Dynamics of Self Rated Health and Selective Mortality*

Florian Heiss

University of Munich, Department of Economics,

Ludwigstr. 28 RG, 80539 Munich, Germany

Phone: +49(621)2180-6291, Fax: +49(621)2180-3954

florian.heiss@lrz.uni-muenchen.de

July 28, 2008

Self rated health status (SRHS) is one of the most frequently used health measures in empirical health economics. This paper documents the temporal correlation structure of SHRS in the first seven waves of the Health and Retirement Study (HRS). It argues that commonly used panel data models including random effects and/or state dependence are unable to explain the findings and suggests a parsimonious alternative with an autocorrelated latent health component. Furthermore, the problem of survivorship bias is accounted for in a joint model of SRHS and mortality. Simulations show that this model succeeds well in explaining the empirical facts.

*The author would like to thank Axel Börsch-Supan, Daniel McFadden, Michael Hurd, Arthur van Soest, Joachim Winter, and David Wise for valuable discussion, comments, suggestions, and insights from joint research.

1 Introduction

Self rated health status (SRHS) is one of the most frequently studied measures of individual health. While it is a subjective evaluation of the own health, it does have considerable objective content such as predictive power for mortality. Furthermore, it is very easy to collect and therefore available in many large scale surveys.

Panel data permit a much richer analysis of the determinants of health than pure cross-sectional data and allow to study and account for unobserved heterogeneity and changes over time. Typical empirical studies of SHRS estimate a model with a simple specification for the correlation over time on a data set of the living respondents. This paper discusses two issues regarding panel data models of SRHS and suggests a parsimonious model structure to overcome the problems of the commonly used approaches.

The first issue concerns the modeling of the intertemporal correlation structure. While it is obvious that individual SRHS is positively correlated over time conditional on observable characteristics, the strategies to explain and account for this fact vary in the literature. Contoyannis, Jones and Rice (2004) carefully discuss classical panel data models of SRHS. They distinguish state dependence from time-constant unobserved heterogeneity. State dependence refers to a causal effect of past outcomes on current outcomes. These kinds of models are common in labor economics, see for example Card and Sullivan (1988). While it is plausible to assume a causal link from yesterday's labor market status for example through human capital accumulation, this seems more questionable for SRHS: Where on the scale a respondent made a tick in a previous wave does most likely not causally affect the results of a later survey. Unobserved heterogeneity refers to individual differences not captured by observable characteristics. While the presence of unobserved heterogeneity in SRHS models is very plausible, it is not so clear while the idiosyncratic health differences should be constant over time.

A closer look at the first seven waves of the Health and Retirement Study (HRS) reveals that the dynamic pattern of SRHS is incompatible with a combination of time-constant

unobserved heterogeneity and (low-order) state dependence. This paper suggests a somewhat different model specification which explains the intertemporal correlation pattern with state dependence of an underlying latent health status instead of SRHS itself. It can also be interpreted as unobserved heterogeneity which is correlated but not necessarily constant over time.

The second issue discussed in this paper is selective mortality or survivorship bias (Jones, Koolman and Rice 2006). Health is obviously only observed for living respondents. Since SRHS is highly correlated over time and a strong predictor of mortality, cross-sectional age profiles necessarily confound effects of age on health with an increasing selection of the relatively healthy survivors. The same holds for any potential determinant of health such as socio-economic status or policy interventions: Increases of health are potentially confounded by changes of the composition of survivors. To account for survivorship bias, this paper proposes a joint dynamic model of SRHS and mortality and estimates it on the HRS sample. This allows to disentangle effects on health from selection effects through mortality. Unlike the approach of inverse probability weighting used by Contoyannis et al. (2004) and Jones et al. (2006), the joint models allows for selection on unobservables which affect both SRHS and mortality. Simulations show that the model captures both the SRHS dynamics and mortality selection observed in the data well.

The paper is structured as follows. Section 2 discusses SRHS, introduces the sample used for the empirical analyzes, and provides descriptive analyzes with a focus on SRHS and mortality. Section 3 discusses how to reconcile the findings with econometric panel data models. Based on this discussion, a joint model for SRHS and mortality is suggested. Results and simulations are presented in Section 4 and section 5 concludes.

2 SRHS: Background and Descriptives

2.1 Self-Rated Health Status

Self rated health status (SRHS) is collected in many surveys. The wording of the question used in the Health and Retirement Study (HRS) is "Would you say your health is excellent, very good, good, fair, or poor?" Obviously, SRHS is a very subjective measure and prone to both random measurement error and potentially systematic reporting biases. Crossley and Kennedy (2002) find that many respondents report a different level of SRHS when answering the same question twice within one survey. Lindeboom and van Doorslaer (2004) discuss the effects of different scales of reference for self-reported measures. A different strand of literature studies the effect of health on labor supply and retirement decisions and finds effects of these decisions on the reporting of SRHS – an effect called justification bias, see for example Disney, Emmerson and Wakefield (2006).

On the other hand, more objective health measures such as the prevalence of chronic conditions are error-ridden as well as long they are reported by the respondent, see Baker, Stabile and Deri (2004) and Butler, Burkhauser, Mitchell and Pincus (1987). Despite its obvious drawbacks, SRHS has been found a useful and powerful measure. It maps the high-dimensional and complex concept of health into one dimension using individual perceptions and judgments. And it has been found to have real content, for example it is a strong predictor of other health measures (Gerdtham, Johannesson, Lundberg and Isacson 1999) and mortality (van Doorslaer and Gerdtham 2003).

With all its advantages and disadvantages, SRHS is one of the most widely studied measures of individual health in empirical health economics. Examples are the papers of Adams, Hurd, McFadden, Merrill and Ribeiro (2003), Deaton and Paxson (1998), and Smith (1999).

2.2 Sample

The data used for the empirical example are from the Health and Retirement Study (HRS).[1] For the analyzes presented here, the RAND HRS Data File (Version G) is used. It is a user-friendly data set produced by the RAND Center for the Study of Aging, with funding from the National Institute on Aging and the Social Security Administration.[2]

The HRS contains data on different cohorts of elderly Americans which entered the panel survey at different points in time. Table 5 gives an overview these cohorts and shows the wave of the first interview, the number of individuals, and the number of observations and deaths in the working sample. It includes all individuals aged 50 or older in the initial interview and excludes individuals with missing information on one of the key variables used in this study. The distribution of these variables in the sample is shown in Table 2.

2.3 SRHS dynamics and mortality

The key issue studied in this paper is the dynamics of the self rated health status (SRHS). Table 3 shows the transitions between two adjacent waves in the HRS data. SRHS is highly persistent – 47 percent of the respondents report the same level of health and more than 86 percent do not change their report by more than one "unit". At the same time, SRHS is a strong predictor of mortality. A respondent who reported poor SRHS in one wave has ten times the risk to die before the next wave compared to someone who reported excellent SRHS.

Current SRHS has predictive power not only for the next wave, but also for the longer future. Figure 1(a) plots the share of the deceased during the next waves by SRHS in wave 1. While only 6.75% of the respondents who reported excellent SRHS in wave 1 have died by wave 7, this is true for 43.65% who reported poor SRHS in wave 1. The

[1] The HRS is sponsored by the National Institute of Aging (grant number NIA U01AG009740) and conducted by the University of Michigan.

[2] See `http://www.rand.org/labor/aging/dataprod` for details.

mortality risk differs in all future waves, so the lines keep diverging instead of becoming parallel at some point.

Figure 1(b) shows the share of respondents in poor or fair health among those who survive all seven waves by initial SRHS. The effect of SRHS in wave 1 decreases slowly over time but it has still strong predictive power for SRHS in wave 7.

The age profile of the share of respondents who report poor or fair SRHS is plotted as the solid line in Figure 2. It rises with age except for the very old. As shown above, SRHS is highly correlated over time and a strong predictor of mortality. The age profile obviously only represents the living respondents, so it is a compound effect of individual development and selection due to differential mortality. The shorter lines in Figure 2 illustrate this effect. They show the age profile only for those respondents who are are observed living at least at age 60, 70, 80, 90, and 95. This takes out the selection effect from the first appearance in the sample through the respective age. These lines are much steeper and represent the pure effect of aging, whereas the difference to the solid line represents the selective mortality.

As already shown, SRHS in wave 1 has explanatory power for both SRHS and mortality in wave 7. The question Table 4 addresses is: does this explanatory power vanish once shorter lags and other variables are controlled for? It shows results from a reduced-form regression of SRHS and mortality on a set of sociodemographic variables and different numbers of lags of SRHS.[3]

The striking result is that even if six lags of SRHS are included so that only the seventh wave for the original HRS cohort is analyzed, the magnitude of the coefficients decreases with the lag length but all lags have significant explanatory power on their own. For mortality, the picture looks similar, but there are only 235 deaths observed between waves 6 and 7, so hardly any coefficients are estimated significantly different

[3]The lagged SRHS values are entered linearly in the results shown in the table. If instead four dummy variables are included for each lag, the results are harder to read and display in one table, but lead to the same qualitative findings. They can be requested from the author.

from zero. This is especially the case for the highly correlated lagged values of SRHS. With three lags, all previous SRHS states have a highly significant predictive power for mortality which decreases in magnitude with increasing lag length. The next section discusses how all these findings can be explained by a parsimonious model.

3 Models of SRHS dynamics and mortality

This section discusses how the dynamics of SRHS and mortality can be modeled in a way which is consistent with the descriptive findings of section 2. Section 3.1 ignores mortality for a moment and looks at SRHS of the survivors alone and section 3.2 adds mortality to the preferred model.

3.1 SRHS dynamics

To use a consistent notation, assume that SRHS is determined by a standard ordered response model. Individual $i = 1, ..., N$ reports a SRHS $y_{it} \in \{1, ..., 5\}$ at time $t = 1, ..., T$ if a latent continuous variable y_{it}^* is between two thresholds:

$$y_{it} = j \quad \Leftrightarrow \quad \alpha_{j-1} \leq y_{it}^* < \alpha_j \quad \text{with } 1 \leq j \leq 5, \tag{1}$$

where $\alpha_0 = -\infty$, $\alpha_5 = \infty$, and α_1 through α_4 are unknown model parameters.[4]

Let $\mathbf{x}_{it}$ denote a vector of strictly exogenous covariates and u_{it} a random error term which is independent over time and individuals and independent of $\mathbf{x}_{it}$. Consider four common types of models for y_{it}^* as discussed by Contoyannis et al. (2004):

(1) Independent $\quad y_{it}^* = \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}$

(2) State Dependence $\quad y_{it}^* = \mathbf{x}_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + u_{it}$

(3) Random effect (RE) $\quad y_{it}^* = \mathbf{x}_{it}\boldsymbol{\beta} + a_i + u_{it}$

(4) State Dependence & RE $\quad y_{it}^* = \mathbf{x}_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + a_i + u_{it}$

[4] The model could be generalized by allowing these thresholds to depend on observed characteristics of the respondents (Lindeboom and van Doorslaer 2004).

The independent model (1) implies that conditional on $\mathbf{x}_{it}$, there is no explanatory power of past values of SRHS which clearly violates the findings in Table 4, at least with the set of covariates used there.

Model (2) with state dependence assumes that the latent variable y_{it}^* is affected by the previous outcome $y_{i,t-1}$.[5] This implies that the first lag of SRHS has explanatory power for today's SRHS, but not the longer lags. The findings in Table 4 disagree with that. Of course, this model can be changed to include higher-order lags, but the results show that one would need at least six lags and there is no indication that the seventh lag would be insignificant if it were available. Intuitively, this model is not very plausible either. If the dependent variable were labor force participation, previous outcomes could affect today's outcomes causally by signaling or human capital depreciation. With SRHS, it is much less clear why the tick I made in one of five boxes in the past should causally affect my current health or SRHS.

The random effects model (3) is more plausible in this sense. It assumes that there is unobserved heterogeneity represented by the random variable a_i which is usually assumed to be independent of $\mathbf{x}_{it}$ as well as u_{it}. Individuals differ by factors beyond the ones captured by the covariates. This creates an additional correlation of the outcomes. In reduced-form regressions like those in Table 4, lagged outcomes are significant predictors since they contain information on a_i. The i.i.d. error term u_{it} might reflect temporary health shocks like a cold or just random influences on the response process like the current mood. They make the signal of lagged SRHS noisy so that all lags have additional explanatory power. But since the random effects are assumed to be constant over time, each lag tends to have the same amount of information and so the predictive power of all lags should not systematically vary. This contradicts the finding in Table 4 that the explanatory power decreases with the lag length.

[5] Typically, one would actually include dummy variables for the different lagged outcomes of $y_{i,t-1}$ instead of a linear term.

Specification (4) is the preferred model of Contoyannis et al. (2004). The combination of state dependence and a random effect has the advantage that it introduces additional flexibility in the correlation structure over time. Compared to the RE model, the first lag is allowed to have additional predictive power due to the state dependence. While this fits the observed pattern in Table 4 better than the previous models, it still cannot capture the decreasing explanatory power of the higher lags.

Instead of the models discussed so far, consider a model with an AR(1) error component instead of the random effect:

$$\begin{aligned} y_{it}^* &= \mathbf{x}_{it}\boldsymbol{\beta} + a_{it} + u_{it} \\ a_{it} &= \rho a_{i,t-1} + e_{it}. \end{aligned} \tag{2}$$

For the random process a_{it} to be stationary, assume that the i.i.d. shocks e_{it} have a variance of $(1-\rho^2)\sigma_a^2$, where σ_a^2 is the variance of a_{it}. This unobserved stochastic process can be interpreted as a latent health component. Similar to the RE model, this model implies that lagged SRHS values $y_{i,t-s}$ predict outcomes because they contain information on latent health in $t-s$, but the information becomes noisier with an increasing lag length because of the accumulated shocks $e_{i,t-s}$ through e_{it}. This can explain the decreasing explanatory power found in Table 4. Note that the random effects model emerges as a special case of this model with $\rho = 1$.

Stern (1994) discusses a method of simulated moments estimator for a similar model structure in which y_{it}^* depends on lagged values $y_{i,t-1}^*$ and gives SRHS as an example for which it is appropriate. For maximum likelihood estimation, the unobserved health component has to be integrated out of the conditional outcome probability numerically just as for nonlinear RE models. While a_i and therefore the integral is univariate for RE models, the sequence $a_{i1}, \ldots, a_{iT}$ in this model is T-dimensional. Define $\mathbf{x}_i = [\mathbf{x}_{i1}, ..., \mathbf{x}_{iT}]$,

$\mathbf{y}_i = [y_{i1}, ..., y_{iT}]$ and $\mathbf{a}_i = [a_{i1}, ..., a_{iT}]$. Conditional on $\mathbf{x}_i$ and $\mathbf{a}_i$, the individual outcomes are independent over time. The conditional outcome probabilities are

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, a_i) = \prod_{t=1}^{T} \Pr(y_{it}|\mathbf{x}_{it}, a_{it}) \tag{3}$$

with $\Pr(\mathbf{y}_{it}|\mathbf{x}_{it}, a_{it})$ representing standard ordered logit probabilities for y_{it} with a_{it} as additional regressors. The evaluation of the likelihood function requires integrating the sequence of latent states out of this expression:

$$\Pr(\mathbf{y}_i|\mathbf{x}_i) = \int \cdots \int \Pr(\mathbf{y}_i|\mathbf{x}_i, a_i) f(\mathbf{a}_i|\mathbf{x}_i)\, da_{i1} \cdots da_{iT}, \tag{4}$$

where $f(\mathbf{a}_i|\mathbf{x}_i)$ is the joint p.d.f. of $\mathbf{a}_i$. The easiest approach to approximate this integral is simulation: Draw a number of vectors $\mathbf{a}_i^1, ..., \mathbf{a}_i^R$ from this joint distribution, calculate $\Pr(\mathbf{y}_i|\mathbf{x}_i, a_i^r)$ for $r = 1, ..., R$ and average the results. For a thorough discussion of estimation using approximation by simulation, see Hajivassiliou and Ruud (1994). An alternative is to use multivariate numerical integration algorithms such as Gaussian quadrature on sparse grids (Heiss and Winschel 2008).

Nonlinear Kalman filters represent an alternative method for approximation the likelihood function of models like the one at hand. Heiss (2008) discusses such an algorithm which is suitable for our purpose. It expresses the joint probability as the product of conditional time-specific probabilities

$$\Pr(\mathbf{y}_i|\mathbf{x}_i) = \Pr(y_{i1}|\mathbf{x}_i) \Pr(y_{i2}|\mathbf{x}_i, y_{i1}) \cdots \Pr(y_{iT}|\mathbf{x}_i, y_{i1}, y_{i2}, \ldots, y_{i,T-1}) \tag{5}$$

and approximates each of these terms in a sequential fashion using reweighted Gaussian quadrature. This algorithm can be a very accurate alternative to simulation (Heiss 2008).

Table 5 shows maximum likelihood results for the independent, random effects, and latent AR(1) models for the HRS sample, separately by gender. The models are se-

quentially nested: The independent model follows from the RE and AR(1) model with $\sigma = 0$ and the RE is a special case of the AR(1) model with $\rho = 1$. Both restrictions are clearly rejected by Wald and likelihood ratio tests, so the AR(1) model fits the data significantly better than the other two models.

In the AR(1) model, the idiosyncratic unobserved health component a_{it} has an estimated standard deviation of 3.0 and 2.9 for females and males, respectively. These are large numbers compared to the i.i.d. error u_{it} which by normalization of the underlying ordered logit model has a standard deviation of $\frac{\pi}{\sqrt{3}} \approx 1.8$. The estimated values of ρ are quite high with values of .941 and .946 for females and males, respectively, but significantly smaller than 1. This can be easily seen from likelihood ratio tests: The random effects model is clearly rejected with LR test statistics of 949.8 and 779.8 for females and males, respectively. While the unobserved idiosyncratic health component is highly correlated over time, it is not constant as the RE model assumes.

Table 6 shows results for the models (2) and (4) which include lagged dependent variables as regressors. Because no lagged values are known for the first observation of each individual, the estimates are based on the likelihood contributions conditional on y_{i1} for all individuals. It is well known that this together with the RE included in model (4) leads to inconsistent parameter estimates due to the initial conditions problem (Heckman 1981). The distribution of the RE depends on y_{i1} and standard random effects software ignores this dependence. As the results reported in Contoyannis et al. (2004), the results for model (4) are based on the approach of Wooldridge (2005): Instead of the usual RE assumption on the distribution of the RE, assume that the RE a_i conditional on $y_{i1} = j$ is normally distributed with mean μ_j and variance σ_a^2, so the conditional mean depends on the first observation. While this assumption is somewhat *ad hoc*, it is very convenient: the model is equivalent to a RE model with dummy variables for the initial value of the dependent variable as additional regressors.

In order to allow a comparison, Table 6 also shows results of the same latent AR(1) model as discussed above, but where the estimation is also conditional on the initial observations. To see how this conditioning is implemented, consider the joint outcome probability conditional on y_{i1}

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, y_{i1}) = \int \cdots \int \Pr(y_{i_2}, \ldots y_{i_T}|\mathbf{x}_i, \mathbf{a}_i) f(\mathbf{a}_i|\mathbf{x}_i, y_{i1})\, da_{i1} \cdots da_{iT} \tag{6}$$

because conditional on the latent health states, the dependent variables are independent over time. Note that the conditional density of the latent health process is

$$f(\mathbf{a}_i|\mathbf{x}_i, y_{i1}) = f(a_{i1}|\mathbf{x}_i, y_{i1}) f(a_{i2} \ldots a_{iT}|a_{i1}). \tag{7}$$

The joint density conditional on the initial value a_{i1} is implied by the AR(1) process assumed for latent health. The conditional distribution of a_{i1} can be expressed by Bayes' rule as

$$f(a_{i1}|\mathbf{x}_i, y_{i1}) = f(a_{i1}|\mathbf{x}_i) \frac{\Pr(y_{i1}|\mathbf{x}_i, a_{i1})}{\Pr(y_{i1}|\mathbf{x}_i)}. \tag{8}$$

This suggests to approximate the likelihood contribution using importance sampling: Given a number of R draws $\{\mathbf{a}_i^1, \ldots, \mathbf{a}_i^R\}$ from the joint distribution characterized by $f(\mathbf{a}_i)$, calculate $w_i^r = \frac{\Pr(y_{i1}|\mathbf{x}_i, a_{i1}^r)}{\sum_{s=1}^R \Pr(y_{i1}|\mathbf{x}_i, a_{i1}^s)}$ for all $r = 1, \ldots, R$. The simulated likelihood contribution is then

$$\tilde{\Pr}(\mathbf{y}_i|\mathbf{x}_i, y_{i1}) = \sum_{r=1}^{R} w_i^r \Pr(y_{i_2}, \ldots, y_{i_T}|\mathbf{x}_i, \mathbf{a}_i). \tag{9}$$

The same reweighting approach works accordingly if sequential Gaussian quadrature is used instead of simulation.

As the results in Table 6 show, the pure state dependency model (2) is clearly rejected by an LR test in favor of model (4) involving both state dependence and a random effect. These models are not nested within the latent AR(1) model, so a straightforward test

is infeasible. However, the latter model has both a better fit as measured by the log likelihood value and fewer parameters than the other models, so it would be favored by any likelihood-based criterion such as AIC or BIC.

The model parameters are not easy to interpret directly except for the signs and relative magnitudes. Before turning to substantive interpretations and simulations in section 4, the next section discusses how to add selective mortality to the latent AR(1) model.

3.2 SRHS and mortality

The fact that healthy individuals have a higher life expectancy creates a selection effect which tends to attenuate true health differentials if it is ignored. This effect was demonstrated for the age profile of SRHS in section 2, but it works also for other effects. For example the impact of a chronic disease on SRHS is underestimated if the selection through mortality is stronger for the chronically ill. And the impact of policy interventions on health is underestimated if its effect on mortality is ignored.

Contoyannis et al. (2004) and Jones et al. (2006) attempt to correct for survivorship bias in a SRHS model using an inverse probability weighting estimator for a pooled ordered probit model. This approach has the disadvantages that it cannot account for unobserved heterogeneity in the SRHS model and that selection is assumed to be random conditional on covariates. In this section, a joint model of SRHS and mortality is presented which allows for the unobserved health component as discussed in the previous section. In addition to covariates, mortality is also allowed to be driven by this latent health variable.

Let d_{it} denote an indicator which has the value 1 if individual i dies at time t and 0 otherwise. Consider the following model which adds mortality to the AR(1) model of SRHS in equations (1) and (2).

$$\begin{aligned}
y^*_{it} &= \mathbf{x}_{it}\boldsymbol{\beta}^y + \sigma^y a_{it} + u^y_{it} \\
d^*_{it} &= \mathbf{x}_{it}\boldsymbol{\beta}^d + \sigma^d a_{it} + u^d_{it} \\
y_{it} &= j \Leftrightarrow \alpha_{j-1} \leq y^*_{it} < \alpha_j \quad \text{with } 1 \leq j \leq 5 \\
d_{it} &= \mathbf{1}[d^*_{it} > \alpha^d] \\
a_{it} &= \rho a_{i,t-1} + e_{it} \\
a_{it} &\sim N(0,1), \quad e_{it} \sim N(0, 1-\rho^2)
\end{aligned} \tag{10}$$

The SRHS part of this model corresponds to the AR(1) model discussed in section 3.1. Mortality is modeled as a binary outcome which depends not only on the same set of covariates as SRHS but also on the same latent health process as SRHS. This generates a correlation of SRHS and mortality conditional on covariates and allows to capture selective mortality beyond the observed differences. While even extremely healthy individuals (in terms of covariates $\mathbf{x}_{it}\boldsymbol{\beta}^d$ and the latent health component a_{it}) can die at any time if they draw an extreme shock u^d_{it} (say a car accident), unhealthy individuals can die from much less severe shocks.

Note that y_{it} is only observed if $d_{it} = 0$. Therefore, the correlation between the remaining transitory errors u^y_{it} and u^d_{it} is not identified and is assumed to be zero. This does not affect the intertemporal correlations since these shocks are assumed to be independent over time. This allows a straightforward calculation of outcome probabilities given values of a_{it}. The conditional mortality risk is given by its binary choice probability. The joint probability of survival and SRHS of j is the product of the binary and the ordered choice probabilities. Similar models have proved to be useful for the study the trajectories of health and disability in old age (Heiss, Börsch-Supan, Hurd and

Wise 2007a) and for the modeling of the dynamics of prescription drug use (Heiss, McFadden and Winter 2007b).

The likelihood function can be approximated in the same fashion as for the SRHS model of section 3.1. But the mortality selection has worked before the initial sampling. If a respondent appears in the sample for the first time at age 90, she belongs to a relatively healthy part of the population. This has to be accounted for when calculating the likelihood function. Assuming no mortality selection before age 50, we have to condition on the outcome of no mortality until the age of initial sampling. This is the same problem as conditioning on the initial SRHS outcome as it was discussed above and the same strategy can be used. Implementations of this type of model in the software package Stata can be requested from the author.

4 Results and Simulations

Parameter estimates for the joint model of SRHS and mortality presented in section 3.2 are shown in Table 7. The latent health state a_{it} enters both the SRHS and the mortality equation highly significantly for females and males. This captures the fact that even when controlling for observable characteristics, these outcomes are correlated due to the common influence.

Because the estimated parameters have no direct quantitative interpretation, Table 8 presents estimated average partial effects of the explanatory variables on the probability of poor or fair SRHS. The first two columns show the results of the independent and AR(1) SRHS models without mortality correction. The other columns are results from the joint model and disentangle direct effects on SRHS and selection effects. The third and fourth column are the partial effects on the mortality risk over the next two years and on poor or fair SRHS, respectively. The fifth column presents the estimated survivorship bias if selective mortality is ignored. The true effect on SRHS and the selection effect are combined in the last column which now presents the estimated difference of the share

poor or fair health among the survivors. As can be expected, this is pretty much in line with the results from the models without mortality correction.

The fact that *ceteris paribus* fewer females are in poor or fair SRHS is pretty much offset by the lower mortality of women so that in the surviving population, SRHS hardly differs by gender. Selective mortality attenuates the SRHS differences by education since the low educated have both worse SRHS and a higher mortality risk. The highest differences can be seen in the age effects. Since both SRHS and mortality selection increase steeply with age, poor or fair SRHS increases much slower if only the surviving are analyzed and the survivorship bias is ignored as has already been observed in Figure 2.

Of course, these results only account for the survivorship bias, not for differences in reporting styles. To add these effects, similar approaches as discussed in the literature such as vignettes (Kapteyn, Smith and Van Soest 2007) could added to the model, but this is beyond the scope of this paper.

The joint model of SRHS and mortality is very parsimonious – in addition to the slope parameters of the explanatory variables, only three parameters capture the dynamics of SRHS and its relation to mortality. The following simulation exercises allow to check whether it is able to explain the full dynamics of the seven SRHS waves. The simulations are based on 1000 artificial data sets. Each of those consists of individuals with the same socio-economics as the initial HRS sample. For each individual, a latent health, SRHS, and mortality process is simulated through wave 7 according to the model with the estimated parameters.

Table 9 repeats the descriptive regression of Table 4 (third column) of wave 7 SRHS on socio-demographics and six lags of SRHS. Besides the results for the HRS sample, these regressions were performed on each of the 1000 artificial data sets generated according to the model with a latent AR(1) health component. The table shows means, standard deviations, the first, and the 99^{th} percentile of these 1000 point estimates. All HRS

coefficients are well within these intervals except for the coefficient of the first lag which is slightly higher in the HRS sample than for the artificial data.

Table 10 shows results of a similar exercise. It repeats the regression of mortality on socio-demographics and three lags of SRHS from Table 4 (fifth column). The coefficients of the socio-demographic variables agree very well using the HRS and the artificial samples. The coefficients of lagged SRHS only agree qualitatively. The explanatory power of the first lag is higher and then decreases faster in the HRS sample compared to the data set generated by the estimated model. But the model is able to reproduce the qualitative structure of significant but decreasing predictive power of lagged SRHS.

One possible explanation would be that there are actually two latent processes: One encompasses deadly health problems and affects both SRHS and mortality risk. A second one could be chronic but not deadly health problems and/or response styles which only affect SRHS but not mortality. Such a generalized model would allow to more flexibly explain the observed data.

Another way to compare the HRS and simulated samples is to reproduce Figure 2 for the latter. Figure 3 shows the age profile of the respondents who (are simulated to) survive to the corresponding age. The profiles look very similar. Notably, the model is able to reproduce the decreasing share of living individuals in poor or fair SRHS beyond age 90 which it explains by survivorship bias. The level in the very high ages is somewhat lower in the simulated sample, but sample sizes are very low in this area.

The interesting question is whether the model can also explain the decomposition of the pure age and the selection effect depicted in Figure 2. Figure 4 repeats these results and adds the simulated counterparts. To facilitate the comparison, the simulated SRHS levels are adjusted for the differences seen in Figure 3. The results agree strikingly well except for a few outliers where sample sizes are small in the original HRS sample.

Overall the simulation results suggest that even the very simple and parsimonious model discussed in the previous section can explain most of the dynamics of SRHS and mortality seen in the first seven waves of the HRS.

5 Conclusions

Self rated health status (SRHS) is a frequently used measure of individual health in survey data. Despite its subjectiveness, it has considerable real content. For example it is a strong predictor of mortality. This leads to a potential survivorship bias – effects of all kinds of determinants of health are confounded by their effect on selective mortality.

This paper documents the dynamics of SRHS and mortality in the first seven waves of the Health and Retirement Study (HRS). It suggests that the findings are inconsistent with typically applied econometric models and suggests a parsimonious alternative. Simulations show that this model succeeds much better in capturing the dynamic structure of the data. They agree with all correlation patterns qualitatively and in most cases also quantitatively.

The strongest survivorship bias is found for the age profile, but also the SRHS differences by gender and education are attenuated by selective mortality.

References

Adams, Peter, Michael D. Hurd, Daniel McFadden, Angela Merrill, and Tiago Ribeiro (2003) 'Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status.' *Journal of Econometrics* 112, 3–56

Baker, Michael, Mark Stabile, and Catherine Deri (2004) 'What do self-reported, objective, measures of health measure?' *Journal of Human Resources* 39(4), 1067–1093

Butler, J. S., Richard V. Burkhauser, Jean M. Mitchell, and Theodore P. Pincus (1987) 'Measurement error in self-reported health variables.' *Review of Economics & Statistics* 69, 644–650

Card, David, and Daniel Sullivan (1988) 'Measuring the effect of subsidized training programs on movements in and out of employment.' *Econometrica* 56(3), 497–530

Contoyannis, Paul, Andrew M Jones, and Nigel Rice (2004) 'The dynamics of health in the British Household Panel Survey.' *Journal of Applied Econometrics* 19, 473 – 503

Crossley, Thomas F., and Steven Kennedy (2002) 'The reliability of self-assessed health status.' *Journal of Health Economics* 21, 643–658

Deaton, Angus S., and Christina H. Paxson (1998) 'Aging and inequality in income and health.' *American Economic Review* 88(2), 248–253

Disney, Richard, Carl Emmerson, and Matthew Wakefield (2006) 'Ill health and retirement in britain: A panel data-based analysis.' *Journal of Health Economics* 25, 621–649

Gerdtham, U.-G., M. Johannesson, L. Lundberg, and D. Isacson (1999) 'A note on validating Wagstaff and van Doorslaer's health measure in the analysis of inequalities in health.' *Journal of Health Economics* 18(1), 117–124

Hajivassiliou, Vassilis A., and Paul A. Ruud (1994) 'Classical estimation methods for LDV models using simulation.' In *Handbook of Econometrics Vol. 4,* ed. Robert F. Engle and Daniel L. McFadden (New-York: Elsevier) pp. 2383–2441

Heckman, James J. (1981) 'The incidental parameters problem and the problem of initial conditions in estimating a discrete time - discrete data stochastic process.' In *Structural Analysis of Discrete Data and Econometric Applications,* ed. Charles F. Manski and Daniel McFadden (Cambridge, Mass.: MIT Press) pp. 179–195

Heiss, Florian (2008) 'Sequential numerical integration in nonlinear state space models for microeconometric panel data.' *Journal of Applied Econometrics* 23(3), 373–389

Heiss, Florian, and Viktor Winschel (2008) 'Likelihood approximation by numerical integration on sparse grids.' *Journal of Econometrics* 144(1), 62–80

Heiss, Florian, Axel Börsch-Supan, Michael Hurd, and David Wise (2007a) 'Pathways to disability: Predicting health trajectories.' In *Health at Older Ages: The Causes and Consequences of Declining Disability Among the Elderly,* ed. David Cutler and David Wise (University of Chicago Press). forthcoming

Heiss, Florian, Daniel McFadden, and Joachim Winter (2007b) 'Mind the gap! Consumer perceptions and choices of Medicare Part D prescription drug plans.' NBER Working Paper No. 13627

Jones, Andrew M., Xander Koolman, and Nigel Rice (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* 169(3), 543–569

Kapteyn, Arie, James P. Smith, and Arthur Van Soest (2007) 'Vignettes and self-reports of work disability in the United States and the Netherlands.' *American Economic Review* 97(1), 461–473

Lindeboom, Maarten, and Eddy van Doorslaer (2004) 'Cut-point shift and index shift in self-reported health.' *Journal of Health Economics* 23, 1083–1099

Smith, James P. (1999) 'Healthy bodies and thick wallets: The dual relation between health and economic status.' *Journal of Economic Perspectives* 13(2), 145–166

Stern, Steven (1994) 'Two dynamic discrete choice estimation problems and simulation method solutions.' *Review of Economics and Statistics* 76, 695–702

van Doorslaer, Eddy, and Ulf-G. Gerdtham (2003) 'Does inequality in self-assessed health predict inequality in survival by income? evidence from swedish data.' *Social Science & Medicine* 57(9), 1621–1629

Wooldridge, Jeffrey M. (2005) 'Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity.' *Journal of Applied Econometrics* 20, 39–54

Tables

Table 1: Sample: Cohorts

		First		Obs./indiv.		Total number of obs.	
Cohort*	Birth year	wave	Indiv.	Alive	Deaths [%]	Alive	Deaths
AHEAD	before 1924	2	8,294	4.0	54.06	33,073	4,484
CODA	1924–1930	4	2,380	3.4	17.35	8,115	413
HRS**	1931–1941	1	13,131	5.4	14.71	71,261	1,932
WB	1942–1947	4	2,531	3.4	3.63	8,641	92
EBB	1948–1953	7	2,777	1.0	0.00	2,777	0
Total			29,113	4.3	23.77	123,867	6,921

*: AHEAD = "The Study of Assets and Health Dynamics Among the Oldest Old", CODA = "Children of Depression Age", HRS = Initial "Health and Retirement Study" cohort, WB = "War Babies", EBB = "Early Baby Boomers"

**: The HRS row includes 113 "HRS/AHEAD overlap cases" who were born before 1924 but included in the initial HRS interview (wave 1).

Table 2: Sample: Descriptive Statistics

		Observations	Percent
Total		123,867	100.0
Self-rated health	excellent	16,672	13.5
	very good	33,681	27.2
	good	37,541	30.3
	fair	23,908	19.3
	poor	12,065	9.7
Gender	male	53,526	43.2
	female	70,341	56.8
Age	50≤age≤60	40,031	32.3
	60<age≤70	38,731	31.3
	70<age≤80	28,354	22.9
	80<age≤90	14,270	11.5
	90<age	2,481	2.0
Ethnicity	caucasian	93,744	75.7
	hispanic	9,947	8.0
	nonwhite, non-hispanic	20,176	16.3
Education	less than high school	40,724	32.9
	high school	38,661	31.2
	more than high school	44,482	35.9

Table 3: Wave-by-wave transitions: SRHS and mortality

		next wave					
SRHS	observations	1. excellent	2. v.good	3. good	4. fair	5. poor	dead
1. excellent	13,685	48.4	33.4	12.2	3.1	1.0	2.0
2. very good	27,036	12.5	49.9	27.0	6.5	1.6	2.5
3. good	29,728	3.7	20.9	48.0	18.4	4.2	4.8
4. fair	18,868	1.4	6.3	23.4	43.7	14.9	10.3
5. poor	9,752	0.6	1.8	7.1	23.7	44.2	22.6
total	99,069	11.5	25.9	28.6	18.4	9.0	6.6

Table 4: SRHS and mortality: reduced-form regressions on lagged SRHS

	SRHS, ordered logit			Mortality, binary logit		
female	-0.024	-0.087**	-0.100*	-0.464**	-0.464**	-0.779**
	(0.012)	(0.017)	(0.048)	(0.028)	(0.038)	(0.150)
age spline, 50-59	0.016**	-0.057**		0.014	-0.059	
	(0.004)	(0.013)		(0.016)	(0.050)	
age spline, 60-69	0.010**	0.009**	0.016	0.076**	0.060**	0.024
	(0.002)	(0.003)	(0.010)	(0.007)	(0.010)	(0.034)
age spline, 70-79	0.031**	0.026**	0.038*	0.077**	0.088**	0.050
	(0.003)	(0.004)	(0.015)	(0.006)	(0.007)	(0.035)
age spline, 80-89	0.017**	0.018**	-0.013	0.090**	0.088**	0.158
	(0.004)	(0.006)	(0.071)	(0.006)	(0.007)	(0.095)
age spline, 90+	-0.011	-0.008	-0.907**	0.120**	0.123**	
	(0.013)	(0.020)	(0.145)	(0.011)	(0.014)	
non-white	0.253**	0.140**	0.039	0.064	0.010	-0.072
	(0.017)	(0.024)	(0.067)	(0.038)	(0.053)	(0.175)
hispanic	0.310**	0.184**	0.203*	-0.371**	-0.428**	-0.534*
	(0.023)	(0.034)	(0.095)	(0.056)	(0.078)	(0.270)
education < HS	0.294**	0.165**	0.185**	0.038	0.069	-0.014
	(0.016)	(0.022)	(0.065)	(0.034)	(0.047)	(0.175)
education > HS	-0.210**	-0.100**	0.004	-0.037	0.023	0.044
	(0.015)	(0.020)	(0.055)	(0.037)	(0.050)	(0.180)
SRHS, lag 1	1.448**	1.001**	1.029**	0.685**	0.577**	0.927**
	(0.010)	(0.014)	(0.042)	(0.015)	(0.026)	(0.119)
SRHS, lag 2		0.574**	0.532**		0.126**	0.042
		(0.014)	(0.040)		(0.024)	(0.103)
SRHS, lag 3		0.411**	0.352**		0.074**	0.065
		(0.012)	(0.039)		(0.023)	(0.095)
SRHS, lag 4			0.263**			-0.185*
			(0.038)			(0.092)
SRHS, lag 5			0.083*			0.004
			(0.036)			(0.088)
SRHS, lag 6			0.130**			0.147
			(0.033)			(0.082)
Individuals	23,694	17,818	6,719	25,988	20,008	7,142
Observations	92,413	47,238	6,719	104,190	52,845	7,142
Deaths				6,525	3,603	235
Log likelihood	-111,015.8	-53,289.5	-7,110.9	-19,837.9	-10,651.6	-885.4

Standard errors (clustered by respondent) in parentheses.

*/**: different from 0 at 5%/1% significance level.

Table 5: SRHS models: Parameter estimates

	females			males		
	(1) Indep.	(3) RE	AR(1)	(1) Indep.	(3) RE	AR(1)
age spline, 50-59	0.032**	0.081**	0.091**	0.041**	0.096**	0.106**
	(0.003)	(0.005)	(0.006)	(0.004)	(0.006)	(0.007)
age spline, 60-69	0.021**	0.046**	0.046**	0.019**	0.059**	0.059**
	(0.003)	(0.004)	(0.005)	(0.003)	(0.004)	(0.006)
age spline, 70-79	0.040**	0.086**	0.098**	0.048**	0.095**	0.108**
	(0.003)	(0.005)	(0.006)	(0.004)	(0.006)	(0.007)
age spline, 80-89	0.036**	0.109**	0.115**	0.023**	0.102**	0.106**
	(0.005)	(0.006)	(0.008)	(0.006)	(0.009)	(0.011)
age spline, 90+	-0.036**	0.041**	0.045*	-0.093**	-0.066*	-0.080*
	(0.012)	(0.014)	(0.018)	(0.020)	(0.028)	(0.034)
non-white	0.656**	1.191**	1.351**	0.476**	0.825**	0.936**
	(0.019)	(0.057)	(0.065)	(0.023)	(0.063)	(0.072)
hispanic	0.764**	1.341**	1.518**	0.478**	0.837**	0.935**
	(0.026)	(0.077)	(0.088)	(0.030)	(0.083)	(0.095)
education < HS	0.740**	1.131**	1.327**	0.544**	0.830**	0.992**
	(0.018)	(0.053)	(0.061)	(0.021)	(0.060)	(0.069)
education > HS	-0.446**	-0.760**	-0.877**	-0.507**	-0.844**	-0.980**
	(0.017)	(0.052)	(0.060)	(0.019)	(0.059)	(0.068)
σ		2.453**	2.978**		2.365**	2.901**
		(0.019)	(0.028)		(0.021)	(0.032)
ρ			0.946**			0.941**
			(0.002)			(0.002)
Individuals	16,273	16,273	16,273	12,840	12,840	12,840
Observations	70,341	70,341	70,341	53,526	53,526	53,526
Log likelihood	-102,265.5	-86,756.3	-86,281.4	-78,601.7	-67,827.5	-67,437.6

Standard errors (clustered by respondent) in parentheses.

*/**: different from 0 at 5%/1% significance level.

Table 6: SRHS models with state dependence

	Females			Males		
	(2) State Depend.	(4) State Dep.+RE	Latent AR(1)	(2) State Depend.	State Dep.+RE	Latent AR(1)
age spline, 50-59	0.009	0.048**	0.111**	0.032**	0.079**	0.162**
	(0.005)	(0.006)	(0.011)	(0.006)	(0.008)	(0.013)
age spline, 60-69	0.012**	0.021**	0.063**	0.008*	0.026**	0.081**
	(0.003)	(0.004)	(0.006)	(0.003)	(0.004)	(0.007)
age spline, 70-79	0.029**	0.038**	0.115**	0.036**	0.044**	0.126**
	(0.003)	(0.005)	(0.007)	(0.004)	(0.006)	(0.008)
age spline, 80-89	0.019**	0.055**	0.124**	0.015*	0.049**	0.129**
	(0.005)	(0.008)	(0.009)	(0.007)	(0.010)	(0.012)
age spline, 90+	0.002	-0.010	0.061**	-0.048*	-0.064	-0.052
	(0.015)	(0.023)	(0.019)	(0.022)	(0.035)	(0.036)
non-white	0.279**	0.317**	0.878**	0.213**	0.263**	0.704**
	(0.021)	(0.040)	(0.117)	(0.027)	(0.049)	(0.127)
hispanic	0.335**	0.425**	1.169**	0.281**	0.443**	1.230**
	(0.030)	(0.058)	(0.162)	(0.035)	(0.067)	(0.165)
education < HS	0.326**	0.349**	1.162**	0.234**	0.236**	0.833**
	(0.021)	(0.038)	(0.110)	(0.024)	(0.043)	(0.121)
education > HS	-0.207**	-0.260**	-0.761**	-0.239**	-0.295**	-0.815**
	(0.019)	(0.034)	(0.110)	(0.023)	(0.039)	(0.119)
SRHS (lag) exc.	-1.570**	-0.610**		-1.379**	-0.546**	
	(0.038)	(0.041)		(0.040)	(0.044)	
SRHS (lag) good	1.369**	0.518**		1.251**	0.457**	
	(0.025)	(0.030)		(0.028)	(0.034)	
SRHS (lag) fair	2.848**	1.122**		2.660**	1.075**	
	(0.034)	(0.046)		(0.039)	(0.052)	
SRHS (lag) poor	4.443**	1.865**		4.369**	1.968**	
	(0.051)	(0.070)		(0.061)	(0.084)	
SRHS (1st) exc.		-1.257**			-1.065**	
		(0.051)			(0.054)	
SRHS (1st) good		1.003**			0.917**	
		(0.041)			(0.045)	
SRHS (1st) fair		2.126**			1.914**	
		(0.058)			(0.066)	
SRHS (1st) poor		3.319**			3.057**	
		(0.090)			(0.103)	
cut point 1	-1.539**	-1.771**	-2.588**	-1.294**	-1.354**	-1.797**
	(0.043)	(0.059)	(0.163)	(0.056)	(0.077)	(0.189)
cut point 2	0.738**	1.016**	0.730**	0.789**	1.164**	1.217**
	(0.041)	(0.059)	(0.162)	(0.055)	(0.077)	(0.189)
cut point 3	2.770**	3.484**	3.641**	2.790**	3.569**	4.080**
	(0.044)	(0.063)	(0.163)	(0.058)	(0.082)	(0.191)
cut point 4	4.807**	5.916**	6.479**	4.788**	5.922**	6.838**
	(0.050)	(0.071)	(0.166)	(0.063)	(0.091)	(0.195)
σ		1.246**	2.888**		1.196**	2.768**
		(0.026)	(0.030)		(0.030)	(0.033)
ρ			0.945**			0.939**
			(0.002)			(0.003)
Individuals	13,634	13,634	13,634	10,316	10,316	10,316
Observations	54,068	54,068	54,068	40,686	40,686	40,686
Parameters	17	22	15	17	22	15
Log likelihood	-64,453.4	-62,325.6	-62,228.7	-49,756.4	-48,293.8	-48,180.5

Standard errors (clustered by respondent) in parentheses.
*/**: different from 0 at 5%/1% significance level.

Table 7: Joint SRHS & mortality model: Parameter estimates

	females		males	
	SRHS	mortality	SRHS	mortality
age spline, 50-59	0.101**	0.049*	0.122**	0.035
	(0.006)	(0.022)	(0.008)	(0.022)
age spline, 60-69	0.065**	0.109**	0.082**	0.097**
	(0.006)	(0.010)	(0.006)	(0.009)
age spline, 70-79	0.136**	0.103**	0.159**	0.115**
	(0.006)	(0.008)	(0.008)	(0.008)
age spline, 80-89	0.168**	0.140**	0.175**	0.136**
	(0.008)	(0.008)	(0.011)	(0.009)
age spline, 90+	0.155**	0.167**	0.051	0.155**
	(0.018)	(0.013)	(0.034)	(0.021)
non-white	1.477**	0.457**	1.040**	0.316**
	(0.074)	(0.057)	(0.084)	(0.061)
hispanic	1.459**	-0.012	0.858**	-0.105
	(0.099)	(0.089)	(0.109)	(0.091)
education $<$ HS	1.575**	0.458**	1.215**	0.267**
	(0.070)	(0.053)	(0.082)	(0.057)
education $>$ HS	-0.941**	-0.236**	-1.109**	-0.346**
	(0.066)	(0.058)	(0.078)	(0.060)
latent health (σ)	3.172**	0.942**	3.152**	0.929**
	(0.031)	(0.028)	(0.036)	(0.029)
ρ	0.949**		0.944**	
	(0.002)		(0.002)	
Individuals	16,337		12,857	
Observations (mortality)	79,460		61,695	
Observations (SRHS)	70,341		53,526	
Log likelihood	-97,129.5		-77,676.6	

Table 8: Average partial effects [percentage points]

	SRHS models		SRHS & mortality model			
	Indep.	AR(1)	Effect on		Selection	SRHS &
			Mortality	SRHS	effect	selection
Gender: reference = male						
Female	0.58	-0.57	-2.07	-2.64	2.55	-0.08
Race: reference = non-hispanic white						
Non-white	11.51	11.53	1.43	13.26	-1.93	11.33
Hispanic	12.86	12.53	-0.19	10.36	2.18	12.54
Education: reference = high school degree						
Education < HS	14.04	12.28	1.37	15.49	-1.82	13.67
Education > HS	-8.06	-8.26	-0.93	-10.82	2.56	-8.26
Age: increase by two years, from current age						
50-59	1.10	1.50	0.13	1.98	-0.79	1.18
60-69	0.80	1.06	0.53	2.91	-2.28	0.63
70-79	1.75	2.16	1.18	7.94	-5.87	2.07
80-89	1.20	2.33	2.59	13.95	-12.28	1.67
90+	-2.32	0.18	2.14	24.55	-28.43	-3.87
Total	1.09	1.58	0.82	5.46	-4.30	1.17

Table 9: Reduced-form SRHS regression: HRS vs. simulated samples

	HRS sample		Simulated data		
	est.	SE	mean	SD	[1%; 99%]
female	-0.099	(0.048)	-0.060	(0.046)	[-0.172; 0.054]
age spline, 60-69	0.016	(0.010)	0.002	(0.010)	[-0.022; 0.026]
age spline, 70-79	0.040	(0.015)	0.042	(0.015)	[0.007; 0.078]
age spline, 80-89	-0.052	(0.080)	-0.005	(0.063)	[-0.153; 0.158]
non-white	0.040	(0.067)	0.129	(0.066)	[-0.019; 0.285]
hispanic	0.197	(0.095)	0.133	(0.091)	[-0.061; 0.364]
education $<$ HS	0.184	(0.065)	0.139	(0.062)	[-0.006; 0.283]
education $>$ HS	0.004	(0.055)	-0.111	(0.057)	[-0.237; 0.012]
SRHS, lag 1	1.029	(0.042)	0.822	(0.031)	[0.752; 0.892]*
SRHS, lag 2	0.531	(0.040)	0.493	(0.031)	[0.421; 0.562]
SRHS, lag 3	0.351	(0.039)	0.300	(0.031)	[0.230; 0.375]
SRHS, lag 4	0.263	(0.038)	0.190	(0.033)	[0.116; 0.268]
SRHS, lag 5	0.084	(0.036)	0.125	(0.031)	[0.050; 0.195]
SRHS, lag 6	0.129	(0.033)	0.096	(0.030)	[0.026; 0.170]

*: Est. parameter for HRS $\notin$ simulated [1%;99%] interval

Table 10: Reduced-form mortality regression: HRS vs. simulated samples

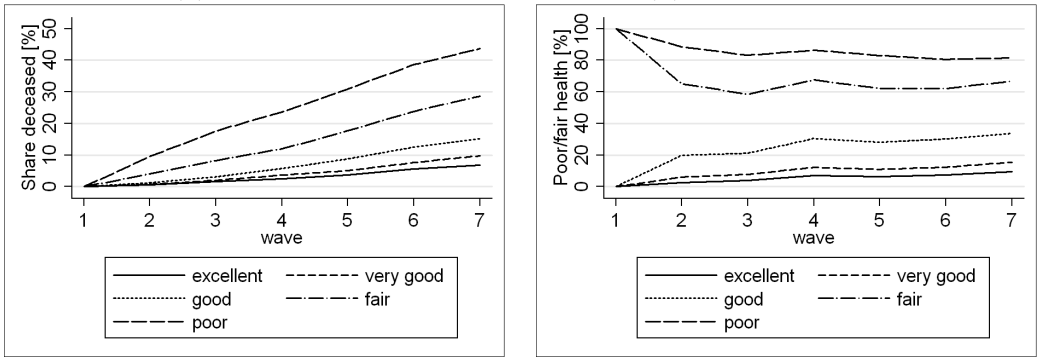
	HRS sample		Simulated data			
	est.	SE	mean	SD	[1%;	99%]
female	-0.464	(0.038)	-0.453	(0.036)	[-0.535;	-0.370]
age spline, 50-59	-0.059	(0.050)	0.020	(0.050)	[-0.093;	0.136]
age spline, 60-69	0.060	(0.010)	0.081	(0.010)	[0.059;	0.106]
age spline, 70-79	0.088	(0.007)	0.076	(0.007)	[0.058;	0.093]
age spline, 80-89	0.088	(0.007)	0.091	(0.008)	[0.074;	0.110]
age spline, 90+	0.123	(0.014)	0.118	(0.013)	[0.090;	0.147]
non-white	0.010	(0.053)	0.083	(0.048)	[-0.025;	0.199]
hispanic	-0.428	(0.078)	-0.307	(0.072)	[-0.468;	-0.141]
education $<$ HS	0.069	(0.047)	0.025	(0.044)	[-0.088;	0.120]
education $>$ HS	0.023	(0.050)	-0.036	(0.048)	[-0.151;	0.067]
SRHS, lag 1	0.577	(0.026)	0.344	(0.021)	[0.289;	0.389]*
SRHS, lag 2	0.126	(0.024)	0.224	(0.021)	[0.175;	0.274]*
SRHS, lag 3	0.074	(0.023)	0.169	(0.022)	[0.116;	0.217]*

*: Est. parameter for HRS $\notin$ simulated [1%;99%] interval

Figures

Figure 1: Mortality and SRHS by SRHS in wave 1

(a) mortality (b) SRHS of survivors



Original HRS cohort

Figure 2: SRHS by age and survivor status at different ages

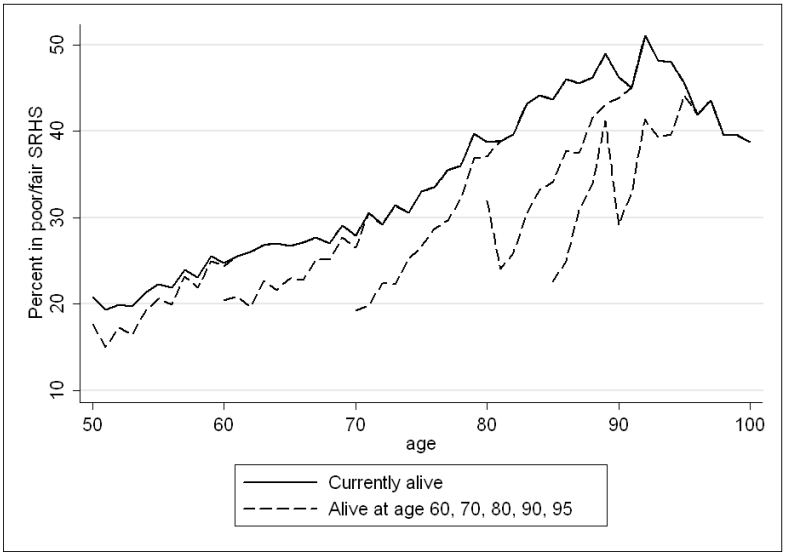


Figure 3: Age profiles of poor/fair SRHS: HRS vs. simulation

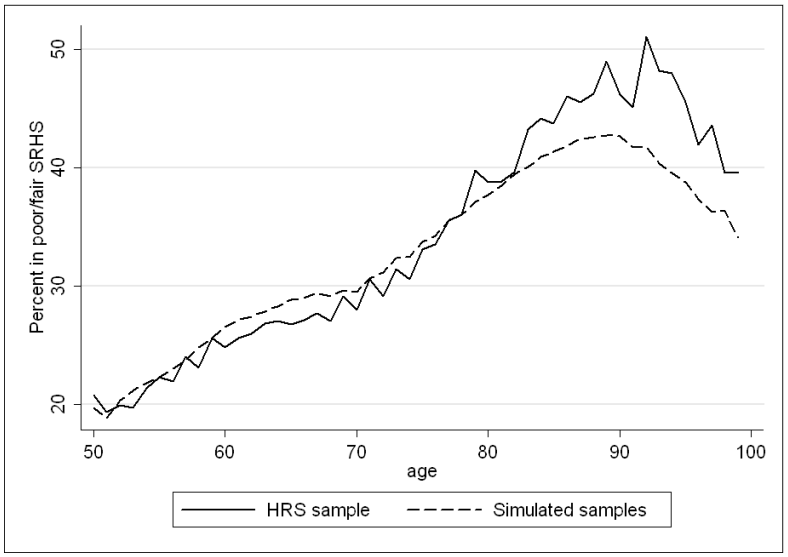


Figure 4: Simulation: SRHS by age and survivor status at different ages