS is a representative large scale rotating panel for the noninstitutionalized population which started in 1996. Once a person participates in the LFS, educational attainment is

[11] In a previous version of the paper we did use the 'SRH change' to clean the 'SRH level' for measurement error. We referred to this cleaned measure of SRH as the 'corrected SRH'. The decision rules for this correction are somewhat arbitrary but lead to very similar results as using both the SRH level and SRH change in the health measurement model. In brief, we constructed this 'corrected' SRH measure as follows: first, we assumed that in the first period we observe a persons' 'true' health status. Next, we assessed whether the health status of the same persons improved, stayed the same, or degenerated in the subsequent period. We compared this change in the level of SRH with the self-reported change in health. If the change in the reported level of health did not correspond with the reported change in health, we modified this measure.

registered. Furthermore, it is updated if a higher educational level is registered in a subsequent survey. The LFS does not sample institutionalized persons, such as individuals in nursing homes. However, educational attainment is available for this group if observed in the LFS before a person permanently moves to a nursing home. As a result, educational attainment is also available for many persons who stay in a nursing home (although these people are not included in the analysis).

We link administrative records on health (which we describe in the next paragraph) and on income, wealth and education to the Municipal Population Register on the basis of a unique personal identifier. Wealth and income data are based on the national tax register and on data from banks, which are available for the whole population.

For a small number of individuals we are unable to link the data records on income (0.77% of the sample). We have checked whether this is related to the person's age or health status (i.e. proximity to death), but this seems not to be the case. We drop these observations from our sample.

From the administrative records on income and wealth we create a variable measuring total household wealth (net worth), a variable measuring net household income, a dummy for home ownership, variable measuring household size and dummy variables for labor market status. Educational attainment refers to the highest level of completed education, according to the Standard Classification of Education (SOI). In the analysis we distinguish between three groups: lower education (primary education or first stage secondary education), intermediate education (second stage secondary education) and higher education (University Bachelor or University Master or higher).

Again, we select all individuals age 15 and older since we are interested in the working age population. In addition we exclude persons from the year of entering a nursing home to make the sample comparable with the LISS survey. Table 2 shows the sample statistics.

3.3 Administrative medical data

In the analysis we use dichotomous indicators of having a medical condition in a specific year. We derive these medical conditions from two sources: (1) the use of prescription medication, and (2) the main diagnosis responsible for hospitalization. The data about prescription medication is administered by the National Health Care Institute (in Dutch: Zorginstituut Nederland). In the data set the dispensed drugs is classified by the Anatomical therapeutic chemical (ATC) code.

With this code we identify the presence of specific medical conditions. We use the same mapping between a specific substance and medical condition as Lamers and van Vliet (2004) and Chini et al. (2011). For example, the ATC-code for insulin is 'A10A' which is used medically to treat (some forms) of diabetes.[12] The derived objective conditions are mainly chronical. We select a set of health conditions for which self-reported information is also available in the Liss survey.

The main diagnosis responsible for hospitalization is based on the Tenth edition of the International Classification of Diseases, ICD10, derived from the hospital discharge register (in Dutch: Landelijke Medische Registratie, LMR). The LMR contains data about hospital admission (inpatient stays) and covers all general and university hospitals and most specialized hospitals.[13] We use the data from 2007-2010 to identify whether a person is diagnosed with cancer since we cannot derive this information from the prescription medication data.[14] In addition to the indicators of having a medical condition we create three indicator variables of medical utilization: (1) hospital admission (2) prescription drug use, and (3) receiving care at home. The data set on the use of long-term care is provided by CAK (in Dutch: Centraal Administratie Kantoor).

The LISS survey asks respondents, in addition to SRH, whether they are currently taking medicine at least once a week for a specified condition and whether the physician has told them that they suffer from a specific disease last year. We use this information to create an indicator variable of having a 'self-reported' medical conditions.

Table 3 provides an overview of the prevalence of medical utilization and the prevalence of medical conditions for respondents of the LISS survey. Comparing the self-reported information (LISS) to administrative records suggests that respondents tend to underreport medical conditions such as mental problems while for cardiac diseases and diabetes the prevalence is about the same. Bharadwaj et al. (2015) report similar discrepancies in the United States which they contribute to stigma about mental illness. Another reason might be that the households are not taking medication as prescribed for some diseases.

[12]Table 8 in the appendix describes the exact mapping of diseases to chronical conditions.

[13]The LMR covers approximately 88% of all inpatient hospital stays (Van der Laan, 2013).

[14]The analysis is primarily based on the prescription medication data. We have identified the same group of medical conditions using the hospital data as the group of conditions derived from prescription medication. We have used this additional information as a sensitivity check. In that case we classify a person as having a disease if a condition is observed in either one of both sources.

4 Results

4.1 Health index

Table 11 shows the estimation results of the health index model. As we explained in section 2 these estimates are obtained by first estimating for each demographic group a multivariate ordered probit model in order to obtain estimates for some auxiliary parameters and then perform sequentially some minimum distance estimation (MDE) steps. In the first MDE step we impose for each demographic group the restriction that the β and threshold parameters are constant over time. The results of the goodness of fit tests indicate that it is allowed to impose these restrictions.

The second MDE step hinges on the hypothesis that all parameters of the health index model except the threshold parameters do not vary across the demographic groups which we consider in this study. The goodness-of-fit test statistic indicates that this hypothesis should be rejected at any reasonable significance level. Nonetheless, we choose to impose these restrictions because we can then construct a 'unique' health index which does not depend on a chosen reference group. In the last MDE step we estimate the parameters γ and σ_c^2 from the autocorrelation coefficients (cf. the system of equations 4). The goodness of fit test indicates that the stochastic part of the health index model can be effectively modeled by means of a random individual effect and an AR(1) distributed idiosyncratic error term.

We first consider the model using only information on the level of SRH and using health measures from the administrative records. In the first column of table 11 the SRH level is the dependent variable. The estimated coefficients, of the impact of a disease on SRH, are displayed in order. All coefficients have the expected negative sign and are highly significant. It is informative to compare our results with other studies to interpret the magnitude of the effect of specific chronic illnesses on subjective health. For diabetes, we find a relative large negative effect in comparison with other studies (see e.g. Sprangers et al., 2000 and Gilliam, 2003), and for rheumatic conditions a relative small association. For the other conditions we find a similar ranking of the coefficients.

The estimates of the threshold parameters (reported in table 11 of the online appendix) suggest that reporting behavior differs significantly across the four demographic groups which we distinguish in this study (cut-point shifts). If we only compare males and females born before 1945 with the same health index y_{it}^* it turns out that males are more positive about their health

status than females: for instance, elderly females are more likely to report 'moderate' or 'poor' health than elderly males. The younger individuals (born after 1944) has a higher tendency to report that their health is 'very good' or 'excellent' than the older generation. It should be stressed again that all the findings on reporting behavior hinges on the assumption that the β-parameters of the health index (cf. equation 1) do not differ across demographic groups.

The estimates of the parameters σ_c^2 and γ imply that correlation between y_{it}^* and y_{it-1}^* is equal to $(1-\gamma)\cdot\sigma_c^2+\gamma = 0.73$. This first autocorrelation coefficient is not that large: if we use these estimates to impute values of the health index in the administrative data set (cf. equation 5) we find that health evolves over the life cycle in a rather erratic way (i.e. big upward and downward shocks in the value of the health index). That is the reason that we also constructed a health measurement model which uses both the SRH level and change to account for reporting error. We describe this 'Level & Change-model' in detail in the online appendix A.2.

Column three of table 11 reports the estimates of the 'Level & Change' health index model. It turns out that in this case the estimates of σ_c^2 and γ are completely different: $\hat{\sigma}_c^2 = 0.004$ and $\hat{\gamma} = 0.923$. We obtain a larger estimate for the first order autocorrelation coefficient of the health index y_{it}^*: 0.9 versus 0.7. For both models, objective medical conditions, such as having diabetes, affect SRH in a similar way. For cancer we derive a somewhat higher coefficient in the Level & Change-model'. This is consistent with the higher coefficient for cancer in column 2 of the table where we use the SRH change instead of the level. This suggests that both measures of SRH indeed contain somewhat different information about a persons latent health and that it is important to incorporate this information.

Next, we analyze whether the results change if we estimate the index solely on the basis of self-reported medical conditions as reported in the LISS survey. We therefore substitute the disease groups as observed in the administrative data (column 1-3 of table 11) by self-reported medical conditions in the LISS survey. Column 4 of the table shows that there are noteworthy differences in most coefficients. This suggests that we should take measurement error issues seriously.

To examine the fit of the health measurement model we use the estimated coefficients to predict the transition probabilities for the LISS sample. The statistics in the remainder of this paper are based on the 'Level & Change-model' unless otherwise stated.

We calculate these in-sample predictions as follows for the SRH level:

$$\Pr\left(SRH_{it} = l | SRH_{it-1} = k, x_{it}, x_{it-1}\right) =$$
$$\Pr\left(\hat{\lambda}^g_{l-1} - x^{'}_{it}\hat{\beta} < \varepsilon_{it} < \hat{\lambda}^g_{l} - x^{'}_{it}\hat{\beta} | \hat{\lambda}^g_{k-1} - x^{'}_{it-1}\hat{\beta} < \varepsilon_{it-1} < \hat{\lambda}^g_{k} - x^{'}_{it-1}\hat{\beta}\right) =$$
$$\frac{P(\hat{\lambda}^g_{l-1} - x^{'}_{it}\hat{\beta} < \varepsilon_{it} < \hat{\lambda}^g_{l} - x^{'}_{it}\hat{\beta}, \hat{\lambda}^g_{k-1} - x^{'}_{it-1}\hat{\beta} < \varepsilon_{it-1} < \hat{\lambda}^g_{k} - x^{'}_{it-1}\hat{\beta}, \hat{\rho}_{12})}{P(\hat{\lambda}^g_{k-1} - x^{'}_{it-1}\hat{\beta} < \varepsilon_{it-1} < \hat{\lambda}^g_{k} - x^{'}_{it-1}\hat{\beta})}, \; k, l = 1, \ldots, 5$$

where the first-order autocorrelation coefficient equals $\hat{\rho}_{12} = (1 - \hat{\gamma})\,\hat{\sigma}^2_c + \hat{\gamma}$.

The first panel of table 5 shows the empirical transition probabilities and the second panel shows the in-sample predictions. A comparison of panel A and B shows that the predicted and empirical transition probabilities agree reasonably well although the health measurement model assigns somewhat higher probabilities to the off-diagonal elements of the transition matrix. The bottom panel of the table shows the predicted transition probabilities using both the SRH change and level. The results show that transitions occur less often using both measures simultaneously in a health measurement model.

4.2 Persistence in health

Table 6 shows transition probabilities for predicted health in the administrative sample by educational attainment for the SRH level model. We construct predicted health status on the basis of the estimated thresholds. We use one threshold as a reference group for all individuals in the sample. The table shows that lower educated people are much more likely to stay in poor health than higher educated people (university bachelor or master).[15]

Table 9 in the online appendix shows the transition matrix by income quintile. We observe a higher persistence of staying in poor health for low income households than for their high income counterparts. We do not observe important differences in health persistence if we stratify the sample by wealth quintile (Table 9).[16]

4.3 Evolution of health status over the life-cycle

To describe the evolution of health over the life cycle we estimate fixed effects models on the administrative sample. The fixed effect captures unobserved time-invariant individual effects,

[15] The results for the SRH 'level & change' model show a higher level of persistence for both lower and higher educated people. This is because we assume homogeneity in the persistence of unobserved health shocks (i.e. they are independent from observed characteristics, such as age, gender or education).

[16] We notice that the level of income and wealth also differs by age-group.

such as cohort effects. The model contains a dummy for every age and is estimated separately by gender and educational level. As we estimate a fixed effect model we can only interpret the slope, or the evolution, in health status as people age, and not the difference in the (initial) level of health. For an easier comparison we let all figures start at zero.

Figure 2 reports the estimated age pattern for both males and females. Health deteriorates with age; as from age 50 we observe that health starts to decline at a faster rate for both males and females. For both males and females we observe that health declines at a similar pace up to about age 60. For men, we observe a further increase in the rate of deterioration in health as from age 60. As a results, there is about a 0.5 standard deviation difference in health between men and women at age 95, which corresponds to having diabetes.

Figure 3 reports the health pattern for different levels of education for males. We observe strong differences in the age gradient by level of education. For lower educated males we observe a relative high gradient already from age 25 onwards; as from age 50 the level of health starts to decline at an even faster rate. For males with an intermediate level of education, the level of health declines slowly up to age 55, after that age the deterioration in health steadily start to accelerate. For highly educated males we also observe a relatively flat age gradient, after age 55 health starts to decline more rapidly. The gradient is however flatter than for males with an intermediate level of education.

Figure 4 stratifies the health profile for females by education. There are three important differences compared to the health profiles for males: First, for lower educated women, the gradient is less steep than for men. Second, for higher and intermediate levels of education the gradient is very similar to men, but the gradient stays relatively linear up to age 70; thereafter health starts to decline at a somewhat faster rate (but not as fast as for men). Finally, also for lower educated women, the increase in the deterioration of health starts at a later age than for lower educated men and the gradient is less steep than for lower educated men.

These figures show two broad patterns. First, there is a strong gradient in education. Second, we observe a gender and education difference in the timing when the rate of worsening in health speeds-up. This moment arrives earlier for males and persons with a lower level of education. Because of using objective health measures we are less prone to justification bias, reporting bias and differences in health perception among age groups. When we compare the results of our health measurement model in figures 3 and 4 with SRH in figure 1 we find faster health deterioration for older persons when using objective health.

Figure 5 shows the standard deviation of predicted health changes over 5 year, by age and education level. We find that the distribution of health changes is wider for older than for younger age groups. This may be a combination of age and cohort effects.[17] Standard deviations are also higher for people with a low education level (except for the youngest age group).

Do education and economic resources reduce the risk to get in poor health and how does this differ by gender? Table 7 gives results (associations) from a regression of health estimated separately for men and women. The different models include dummies for the level of educational attainment (the reference category is lower education), income and wealth quantiles (the bottom quantile is the reference category), a homeowner dummy, a dummy variable for being married and a variable measuring household size. All estimated models account for age and year effects as well.

Women are on average in worse health than men. Women's health benefits more from higher education than men. This even holds after controlling for economic resources. The results also suggests that economic variables as income and wealth are more protective for women's health than for men's health. A possible explanation is that poor health is in particular detrimental for household income when this affects the earning capacity of the main earner; women more often work part-time than men. Indeed when we account for labor market status the association between income and health disappears for a large extent for men and less for women (see also Case and Deaton, 2005)

5 Conclusions

We construct a health measurement model where we combine survey data on SRH linked to a rich set of health measures from medical records. The estimated health model allows us to predict health for the population at large. We thereby account for unobserved heterogeneity and the persistence in unobserved health shocks by exploiting that we have panel data on SRH at our disposal.

To account for inconsistent reporting patterns in SRH over time we propose a model using two measures of SRH: the level of self reported health and the self-reported health change. We show that using both measures of SRH substantially increases the estimated persistence in health.

[17] Van Kippersluis et al. (2009) find that self assessed health is more dispersed among older generations

We use predicted health to study the evolution of health as individuals age and the interaction with economic variables and education. We find that people of low socio-economic status are more likely to stay in poor health. Studies using SRH usually find a weaker pattern.

The age at which health starts to decline at a greater rate arrives earlier for males and low educated individuals. Finally, we show that women on average are in worse health than men due to a higher prevalence of chronic diseases that have a relative detrimental effect on health. Women's health seems to benefit more from having higher education than men. We also provide evidence that income and wealth are protective of health over and above education. Since a woman's health deteriorates at a lower rate over the life-cycle then the health of men, their health status converges.

These stylized facts are able to explain the variation in the decline in health status for different socio-economic groups which is also reported in other studies using SRH (e.g. Case and Deaton, 2005). However, in this study we find for our objective health measure a somewhat higher level of decay for older persons than found for SRH. This may be due to the fact that, compared to SRH, our objective health measure is less likely to suffer from justification bias, reporting bias and differences in health perception among age groups.

References

Baker, M., Stabile, M., Deri, C., 2004. What do self-reported, objective, measures of health measure? Journal of Human Resources 39 (4), 1067–1093.

Benitez-Silva, H., Ni, H., 2008. Health status and health dynamics in an empirical model of expected longevity. Journal of Health Economics 27, 564–584.

Bharadwaj, P., Pai, M. M., Suziedelyte, A., June 2015. Mental health stigma. Working Paper 21240, National Bureau of Economic Research.

Bound, J., 1991. Self-reported versus objective measures of health in retirement models. The Journal of Human Resources 26 (1), pp. 106–138.

Bound, J., Schoenbaum, M., Stinebrickner, T. R., Waidmann, T., 1999. The dynamic effects of health on the labor force transitions of older workers. Labour Economics 6 (2), 179 – 202.

Cameron, A. C., Trivedi, P., 2005. Microeconometrics : Methods and Applications. Cambridge University Press.

Case, A., Deaton, A. S., August 2005. Broken Down by Work and Sex: How Our Health Declines. University of Chicago Press, pp. 185–212.

Chini, F., Pezzotti, P., Orzella, L., Borgia, P., Guasticchi, G., 2011. Can we use the pharmacy data to estimate the prevalence of chronic conditions? a comparison of multiple data sources. BMC Public Health 11 (1), 688.

Contoyannis, P., Jones, A. M., Rice, N., 2004. The dynamics of health in the British Household Panel Survey. Journal of Applied Econometrics 19 (4), 473–503.

Crossley, T. F., Kennedy, S., 2002. The reliability of self-assessed health status. Journal of Health Economics 21 (4), 643 – 658.

Cutler, D. M., Richardson, E., Keeler, T. E., Staiger, D., 1997. Measuring the health of the U.S. population. Brookings Papers on Economic Activity. Microeconomics 1997, pp. 217–282.

Das, M., Couper, M. P., 2014. Optimizing opt-out consent for record linkage. Journal of Official Statistics 30 (3), 479–497.

De Nardi, M., French, E., Jones, J., 2010. Why do the elderly save? The role of medical expenses. Journal of Political Economy 118 (1), 39–75.

Erdogan-Ciftci, E., Van Doorslaer, E., Bago d'Uva, T., Van Lenthe, F., 2010. Do self-perceived health changes predict longevity? Social Science and Medicine 71, 1981–1988.

Gilliam, F., 2003. The impact of epilepsy on subjective health status. Current Neurology and Neuroscience Reports 3 (4), 357–362.

Heiss, F., Venti, S. F., Wise, D. A., July 2014. The persistence and heterogeneity of health among older Americans. Working Paper 20306, National Bureau of Economic Research.

Johnston, D. W., Propper, C., Shields, M. A., 2009. Comparing subjective and objective measures of health: Evidence from hypertension for the income/health gradient. Journal of Health Economics 28 (3), 540 – 552.

Jones, A. M., Koolman, X., Rice, N., 2006. Health-related non-response in the British Household Panel Survey and European Community Household Panel: Using inverse-probability-weighted estimators in non-linear models. Journal of the Royal Statistical Society. Series A (Statistics in Society) 169 (3), pp. 543–569.

Jürges, H., 2007. True health vs response styles: exploring cross-country differences in self-reported health. Health Economics 16 (2), 163–178.

Kapteyn, A., Meijer, E., June 2013. A Comparison of Different Measures of Health and their Relation to Labor Force Transitions at Older Ages. University of Chicago Press, pp. 115–150.

Lamers, L. M., van Vliet, R. C., 2004. The pharmacy-based cost group model: validating and adjusting the classification of medications for chronic conditions to the dutch situation. Health Policy 68 (1), 113 – 121.

Lange, F., McKee, D., 2012. The evolution of latent health over the life course. Tech. rep., Yale University.

Lindeboom, M., van Doorslaer, E., 2004. Cut-point shift and index shift in self-reported health. Journal of Health Economics 23 (6), 1083 – 1099.

Little, R. J. A., Rubin, D. B., Little, R. J. A., Rubin, D. B., 2002. Single Imputation Methods. John Wiley & Sons, Inc., pp. 59–74.

McFadden, E., Luben, R., Bingham, S., Wareham, N., Kinmonth, A.-L., Khaw, K.-T., 2009. Does the association between self-rated health and mortality vary by social class? Social Science & Medicine 68 (2), 275 – 280.

Poterba, J. M., Venti, S. F., Wise, D. A., September 2010. The asset cost of poor health. Working Paper 16389, National Bureau of Economic Research.

Roodman, D., 2011. Fitting fully observed recursive mixed-process models with cmp. Stata Journal 11 (2), 159–206(48).

Ross, C. E., Masters, R. K., Hummer, R. A., 2012. Education and the gender gaps in health and mortality. Demography 49 (4), 1157–1183.

Shulman, L. M., Pretzer-Aboff, I., Anderson, K. E., Stevenson, R., Vaughan, C. G., Gruber-Baldini, A. L., Reich, S. G., Weiner, W. J., 2006. Subjective report versus objective measurement of activities of daily living in Parkinson's disease. Movement Disorders 21 (6), 794–799.

Sprangers, M. A., de Regt, E. B., Andries, F., van Agt, H. M., Bijl, R. V., de Boer, J. B., Foets, M., Hoeymans, N., Jacobs, A. E., Kempen, G. I., Miedema, H. S., Tijhuis, M. A., de Haes, H. C., 2000. Which chronic conditions are associated with better or poorer quality of life? Journal of Clinical Epidemiology 53 (9), 895 – 907.

Van der Laan, J., 2009. Representativity of the LISS Panel, Statistics Netherlands discussion paper (09041).

Van der Laan, J., 2013. Quality of the Dutch Medical Registration (LMR) for the calculation of the Hospital Standardised Mortality Ratio. Discussion paper 201308, Statistics Netherlands, Statistics Netherlands.

Van Doorslaer, E., Jones, A. M., 2003. Inequalities in self reported health: validation of a new approach to measurement. Journal of Health Economics 22, 61–87.

Van Kippersluis, H., O'Donnell, O., Van Doorslaer, E., Van Ourti, T., 2010. Socioeconomic differences in health over the life cycle in an egalitarian country. Social Science and Medicine 70, 428–438.

Van Kippersluis, H., Van Ourti, T., O'Donnell, O., Van Doorslaer, E., 2009. Health and income across the life cycle and generations in europe. Journal of Health Economics 28, 818–830.

Tables and figures

Figure 1: Self reported health by education

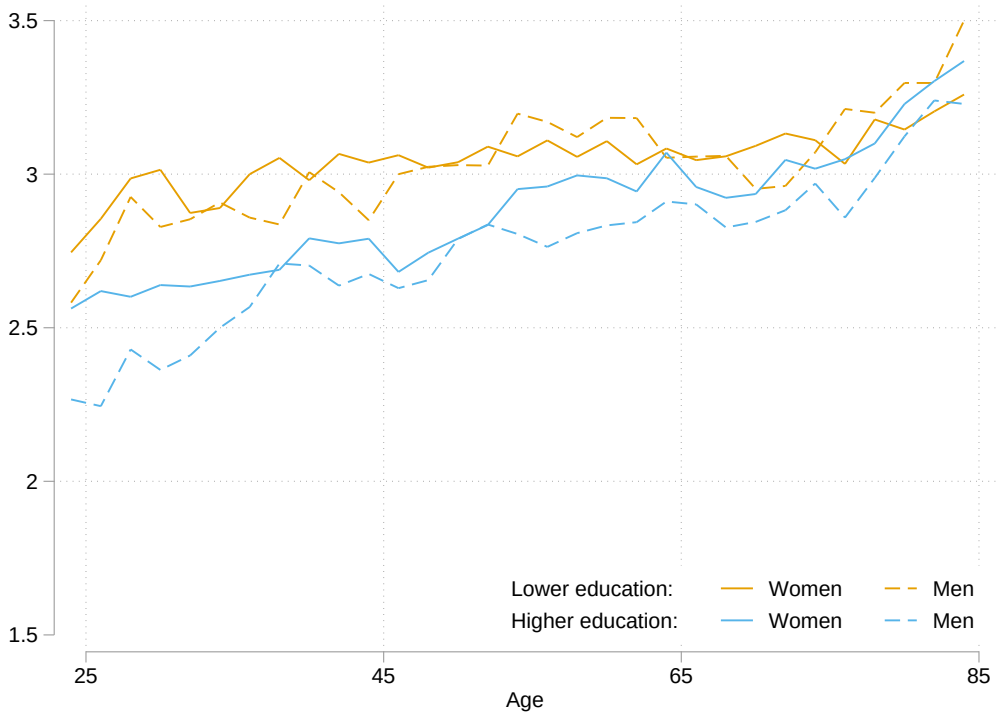

Figure 2: Predicted health by age and gender

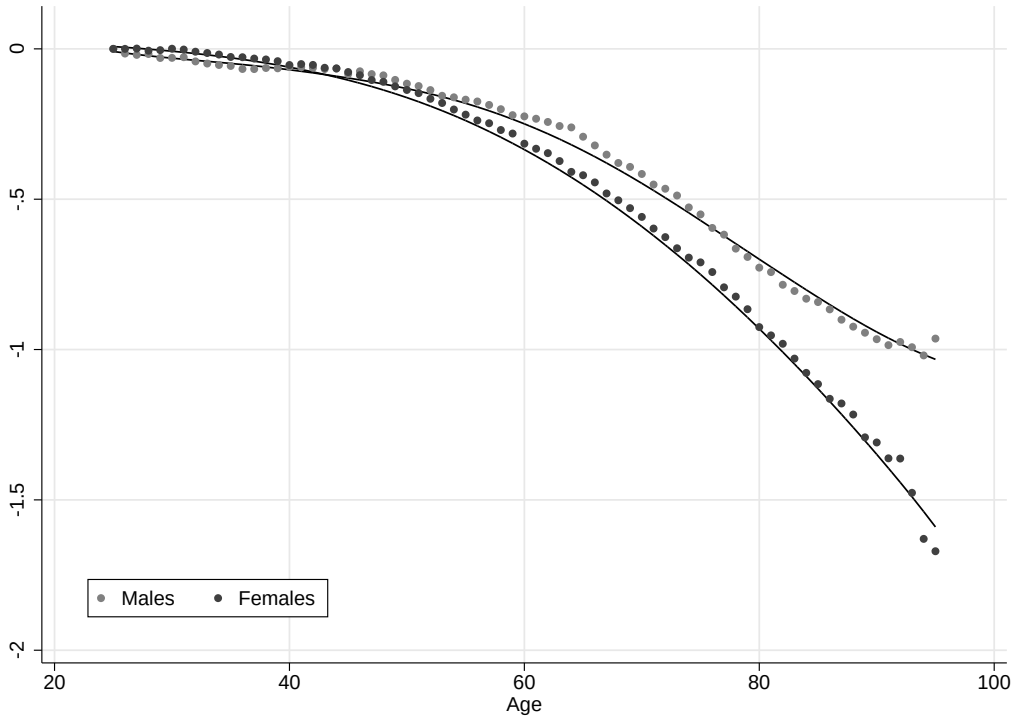

Figure 3: Predicted health by age and education - males

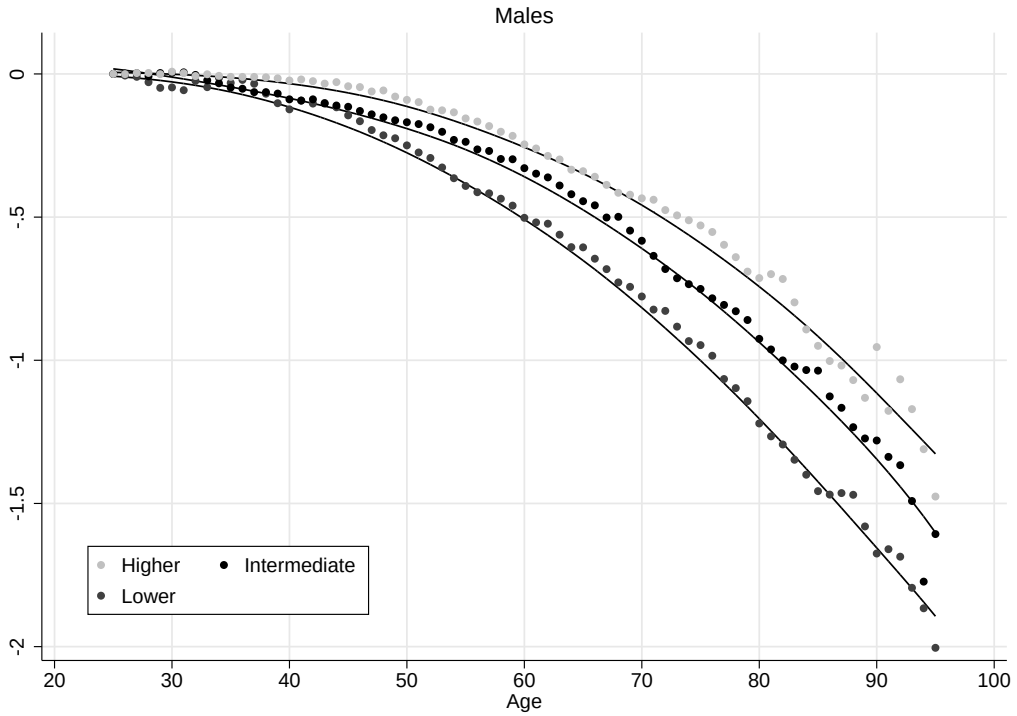

Figure 4: Predicted health by age and education - females

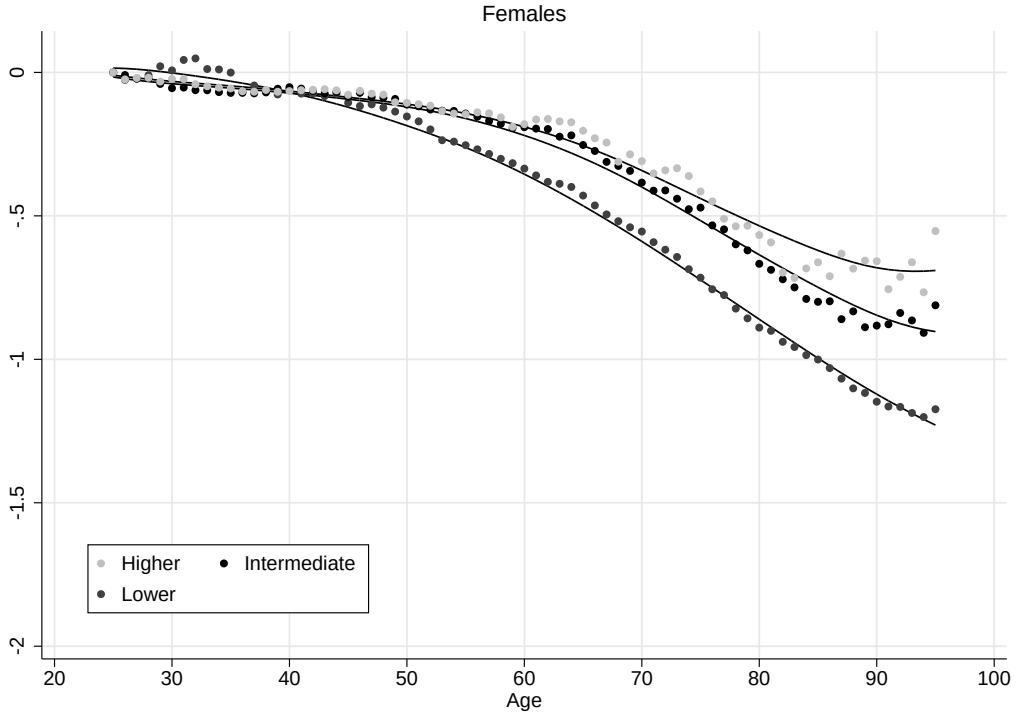

Figure 5: Standard deviation of predicted health by age and education

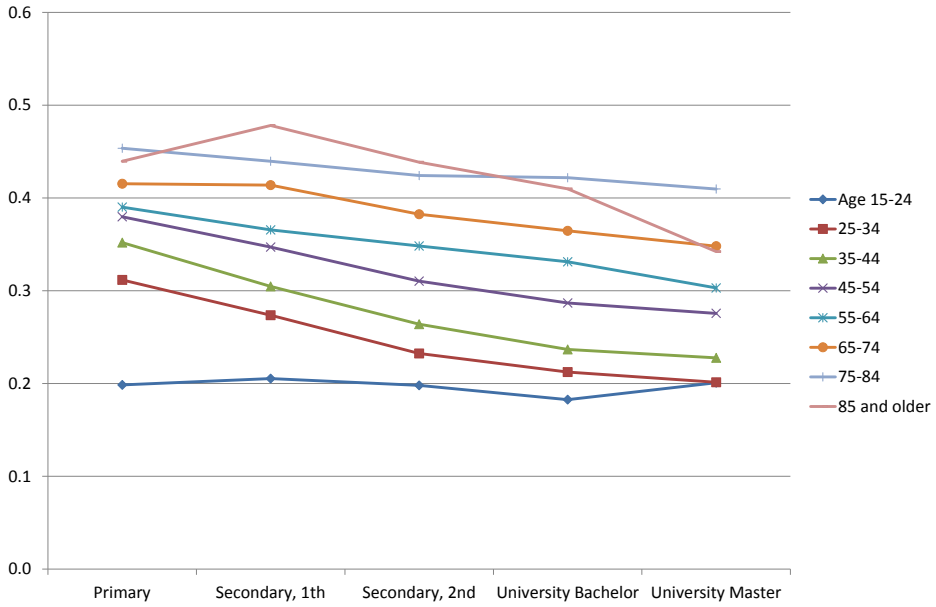

Table 1: Self-reported health for observations which can and cannot be linked to administrative records, $N = 24,486$

Self-reported health	Linked data	No linkage
Poor	0.9	1.4
Moderate	14.0	14.4
Good	60.1	59.6
Very good	19.5	18.9
Excellent	5.4	5.8

Table 2: Summary statistics, 2006, $N = 163,695$

	Mean	SD	
Age	45.167	18.716	
Married	0.492	0.500	
Female	0.505	0.500	
Household size	2.709	1.400	
Primary education	0.032	0.176	
Secondary education, 1st stage	0.185	0.389	
Secondary education, 2nd stage	0.338	0.473	
University bachelor	0.278	0.448	
University master	0.167	0.373	
Owner occupied house	0.642	0.480	
Employed	0.579	0.494	
Job seeker	0.015	0.120	
Exempted from job seeking	0.020	0.139	
(Partial) disabled	0.031	0.172	
Retired	0.179	0.383	
Student	0.092	0.289	
No paid work	0.085	0.279	
Yearly net household income	35,395	25,829	
Total household wealth	202,004	673,287	
Distribution	p25	p50	p75
Yearly net household income	20,994	32,371	44,792
Total household wealth	9,724	90,494	244,892

Table 3: Prevalence of hospital care, home care, prescription drug use and chronic conditions, 2007-2010

	Men				Women			
	Born ≥ 1945		Born < 1945		Born ≥ 1945		Born < 1945	
	Liss	Admin	Liss	Admin	Liss	Admin	Liss	Admin
High blood cholesterol	6.87	8.40	32.16	34.15	3.75	4.28	22.72	25.26
Cardiovascular disease	10.42	13.15	46.88	51.51	10.15	13.49	43.30	48.39
Hypertension / cardiac disease	9.93	11.35	43.46	42.17	9.61	11.76	41.71	42.02
Coronary disease	1.51	5.41	11.78	33.99	0.98	3.59	4.77	23.13
Rheumatism / osteoporosis	3.63	19.49	7.74	21.88	8.14	25.84	22.01	37.58
Rheumatism	3.39	18.94	7.45	20.66	7.11	24.72	16.58	29.08
Osteoporosis	0.40	0.87	0.48	2.18	1.32	2.01	7.03	13.36
Cataract / glaucoma	0.73	7.26	6.78	14.92	0.56	8.85	10.29	23.32
Diabetes	3.03	3.27	9.81	10.52	1.61	2.05	8.56	8.32
Peptic Ulcers	4.72	10.38	12.12	26.13	4.85	12.97	14.27	31.88
Chronic bronchitis / Asthma	4.39	15.64	5.82	22.68	5.16	21.37	6.09	23.32
Depression and anxiety	3.34	6.82	2.36	7.43	4.69	10.45	4.39	14.39
Malignancies	0.48	0.31	4.28	2.23	0.91	0.57	2.19	1.82
Medication use	32.98	59.67	73.41	86.51	41.65	82.54	74.59	89.92
Home care	1.35	0.53	3.22	1.38	2.20	0.25	8.20	3.04
Hospital stay	7.88	6.95	13.59	17.90	11.26	10.29	12.64	17.61
Length of stay (days)	0.48	0.26	0.74	0.95	0.48	0.35	0.68	0.89
N	5824		2083		7293		1828	

This table shows the prevalence of hospital care, home care, prescription drug use and medical conditions for the LISS sample for the years 2007-2010 (N=17,114). 'Admin.' indicates the prevalence of a health measure as measured in the administrative data and 'Liss' indicates the self-reported prevalence of a health measure in the LISS survey.

Table 4: Estimation results of the health index model, $N = 16,720$

	Admin Level		Admin Change		Admin Level & Change		Liss Level & Change	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Diabetes	-.567	.046	-.479	.045	-.488	.040	-.335	.041
Depression and anxiety	-.324	.023	-.257	.020	-.281	.017	-.350	.033
Cardiovascular disease	-.290	.022	-.223	.020	-.256	.017	-.248	.023
Malignancies	-.270	.066	-.406	.058	-.383	.052	-.735	.045
Peptic Ulcers	-.216	.018	-.129	.016	-.198	.014	-.205	.026
High blood cholesterol	-.203	.029	-.178	.027	-.204	.025	-.149	.026
Home care	-.150	.066	-.141	.057	-.325	.048	-.187	.032
Medication use	-.114	.015	-.049	.013	-.070	.012	-.239	.016
Rheumatism / osteoporosis	-.094	.014	-.068	.012	-.080	.010	-.289	.024
Cataract / glaucoma	-.079	.019	-.052	.016	-.040	.014	-.165	.045
Chronic bronchitis / Asthma	-.071	.016	-.050	.013	-.054	.012	-.103	.031
Hospital stay	-.063	.018	-.038	.016	-.069	.014	-.228	.014
Length of stay (days)	-.012	.002	-.008	.002	-.011	.001	-.001	.001
γ	.269	.022			.923	.021	.722	.131
σ_ε^2	.624	.010			.004	.190	.580	.088
σ_ζ^2					.789	.189	.273	.073
σ_d^2					.211	.008	.284	.022

For the threshold parameters we distinguish on basis of the variables 'gender' and 'year-of-birth' four different demographic groups ($G = 4$): (1) males born before 1945, (2) males born after 1945 (3) females born before 1945, (4) females born after 1945 In column (4) we replace the objective health measures by self-reported health measures as reported in the LISS survey.

Table 5: In-sample predictions: 'Level' and 'Level & change' SRH, $N = 16,720$

$t-1 \setminus t$	Poor	Moderate	Good	Very good	Excellent
Panel A: Empirical SRH level					
Poor	46.8	46.8	6.4	0.0	0.0
Moderate	3.3	59.5	36.3	0.9	0.1
Good	0.2	8.3	78.6	11.7	1.3
Very good	0.1	1.6	37.3	51.6	9.5
Excellent	0.2	0.7	17.6	37.1	44.5
Panel B: In-sample predictions SRH level					
Poor	34.2	60.2	5.6	0.0	0.0
Moderate	3.8	54.1	41.7	0.5	0.0
Good	0.1	10.8	74.4	13.5	1.2
Very good	0.0	0.4	42.8	43.7	13.1
Excellent	0.0	0.0	12.6	42.1	45.2
Panel C: In-sample predictions SRH 'Level & change'					
Poor	62.3	37.6	0.0	0.0	0.0
Moderate	2.9	73.5	23.6	0.0	0.0
Good	0.0	6.4	85.4	8.1	0.1
Very good	0.0	0.0	25.5	64.9	9.6
Excellent	0.0	0.0	0.6	29.9	69.6

The predictions are based upon the coefficients of the level-change model.

Table 6: Health persistence by Educational Attainment, SRH 'level' model, N=163,695

	$t \setminus t+1$	Poor	Moderate	Good	Very good	Excellent
Lower	Poor	36.7	55.6	7.6	0.0	0.0
	Moderate	5.1	50.5	43.8	0.6	0.0
	Good	0.2	12.0	74.2	12.5	1.2
	Very good	0.0	0.7	48.3	39.9	11.1
	Excellent	0.0	0.1	18.5	43.9	37.6
Higher	Poor	31.2	60.4	8.5	0.0	0.0
	Moderate	3.1	47.7	48.7	0.6	0.0
	Good	0.1	10.0	74.9	13.7	1.3
	Very good	0.0	0.5	46.8	40.8	11.9
	Excellent	0.0	0.0	16.8	44.6	38.6

Table 7: Estimation results, dependent variable: predicted health, N=163,695

	Men	Women	Men	Women	Men	Women
Intermediate education	0.044 ***	0.085 ***	0.026 ***	0.064 ***	0.022 ***	0.063 ***
	(0.005)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Higher education	0.064 ***	0.120 ***	0.034 ***	0.086 ***	0.029 ***	0.083 ***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Married	0.005	0.023 ***	-0.003	0.004	-0.004	0.004
	(0.004) *	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Household size	0.003	0.010 **	-0.005 ***	0.001	-0.004 **	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Homeowner			0.011 **	0.020 ***	0.008 *	0.017 ***
			(0.005)	(0.005)	(0.005)	(0.005)
2nd income quintile			0.011 **	0.026 ***	0.005	0.022
			(0.006)	(0.005)	(0.006)	(0.005)
3th income quintile			0.023 ***	0.014 **	0.012 **	0.008 ***
			(0.006)	(0.006)	(0.006)	(0.006)
4th income quintile			0.027 ***	0.030 ***	0.013 **	0.022 ***
			(0.006)	(0.006)	(0.006)	(0.006)
5th income quintile			0.024 ***	0.036 ***	0.007	0.025 ***
			(0.007)	(0.007)	(0.007)	(0.007)
2nd wealth quintile			-0.001	0.009 *	-0.005	0.005
			(0.006)	(0.005)	(0.006)	(0.005)
3th wealth quintile			0.052 ***	0.041 ***	0.049 ***	0.037 ***
			(0.006)	(0.006)	(0.006)	(0.006)
4th wealth quintile			0.056 ***	0.062 ***	0.053 ***	0.057 ***
			(0.006)	(0.006)	(0.006)	(0.006)
5th wealth quintile			0.085 ***	0.088 ***	0.082 ***	0.082 ***
			(0.006)	(0.007)	(0.007)	(0.007)
Retired					-0.040 ***	-0.014
					(0.010)	(0.010)
Unemployed					-0.023 **	-0.037 ***
					(0.011)	(0.010)
Disabled					-0.247 ***	-0.203 ***
					(0.010)	(0.009)
Self-employed					-0.003	0.027 ***
					(0.005)	(0.007)
Other					0.009	-0.007
					(0.009)	(0.006)
Constant	-0.132 ***	-0.225 ***	-0.144 ***	-0.229 ***	-0.133 ***	-0.214 ***
	(0.013)	(0.013)	(0.014)	(0.013)	(0.015)	(0.014)

Significant at the *** 1%; ** 5%; * 10% level. Includes age dummies.

A Online appendix

A.1 Online tables

Table 8: Mapping prescription drugs to chronic conditions

Chronic disease	ATC-code in medical register
Coronary disease	B01A, C04A
Cardiac disease	C01, C03C
Hypertension	C02, C03A, C07, C08, C09A,B
Rheumatic conditions	H02, M01, M02
High blood cholesterol	C10A
Glaucoma and cataract	S01E
Peptic Ulcers	A02A, A02B
Chronic bronchitis, asthma	R03
Anxiety and depression	N05B, N06A
Osteoporosis	M05

Table 9: Health persistence by 2006 Income quintile, SRH 'level' model, N=163,695

	$t \setminus t+1$	Poor	Moderate	Good	Very good	Excellent
Q1	Poor	39.5	53.0	7.5	0.0	0.0
	Moderate	6.1	51.8	41.6	0.5	0.0
	Good	0.3	13.2	73.6	11.9	1.1
	Very good	0.0	0.9	48.4	39.9	10.9
	Excellent	0.0	0.1	18.2	44.1	37.6
Q2	Poor	36.9	55.9	7.2	0.1	0.0
	Moderate	5.1	50.8	43.5	0.6	0.0
	Good	0.2	12.4	74.2	12.1	1.1
	Very good	0.0	0.7	49.2	39.2	10.9
	Excellent	0.0	0.1	18.9	43.9	37.2
Q3	Poor	33.3	57.7	8.8	0.1	0.0
	Moderate	3.9	49.1	46.4	0.6	0.0
	Good	0.1	10.9	74.7	13.0	1.2
	Very good	0.0	0.5	47.2	40.6	11.7
	Excellent	0.0	0.0	17.4	45.1	37.5
Q4	Poor	32.9	60.0	7.2	0.0	0.0
	Moderate	3.6	47.5	48.3	0.6	0.0
	Good	0.1	10.3	74.6	13.7	1.4
	Very good	0.0	0.6	47.5	40.4	11.6
	Excellent	0.0	0.1	17.8	44.0	38.1
Q5	Poor	28.1	62.5	9.4	0.0	0.0
	Moderate	3.1	48.2	48.1	0.6	0.0
	Good	0.1	10.0	75.0	13.6	1.3
	Very good	0.0	0.5	46.8	40.8	11.9
	Excellent	0.0	0.0	17.1	43.7	39.1

Table 10: Health persistence by 2006 Wealth quintile, SRH 'level' model, N=163,695

	$t \setminus t+1$	Poor	Moderate	Good	Very good	Excellent
Q1	Poor	36.5	54.7	8.8	0.0	0.0
	Moderate	4.6	49.0	45.8	0.7	0.0
	Good	0.2	11.0	74.5	13.1	1.2
	Very good	0.0	0.6	47.3	40.7	11.4
	Excellent	0.0	0.0	17.5	44.5	38.0
Q2	Poor	39.3	54.0	6.7	0.1	0.0
	Moderate	5.2	50.4	43.8	0.6	0.0
	Good	0.2	11.9	74.2	12.5	1.3
	Very good	0.0	0.7	47.2	40.4	11.7
	Excellent	0.0	0.1	17.7	44.3	38.0
Q3	Poor	29.9	61.1	8.9	0.1	0.0
	Moderate	3.9	49.0	46.6	0.5	0.0
	Good	0.1	10.7	74.4	13.5	1.3
	Very good	0.0	0.6	47.7	40.2	11.5
	Excellent	0.0	0.0	18.0	45.2	36.8
Q4	Poor	35.5	55.7	8.8	0.0	0.0
	Moderate	4.2	49.6	45.7	0.5	0.0
	Good	0.2	11.4	74.6	12.6	1.3
	Very good	0.0	0.6	48.1	39.9	11.4
	Excellent	0.0	0.1	18.1	44.9	36.9
Q5	Poor	34.4	59.2	6.4	0.0	0.0
	Moderate	4.4	50.1	44.9	0.6	0.0
	Good	0.2	11.5	74.5	12.7	1.2
	Very good	0.0	0.6	48.5	40.0	11.0
	Excellent	0.0	0.1	17.7	42.0	40.2

Table 11: Estimation results of the health index model - threshold parameters, $N = 16,720$

	Admin		Admin		Admin		Liss	
	Level		Change		Level & Change		Level & Change	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
cutlevl 1	-3.117	.081			-3.090	.077	-3.110	.077
cutlevl 2	-1.598	.041			-1.525	.039	-1.517	.037
cutlevl 3	.455	.038			.519	.036	.479	.035
cutlevl 4	1.444	.051			1.485	.049	1.467	.049
cut difl 1			-2.312	.092	-2.381	.068	-2.323	.062
cut difl 2			-.940	.040	-.957	.028	-.887	.027
cut difl 3			1.368	.047	1.362	.032	1.338	.031
cut difl 4			2.075	.075	2.049	.051	2.124	.058
cutlev2 1	-2.826	.049			-2.796	.048	-2.780	.054
cutlev2 2	-1.441	.024			-1.380	.023	-1.408	.025
cutlev2 3	.338	.020			.382	.019	.363	.019
cutlev2 4	1.317	.024			1.318	.024	1.278	.025
cut dif2 1			-2.475	.060	-2.530	.043	-2.419	.049
cut dif2 2			-1.088	.024	-1.094	.017	-1.095	.020
cut dif2 3			1.059	.023	1.045	.016	1.049	.020
cut dif2 4			1.927	.037	1.883	.026	1.906	.033
cutlev3 1	-3.476	.119			-3.492	.117	-2.875	.054
cutlev3 2	-1.348	.040			-1.257	.038	-1.081	.025
cutlev3 3	.539	.040			.635	.038	.518	.019
cutlev3 4	1.498	.056			1.566	.053	1.322	.025
cut dif3 1			-2.081	.079	-2.159	.056	-2.167	.049
cut dif3 2			-.826	.041	-.829	.028	-.803	.020
cut dif3 3			1.311	.048	1.306	.032	1.257	.020
cut dif3 4			2.006	.074	2.003	.050	1.940	.033
cutlevl 1	-2.814	.043			-2.793	.043	-2.819	.045
cutlevl 2	-1.381	.023			-1.298	.022	-1.298	.022
cutlevl 3	.450	.020			.498	.019	.515	.018
cutlevl 4	1.409	.025			1.414	.024	1.405	.024
cut dif4 1			-2.261	.044	-2.327	.032	-2.299	.036
cut dif4 2			-1.042	.020	-1.045	.014	-1.016	.017
cut dif4 3			1.004	.020	.994	.014	1.000	.017
cut dif4 4			1.891	.033	1.851	.023	1.890	.028

For the threshold parameters we distinguish on basis of the variables 'gender' and 'year-of-birth' four different demographic groups ($G = 4$): (1) males born before 1945, (2) males born after 1945 (3) females born before 1945, (4) females born after 1945 In column (4) we replace the objective health measures by self-reported health measures as reported in the LISS survey.

A.2 Modelling persistence in a 'level & and change' model

- Let y_{it} be the answer to the SRH question. Its distributional conditional upon $\boldsymbol{x}_{it}$ is of the ordered probit type where y_{it}^* denotes the associated latent variable. We postulate the following model for y_{it}^*:

$$y_{it}^* = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \tag{6}$$

The error term ε_{it} can be decomposed into a random effect c_i, an idiosyncratic error term ζ_{it} and a measurement error ξ_{it},i.e.:

$$\varepsilon_{it} = c_i + \zeta_{it} + \xi_{it}$$

We make the following assumption about the components of ε_{it}:

1. c_i, ζ_{it} and ξ_{it} are mutually independent and independent of $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}', \ldots, \boldsymbol{x}_{iT}')'$
2. $\text{var}(\varepsilon_{it}) = 1$
3. $c_i \sim NID(0, \sigma_c^2)$
4. The measurement errors $\xi_{it},\ t = 1, \ldots, T$ are serially independent and independent of c_i and ζ_{it}. Moreover, $\xi_{it} \sim NID(0, \sigma_\xi^2)$.

$$\zeta_{it} = \gamma\zeta_{it-1} + \upsilon_{it},\ 0 \leq |\gamma| < 1 \tag{7}$$

We assume that $\upsilon_{it}|\boldsymbol{x}_i \sim NID(0, \sigma_\upsilon^2)$ Equation (7) implies that $\zeta_{it}|\boldsymbol{x}_i \sim NID\left(0, \frac{\sigma_\upsilon^2}{1-\gamma^2}\right)$.

- These assumption imply that

$$\text{var}(\varepsilon_{it}) = \sigma_c^2 + \sigma_\zeta^2 + \sigma_\xi^2 = \sigma_c^2 + \frac{\sigma_\upsilon^2}{1-\gamma^2} + \sigma_\xi^2 = 1 \tag{8a}$$

$$\text{cov}(\varepsilon_{it}, \varepsilon_{i\tau}) = \sigma_c^2 + \gamma^{t-\tau}\sigma_\zeta^2 = \sigma_c^2 + \gamma^{t-\tau}\frac{\sigma_\upsilon^2}{1-\gamma^2},\quad t > \tau \tag{8b}$$

- latent difference model

$$dy_{it}^* = \Delta\boldsymbol{x}_{it}'\boldsymbol{\beta} + u_{it} \tag{9}$$

The error term u_{it} can be decomposed into an idiosyncratic error term $\Delta\zeta_{it}$ (ζ is defined in equation (7)) and a measurement error θ_{it}. We make the following assumption about (the components of) u_{it}

1.

$$u_{it} = \Delta\zeta_{it} + \theta_{it} = (\gamma - 1)\zeta_{it-1} + \upsilon_{it} + \theta_{it} \tag{10}$$

2. $\text{var}(u_{it}) = 1$
3. The measurement error θ_{it} can be decomposed into a random effect d_i and a idiosyncratic serially independent error term η_{it}:

$$\theta_{it} = d_i + \eta_{it}$$

We allow for a random effect d_i in the measurement error because the reduced for evidence suggests $\text{cov}(u_{it}, u_{i\tau}) > 0, t \neq \tau$. Without a random effect d_i, which might capture recall bias, we do not obtain a positive covariance unless we allow for a positive correlation between the measurement errors xi_{it} and η_{it}. Given this assumption equation (10) can be rewritten as follows:

$$u_{it} = d_i + (\gamma - 1)\zeta_{it-1} + \upsilon_{it} + \eta_{it} \tag{11}$$

4. $\text{cov}(\eta_{it}, \xi_{i\tau}) = \text{cov}(c_i, d_i) = \text{cov}(\eta_{it}, d_i) = \text{cov}(\eta_{it}, c_i) = \text{cov}(\xi_{it}, d_i) = 0$ conditional upon $\boldsymbol{x}_i$.
5. $d_i|\boldsymbol{x}_i \sim NID(0, \sigma_d^2)$, $\eta_{it} \sim NID(0, \sigma_\eta^2)$.

These assumption imply that $(t > \tau)$

$$\text{var}(u_{it}) = \sigma_d^2 + (\gamma - 1)^2 \frac{\sigma_\upsilon^2}{1 - \gamma^2} + \sigma_\upsilon^2 + \sigma_\eta^2 = \sigma_d^2 + 2(1 - \gamma)\sigma_\zeta^2 + \sigma_\eta^2 = 1 \tag{12a}$$

$$\text{cov}(u_{it}, u_{i\tau}) = \sigma_d^2 - \gamma^{t-1-\tau}(\gamma - 1)^2\sigma_\zeta^2 = \sigma_d^2 + \frac{\gamma^{t-1-\tau}(\gamma - 1)\sigma_\upsilon^2}{1 + \gamma} \tag{12b}$$

$$\text{cov}(u_{it}, \varepsilon_{it}) = \frac{\sigma_\upsilon^2}{1 + \gamma} = (1 - \gamma)\sigma_\zeta^2 \tag{12c}$$

$$\text{cov}(u_{i\tau}, \varepsilon_{it}) = \text{cov}(d_i + \zeta_{i\tau} - \zeta_{i\tau-1} + \eta_{i\tau}, c_i + \zeta_{it} + \xi_{it}) = \gamma^{t-\tau}(1-\gamma)\sigma_\zeta^2 = \gamma^{t-\tau}(1-\gamma)\sigma_\zeta^2 > 0 \tag{12d}$$

$$\text{cov}(u_{it}, \varepsilon_{i\tau}) = \text{cov}(d_i + \zeta_{it} - \zeta_{it-1} + \eta_{it}, c_i + \zeta_{i\tau} + \xi_{i\tau}) = \gamma^{t-1-\tau}(\gamma - 1)\sigma_\zeta^2 < 0 \tag{12e}$$