HEDG

HEALTH, ECONOMETRICS AND DATA GROUP

WP 18/26

The effects of self-assessed health: Dealing with and understanding misclassification bias

Linkun Chen; Philip M. Clarke; Dennis J. Petrie and

Kevin E. Staub

August 2018

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

The effects of self-assessed health: Dealing with and understanding misclassification bias*

Linkun Chen[†] University of Melbourne
Philip M. Clarke[‡] University of Oxford
Dennis J. Petrie[§] Monash University
Kevin E. Staub[¶] University of Melbourne

July 13, 2018

Abstract

Categories of self-assessed health (SAH) are often used as a measure of health status. However, the difficulties with measuring overall health mean that the same individual may select into different SAH categories even though their underlying health has not changed. Thus, their observed SAH may involve misclassification, and the chance of misclassification may differ across individuals. As shown in this paper, if neglected, misclassification can lead to substantial biases in not only the estimation of the effects of SAH on outcomes, but also on the effects of other variables of interest, such as education and income. This paper studies nonlinear regression models where SAH is a key explanatory variable, but where two potentially misclassified measures of SAH are available. In contrast to linear regression models, the standard approach of using one SAH measure as an instrumental variable for the other cannot produce consistent estimates. However, we show that the coefficients can be identified from the joint distribution of the outcome and the two misclassified measures without imposing additional structure on the misclassification, and we propose simple likelihood-based approaches to estimate all parameters consistently via a convenient EM algorithm. The estimator is applied to data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey, where we exploit the natural experiment that in some waves individuals were asked the same question about their health status twice, and almost 30% of respondents change their SAH response. We use the estimator to (*i*) obtain the first reliable estimates of the relationship between SAH and long-term mortality and morbidity, and to (*ii*) document how demographic and socio-economic determinants shape patterns of misclassification of SAH.

Keywords: Misreporting; measurement error; multinomial regressor; discrete and limited dependent variables; subjective health; mortality; chronic conditions.
JEL classification: C35; I12.

Acknowledgements: We thank Denzil Fiebig, Bill Griffiths, Joe Hirschberg, Maarten Lindeboom, Jenny Lye, Frank Windmeijer, the participants of the Asian Meeting of the Econometric Society (Hong Kong), the China Meeting of the Econometric Society (Wuhan), the International Association for Applied Econometrics conference (Sapporo), the Australian Health Economics Society conference (Freemantle, WA), the Health and Wellbeing Workshop (Werribee, VIC) and seminar participants at Erasmus University for helpful comments. Staub acknowledges support from the Australian Research Council through grant DE170100644. Petrie acknowledges support from the Australian Research Council through grant DE150100309. Alex Ballantyne and Edwin Chan provided excellent research assistance.

[†]Melbourne School of Population and Global Health, 207 Bouverie St, The University of Melbourne 3010 VIC, Australia

[‡]Health Economics Research Centre, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK. E-mail: philip.clarke@unimelb.edu.au

[§]Centre for Health Economics, Monash Business School, 15 Innovation Walk, Monash University 3800 VIC, Australia. E-mail: dennis.petrie@monash.edu

[¶]Department of Economics, 111 Barry St, The University of Melbourne, 3010 VIC, Australia. E-mail: kevin.staub@unimelb.edu.au

1 Introduction

Self-reported health (SAH) is one of the most frequently used health measures in population and health-related surveys as well as in social science research (Au & Johnston, 2014). It is often asked as a simple question, for instance, "in general, how would you rate your health?", and respondents can select from a few categories such as excellent, very good, good, fair or poor. Among all empirical studies that involve the use of SAH, a particular field that is of great importance is the role of SAH in predicting on mortality. Studies have found that SAH can predict survival within the entire population (Idler & Angel, 1990; Kaplan *et al.*, 2007; Doiron *et al.*, 2015), the working age population (Miilunpalo *et al.*, 1997), the elderly population (DeSalvo *et al.*, 2005; McCallum *et al.*, 1994; Mossey & Shapiro, 1982), and other representative community samples (DeSalvo *et al.*, 2006; Idler & Benyamini, 1997), even after controlling for other specific illnesses, comorbidities and disability.

While there is strong evidence that poor compared to excellent self-reported health is associated with increased mortality, most of the above-mentioned studies do not take into consideration the possible measurement error or misclassification in SAH, which in turn will bias the estimated odds ratios of people in different health-rating categories (Baker *et al.*, 2004; Butler *et al.*, 1987; Clarke & Ryan, 2006; Black *et al.*, 2016; Crossley & Kennedy, 2002). In this paper, we use the availability of a repeated measure of self-perceived health to estimate models of future morbidity and mortality that adjust for the potential misclassification of SAH. Our analysis takes advantage of the Household, Income and Labour Dynamics in Australia (HILDA) Survey, which records the same individual's SAH responses in two different but similar questionnaires in the same wave, allowing us to quantify the extent of measurement error in SAH in both questionnaire modes. We document response changes to the SAH questions in waves where it was asked twice in two different modes, by face-to-face or over-the-phone interview, and on a self-completion questionnaire.

In the presence of classical measurement error in a regressor, the instrumental variable method where one mismeasured variable is used to instrument for another mismeasured variable can resolve the issue of mismeaserument and obtain an unbiased estimate of the effect of the correctly measured variable. However, the measurement error in *categorical variables* such as SAH is not classical because the error in one misclassified variable will be correlated with the error in the other misclassified variable. Moreover, health outcome models are often nonlinear, such as discrete choice, duration or count models. Thus, conventional tools such as two stage least squares are not applicable. We build on recent advances in the econometric literature on misclassification (Kane *et al.*, 1999; Mahajan, 2006; Lewbel, 2007; Hu & Schennach, 2008; Hu, 2008; Lindeboom & Kerkhofs, 2009; Battistin *et al.*, 2014; Gosling & Saloniki, 2014) and show that under suitable independence conditions the coefficients of interest in the most commonly used nonlinear parametric models are identified from the joint distribution of the outcome and the two misclassified measures. Our approach is closely related to Hu (2008) and Battistin *et al.* (2014) in that the identification leads to a finite mixture type model. With the exception of Hu (2008) and Lindeboom & Kerkhofs (2009), this literature considers misreporting only in a binary regressor, rather than a general categorical regressor. We further innovate in a number of ways. Our application is the first in the literature to allow for flexible effects of a categorical regressor across individual

characteristics (interaction terms with the unobserved true health categories). Moreover, we are the first to consider the estimation of a system of outcomes, rather than a single univariate outcome, and this has advantages as it can significantly reduce small sample bias and improve efficiency. In our application, for instance, we will jointly estimate the relationship between SAH and both mortality and, conditional on survival, morbidity. We show how to implement these methods with simple likelihood-based estimators via convenient and robust Expectation-Maximisation (EM) algorithms. In particular, we show that using a penalised likelihood approach can further reduce potential low power issues. While the approach we pursue is parametric in the outcome, it is extremely flexible in terms of allowing almost any patterns of misclassification and where the level of misclassification may differ across individual characteristics.

The estimation of models that correct for misclassification in categorical regressors is notoriously difficult as there can be relatively many misclassification parameters and, for some of the misclassification errors, the information in the data available might be scarce. Almost nothing is known about the finite sample performance of these estimators. To study this, we conduct an extensive Monte Carlo simulation. The results of the simulation indicate that for more demanding misclassification models with sample sizes of 1,000 observations, estimates can exhibit substantial small sample bias. For sample sizes of 10,000 observations, the coefficients of interest are estimated with essentially zero bias. The results also show that using multivariate outcomes can reduce small sample biases significantly due to both outcomes providing additional information which improves the precision with which the misclassification is estimated. A second goal of the simulation is to compare the performance of the consistent misclassification estimator to other potential approaches that are commonly employed by practitioners. First, we compare it to ad-hoc measures to reduce misclassification bias such as using the average of the two SAH measures as a regressor, or restricting the estimation to individuals who reported the same SAH twice. We also compare estimators that use instrumental variables and control function approaches to deal with misclassification. While these approaches are, in general, inconsistent for non-linear models with discrete endogenous regressors, the approach of including first-stage predictions into the second stage is used often by practitioners, nevertheless. The control function approach has been shown to be consistent for some specific class of models (Terza *et al.*, 2008), and its use has been advocated as means to reduce endogeneity bias in some contexts (Basu & Coe, 2015; Wooldridge, 2014). However, our simulation results suggest that their use cannot be recommended, in general, for the case of endogeneity stemming from misclassification bias.

In our application using the Household, Income and Labour Dynamics in Australia (HILDA) Survey, we estimate the relationship between SAH, measured in 2001, on the 15-year mortality probability and the development of chronic health conditions in 2016 using a sample of 12,908 individuals. SAH contains five categories and is measured twice in 2001. Around 30 percent of respondents do not answer with the same category. Our results represent the first reliable estimates of the mortality rate differences between categories, which corrects for the impact of this misreporting on the effects of SAH. We find that, compared to the naïve approach using one SAH measure, the corrected estimates are between 10 to 20 percent larger. However, in models where the effect of health is heterogeneous with respect to key regressors, differences can range from 30 to almost 100 percent. We find strong evidence for the

presence of misreporting, and for heterogeneity in misreporting behaviour across different population subgroups. In particular, older individuals and low-income individuals tend to overstate their true health status. Highly-educated individuals also overstate health, while individuals with low education levels tend to understate it.

In Section 2 we introduce the econometric models and estimators in the setting of a logit model with a binary health indicator. Section 3 contains simulation results, Section 4 the application, and Section 5 some concluding remarks.

2 Methods

To fix ideas, introduce notation, and give an intuition about the identification of the model, this section starts out by discussing a minimal example of a logit model with a binary potentially misreported regressor. The general model and our estimation approach is then introduced in Section 2.2, and we discuss potential ways of improving the estimation by increasing statistical power in Section 2.3.

2.1 A logit model with misreported binary health

Consider the estimation of a simple logit model for mortality, an outcome we will use in our application in Section 4. The outcome y_i equals 1 if individual i is alive 15 years after the initial survey, and 0 otherwise. We are interested in how health, h_i^*, at the time of the initial survey, is related to mortality y_i. For now, let health be a binary variable: $h_i^*=1$ indicates that individual i is in good health; and $h_i^*=0$ that i is in bad health. The key feature of the models we consider is that the true health status, h_i^*, is unknown; what is known instead is an individual's self-reported health status, and this reported health might be misclassified. Each individual reports his health twice, thus providing two potentially misclassified measures of his health status. True health is related to mortality as follows:

$$y_i = \mathbb{1}(\alpha\, h_i^* + \beta_0 + \varepsilon_i > 0), \quad i = 1, \ldots, N, \tag{1}$$

where a key object of interest is the unknown scalar coefficient α. The model includes a constant, β_0, which we will later generalise to a $K \times 1$ vector of covariates $\boldsymbol{x}_i$ with conforming parameter vector $\boldsymbol{\beta}$. The term ε_i is an IID logistically distributed idiosyncratic error. Thus, the probability of survival as a function of health status is

$$P(y_i = 1|h_i^*) = \frac{\exp(\alpha h_i^* + \beta_0)}{1 + \exp(\alpha h_i^* + \beta_0)} \equiv \Lambda(\alpha h_i^* + \beta_0). \tag{2}$$

If h_i^* were observed, (2) would serve as the basis for a standard logit estimation; but since h_i^* is unobserved, this is infeasible. Instead, we consider conditions under which we can estimate α and β_0 by using two potentially misclassified health measures.

Let the measures of reported health be denoted as h_{1i} and h_{2i}, corresponding to the first and second response of the individuals, respectively. We define the following misclassification probabilities—i.e.,

conditional probabilities of misreporting true health—as

$$\delta^m_{0|1} = P(h_{mi} = 0|h^*_i = 1) \quad \text{and} \quad \delta^m_{1|0} = P(h_{mi} = 1|h^*_i = 0), \quad \text{for } m = 1, 2. \tag{3}$$

We denote the distribution of the true health status as

$$P(h^*_i = 1) \equiv \pi. \tag{4}$$

The marginal distributions of the observed health measures can then be expressed as functions of the parameters defined in equations (3) and (4):

$$P(h_{mi} = 1) = \pi_i(1 - \delta^m_{0|1}) + (1 - \pi_i)\delta^m_{1|0}. \tag{5}$$

Not observing h^*_i, we will identify and estimate the parameters of the outcome equation (2) using the structure provided by equations (2) and (3), and the data (y_i, h_{1i}, h_{2i}). Specifically, we will use the structure to derive the joint distribution of y_i, h_{1i}, h_{2i}. To be able to do this, we need to make some assumptions about the relationship between the health measures and the outcome beyond the model structure we just defined:

> Independence Assumption (IA): Conditional on the true health status h^*_i, the reported measures, h_{1i} and h_{2i}, are independent of each other and of the outcome, y_i.

The joint distribution of outcome and the two misreported health measures consists of the eight probabilities $P(y_i{=}r_0,\, h_{1i}{=}r_1,\, h_{2i}{=}r_2 \,|\, \boldsymbol{x}_i) \equiv F(r_0, r_1, r_2)$, where $r_0 \in \{0, 1\}$, $r_1 \in \{0, 1\}$, $r_2 \in \{0, 1\}$. Then,

$$\begin{aligned} F(r_0, r_1, r_2) &= \pi\, F(r_0, r_1, r_2|h^*_i{=}1) \quad + \quad (1 - \pi)\, F(r_0, r_1, r_2|h^*_i{=}0) \\ &= \pi\, F(r_0|h^*_i{=}1)\, F(r_1|h^*_i{=}1)\, F(r_2|h^*_i{=}1) \\ &\quad + (1 - \pi)\, F(r_0|h^*_i{=}0)\, F(r_1|h^*_i{=}0)\, F(r_2|h^*_i{=}0), \end{aligned} \tag{6}$$

where

$$\begin{aligned} F(r_m|h^*_i = 1) &= (\delta^m_{0|1})^{1-r_m}\, (1 - \delta^m_{0|1})^{r_m}, \\ F(r_m|h^*_i{=}0) &= (\delta^m_{1|0})^{r_m} \quad (1 - \delta^m_{1|0})^{1-r_m}, \\ F(r_0|h^*_i{=}1) &= \Lambda(\alpha + \beta_0)^{r_0}\, \left(1 - \Lambda(\alpha + \beta_0)\right)^{1-r_0}, \\ F(r_0|h^*_i{=}0) &= \Lambda(\beta_0)^{r_0}\, \left(1 - \Lambda(\beta_0)\right)^{1-r_0}. \end{aligned}$$

The second equality in (6) follows from the independence assumption (IA). To see an example of one of the expressions in (6), consider $F(1, 1, 1)$:

$$\begin{aligned} F(1,1,1) &= P(y_i = 1, h_{1i} = 1, h_{2i} = 1 \,|\, \boldsymbol{x}_i) = \pi F(1,1,1|h^*_i = 1) + (1 - \pi)F(1,1,1|h^*_i = 0) \\ &= \pi\, \Lambda(\alpha + \beta_0)\, (1 - \delta^1_{0|1})\, (1 - \delta^2_{0|1}) \quad + \quad (1 - \pi)\, \Lambda(\beta_0)\, \delta^1_{1|0}\, \delta^2_{1|0}. \end{aligned}$$

The model fulfils a necessary condition for identification since the data provides seven linearly independent quantities $F(r_0, r_1, r_2)$, which we can map to the seven parameters of the model: $\alpha, \beta_0, \pi, \delta^1_{0|1}, \delta^1_{1|0}, \delta^2_{0|1}, \delta^2_{1|0}$. To obtain a unique solution and identify the parameters, we require for each measure that the probability of reporting truthfully be greater than the probability of misreporting:

$$\delta^m_{1|1} > \delta^m_{0|1} \qquad \text{and} \qquad \delta^m_{0|0} > \delta^m_{1|0}, \tag{7}$$

which amounts to assuming that $\delta^m_{0|1}, \delta^m_{1|0} < 0.5$.[1] With this condition, we rule out the mirror solution in which probabilities of misreporting and correctly reporting are switched.

Thus, under IA and the no-mirror-solution condition, the system is just-identified, paving the way for estimation. If only one health measure, say h_{1i}, was available, the joint distribution (y_i, h_{1i}) would consist of three independent probabilities. However, there would be five parameters to estimate— $\pi, \delta^1_{0|1}, \delta^1_{1|0}, \alpha, \beta_0$— and the system would be under-identified. Similarly, with two health measures but without the outcome y_i it would also be impossible to identify the misclassification probabilities. There would only be the three independent probabilities of the joint distribution of (h_{1i}, h_{2i}) to estimate the four parameters $\delta^1_{0|1}, \delta^1_{1|0}, \delta^2_{0|1}, \delta^2_{1|0}$ (or five, including π).

Introducing covariates is straightforward. The constant β_0 can be replaced by a linear index $\boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{x}_i$ is a $K \times 1$ vector of covariates with conforming coefficient vector $\boldsymbol{\beta}$. The joint distribution in (6) and the corresponding expressions are then simply to be taken conditional on $\boldsymbol{x}_i$. The number of parameters to be estimated is then $6+K$ (the five probabilities $\pi, \delta^1_{0|1}, \delta^2_{0|1}, \delta^1_{1|0}, \delta^2_{1|0}$, the key parameter of interest α, as well as the K elements in $\boldsymbol{\beta}$). In this case the system is over-identified since there will be at least $(1+2^{K-1}) \times 7$ different values of $F(r_0, r_1, r_2|\boldsymbol{x}_i)$, the number $(1+2^{K-1}) \times 7$ corresponding to the minimal case of a constant and $K-1$ linearly independent binary regressors.

Having covariates, it is also possible to revisit the independence assumption. The independence assumption is strong, but it might be reasonable in some contexts. For instance, in the field of health economics, Gosling & Saloniki (2014) invoke it in an application to misreported binary disability status. The independence assumption can easily be violated though. For example, if men and women have different misreporting probabilities the assumption does not hold because the two misreported measures will be dependent through the impact of gender. Thus, a way to weaken this assumption is to explicitly make the misclassification probabilities dependent on x_i and only require independence to hold conditional on some x_i.

> Conditional Independence Assumption (CIA): Conditional on the true health status h^*_i and on observed variables $\boldsymbol{x}_i$, the reported measures, h_{1i} and h_{2i}, are independent of each other and of the outcome, y_i.

Is the model still identified under CIA and the no-mirror-solution condition (7)? Consider first the case of discrete regressors $\boldsymbol{x}_i$. In this case, we know from above that we could identify the parameters for each subsample defined by one particular set of values of $\boldsymbol{x}_i$. Thus, the identification of the model

[1] See Hu (2008) for alternative identifying assumptions.

under CIA is equivalent to the identification of each subsample under the constraint that α is the same across subsamples.

When some of the regressors are continuous, the model is no longer identified without further assumptions. One way to proceed is to use the framework of Hu (2008), where semiparametric identification is achieved under quite general additional assumptions on the misclassification. However, having such a flexible framework for the misclassification also has the tradeoff of increasing small sample bias and becoming computationally intensive when there are multiple regressors over which the level of misclassifications may vary. Instead, we proceed by using a more parametric, and hence restrictive, approach for the misclassification, which has the advantage of reducing the impact of small sample bias and easily being able to incorporate many regressors in the misclassification equations. With our approach one can also easily increase the flexibility of the parametric form in the misclalssification equations to test the sensitivity of the results to a particular functional form. We assume that the misclassification probabilities are known functions of the regressors,

$$\delta^m_{0|1} = \Lambda\big(-\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}^m_{0|1})\big), \qquad \text{and} \qquad \delta^m_{1|0} = \Lambda\big(-\exp(\boldsymbol{x}_i'\gamma^m_{1|0})\big). \tag{8}$$

The logistic function coupled with the negative exponential function in (8) enforces the (0,0.5)-bounds under the no-mirror-solution condition on the misclassification probabilities. Similarly, we assume that true health is also a known function of the regressors,

$$\pi_i \equiv P(h^*_i = 1|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\eta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\eta})}. \tag{9}$$

That is, we assume true health conditional on covariates to be of the logit form.

2.2 The general model and its estimation

The general model with categorical health

For our application, we are interested in extending the framework in two important directions. First, we want to be able to deal flexibly with categorical health measures with more than two categories. Second, we want to be able to model interaction effects in unobserved health.

For concreteness, let us assume health has five outcomes, $h^*_i \in \{0,1,2,3,4\}$, as this is the case in our application of Section 4. As before, two potentially misclassified measures, h_{1i}, h_{2i} are observed. The model is now

$$y_i = \mathbb{1}(\boldsymbol{d}^{*'}_i\boldsymbol{\alpha} + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0), \quad i = 1,\ldots,N \tag{10}$$

with $\boldsymbol{\alpha} = (\alpha_1,\alpha_2,\alpha_3,\alpha_4)'$ and $\boldsymbol{d}^*_i = (d^*_{1i}, d^*_{2i}, d^*_{3i}, d^*_{4i})'$. The elements of the latter are indicators of a particular true health status:

$$d^*_{ji} = \mathbb{1}(h^*_i = j), \quad \text{for } j = 1,2,3,4.$$

I.e. there are $4+K$ parameters from the outcome equation (10). There are now twenty misreporting probabilities per measure h_{mi}, $m=1,2$, which we denote as

$$\delta^m_{k|j} = P(h_{mi}=k|h^*_i=j) \quad \forall j,k=0,1\ldots,4, \text{ and } j\neq k.$$

In addition, there are four parameters of the distribution of the true health status, $\boldsymbol{\pi}=(\pi_1,\pi_2,\pi_3,\pi_4)'$, where $\pi_j \equiv P(h^*_i=j)$. All in all, these are 44 parameters. Thus, the grand total is $48+K$ parameters to be estimated. Without covariates, for instance, that is 49 parameters.

As before, we can base identification and estimation on the joint distribution of (y_i,h_{1i},h_{2i}). The joint probabilities $P(y_i=r_0,\, h_{1i}=r_1,\, h_{2i}=r_2\,|\,\boldsymbol{x}_i) \equiv F(r_0,r_1,r_2)$ are now defined for $r_0\in\{0,1\}$, $r_1\in\{0,1,2,3,4\}$, $r_2\in\{0,1,2,3,4\}$. Thus, the joint distribution has $2\times5\times5=50$ support points, of which the last one is not linearly independent. The other 49 points will provide the necessary equations to identify the 49 parameters in the case without covariates. With covariates, the system is overidentified as before. The pendant to equation (6) in the ordinal case and other details of the general model are given in the appendix (Appendix A.1).

The condition to avoid mirror solutions in the case with multiple categories of SAH is that the probability of truthfully reporting a health level j ($\delta^m_{j|j}$) is larger than any probability of misreporting it:

$$\delta^m_{j|j} > \delta^m_{k|j}, \quad \forall j,k, \tag{11}$$

which is a generalisation of the condition (7) for the case of two categories. To implement this constraint in the estimation, we use multinomial logit-based expressions similar to the ones above (see Appendix A.1).

The model with a categorical regressor (10) is an important generalisation with respect to the binary case. The only similar application of a model for categorical regressors in the literature (Hu, 2008) imposes linearity in the effect of the categorical regressor, whereas (10) allows the effect of h^*_i to be completely flexible. With our proposed approach we can even go one step beyond and also accommodate interaction terms between all or some of the regressors $\boldsymbol{x}_i$ and unobserved health. The model with interactions in true categorical health status is

$$y_i = \mathbb{1}\left(\sum_{j=1}^{J} d^*_{ji}\alpha_j + \sum_{j=1}^{J} d^*_{ij}x_{ki}\alpha_{j,x} + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0\right), \quad i=1,\ldots,N, \tag{12}$$

for some variable of interest x_{ki} such as education. To gain an intuition for the identification of the interaction effect, we imagine again that the regressors are discrete and that $\boldsymbol{x}_i$ is fully saturated. In that case, the interaction effects are obtained by simply estimating the model separately by each subsample without imposing the restriction that the slopes on $\boldsymbol{x}_i$ be the same across subsamples.

Before discussing estimation of the model, we note that the approach, which we have presented for a logit binary outcome, can be extended to many common nonlinear models that follow the form

$$f(y_i|h^*_i,\boldsymbol{x}_i) = g(\boldsymbol{d}^{*'}_i\boldsymbol{\alpha} + \boldsymbol{x}_i'\boldsymbol{\beta};\, \omega), \tag{13}$$

where $f(\cdot|\cdot)$ is a functional of the conditional distribution of y_i given true health h_i^* (captured by the indicator variables in $\boldsymbol{d}_i^*$) and $\boldsymbol{x}_i$, and $g(\cdot)$ is a known nonlinear function, which might include ancillary parameters ω. Typical examples for $f(y_i|h_i^*, \boldsymbol{x}_i)$ include it being a survival rate (probability) as in this application, the time until developing a health condition (hazard rate), the number of doctor visits (count), or expenditures for health care (nonlinear expectation). Appendix A.2 presents details for Poisson count and Weibull duration models.

Estimating the model via the EM algorithm

The model can be estimated in a number of ways based on based on the joint distribution function (6), which takes the form of a finite mixture (FM) or latent class model. Appendix A.3 presents a GMM estimator. We found that, especially in models with categorical health and several regressors, maximum likelihood estimation via the Expectation-Maximisation (EM) algorithm was substantially faster and more stable, making it our estimator of choice.

The maximum likelihood estimator of the model is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}; y_i, h_{1i}, h_{2i}, \boldsymbol{x}_i) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \sum_{r_0} \sum_{r_1} \sum_{r_2} I_i^{r_0 r_1 r_2} \ln(F(r_0, r_1, r_2)), \tag{14}$$

where $\boldsymbol{\theta}$ collects all the parameters: $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\gamma}_{k|j}^m$ for $m = 1, 2$ and $j \neq k$. Maximisation can, in principle, proceed by a standard Newton-Raphson procedure. However, as in the case of GMM, finite mixture models can be difficult to estimate by direct maximum likelihood. We obtain the parameters via the EM algorithm, a more stable alternative, by iterating between the maximisation or M-step, and the expectation or E-step. The nth iteration of the M-step is

$$\hat{\boldsymbol{\theta}}^n = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \tilde{\ell}_i(\boldsymbol{\theta}; y_i, h_{1i}, h_{2i}, \boldsymbol{x}_i, \hat{w}_i^n), \tag{15}$$

where

$$\tilde{\ell}_i(\cdot) = \tag{16}$$
$$\sum_{j=0}^{4} \hat{w}_{ji}^n \left(\ln F(y_i|h_i^*=j, \boldsymbol{x}_i) + \ln F(h_{1i}|h_i^*=j, \boldsymbol{x}_i) + \ln F(h_{2i}|h_i^*=j, \boldsymbol{x}_i) + \ln \pi_{ji} - \ln \hat{w}_{ji}^n \right),$$

and all $F(\cdot|\cdot)$ correspond to terms like those defined in (6), and the $\hat{w}_{ji}^n$ are estimates of the posterior probabilities $P(h^* = j|y_i, h_{1i}, h_{2i}, \boldsymbol{x}_i)$. In the $(n{+}1)$th iteration of the E-step, we update these posterior probabilities as follows:

$$\hat{w}_{ji}^{n+1} = \frac{\hat{\pi}_{ji}^n \, \hat{F}^n(y_i|h_i^*=j) \, \hat{F}^n(h_{1i}|h_i^*=j) \, \hat{F}^n(h_{2i}|h_i^*=j)}{\sum_{j=0}^{4} \hat{\pi}_{ji}^n \, \hat{F}^n(y_i|h_i^*=j) \, \hat{F}^n(h_{1i}|h_i^*=j) \, \hat{F}^n(h_{2i}|h_i^*=j)}, \tag{17}$$

where all $\hat{F}^n(\cdot|\cdot)$ correspond to terms similar to the ones in (6) and are evaluated at $\hat{\boldsymbol{\theta}}^n$.

The increased stability and speed of EM comes from the fact that, first, as opposed to the likelihood $\ell_i(\cdot)$, in $\tilde{\ell}_i(\cdot)$ of the M-step, the logarithm goes through the sum of the finite mixture components of the

joint distribution $F(y, h_1, h_2)$; and, second, these components depend on separate sets of parameters, meaning that each can be estimated separately: the first term in the parentheses, $F(y_i|h_i^* = j, \boldsymbol{x}_i)$ is a function only of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$; the second and third are functions of all the $\boldsymbol{\gamma}^1_{k|j}$ and $\boldsymbol{\gamma}^2_{k|j}$ vectors (with $j \neq k$), respectively; and $\pi_i = \pi(\boldsymbol{x}_i)$ is a function only of $\boldsymbol{\eta}$.

2.3 Improving finite sample performance

A potential issue in the estimation of models with a flexible parametrisation of the misclassification system as proposed here is that in finite samples there might be low statistical power to estimate the misclassification probabilities given that they (i) depend on potentially many parameters (if the dimension of $\boldsymbol{x}_i$ is large), (ii) are potentially small, and (iii) are identified from potentially low frequency cells of the joint distribution of $(y_i, h_{1i,h_{2i}})$.

When there is low statistical power, the likelihood function may not be maximised at the mean of the likelihood function and thus the estimated parameters will not be normally distributed around the true parameter values but instead be biased. This bias in the misclassification probabilities then in turn biases the estimated parameters of the outcome equation. In the most extreme case, the likelihood function may be maximised for the sample at hand when misclassification is set to zero, which may manifest itself as a convergence problem in the maximum likelihood procedure as parameters will tend to infinity.

To overcome convergence such issues and reduce the small sample/low statistical power bias we suggest implementing a penalised likelihood estimation, which rules out extreme misclassification probabilities. For each of the ten components in (16) related to the misclassification probabilities (that is, $\ln F(h_{mi}|h_i^* = j, \boldsymbol{x}_i)$ for $m = 1, 2$ and $j = 0, \ldots, 4$), we add a ridge penalty to their objective function:

$$\hat{\boldsymbol{\gamma}}^m_j = \arg\max_{\boldsymbol{\gamma}^m_j} \sum_i \ln F(h_{mi}|h_i^* = j, \boldsymbol{x}_i) - \frac{t}{N}\, \boldsymbol{\gamma}^{m\prime}_j \boldsymbol{\gamma}^m_j \tag{18}$$

$$= \arg\max_{\boldsymbol{\gamma}^m_j} \sum_i \hat{w}_{ji} \left(\sum_k \mathbb{1}(h^m = k) \ln \delta^m_{k|j}(\boldsymbol{x}_i, \boldsymbol{\gamma}^m_{k|j}) \right) - \frac{t}{N}\, \boldsymbol{\gamma}^{m\prime}_j \boldsymbol{\gamma}^m_j; \tag{19}$$

where $\boldsymbol{\gamma}^m_j$ contains all the parameter vectors $\gamma^m_{k|j}$ for a given j and m such that $j \neq k$. The scalar t is a tuning parameter which determines the weight given to the penalty. However, as with all penalised likelihood estimations, by penalising is introducing bias in the other direction such that a too harsh a penalty (too large t) may increase bias.

Apart from penalisation, a second possible avenue for reducing low power issues is using more than one outcome variable, say $\boldsymbol{y}_i = (y_{1i}, y_{2i})$. If more than one possible outcome which is dependent on true SAH is available, then joint estimation of the outcomes can be beneficial for the accuracy of the estimation and minimising bias in small samples. We propose pooling outcomes and treating them as independent but potentially correlated. The connection between the two (or more) models is that the true unobserved SAH is obviously the same for each observation across both models and thus the misclassification parameters are also the same, which can be imposed as a restriction to

reduce the loss of degree of freedoms relative to the case of separate estimation.[2] Adapting the EM algorithm is straightforward. In equations (15)-(17), the terms $F(y_i|h_i^* = j, \boldsymbol{x}_i)$ are simply replaced by $F(y_{1i}|h_i^* = j, \boldsymbol{x}_i)F(y_{2i}|h_i^* = j, \boldsymbol{x}_i)$.

3 Monte Carlo experiments

While the estimators presented in the last section are consistent if the assumptions are met, they are biased in finite samples. The estimation of the many parameters relating to the misreporting probabilities, for instance, can pose a challenge in practice. We examine the finite sample performance of these estimators in a Monte Carlo simulation study. The interest lies primarily in the quality of the estimates of α and $\boldsymbol{\beta}$, the parameters of the outcome equation. To benchmark the performance of the estimators that adjust for misclassification of SAH, we compare their performance to the ideal estimator that uses the true SAH status, and which is infeasible in practice as true SAH is unobserved. On the other end of the spectrum, we compare the performance of the proposed estimators to the naïve estimators that just use either the first or the second observed misreported SAH measure.

We start by examining four potential competitor estimators, which address the misclassification in an ad-hoc way and are sometimes encountered in the literature. We then consider our proposed finite mixture (FM) estimator in more detail. We are particularly interested in the possibility of improving FM's performance by using a penalised FM (PFM) variant. We consider how estimating interaction effects, facing an increased number of categories of health, or jointly estimating two outcome variables affect the finite sample performance of the FM and PFM estimators, and we explore how the estimators perform under misspecification of the misclassification system.

3.1 Simulation design

The baseline design we use is a simple data generating process (DGP) with a single, uniformly-distributed covariate and a binary health indicator. That is, $\boldsymbol{x}_i = (1, x_i)$, where $x_i \sim U(0, 1)$, and true health h_i^* is drawn from a Bernoulli distribution with probability π_i. We draw ε_i from a logistic distribution; survival status y_i (=1 if alive) is then generated as

$$y_i = \mathbb{1}\left(\alpha\, h_i^* + \beta_0 + \beta_1 x + \varepsilon_i > 0\right). \tag{20}$$

We use the four misreporting probabilities $\delta^1_{0|1}$, $\delta^2_{0|1}$, $\delta^1_{1|0}$ and $\delta^2_{1|0}$ to generate the two reported health measures h_{1i} and h_{2i}. Specifically, for observations with $h_i^* = 1$ we draw h_{mi} from a Bernoulli distribution with probability $1 - \delta^m_{0|1}$; and for observations with $h_i^* = 0$ we draw h_{mi} from a Bernoulli distribution with probability $\delta^m_{1|0}$. Thus, jointly, the four misreporting probabilities, the parameter governing the distribution of true health, and the parameters of the outcome equation α, β_0, β_1 determine endogenously the distribution of the survival outcome y_i, and the distribution of the reported

[2] In the EM algorithm both outcomes are also used to estimated the posterior probabilities of the true SAH category. Note, we are not proposing a seemingly-unrelated-regression-type approach that exploits efficiency gains through correlated errors in the outcomes.

health measures h_{mi}. The parameter values are specified as $\alpha = 1$, $\beta_0 = 0$ and $\beta_1 = 1$. Misreporting probabilities are parametrised as

$$\delta^m_{k|j} = \Lambda\big(-\exp(\gamma^m_{k|j}\text{const} + \gamma^m_{k|j}\text{slope}\, x_i)\big), \quad m = 1, 2, \quad j \neq k = 0, 1,$$

with all four slope parameters $\gamma^m_{k|j}$slope $= 1$, and the four constants $\gamma^1_{0|1}$const=-0.25, $\gamma^2_{0|1}$const=-0.75 , $\gamma^1_{1|0}$const=0, and $\gamma^2_{1|0}$const=-0.5. The distribution of h^*_i is given by

$$\pi_i = \Lambda(\eta_0 + \eta_1 x_i),$$

with η_1=1.5 and η_0=-0.1342.

The simulation DGP implies that the marginal probability of being in good health is $P(h^* = 1) = 0.7$. Similar to the survey data used in our application, the reported health measures in this DGP have distributions which are broadly but not exactly similar to each other— $P(h_1 = 1) = 0.61$ and $P(h_2 = 1) = 0.57$—while at the same time there is a substantial share of conflicting answers: $P(h_1 \neq h_2)$=0.37. The average misreporting probabilities are about 0.21 ($\delta^1_{0|1}$), 0.31 ($\delta^2_{0|1}$)), 0.16 ($\delta^1_{1|0}$)) and 0.26 ($\delta^2_{1|0}$)).

We use two sample sizes, $N = \{1000; 10000\}$ and replicate the estimations 500 times.

3.2 Simulation results

To benchmark the performance of the estimators that adjust for misclassification of SAH, throughout the simulations we compare their performance to the ideal estimator that uses the true SAH status, and which is infeasible in practice as true SAH is unobserved. In tables, we denote this estimator as "h^*_i". On the other end of the spectrum, we compare the performance of the proposed estimators to the naïve estimator that just uses one observed misreported SAH measure. In practice, we use the first measure for this estimator, and consequently we denote it in tables as "h_1".

We begin our simulation experiments by assessing the performance of four potential competitor estimators, which address the misclassification in an ad-hoc way. First, we experiment by using the average of the two SAH measures as the regressor in the models ("$\bar{h}$"). If the measurement error were classical, this approach would produce an (unbiased) SAH measure with less measurement error, thus mitigating some of the bias. A second simple ad-hoc way of addressing the misclassification is to drop all individuals from the estimation sample whose second response to the SAH question is different from the first ("$\bar{\bar{h}}$"). This leaves a sample of individuals with what sometimes is called "consistent responses". It is clear that this is also a procedure leading to biased estimates, since some of the individuals in such a sample will have misreported their SAH status twice. Moreover, this procedure results in a reduced sample size and, therefore, less precise estimates. Nevertheless, similar to the averaging of the SAH responses, the severity of the misclassification problem might be mitigated by this approach.

The last two estimators included in the simulation correspond to approaches that mimic two-stage least squares in linear models. They consist in using one SAH measure as an instrument for the other. Both estimators use the same first stage in which one SAH measure is regressed on the other. The first of

Table 1: Simulation results: Ad-hoc misclassification approaches from the literature

		h^*	h_1	$\bar{h}$	$\bar{\bar{h}}$	$\hat{h}_1$	$\hat{e}_1$
$N = 1,000$							
$\hat{\alpha}$	Bias	0.004	-0.459	-0.259	-0.243	0.632	0.675
	RMSE	0.152	0.482	0.321	0.307	0.953	0.989
$\hat{\beta}$ const	Bias	-0.007	0.258	0.162	0.163	-0.262	-0.276
	RMSE	0.167	0.303	0.234	0.265	0.452	0.465
$\hat{\beta}$ slope	Bias	0.014	0.169	0.152	0.054	-0.138	-0.134
	RMSE	0.271	0.317	0.308	0.346	0.359	0.359
$N = 10,000$							
$\hat{\alpha}$	Bias	0.002	-0.457	-0.259	-0.245	0.602	0.643
	RMSE	0.050	0.460	0.268	0.253	0.647	0.686
$\hat{\beta}$ const	Bias	-0.002	0.260	0.165	0.158	-0.245	-0.258
	RMSE	0.049	0.265	0.172	0.171	0.272	0.284
$\hat{\beta}$ slope	Bias	0.003	0.156	0.141	0.057	-0.144	-0.140
	RMSE	0.084	0.177	0.164	0.120	0.179	0.176

these estimators then includes the first-stage predictions as the regressor in the outcome model ("$\hat{h}_1$"). This approach is inconsistent, in general, for nonlinear models, but it is often applied by practitioners. The second estimator includes the first-stage residuals as an additional regressor along the mismeasured SAH response in the outcome model ("$\hat{e}_1$"). This is a version of the control function approach and is valid for nonlinear models under certain conditions. In general, for instance, the endogenous regressor (here, SAH) needs to be continuous. There are, however, specific forms of endogeneity under which the control function approach is consistent with a discrete endogenous regressor (see, for instance, the setup used in Terza *et al.*, 2008). And even when it is inconsistent, the control function approach has been advocated as a potentially useful remedy that might not cure the problem but reduce it in some circumstances (Basu & Coe, 2015; Wooldridge, 2014).[3]

The results in Table 1 show that the infeasible estimator in column "h^*" is virtually unbiased. The naïve estimator which uses the misreported SAH measures, depicted in column "h_1"', is severely biased. The average estimate of α is about 45 percent below its true value of 1 in both sample sizes, illustrating the pernicious effects of misreporting. The following two columns show the results obtained by using the two common *ad hoc* fixes for reducing misreporting bias, averaging the two available measures, and keeping only observations with the same reported SAH across both measures. The bias in the estimated α is about -75 percent for both estimators. Thus, these procedures not only fail to improve over the estimation using a single reported measure, but they even worsen the bias.

The columns "$\hat{h}_1$" and "$\hat{e}_1$" report the results for the possible *ad hoc* methods related to IV estimation. All estimated parameters, including the slope of x, are very distorted overestimating the true value on average by about 63 and 67 percent. Thus, such approaches, while well-suited to measurement error in linear models, cannot be recommended as solutions to the measurement error problem at hand. We

[3]The control function approach might also be useful if the focus is on testing rather than estimation. Some tests might be valid even when the estimator is inconsistent (Wooldridge, 2014; Staub, 2009).

Table 2: Simulation results: Baseline DGP

		$N = 1,000$				$N = 10,000$			
		h^*	h_1	FM	PFM	h^*	h_1	FM	PFM
$\hat{\alpha}$	Bias	0.004	-0.459	0.041	0.018	0.002	-0.457	0.008	0.005
	RMSE	0.152	0.482	0.286	0.267	0.050	0.460	0.085	0.083
$\hat{\beta}$ const	Bias	-0.007	0.258	0.004	0.073	-0.002	0.260	-0.010	-0.000
	RMSE	0.167	0.303	0.349	0.247	0.049	0.265	0.118	0.098
$\hat{\beta}$ slope	Bias	0.014	0.169	0.017	-0.012	0.003	0.156	0.012	0.021
	RMSE	0.271	0.317	0.457	0.306	0.084	0.177	0.154	0.124
$\hat{\eta}$ const	Bias			-0.127	-0.313			0.036	0.003
	RMSE			1.142	0.591			0.393	0.289
$\hat{\eta}$ slope	Bias			-0.013	-0.002			-0.064	-0.120
	RMSE			1.560	0.552			0.517	0.363
$\hat{\gamma}^1_{1\|0}$ const	Bias			-0.005	-0.064			0.039	0.034
	RMSE			1.649	0.349			0.373	0.252
$\hat{\gamma}^1_{1\|0}$ slope	Bias			-0.272	-0.624			0.010	-0.199
	RMSE			5.787	0.735			0.612	0.424
$\hat{\gamma}^2_{1\|0}$ const	Bias			-0.218	0.149			-0.012	0.048
	RMSE			1.560	0.351			0.323	0.227
$\hat{\gamma}^2_{1\|0}$ slope	Bias			-0.027	-0.538			0.022	-0.156
	RMSE			9.004	0.672			0.547	0.380
$\hat{\gamma}^1_{0\|1}$ const	Bias			0.109	0.224			-0.010	0.007
	RMSE			0.964	0.395			0.249	0.195
$\hat{\gamma}^1_{0\|1}$ slope	Bias			0.063	-0.162			0.028	0.031
	RMSE			1.342	0.379			0.337	0.246
$\hat{\gamma}^2_{0\|1}$ const	Bias			0.024	0.388			-0.029	0.044
	RMSE			0.770	0.482			0.235	0.194
$\hat{\gamma}^2_{0\|1}$ slope	Bias			0.100	-0.342			0.049	-0.016
	RMSE			1.056	0.489			0.281	0.221

see that for all these four *ad-hoc* approaches the estimated root mean squared error (RMSE) is driven primarily by the bias. As these biases do not vanish with larger sample sizes, the RMSE approach the bias as variances shrink with increasing N.

In Table 2 we present estimates from the proposed finite mixture (FM) estimator, as well as its variant, the penalised finite mixture (PFM) estimator, for the same set of replications as in Table 1. For reference, we have reprinted the infeasible (h_i^*) and naïve (h_1) estimators. The FM estimator in samples of N=1,000 is able to greatly reduce the bias from h_1 from 46 to 4 percent for α. In samples of N=10,000, the bias is less than 1 percent. The RMSE in the DGP with N=1,000 is about twice as large as that of the infeasible estimator. The other parameters of the outcome model, β_0 and β_1, are estimated similarly well.

However, at N=1,000, there are larger biases, ranging up to about 20 percent, for the parameters of the misclassification system; and even when the biases are small, the RMSE can still be substantial. It is for this issue that we see the advantages of the PFM estimator most clearly. It achieves reductions

Table 3: Simulation results: DGP with interaction effect in health

		$N = 1,000$				$N = 10,000$			
		h^*	h_1	FM	PFM	h^*	h_1	FM	PFM
$\hat{\alpha}$ const	Bias	-0.004	-0.606	0.234	-0.020	0.008	-0.596	0.020	0.002
	RMSE	0.284	0.669	0.966	0.560	0.090	0.602	0.186	0.177
$\hat{\alpha}$ slope	Bias	0.025	-0.374	-0.251	0.117	-0.011	-0.409	-0.018	0.012
	RMSE	0.560	0.660	1.392	0.966	0.178	0.442	0.295	0.286
$\hat{\beta}$ const	Bias	-0.001	0.326	-0.138	0.055	-0.003	0.324	-0.015	-0.009
	RMSE	0.212	0.380	0.820	0.349	0.063	0.330	0.149	0.131
$\hat{\beta}$ slope	Bias	-0.001	0.508	0.215	-0.007	0.006	0.514	0.018	0.024
	RMSE	0.426	0.643	1.161	0.583	0.129	0.528	0.222	0.203

in the RMSE of these parameters that range from 50 to almost 90 percent. This improvement in the estimation of the misclassification system also translates into uniformly lower RMSE in the estimates of the outcome parameters, and sometimes also in bias reductions. For the estimate of α, for instance, PFM reduces FM's bias of 4 percent to less than 2 percent.

The DGP with a binary outcome, which we chose as our baseline, is the most difficult case for correcting misclassification, as the additional information stemming from the outcome that identifies the whole system is sparsest. Table A1 in the Appendix explores other nonlinear outcome models where there is more information in the dependent variable: counts and durations. In both these cases, the results indicate that FM and PFM perform even better.

Next we make things even harder for the estimators by simulating from a DGP where the impact of SAH on the outcome varies with x. The ability to easily specify interaction effects is a hallmark of our approach. Thus, the outcome equation (20) has been augmented with an interaction effect:

$$y_i = \mathbb{1}\left(\alpha\, h_i^* + \alpha_x h_i^* x_i + \beta_0 + \beta_1 x + \varepsilon_i > 0\right), \tag{21}$$

where α_x is the coefficient on the new interaction between health and x, which in the simulation is set to $\alpha_x = 1$. Table 3 shows the results from this DGP, where for space reasons, only the parameters of the outcome model are depicted. That this is a more challenging DGP can be clearly seen by observing the RMSE at N=1,000 for the infeasible estimator, which almost doubles for the constant in α (and quadruples for the slope in α, i.e. the interaction coefficient) relative to RMSE of α in the baseline.

The FM estimator, while still improving substantially over the naïve approach, displays visible biases. The estimate of both main and interaction effect of SAH have biases in absolute value of about 25 percent with N=1,000. These largely disappear with the larger sample size, where they are only about 2 percent. However, the PFM estimator is able to obtain improved estimates already in the smaller sample size, with biases of about 2 and 12 percent for main effect and interaction, yielding reductions in RMSE of about 50 and 40 percent relative to FM. At N=10,000, however, where FM works well, the advantages of PFM over FM in this DGP are only marginal.

We conclude this section on the main results from the simulation by considering the case of a multi-

Table 4: Simulation results: Multivariate DGP for $\boldsymbol{y} = (y_1, y_2)'$, $N = 1,000$

$\rho =$		1.00	0.75	0.50	0.25	0.00
FM						
$\hat{\alpha}$	Bias	0.059	0.039	0.015	0.007	0.001
	RMSE	0.309	0.289	0.283	0.286	0.281
$\hat{\beta}$ const	Bias	0.007	0.010	0.027	0.033	0.033
	RMSE	0.354	0.326	0.313	0.312	0.304
$\hat{\beta}$ slope	Bias	0.004	0.017	0.008	0.003	0.004
	RMSE	0.474	0.439	0.423	0.423	0.420
PFM						
$\hat{\alpha}$	Bias	0.045	0.029	0.009	0.001	-0.001
	RMSE	0.284	0.265	0.262	0.266	0.263
$\hat{\beta}$ const	Bias	0.044	0.044	0.059	0.064	0.065
	RMSE	0.249	0.233	0.228	0.227	0.230
$\hat{\beta}$ slope	Bias	0.000	0.011	0.002	0.002	-0.000
	RMSE	0.338	0.314	0.318	0.315	0.322

variate outcome. In Table 4 we present some results from estimations with two outcomes. We simulate two binary outcomes from the specification:

$$\begin{aligned} y_{1i} &= \mathbb{1}\left(\alpha\, h_i^* + \beta_0 + \beta_1 x + \varepsilon_{1i} > 0\right) \\ y_{2i} &= \mathbb{1}\left(\alpha\, h_i^* + \beta_0 + \beta_1 x + \varepsilon_{2i} > 0\right). \end{aligned}$$

This is a setup in the vein of "seemingly unrelated regressions". The true coefficients have been specified as having the same values across the two outcome equations, but this is merely for convenience and the estimated coefficients are allowed to vary in estimation (i.e. they are not constrained to be the same across equations). As explained previously, the gain from considering y_1 and y_2 jointly is that, since the parameters of the misclassification probabilities are the same across both outcomes, we are, very loosely speaking, doubling the sample size available to estimate these parameters. The extent in which pooling both outcomes adds information depends on the degree of the dependence between the two errors, ε_1 and ε_2. In the worst case, $\varepsilon_1 = \varepsilon_2$ and joint estimation will bring no advantage. Since the only link between the two outcome equations is through the misreporting probabilities, for all the approaches except our proposed method this multivariate DGP for (y_1, y_2) amounts to separate, equation-by-equation estimation, a case indistinguishable from that of Tables 1and 2. Since the DGP is symmetric for y_1 and y_2, we only present estimates for equation y_1. The table presents results for N=1,000 for the cases where the correlation between the errors ε_1 and ε_2 is equal to 1, 0.75, 0.50, 0.25, and 0.

The case ρ=1 is the same as the baseline, and indeed we get very similar results. As the correlation decreases, the estimators are mostly progressively more successful at reducing the biases in general, although not uniformly (the bias in $\hat{\beta}_0$ increases, for instance). However, the RMSE is reduced in all cases, with the magnitude of the reduction for FM ranging from about 10 to 20 percent. Similar, although often slightly larger reductions in RMSE achieved for the parameters of the misclassification

Table 5: Simulation results: Misspecified functional form of misclassification, N=500

		Scenario 1					*Scenario 2*				
		h^*	h_1	NPIV	FM	PFM	h^*	h_1	NPIV	FM	PFM
$\hat{\alpha}$	Bias	0.012	-0.520	-0.124	0.087	0.070	0.015	-0.491	-0.108	0.062	0.070
	RMSE	0.157	0.538	0.409	0.309	0.375	0.160	0.509	0.318	0.253	0.233
$\hat{\beta}$ const	Bias	0.000	0.275	0.061	-0.012	0.008	-0.001	0.263	0.052	0.006	-0.016
	RMSE	0.104	0.290	0.238	0.138	0.129	0.104	0.279	0.205	0.132	0.137
$\hat{\beta}$ slope	Bias	-0.011	-0.150	-0.138	0.014	0.020	-0.014	-0.094	-0.071	0.015	0.017
	RMSE	0.165	0.210	0.332	0.220	0.230	0.165	0.176	0.307	0.234	0.197

system (see results in Appendix Table A2).

3.3 Misspecification

So far we have evaluated the performance of the FM and PFM estimators in DGPs where they correctly specify the misclassification system. We now evaluate these proposed parametric estimators in a DGP where the misclassification probabilities are misspecified. We use the setup of Hu (2008), and also compare our estimator against the nonparametric instrumental variables (NPIV) estimator introduced in that paper. We have argued that the FM/PFM estimators may have two potential advantages despite the drawback of fully specifying the functional form of the misclassification probabilities and the true health distribution. First, by using flexible specifications of the linear indices $\boldsymbol{x}_i'\gamma_{k|j}^m$, many functional forms may be approximated well. Second, compared to more nonparametric approaches, even if FM/PFM might be biased, they might still be preferable in RMSE. Here, we give some evidence of the second point. That is, we don't explore potential improvements by specifying polynomials of $\boldsymbol{x}_i$ in the linear indices.

The DGP in Hu (2008) is for a probit outcome y_i, a binary misclassified regressor h_i^*, and a normally distributed covariate x_i. Importantly, misclassification does not follow our logit-based functional forms. Rather, it is a partially linear function with kinks (see Hu, 2008, p.45, for details) for details. We adjust our outcome model to be a probit, but leave the misclassification probabilities and π_i as logistic. Table 5 shows our results for FM and PFM from our simulation from this DGP, with N=500 and 200 replications as in the original paper, next to the h_i^*, h_1 and NPIV results from the paper. Scenarios 1 and 2 depicted in the table correspond to two variants of the DGP in which the probabilities $\delta_{0|1}^m$ depend negatively (Scenario 1) or positively (Scenario 2) on the regressor x_i.

While NPIV substantially reduces the bias of the naïve estimator, for instance from about 50 percent to 12 percent for $\hat{\alpha}$ in Scenario 1, FM and PFM reduce the bias even further, and they also have the lowest RMSE of the feasible estimators presented.

Table 6: Simulation results: DGP with multinomial health ($h^* = 0, 1 \ldots, 4$)

		$N = 1,000$				$N = 10,000$			
		h^*	h_1	FM	PFM	h^*	h_1	FM	PFM
$\hat{\alpha}_1$	Bias	0.056	-0.298	0.166	0.105	0.017	-0.293	0.039	0.013
	RMSE	0.392	0.379	0.786	0.738	0.094	0.302	0.175	0.169
$\hat{\alpha}_2$	Bias	0.033	-0.521	0.095	0.057	0.005	-0.534	0.028	0.012
	RMSE	0.301	0.570	0.381	0.574	0.093	0.539	0.157	0.149
$\hat{\alpha}_3$	Bias	0.055	-0.741	0.161	0.147	0.003	-0.754	0.019	0.001
	RMSE	0.307	0.772	0.533	0.612	0.082	0.758	0.155	0.145
$\hat{\alpha}_4$	Bias	-0.123	-0.926	0.078	0.127	0.003	-0.937	0.027	0.010
	RMSE	0.285	0.951	0.453	0.571	0.087	0.940	0.130	0.131
$\hat{\beta}$ const	Bias	-0.026	0.648	-0.123	-0.082	-0.011	0.660	-0.030	-0.009
	RMSE	0.241	0.670	0.419	0.517	0.077	0.662	0.128	0.124
$\hat{\beta}$ slope	Bias	0.110	0.133	0.150	0.039	0.005	0.165	0.005	0.019
	RMSE	0.266	0.264	0.275	0.278	0.071	0.180	0.073	0.077

3.4 Extension to multinomial health

To conclude our simulation study, we present results from a DGP with a discrete SAH measure with five categories, $h^* = 0, \ldots, 4$. We simulate from the following DGP:

$$y_i = \mathbb{1}\left(\alpha_1 h^*_{1i} + \alpha_2 h^*_{2i} + \alpha_3 h^*_{3i} + \alpha_4 h^*_{4i} + \beta_0 + \beta_1 x + \varepsilon_i > 0\right), \tag{22}$$

where we specify $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)' = (0.5, 1.0, 1.5, 2.0)'$. The parameters β_0 and β_1 are set to -1 and 1. We specify the misreporting probabilities as

$$\delta^m_{k|j,i} = \frac{\exp(-\exp(\gamma^m_{k|j}\text{const} + \gamma^m_{k|j}\text{slope}\ x_i))}{1 + \sum_{k \neq j} \exp(-\exp(\gamma^m_{k|j}\text{const} + \gamma^m_{k|j}\text{slope}\ x_i))}, \qquad \text{for } j \neq k,$$

and set all slope parameters equal to 1, $\gamma^m_{k|j}$slope=1, and specify the constants as $\gamma^m_{k|j}\text{const} = 0.25|j-k|$. The marginal distribution of true health is specified as $\boldsymbol{\pi} = (0.10, 0.15, 0.20, 0.25, 0.30)$ by setting

$$\pi_{ji} = \frac{\exp(\eta_j\text{const} + \eta_j\text{slope}\ x_i)}{1 + \sum_{j=1}^{4} \exp(\eta_j\text{const} + \eta_j\text{slope}\ x_i)}, \quad j = 1, 2, 3, 4,$$

with slopes equal to 1.0, 2.0, 2.0 and 2.5, and constant chosen such as to yield the marginal distribution specified above.

This DGP is more challenging not only in that it has more parameters, but also in that misreporting is much more prevalent. About 61 percent of individuals report different values for h_1 and h_2. For roughly half of these, 31 percent, the discrepancy between the first and second SAH measure is 1. Discrepancies of 2, 3, and 4 occur in 18, 9, and 3 percent of individuals. The $\delta^m_{k|j,i}$ vary between about 2 and 20 percent. To the best of our knowledge, this is the first simulation evidence of this type of DGP of a categorical regressor with flexible effects.

The results of the simulation for the parameters of the outcome model are collected in Table 6. Again, that this is a more challenging DGP can be seen in the biases and RMSE that are apparent in the

infeasible estimator. We see that at N=1,000, FM and PFM show some visible biases, in the order of about 8 to 16 percent. However, in the larger sample size these biases have all but disappeared, with the maximum bias in FM being less than 4 percent and that in PFM less than 2 percent.

To summarise, the simulation results in this section illustrated a number of issues which inform our application of the estimation to real world data. First, none of the inconsistent estimators can be recommended in general. Second, the performance of the FM estimator can often be substantially improved, especially in smaller samples by using the penalised version. Third, the performance might also be improved by combining outcomes and estimating them jointly. Fourth, the estimator is able to estimate the effects of interest reliably even under challenging circumstances such as many health categories, interaction effects, and severe misreporting, in samples of about 10,000 observations. In the next section, we will estimate a joint logit-logit model for mortality and morbidity using two five-category SAH measures and a sample of over 12,000 individuals.

4 Estimating the effects of SAH on mortality and morbidity

In this section we present estimates of the association between SAH and two outcomes measured 15 years later: mortality (whether the individual is deceased) and, if the individual is not deceased, whether he or she developed any chronic conditions in the 15-years period. We first present the HILDA data in Section and have a close look at the categorical self-reported health measures in HILDA, for which repeated measures are available in some waves (Section 4.1). We then estimate our joint model and discuss the estimates from the outcome equations in Section 4.2 and the estimates from the misclassification system in Section 4.3.

4.1 Descriptive statistics

The HILDA Survey is a yearly household-based longitudinal survey in Australia that began in 2001 (Summerfield *et al.*, 2014). The survey covers a broad range of social and economic topics such as household formation, income, work and health, and most questions are repeated every year in each wave. Responses of individuals aged 15 or above are published and the non-response reasons are recorded where they are known. Wave 1 of the survey covers a total of 7,682 households and 13,969 responding individuals. These individuals were followed up in the later waves and new household members joining the original sample were also included. A further 2,153 household and 4,009 individuals were added as a top-up sample in 2011. Overall, there are roughly about 13,000 respondents in each wave of the HILDA Survey from 2001 to 2013. In 2014, the survey sample was matched to the National Death Index so that details of individuals' year and age of death are now available in HILDA for all those originally in the survey, including the non-responders.

In waves 1, 9 and 13 of the survey, the SAH question is asked twice for each individual. The question is first asked as a part of the Person Questionnaire that is conducted by an interviewer face-to-face or over

Table 7: Descriptive statistics

Variable	N	Mean	Std.Dev.
Covariates (Wave 1)			
age/10 (years/10)	12,908	0.438	0.176
male (=1, if yes)	12,908	0.470	0.499
education/10 (years/10)	12,908	1.272	0.203
log HH income	12,908	3.135	0.654
chronic condition (=1 if any chronic conditions in 2001)	12,908	0.233	0.423
married (=1, if married or in a relationship)	12,908	0.642	0.479
overseas (=1, if born overseas)	12,908	0.243	0.429
not in labour force (=1, if yes)	12,908	0.344	0.475
unemployed (=1, if yes)	12,908	0.042	0.201
smoker (=1, if current or former smoker)	12,908	0.493	0.500
Outcomes (Wave 16)			
dead (=1, if deceased by 2016)	12,908	0.109	0.312
cond (=1, if any new chronic conditions in 2016 since 2001)	7,340	0.161	0.368

the phone. The SAH question is the first question in the health section, followed by a number of other health-related questions such as long-term conditions and disabilities. We designate this variable as h_1. Respondents are asked to choose their rating on a 5-option scale labelled as "Poor", "Fair", "Good", "Very Good", and "Excellent" and which we code as $0, \ldots, 4$. Then, individuals who have responded to the Person Questionnaire are issued with the Self Completion Questionnaire, which is to be filled in by the respondents themselves and collected by the interviewers after completion. In this questionnaire, the same SAH question is asked again at the beginning of the SF-36 Health Survey. We designate this variable as h_2, and code it in the same way as h_1. The dates of completing both questionnaires are available in HILDA for waves 9 and 13. On average, the questionnaires were completed only 4.8 days and 4.6 days apart in 2009 and 2013, respectively. The median for time between completion of the two questionnaires is 1 day in both survey waves. Since the surveys were taken not too long apart, the likelihood of an actual change in health is fairly low. As a result, we believe that most of the changes to the answers of SAH questions are merely random changes that were unlikely due to changes in their underlying true health status.

Ultimately, we are only interested in h_{1i} and h_{2i} for the first wave in 2001 because we want to study long-term (15-year) mortality. There are 12,908 individuals with responses on h_{1i} and h_{2i}. Descriptive statistics for selected demographic and socio-economic characteristics ofthese individuals are given in Table 7. The top panel of Table 8 reports the joint-distribution (in percent of respondents) from the two SAH questions. About 27.8 percent of respondents changed their health status between h_{1i} and h_{2i}, a finding which is similar to that reported by Clarke & Ryan (2006). It could be that this pattern is specific to the first wave. However, the joint distributions of h_{1i} and h_{2i} in waves 9 (N=11,110) and 13 (N=14,993) are very similar to the one in wave 1 (middle and bottom panels of Table 8), and so is

Table 8: Joint distribution of SRH measures from Personal Questionnaire (h_1) and Self-Completion Questionnaire (h_2) in HILDA

Wave 1, $N = 12,908$			h_2			
h_1	0	1	2	3	4	Total
0	2.65	1.03	0.16	0.04	0.02	3.90
1	0.55	8.94	2.11	0.30	0.02	11.92
2	0.13	2.36	21.64	3.97	0.52	28.62
3	0.04	0.53	7.23	25.67	2.03	35.50
4	0.02	0.09	0.90	5.74	13.33	20.07
Total	3.38	12.95	32.04	35.71	15.92	100.00
Wave 9, $N = 11,110$			h_2			
h_1	0	1	2	3	4	Total
0	2.66	1.12	0.16	0.07	0.01	4.02
1	0.32	8.53	3.31	0.23	0.05	12.44
2	0.08	2.20	23.96	5.25	0.23	31.72
3	0.00	0.23	6.24	27.55	2.05	36.08
4	0.00	0.05	0.53	4.28	10.89	15.74
Total	3.06	12.12	34.20	37.38	13.23	100.00
Wave 13, $N = 14,993$			h_2			
h_1	0	1	2	3	4	Total
0	2.31	1.21	0.19	0.04	0.01	3.75
1	0.54	9.38	3.72	0.34	0.01	14.00
2	0.11	2.40	24.16	4.84	0.36	31.86
3	0.03	0.26	6.71	27.36	2.06	36.42
4	0.00	0.02	0.42	3.93	9.60	13.97
Total	2.98	13.27	35.20	36.50	12.05	100.00

the share of respondents giving different answers for h_1 and h_2: 26.4 and 27.2 percent for waves 9 and 13, respectively.

Although there is a consistent percentage of individuals who revised their health status in each wave, the change was not driven by the same individuals over time. The correlation of switchers (individuals who revised their response) in wave 1 and switchers in wave 9 is only 0.03 while the correlation of switchers in wave 9 and switchers in wave 13 is only 0.05, which means the vast majority of the switchers are actually new switchers from one wave to another. This increases our confidence that switching displays are large amount of randomness.

Given the two questionnaires were completed around the same time for most people in each wave, and the percentage of switchers stays consistent over time, we conjecture that at least one of the SAH measures, if not both, is measured with some error. The marginal distributions of h_1 and h_2 given in Table 8 also reveal that individuals are more likely to select the extreme categories—"poor" (0) and "excellent" (4)—when responding to an interview (h_1) than a written questionnaire (h_2). This may suggest that compared to the self-completion mode the interviewing mode increases the chance

Table 9: Estimation results: Logit models for changes in SAH response: any change (=1 if $h_1 \neq h_2$), up (=1 if $h_1 < h_2$), and down (=1 if $h_1 > h_2$)

Dep. var.	change		up		down	
	(1)	(2)	(3)	(4)	(5)	(6)
age/100	0.69	-0.17	0.75	0.88	0.56	-0.79
	(0.57)	(0.65)	(0.83)	(0.94)	(0.66)	(0.76)
age^2/100	0.16	1.01	-0.36	-0.83	0.36	1.91**
	(0.59)	(0.67)	(0.86)	(0.98)	(0.68)	(0.78)
male	0.05	0.05	0.23**	0.23**	-0.07	-0.08
	(0.04)	(0.04)	(0.06)	(0.06)	(0.05)	(0.05)
education/10	-0.54**	-0.59**	-0.45**	-0.44**	-0.48**	-0.55**
	(0.11)	(0.11)	(0.16)	(0.16)	(0.13)	(0.13)
log HH income	-0.13**	-0.13**	-0.21**	-0.16**	-0.04	-0.07*
	(0.03)	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)
chronic condition		-0.14**		0.27**		-0.39**
		(0.05)		(0.07)		(0.06)
married		0.13**		-0.02		0.19**
		(0.05)		(0.07)		(0.06)
overseas		0.21**		0.22**		0.15**
		(0.05)		(0.07)		(0.05)
not in labour force		0.03		0.11		-0.03
		(0.05)		(0.08)		(0.06)
unemployed		0.05		0.10		0.01
		(0.10)		(0.14)		(0.12)
smoker		0.03		-0.03		0.06
		(0.04)		(0.06)		(0.05)
mean dep. var.	0.278		0.102		0.176	
N	12,908		12,908		12,908	

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$

that individuals misclassify into more extreme categories; or, alternatively, that compared to the self-completion mode the interviewing mode reduces the chance that individuals misclassify into the middle categories. Either or both cases could produce the observed joint distribution.

We begin our empirical investigation by applying a strategy used in the previous literature to characterise the misclassification behaviour of individuals (Black *et al.*, 2016). In Table 9, we present estimates of logit models where the dependent variable is an indicator that an individual gave two conflicting reports of SAH, $\mathbb{1}(h_{1i} \neq h_{2i})$ (Columns 1 and 2), an indicator that they gave a higher SAH in the self-completion questionnaire, $\mathbb{1}(h_{1i} < h_{2i})$ (Columns 3 and 4), and that gave a higher SAH in the personal questionnaire, $\mathbb{1}(h_{1i} < h_{2i})$ (Columns 5 and 6). Pairs of columns show results for a minimal specification based only on age, sex, education and income, and a fuller specification which in addition includes indicators for whether individuals suffered from any chronic conditions in 2001 (*chronic condition*), whether they were married or in a relationship (*married*), whether they were born overseas (*overseas*), whether they were not in the labour force (*not in labour force*), whether they were

unemployed (*unemployed*) and whether they were currently smokers or had been smokers in the past (*smoker*).

The presence of some statistically significant estimates suggest that misclassification is related to covariates. In particular, consistent with the previous literature, low education and income are strongly predictive of giving conflicting reports of health. However, insignificant estimates are harder to interpret. There could be different types of misclassification patterns which 'average out', resulting in the finding of insignificant effects on the variable "change" ($\mathbb{1}(h_{1i} \neq h_{2i})$). For instance, the regressor *male*, has an effect on the dependent variable "up" ($\mathbb{1}(h_{1i} < h_{2i})$) despite not having effect on "change". However, more complex patterns can be completely undetectable with these dependent variables based on the cross-tabulation of h_1 and h_2. By estimating our finite mixture model, we can go beyond these reduced-form patterns in misclassification and instead examine the underlying misclassification probabilities which generate these patterns.

4.2 A joint model for mortality and development of new chronic conditions adjusting for misreported SAH

We estimate a simple model for mortality where the probability of being deceased within 15 years is affected by age, gender, and basic indicators of socio-economic status such as education and household income. Conditional on survival, we also estimate the development of any new chronic conditions that individuals report having 15 years after the initial survey. There were 12,908 individuals in the 2001 survey of which 10.9%) were deceased by 2016. We can obtain information on the development of new chronic conditions in 2013 for 7,340 individuals. Means and standard deviations for these outcome variables are also reported in Table 7, along with those of the covariates. We use the same specifications as before and estimate the two outcomes jointly using the penalised finite mixture estimator, using the same $\boldsymbol{x}_i$ specification to parametrise the misclassification probabilities according to equation (25).

Table 10 contains the estimates of the outcome parameters of our model. We present results from the same two specifications as before: a reduced one which estimates the effect of true SAH on mortality and chronic conditions controlling only for age, sex, education and income; and a more extended one which controls for further socio-economic measures, as well as for health status in 2001 through the presence of chronic conditions, and for risky behaviours through the presence of the smoker status. The shorter specification results are depicted in Columns (1) and (3) for mortality and chronic conditions. We pair each column of results with a column indicating the difference to the corresponding estimates from the naïve estimator which simply uses the first observed measure of SAH, h_1. For the key parameters of interest, the health coefficients α_j, the differences to the naïve approach are larger and more statistically significant for mortality than for the new chronic condition indicator. Interestingly, the differences in Column (2) for the mortality outcomes are increasing, indicating that not only is there a statistically significant bias in the naïve approach, but that that the pattern of the effect of SAH on mortality is also biased. However, despite the large share of individuals with different answers in h_1 and h_2 and, thus, the large implicit potential for bias which we documented in the simulations, the magnitude of the statistically significant biases that we find are moderate, ranging mostly from

Table 10: Estimation results: System Penalised Finite Mixture models for mortality (dead: yes/no) and morbidity (chronic condition: yes/no)

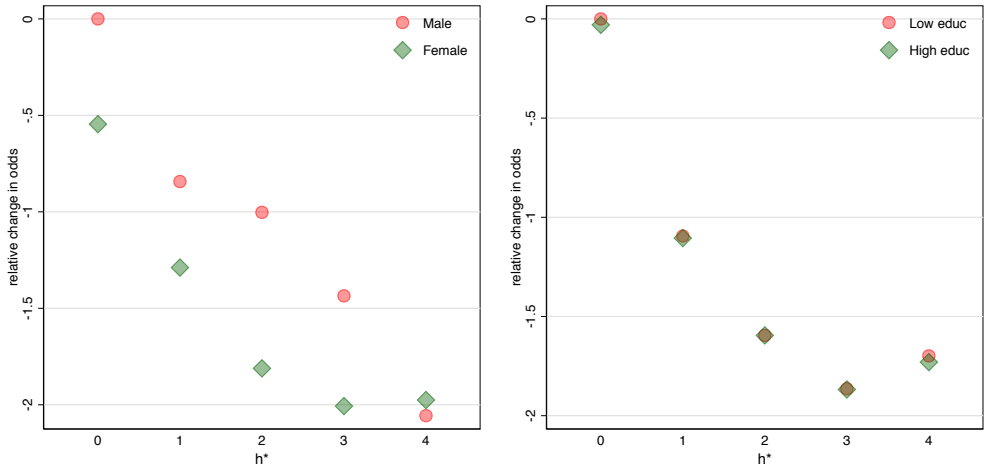
	System PFM				System PFM			
Dep. var./*diff. PFM–naïve*	Dead	*diff.*	Cond.	*diff.*	Dead	*diff.*	Cond.	*diff.*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
α_1	-0.96**	-0.07	-0.23	0.07	-0.80**	-0.06	-0.15	0.01
	(0.14)	(0.06)	(0.17)	(0.09)	(0.14)	(0.04)	(0.17)	(0.08)
α_2	-1.42**	-0.11**	-0.72**	0.03	-1.13**	-0.10**	-0.41**	0.03
	(0.13)	(0.05)	(0.16)	(0.08)	(0.15)	(0.05)	(0.17)	(0.08)
α_3	-1.84**	-0.16**	-1.10**	-0.02	-1.46**	-0.12*	-0.71**	-0.02
	(0.14)	(0.06)	(0.16)	(0.08)	(0.16)	(0.06)	(0.18)	(0.08)
α_4	-2.18**	-0.26**	-1.49**	-0.11	-1.77**	-0.24**	-1.11**	-0.16*
	(0.19)	(0.09)	(0.19)	(0.09)	(0.20)	(0.09)	(0.20)	(0.09)
age/100	-4.88**	-0.28**	5.82**	-0.12	-3.90**	-0.11	6.28**	-0.08
	(1.50)	(0.10)	(1.29)	(0.09)	(1.56)	(0.08)	(1.39)	(0.09)
age^2/100	13.94**	0.16	-2.55*	0.03	13.20**	0.04	-3.08**	0.00
	(1.36)	(0.10)	(1.36)	(0.09)	(1.44)	(0.08)	(1.49)	(0.10)
male	0.65**	-0.00	-0.09	0.01**	0.58**	-0.00	-0.12*	0.01
	(0.08)	(0.01)	(0.07)	(0.00)	(0.08)	(0.00)	(0.07)	(0.00)
education/10	-0.21	0.06**	-0.63**	0.03**	-0.12	0.03**	-0.52**	0.01
	(0.22)	(0.02)	(0.18)	(0.01)	(0.22)	(0.01)	(0.18)	(0.01)
log HH. income	-0.13**	0.03**	-0.23**	0.02**	-0.10*	0.01**	-0.17**	0.01**
	(0.06)	(0.01)	(0.05)	(0.00)	(0.06)	(0.00)	(0.06)	(0.00)
chronic condition					0.27**	-0.03*	0.39**	-0.00
					(0.09)	(0.02)	(0.09)	(0.02)
married					-0.38**	-0.01**	-0.14*	0.00
					(0.08)	(0.00)	(0.08)	(0.00)
overseas					-0.25**	0.01**	-0.05	0.01**
					(0.09)	(0.00)	(0.08)	(0.00)
not in labour force					0.07	-0.02**	0.13	-0.00
					(0.11)	(0.01)	(0.09)	(0.01)
unemployed					0.07	0.01	0.31*	0.01
					(0.25)	(0.01)	(0.17)	(0.01)
smoker					0.60**	0.00	0.29**	0.00
					(0.08)	(0.01)	(0.07)	(0.00)
N	12,908		7,340		12,908		7,340	

Standard errors in parentheses

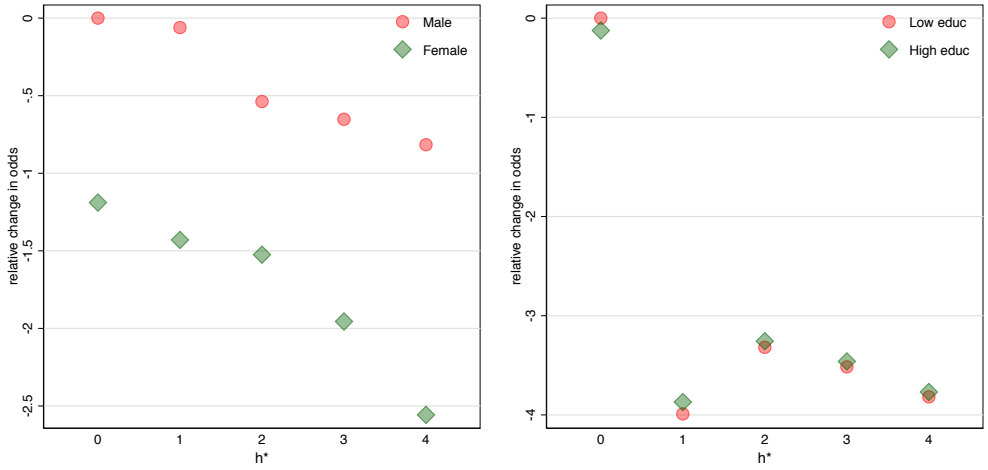
* $p < 0.10$, ** $p < 0.05$

Figure 1: Heterogeneity in the effect of health on morbidity and mortality

(a) Relative change in odds of "dead"



(b) Relative change in odds of "cond"



Notes: Data from HILDA waves 1 and 16 for individuals who responded to SRH questions in wave 1.

about 10 to 20 percent in both $\boldsymbol{\alpha}$ as well as $\boldsymbol{\beta}$.

As a sensitivity analysis, we also estimated specifications where we replaced the continuous variables age, education and household income by sets of dummy variables. The estimation results can be found in Appendix Table A3 (and additional descriptive statistics for the discretised variables in Table A4). We find broadly similar results to the ones in our baseline specification with continuous regressors, althougg differences tend to be somewhat larger.

Finally, we also estimated specifications with interaction effects in true unobserved health. We run four separate specifications where we interacted health with education, household income, sex, and age (results for these specifications are in Table A5 in the appendix). We found little evidence for significant differences to the naïve approach for either income and age. However there where large and significant differences in both mortality and chronic conditions for education, and in mortality for sex. These differences ranged from roughly 30 to 100 percent. To examine the heterogeneity in these variables further, Figure 1 visualises the estimates by comparing the effects for each true health category for

male vs. women, and for low education (11 years) vs high education (18 years). Dots represent relative changes in the odds of either dying (Panel (a)) or, conditional on surviving, developing a chronic condition (Panel (b)) relative to a male in poor health (left-hand-side graphs) or relative to a person with less than 12 years of education and in poor health (right-hand-side graphs). For instance, we see from the top left graph that a female in poor health has, ceteris paribus, about 50 percent lower odds of dying within 15 years than a male in poor health; however, both female and males in excellent health have about the same odds of dying, and these are about 200 percent lower than those of the male in poor health. The heterogeneity pattern between males and females in the effect of health on chronic conditions is different. Here, for all health levels except "excellent", the difference in the odds of dying between males and females are approximately constant and about 100 percent, whereas for "excellent" health the difference is over 150 percent.

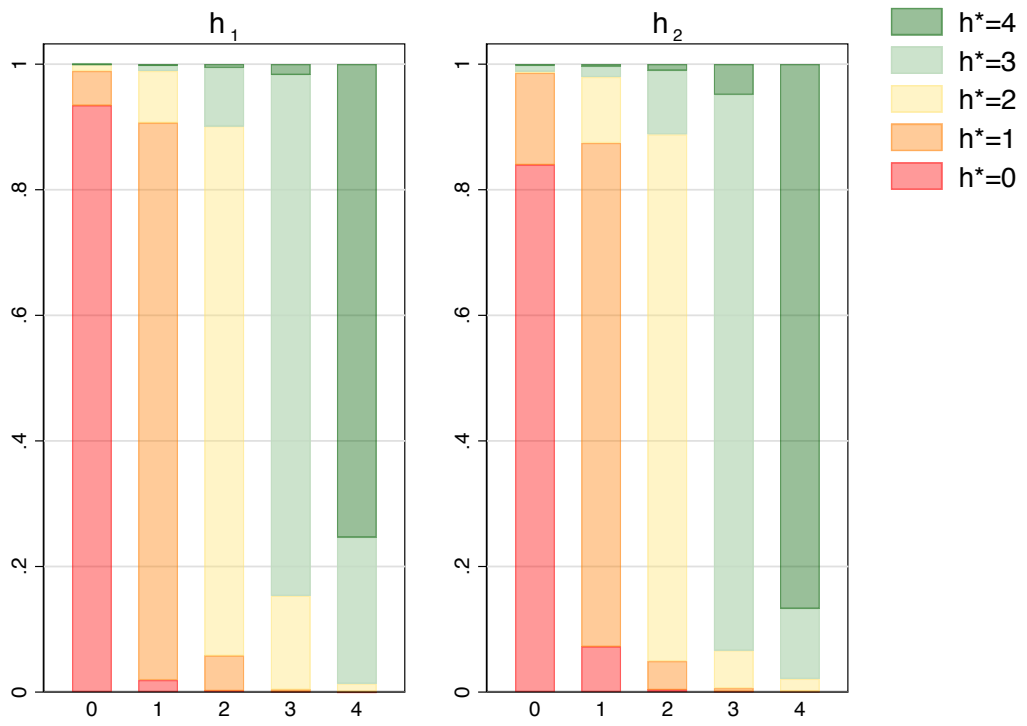
Of particular interest are the results from the graphs that look for heterogeneous responses in education (right-hand-side graphs). These graphs strongly suggest that there is no heterogeneity in the effect of health across education on neither mortality nor on chronic conditions. However, for these results the difference of the PFM estimates to the estimates of the naïve estimator varies strongly with the level of health. Thus, the naïve estimator would have "detected" a pattern of heterogeneity with respect to education in the effect of health on mortality and morbidity. The contrast to the PFM results indicates that in reality these patterns showing up in the naïve estimator are likely biases arising from differences in misreporting across education levels.

Concluding, we found small statistical differences between the naïve and the PFM approach in the baseline specification. The moderate magnitude of these differences suggest that the use of SAH as a control variable might not be compromised despite the large shares of inconsistent answers in SAH. This is a result which might be useful to researchers who rely on including SAH in their empirical analysis as a useful way of addressing omitted variable bias from health status. However, at the same time, the larger biases found in the specifications with interaction effects also showed the limits of what can be learned from SAH without addressing misclassification. For such more nuanced specifications, relying on the naïve approach can lead to substantially biased conclusions

4.3 SAH and misclassification

We use our estimates next to assess the extent of misclassification. First, Figure 2 shows the posterior probabilities of belonging to different true health categories, that is $P(h^* = j|h_1 = k_1, h_2 = k_2, \text{dead}, \text{conditions}, \boldsymbol{x}, \hat{\theta})$, averaged over each reported health category. We already commented on the discrepancy between the two SRH in the best health category, with more people reporting being in "excellent" health in face-to-face interviews (h_1) than when filling out questionnaires privately (h_2). The figure suggests that the share of individuals in true excellent health is lower among responses in the face-to-face interviews than in the privately filled-out questionnaires. Conversely, the share of truthfully reported poor health status is higher in the face-to-face interviews, perhaps because such a health status would also be evident to the interviewer. Both measures look most similar in their composition of the reported middle category.

Figure 2: Posterior probabilities of true health status for each reported health status, $N = 12,908$



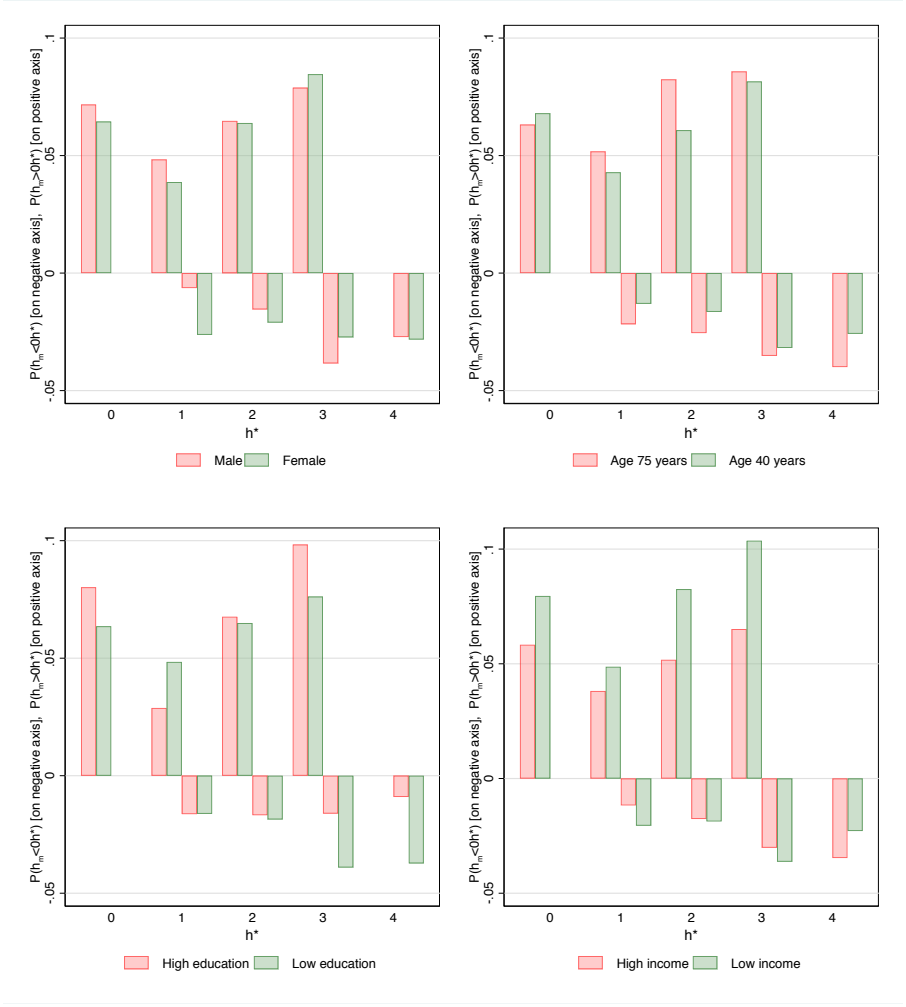
Notes: Data from HILDA waves 1 and 16 for individuals who responded to SRH questions in wave 1.

While Figure 2 considers average probabilities, Figure A1 in the appendix shows that there is considerable heterogeneity in the individual probabilities of reporting health status truthfully. Using box plots, Figure A1 documents the large dispersion in probabilities of reporting true health that exist in the face-to-face SAH question for reported poor health, and in the self-completion SAH question for reported excellent health. The three middle categories tend to be less dissimilar, although we find that the face-to-face interview produces substantially lower probabilities of truthful reporting: For "good" and "very good" health, for instance, the median probability for h_1 is below the first quartile of the h_2 probabilities.

We have so far focussed on the probability of reporting truthfully. Now we take a closer look at the ways people fail to report truthfully. In Figure 3 we show average marginal effects, or rather average discrete effects, of four variables on the probabilities of reporting higher and lower health than the true health status, for each level of true health. True health is on the x axis, and for every category the bars in the upper part of a graph represent the probability of overstating health; and the bars in the lower part of the graph, that of understating it. Overall, reporting better health than one's true health is more prevalent than reporting worse than true health.

The top left graph contrasts these probabilities by sex. The differences are small; men tend to have larger probabilities of reporting better than true health, while women worse than true. The top right graph considers differences by age. The average effect considered here is a change from age 40, which is

Figure 3: Marginal effects of reporting up $(h_m > h^*)$ and down $(h_m < h^*)$, $N = 12,908$



Notes: Data from HILDA waves 1 and 16 for individuals who responded to SRH questions in wave 1.

close to the sample average, to age 75. Older individuals have visibly higher misreporting probabilities across the board; with only the exception of poor health, for which misreporting is slightly higher for the younger age. It is interesting to note that the simple reduced form regressions largely missed the effect of age on misreporting. The bottom left graph contrasts the effect of high education (18 years of schooling) to that of low education (less than 12 years of schooling). While we see that the differences by education are large, the signs are both positive and negative: The highly-educated overstate their true health, while lowly-educated understate it. Finally, contrasting low income (average income in lowest quintile) to high income (average in highest quintile), the bottom right graph suggests that low-income tends to be associated with higher misreporting in both directions, but it is especially related to overstating true health status in all categories (rather than understating it). As a complement to this figure, Appendix Figure A2 shows similar marginal effects but on the probability of reporting truthfully.

5 Conclusions

In this paper, we considered nonlinear regression models where the key regressor is a categorical health variable, and only potentially misclassified measures of SAH are available. We studied the finite sample performance of an estimator that exploits the joint distribution of the outcome and two misclassified health measures, under the assumption that all three variables are independent of each other conditional on true health (that misclassification is not outcome dependent nor systematically related across the two misclassified health measures). The results illustrated the superior performance of this estimator against possible ad-hoc ways of dealing with the misclassification.

Our simulation results can provide guidance to other practitioners working with misclassified categorical regressors. The use of *ad-hoc* methods such as averaging the responses or restricting the sample to individuals with the same responses in both measures cannot be recommended in most cases; nor can the use of two-stage prediction inclusion or residual inclusion. Sample sizes in the order of 10,000 observations seem to be necessary to achieve reliable estimates when using a misclassified regressor with many categories and the dependent variable only has a few outcomes. Using a penalised estimator can visibly improve the performance of the estimator. Using several dependent variables jointly can help reduce finite sample bias. Finite sample bias is also expected to be smaller with dependent variables with more possible outcomes, such as counts or durations, compared to, for instance, binary dependent variables. Finally, in principle, estimates of the misclassification parameters from one study can be used to adjust key outcome parameters from another study using the assumption that the nature of misclassification will stay constant. This might be especially useful for exploring the sensitivity to misclassification in studies which only have one mis-measured SAH measure or have small sample sizes available.

We applied the misclassification estimator to survey data from HILDA, where repeated measures of SAH made the degree of the misclassification problem visible in the high share of respondents exhibiting different answers to the repeated health questions. The estimates we obtained adjusting for misclassification thus represent the first reliable evidence of the association between SAH and long-term mortality and morbidity. Compared to the naïve approach of using observed SAH, the proposed estimator delivered results which indicated an impact of SAH which tended to be somewhat larger and less linear. Large differences were obtained for some specifications with interaction effects, where the conclusions drawn from the analysis would have been different if no adjustment for misclassification were made. The analysis of the estimated misclassification probabilities revealed the pervasiveness of misreporting and a substantial amount of heterogeneity in misreporting linked to observable characteristics.

References

[1] Au, Nicole & David W Johnston, 'Self-assessed health: What does it mean and what does it hide?' *Social Science & Medicine*, **121**, pp. 21–28, *2014*.

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

[3] Basu, Anirban & Norma Coe, '2SLS vs 2SRI: Appropriate methods for rare outcomes and/or rare exposures.' *Unpublished mansucript, University of Washington, Seattle*, *2015*.

[4] Battistin, Erich, Michele De Nadai, & Barbara Sianesi, 'Misreported schooling, multiple measures and returns to educational qualifications.' *Journal of Econometrics*, **181** (2), pp. 136–150, *2014*.

[5] Black, Nicole, David W. Johnston, Michael Shields, & Agne Suziedelite, 'Inconsistent Reporting of Self-Assessed Health: Prevalence by Cognition, Age and Socioeconomic Status.' *Unpublished manuscript, Monash University*, *2016*.

[6] Butler, Joseph S, Richard V Burkhauser, Jean M Mitchell, & Theodore P Pincus, 'Measurement error in self-reported health variables.' *Review of Economics and Statistics*, pp. 644–650, *1987*.

[7] Clarke, Philip M & Chris Ryan, 'Self-reported health: reliability and consequences for health inequality measurement.' *Health Economics*, **15** (6), pp. 645–652, *2006*.

[8] Crossley, Thomas F & Steven Kennedy, 'The reliability of self-assessed health status.' *Journal of Health Economics*, **21** (4), pp. 643–658, *2002*.

[9] DeSalvo, Karen B, Nicole Bloser, Kristi Reynolds, Jiang He, & Paul Muntner, 'Mortality prediction with a single general self-rated health question.' *Journal of General Internal Medicine*, **21** (3), pp. 267–275, *2006*.

[10] ——— , Vincent S Fan, Mary B McDonell, & Stephan D Fihn, 'Predicting mortality and healthcare utilization with a single question.' *Health Services Research*, **40** (4), pp. 1234–1246, *2005*.

[11] Doiron, Denise, Denzil G Fiebig, Meliyanni Johar, & Agne Suziedelyte, 'Does self-assessed health measure health?' *Applied Economics*, **47** (2), pp. 180–194, *2015*.

[12] Gosling, Amanda & Eirini-Christina Saloniki, 'Correction of misclassification error in disability rates.' *Health Economics*, **23** (9), pp. 1084–1097, *2014*.

[13] Hu, Yingyao, 'Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution.' *Journal of Econometrics*, **144** (1), pp. 27–61, *2008*.

[14] ——— & Susanne M Schennach, 'Instrumental variable treatment of nonclassical measurement error models.' *Econometrica*, **76** (1), pp. 195–216, *2008*.

[15] Idler, Ellen L & Ronald J Angel, 'Self-rated health and mortality in the NHANES-I Epidemiologic Follow-up Study..' *American Journal of Public Health*, **80** (4), pp. 446–452, *1990*.

[16] ——— & Yael Benyamini, 'Self-rated health and mortality: a review of twenty-seven community studies.' *Journal of Health and Social Behavior*, pp. 21–37, *1997*.

[17] Kane, Thomas J, Cecilia Elena Rouse, & Douglas Staiger, 'Estimating returns to schooling when schooling is misreported.' *NBER Working Paper Series, Paper 7235*, *1999*.

[18] Kaplan, Mark S, Jean-Marie Berthelot, David Feeny, Bentson H McFarland, Saeeda Khan, & Heather Orpana, 'The predictive validity of health-related quality of life measures: mortality in a longitudinal population-based study.' *Quality of Life Research*, **16** (9), pp. 1539–1546, *2007*.

[19] Lewbel, Arthur, 'Estimation of average treatment effects with misclassification.' *Econometrica*, **75** (2), pp. 537–551, *2007*.

[20] Lindeboom, Maarten & Marcel Kerkhofs, 'Health and work of the elderly: subjective health measures, reporting errors and endogeneity in the relationship between health and work.' *Journal of Applied Econometrics*, **24** (6), pp. 1024–1046, *2009*.

[21] Mahajan, Aprajit, 'Identification and estimation of regression models with misclassification.' *Econometrica*, **74** (3), pp. 631–665, *2006*.

[22] McCallum, John, Bruce Shadbolt, & Dong Wang, 'Self-rated health and survival: a 7-year follow-up study of Australian elderly..' *American Journal of Public Health*, **84** (7), pp. 1100–1105, *1994*.

[23] Miilunpalo, Seppo, Ilkka Vuori, Pekka Oja, Matti Pasanen, & Helka Urponen, 'Self-rated health status as a health measure: the predictive value of self-reported health status on the use of physician services and on mortality in the working-age population.' *Journal of Clinical Epidemiology*, **50** (5), pp. 517–528, *1997*.

[24] Mossey, Jana M & Evelyn Shapiro, 'Self-rated health: a predictor of mortality among the elderly..' *American Journal of Public Health*, **72** (8), pp. 800–808, *1982*.

[25] Staub, Kevin E, 'Simple tests for exogeneity of a binary explanatory variable in count data regression models.' *Communications in Statistics: Simulation and Computation*, **38** (9), pp. 1834–1855, *2009*.

[26] Summerfield, Michelle, Simon Freidin, Markus Hahn, Peter Ittak, Ning Li, Ninette Macalalad, Nicole Watson, Roger Wilkins, & Mark Wooden, 'HILDA User Manual–Release 13.' *Melbourne Institute of Applied Economic and Social Research, University of Melbourne*, *2014*.

[27] Terza, Joseph V, Anirban Basu, & Paul J Rathouz, 'Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling.' *Journal of Health Economics*, **27** (3), pp. 531–543, *2008*.

[28] Wooldridge, Jeffrey M, 'Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables.' *Journal of Econometrics*, **182** (1), pp. 226–234, *2014*.

Appendix

A.1 General model with categorical health

The pendant to equation (6) in the case of an ordinal regressor h_i^* is

$$\begin{aligned} F(r_0, r_1, r_2) &= \sum_{j=0}^{4} p_j^* \, F(r_0, r_1, r_2 | h_i^* = j) \\ &= \sum_{j=0}^{4} p_j^* \, F(r_0 | h_i^* = j) F(r_1 | h_i^* = j) F(r_2 | h_i^* = j), \end{aligned} \tag{23}$$

where, naturally, $\pi_0 = 1 - \sum_{j=1}^{4} \pi_j$.

For $F(r_0|h_i^*)$ we have:

$$\begin{aligned} F(r_0|h_i^* = j) &= \Lambda\left(\alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta}\right)^{r_0} \left(1 - \Lambda\left(\alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)^{1-r_0} \quad \text{for } j = 1, 2, 3, 4 \\ F(r_0|h_i^* = 0) &= \Lambda\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)^{r_0} \left(1 - \Lambda\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)^{1-r_0} \end{aligned}$$

And for $F(r_m|h^*)$:

$$F(r_m|h_i^* = j) = (\delta_{0|j}^m)^{\mathbb{1}(r_m=0)} (\delta_{1|j}^m)^{\mathbb{1}(r_m=1)} (\delta_{2|j}^m)^{\mathbb{1}(r_m=2)} (\delta_{3|j}^m)^{\mathbb{1}(r_m=3)} (\delta_{4|j}^m)^{\mathbb{1}(r_m=4)}, \quad \text{for } j = 1, 2, 3, 4.$$

In this formula, there is always one $\delta_{k|j}^m$ with $j = k$. These are defined as

$$\delta_{j|j}^m = P(h_{mi} = j | h_i^* = j) = 1 - \sum_{k \neq j} \delta_{k|j}^m. \tag{24}$$

The condition to avoid mirror solutions in the case with multiple categories of SAH is that the probability of truthfully reporting a health level j is larger than any probability of misreporting it:

$$\delta_{j|j}^m > \delta_{k|j}^m, \quad \forall j, k,$$

which is a generalisation of the condition (7) for the case of two categories. To implement this constraint in the estimation, we use the following parametrisation of misreporting probabilities:

$$\delta_{k|j}^m = \frac{\exp(-\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{k|j}^m))}{1 + \sum_{k:k\neq j} \exp(-\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{k|j}^m))}, \tag{25}$$

which in turn is a generalisation of (8). With covariates, the multinomial logit model is the natural generalisation of (9), the model for true health conditional on $\boldsymbol{x}_i$:

$$\pi_{j,i} \equiv P(h_i^* = j | \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\eta}_j)}{1 + \sum_{j=0}^{4} \exp(\boldsymbol{x}_i'\boldsymbol{\eta}_j)}. \tag{26}$$

A.2 Counts and durations: Poisson and Weibull models

The proposed approach can be extended to many common nonlinear models that follow the form

$$f(y_i|h_i^*, \boldsymbol{x}_i) = g(\alpha\, h_i^* + \boldsymbol{x}_i'\boldsymbol{\beta};\ \omega), \tag{27}$$

where $f(\cdot|\cdot)$ is a functional of the conditional distribution of y_i given true health status h_i^* and a $K \times 1$ vector of covariates $\boldsymbol{x}_i$, and $g(\cdot)$ is a known nonlinear function, which might include ancillary parameters ω. To avoid notational clutter, h_i^* is binary. Typical examples for $f(y_i|h_i^*, \boldsymbol{x}_i)$ include it being a survival rate (probability), the time until developing a health condition (hazard rate), the number of doctor visits (count), or expenditures for health care (nonlinear expectation).

For instance, if y_i follows a Poisson distribution we might use the specification

$$P(y_i|h_i^*, \boldsymbol{x}_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \qquad \lambda_i = \exp(\alpha\, h_i^* + \boldsymbol{x}_i'\boldsymbol{\beta}), \tag{28}$$

where the left-hand-side of (28) corresponds to $f(\cdot|\cdot)$ and the right-hand-side to $g(\cdot)$. We can use the EM algorithm described in (15)-(17) to estimate this model directly, simply by replacing $F(y_i|h_i^*)$ in those equations by $P(y_i|h_i^*, \boldsymbol{x}_i)$ from (28). Alternatively, one could also base estimation of the Poisson model on its expectation $E(y_i|h_i^*, \boldsymbol{x}_i) = \lambda_i$ and use the GMM approach based on moment conditions

$$E\Big((y_i - \lambda_i)\boldsymbol{x}_i\Big) = 0, \tag{29}$$

where, here, $f(y_i|h_i^*, \boldsymbol{x}_i) = E(y_i|h_i^*, \boldsymbol{x}_i)$ and $g(\alpha\, h_i^* + \boldsymbol{x}_i'\boldsymbol{\beta}) = \lambda_i$.

Similarly, if y_i was a duration and followed a Weibull distribution with parameters λ_i and ω, we could estimate the model using the EM algorithm. The corresponding $F(y_i|h_i^*)$ term in this case would simply be the probability density function

$$f(y_i|h_i^*, \boldsymbol{x}_i) = \lambda_i \omega y_i^{\omega-1} \exp(-\lambda y_i^{\omega}), \qquad \lambda_i = \exp(\alpha\, h_i^* + \boldsymbol{x}_i'\boldsymbol{\beta}). \tag{30}$$

A.3 GMM estimator

To estimate the model by GMM, we use the indicator variables $I_i^{r_0 r_1 r_2}$, defined as

$$I_i^{r_0 r_1 r_2} \equiv \mathbb{1}(y_i = r_0,\ h_{1i} = r_1,\ h_{2i} = r_2),$$

and which are equal to one if all their arguments are true, and equal to zero otherwise. We then base estimation on the $7 \times K$ moment conditions of the form

$$E\big(\ \big[I_i^{r_0 r_1 r_2} - F_i(r_0, r_1, r_2)\big]\, \boldsymbol{x}_i\ \big) = 0, \tag{31}$$

for seven unique values of the triplet (r_0, r_1, r_2) —e.g., (0,0,0), (0,0,1), (0,1,0), etc.—, and where K is the number of regressors in $\boldsymbol{x}_i$ including a constant. (The eighth variable, say I_i^{111}, is linearly dependent of the other seven; as is $F(1,1,1)$ of the other seven $F(r_0, r_1, r_2)$. Thus, the eighth

equation provides no additional information and is discarded.) We obtain, $\hat{\boldsymbol{\theta}}$, an estimate for $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}', p^*, \delta^1_{0|1}, \delta^2_{0|1}, \delta^1_{1|0}, \delta^2_{1|0})$, by solving the GMM minimisation problem

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \boldsymbol{Q}_i(\boldsymbol{\theta})' \boldsymbol{W}_N \boldsymbol{Q}_i(\boldsymbol{\theta}), \tag{32}$$

where the $[7K \times 1]$-vector of moment conditions is

$$\boldsymbol{Q}_i(\boldsymbol{\theta}) = \begin{pmatrix} \left[I_i^{000} - F_i(0,0,0)\right] \boldsymbol{x}_i \\ \left[I_i^{001} - F_i(0,0,1)\right] \boldsymbol{x}_i \\ \left[I_i^{010} - F_i(0,1,0)\right] \boldsymbol{x}_i \\ \left[I_i^{011} - F_i(0,1,1)\right] \boldsymbol{x}_i \\ \left[I_i^{100} - F_i(1,0,0)\right] \boldsymbol{x}_i \\ \left[I_i^{101} - F_i(1,0,1)\right] \boldsymbol{x}_i \\ \left[I_i^{110} - F_i(1,1,0)\right] \boldsymbol{x}_i \end{pmatrix},$$

and $\boldsymbol{W}_N$ is a $[7K \times 7K]$ positive definite weighting matrix with plim $\boldsymbol{W}$. The weighting matrix $\boldsymbol{W}_N$ may be specified as the identity matrix, or estimated in an optimal two-step approach. Note that the i subscript for the joint probabilities $F_i(r_0, r_1, r_2)$ stems from the dependence of these terms on $\boldsymbol{x}_i$.

A.4 Additional results

Table A1: Simulation results: Counts (Poisson) and Durations (Weibull) DGPs

		Poisson, $N = 1,000$				Weibull, $N = 1,000$			
		h^*	h_1	FM	PFM	h^*	h_1	FM	PFM
$\hat{\alpha}$	Bias	-0.004	-0.469	0.001	-0.008	-0.000	-0.469	0.005	-0.007
	RMSE	0.049	0.474	0.073	0.073	0.072	0.470	0.117	0.113
$\hat{\beta}$ const	Bias	0.001	0.197	-0.002	0.005	0.005	0.191	0.008	0.019
	RMSE	0.055	0.210	0.101	0.095	0.076	0.193	0.141	0.117
$\hat{\beta}$ slope	Bias	0.005	0.126	-0.001	0.005	-0.002	0.129	-0.000	0.015
	RMSE	0.058	0.173	0.085	0.077	0.113	0.134	0.198	0.155
$\hat{\eta}$ const	Bias			0.003	0.011			-0.034	-0.065
	RMSE			0.316	0.278			0.514	0.375
$\hat{\eta}$ slope	Bias			0.009	-0.043			0.036	-0.021
	RMSE			0.459	0.403			0.773	0.522
$\hat{\gamma}^1_{1\|0}$ const	Bias			-0.028	0.113			-0.014	0.115
	RMSE			0.483	0.287			0.629	0.317
$\hat{\gamma}^1_{1\|0}$ slope	Bias			0.118	-0.259			0.137	-0.366
	RMSE			0.880	0.482			1.234	0.572
$\hat{\gamma}^2_{1\|0}$ const	Bias			-0.028	0.209			-0.124	0.198
	RMSE			0.448	0.326			0.728	0.338
$\hat{\gamma}^2_{1\|0}$ slope	Bias			0.043	-0.368			0.211	-0.417
	RMSE			0.722	0.549			1.317	0.621
$\hat{\gamma}^1_{0\|1}$ const	Bias			-0.022	0.063			0.024	0.131
	RMSE			0.262	0.217			0.410	0.287
$\hat{\gamma}^1_{0\|1}$ slope	Bias			0.034	-0.084			-0.002	-0.139
	RMSE			0.363	0.299			0.580	0.363
$\hat{\gamma}^2_{0\|1}$ const	Bias			-0.042	0.156			-0.045	0.209
	RMSE			0.366	0.286			0.442	0.334
$\hat{\gamma}^2_{0\|1}$ slope	Bias			0.053	-0.209			0.072	-0.249
	RMSE			0.485	0.390			0.584	0.433

Table A2: Simulation results: Full results—Multivariate DGP $\boldsymbol{y} = (y_1, y_2)'$, $N = 1,000$

		FM					PFM				
$\rho =$		1.00	0.75	0.50	0.25	0.00	1.00	0.75	0.50	0.25	0.00
$\hat{\alpha}$	Bias	0.059	0.039	0.015	0.007	0.001	0.045	0.029	0.009	0.001	-0.001
	RMSE	0.309	0.289	0.283	0.286	0.281	0.284	0.265	0.262	0.266	0.263
$\hat{\beta}$ const	Bias	0.007	0.010	0.027	0.033	0.033	0.044	0.044	0.059	0.064	0.065
	RMSE	0.354	0.326	0.313	0.312	0.304	0.249	0.233	0.228	0.227	0.230
$\hat{\beta}$ slope	Bias	0.004	0.017	0.008	0.003	0.004	0.000	0.011	0.002	0.002	-0.000
	RMSE	0.474	0.439	0.423	0.423	0.420	0.338	0.314	0.318	0.315	0.322
$\hat{\eta}$ const	Bias	-0.145	-0.142	-0.153	-0.151	-0.118	-0.242	-0.223	-0.229	-0.229	-0.216
	RMSE	1.157	1.093	1.038	1.009	0.987	0.652	0.587	0.578	0.554	0.544
$\hat{\eta}$ slope	Bias	-0.009	0.013	0.028	0.059	0.035	-0.047	-0.047	-0.040	-0.035	-0.034
	RMSE	1.562	1.509	1.446	1.400	1.395	0.676	0.643	0.627	0.611	0.613
$\hat{\gamma}^1_{1\|0}$ const	Bias	-0.093	-0.087	-0.078	-0.121	-0.099	-0.072	-0.060	-0.070	-0.071	-0.061
	RMSE	1.601	1.525	1.384	1.385	1.321	0.444	0.412	0.394	0.378	0.372
$\hat{\gamma}^1_{1\|0}$ slope	Bias	0.019	0.025	0.057	0.163	0.214	-0.479	-0.467	-0.454	-0.442	-0.435
	RMSE	2.811	2.756	2.594	2.519	2.477	0.741	0.726	0.706	0.696	0.684
$\hat{\gamma}^2_{1\|0}$ const	Bias	-0.201	-0.267	-0.257	-0.297	-0.216	0.114	0.116	0.117	0.114	0.118
	RMSE	1.645	1.939	1.598	1.998	1.466	0.455	0.414	0.406	0.394	0.387
$\hat{\gamma}^2_{1\|0}$ slope	Bias	0.297	0.368	0.339	0.417	0.312	-0.436	-0.426	-0.434	-0.435	-0.424
	RMSE	2.525	2.767	2.500	2.842	2.371	0.717	0.687	0.685	0.679	0.660
$\hat{\gamma}^1_{0\|1}$ const	Bias	0.094	0.114	0.108	0.096	0.073	0.148	0.131	0.130	0.131	0.125
	RMSE	0.960	0.964	0.894	0.871	0.846	0.454	0.403	0.387	0.380	0.371
$\hat{\gamma}^1_{0\|1}$ slope	Bias	0.066	0.015	0.026	0.027	0.054	-0.048	-0.041	-0.042	-0.045	-0.047
	RMSE	1.310	1.287	1.235	1.209	1.214	0.489	0.459	0.451	0.447	0.443
$\hat{\gamma}^2_{0\|1}$ const	Bias	0.068	0.048	0.038	0.030	0.012	0.291	0.276	0.281	0.276	0.265
	RMSE	0.774	0.710	0.691	0.701	0.640	0.474	0.447	0.440	0.427	0.413
$\hat{\gamma}^2_{0\|1}$ slope	Bias	0.050	0.058	0.063	0.062	0.070	-0.225	-0.219	-0.227	-0.226	-0.220
	RMSE	1.044	0.960	0.950	1.014	0.872	0.497	0.481	0.478	0.470	0.461

Table A3: Estimation results: Specification with discretised continuous variables

	System PFM				System PFM			
Dep. var./*diff. PFM–naïve*	Dead	*diff.*	Cond.	*diff.*	Dead	*diff.*	Cond.	*diff.*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
α_1	-0.93**	-0.06	-0.28*	0.00	-0.78**	-0.08	-0.16	-0.00
	(0.13)	(0.07)	(0.17)	(0.09)	(0.13)	(0.06)	(0.17)	(0.10)
α_2	-1.40**	-0.09	-0.77**	-0.01	-1.12**	-0.14**	-0.44**	0.00
	(0.13)	(0.06)	(0.16)	(0.08)	(0.14)	(0.06)	(0.17)	(0.09)
α_3	-1.88**	-0.20**	-1.14**	-0.05	-1.45**	-0.16**	-0.74**	-0.03
	(0.14)	(0.07)	(0.16)	(0.08)	(0.15)	(0.07)	(0.17)	(0.09)
α_4	-2.13**	-0.21**	-1.56**	-0.16*	-1.72**	-0.25**	-1.14**	-0.17*
	(0.19)	(0.10)	(0.18)	(0.09)	(0.20)	(0.11)	(0.20)	(0.10)
age: 30s	1.09**	0.00	0.54**	0.01	1.27**	-0.00	0.56**	0.01
	(0.25)	(0.01)	(0.12)	(0.01)	(0.26)	(0.01)	(0.13)	(0.00)
age: 40s	1.46**	-0.02*	0.84**	-0.00	1.66**	-0.01*	0.86**	-0.01
	(0.24)	(0.01)	(0.12)	(0.01)	(0.25)	(0.01)	(0.12)	(0.01)
age: 50s	2.19**	-0.06**	1.08**	-0.01	2.41**	-0.03**	1.08**	-0.00
	(0.23)	(0.01)	(0.12)	(0.01)	(0.24)	(0.01)	(0.13)	(0.01)
age: 60s	3.41**	-0.05**	1.43**	-0.03**	3.61**	-0.01	1.43**	-0.02**
	(0.23)	(0.01)	(0.13)	(0.01)	(0.24)	(0.01)	(0.14)	(0.01)
age: 70 plus	5.00**	-0.07**	1.69**	-0.07**	5.16**	-0.04**	1.72**	-0.05**
	(0.23)	(0.01)	(0.16)	(0.02)	(0.24)	(0.01)	(0.17)	(0.01)
male	0.59**	-0.00	-0.12*	-0.00	0.58**	0.00	-0.15**	0.00
	(0.08)	(0.01)	(0.07)	(0.01)	(0.08)	(0.01)	(0.07)	(0.00)
education: year 12	0.11	0.05**	-0.13	0.03**	0.14	0.04**	-0.10	0.02**
	(0.13)	(0.01)	(0.11)	(0.01)	(0.14)	(0.01)	(0.11)	(0.01)
education: certificate	-0.14	0.01	-0.05	0.00	-0.13	0.00	-0.03	-0.01
	(0.09)	(0.01)	(0.08)	(0.01)	(0.09)	(0.01)	(0.08)	(0.01)
education: bachelor	-0.11	0.02*	-0.36**	0.01**	-0.07	0.01	-0.31**	0.00
	(0.13)	(0.01)	(0.10)	(0.01)	(0.13)	(0.01)	(0.11)	(0.01)
HH income, 2nd quint.	-0.17*	-0.01	-0.32**	-0.02**	-0.06	-0.01	-0.25**	-0.01*
	(0.10)	(0.01)	(0.10)	(0.01)	(0.10)	(0.01)	(0.11)	(0.01)
HH income, 3rd quint.	-0.25**	0.00	-0.30**	-0.00	-0.13	-0.00	-0.22**	-0.00
	(0.11)	(0.01)	(0.11)	(0.01)	(0.12)	(0.01)	(0.11)	(0.01)
HH income, 4th quint.	-0.19	0.05**	-0.34**	0.01	-0.03	0.02**	-0.23**	0.00
	(0.12)	(0.01)	(0.11)	(0.01)	(0.12)	(0.01)	(0.11)	(0.01)
HH income, 5th quint.	-0.44**	0.07**	-0.41**	0.04**	-0.26*	0.04**	-0.29**	0.03**
	(0.13)	(0.01)	(0.11)	(0.01)	(0.14)	(0.01)	(0.12)	(0.01)
chronic condition					0.30**	-0.06**	0.40**	-0.02
					(0.09)	(0.02)	(0.09)	(0.02)
married					-0.53**	-0.01*	-0.11	-0.01*
					(0.08)	(0.01)	(0.08)	(0.00)
overseas					-0.24**	0.00	-0.02	0.01*
					(0.08)	(0.01)	(0.08)	(0.01)
not in labour force					0.20*	-0.03**	0.11	-0.00
					(0.11)	(0.01)	(0.09)	(0.01)
unemployed					0.05	-0.03*	0.27	-0.02
					(0.26)	(0.02)	(0.17)	(0.01)
smoker					0.49**	-0.00	0.30**	-0.00
					(0.08)	(0.01)	(0.07)	(0.00)
N	12,908		7,340		12,908		7,340	

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$

Table A4: Descriptive statistics for additional discretised variables

Variable	*N*	Mean	Std.Dev.
Covariates (Wave 1)			
age: 30s (=1 if 30 years≤age< 40 years)	12,908	0.209	0.407
age: 40s (=1 if 40 years≤age< 50 years)	12,908	0.200	0.400
age: 50s (=1 if 50 years≤age< 60 years)	12,908	0.150	0.358
age: 60s (=1 if 60 years≤age< 70 years)	12,908	0.102	0.303
age: 70 plus (=1 if age≥ 70 years)	12,908	0.101	0.301
education: year 12 (=1 if highest education Year 12)	12,908	0.145	0.353
education: certificate (=1 if highest education certificate)	12,908	0.256	0.437
education: bachelor (=1 if highest education bachelor or higher)	12,908	0.178	0.382
HH income, 2nd quint. (=1 if HH income in 2nd quintile)	12,908	0.200	0.400
HH income, 3rd quint. (=1 if HH income in 3rd quintile)	12,908	0.200	0.400
HH income, 4th quint. (=1 if HH income in 4th quintile)	12,908	0.200	0.400
HH income, 5th quint. (=1 if HH income in 5th quintile)	12,908	0.200	0.400

Table A5: Estimation results: System PFM specifications with interactions in health

	Interaction w. education			
	Dead	*diff.*	Cond.	*diff.*
educ	-0.43	-0.11	-3.15**	-1.06*
	(0.76)	(0.14)	(1.07)	(0.55)
α_1: educ	0.28	-0.07	3.49**	1.32*
	(0.90)	(0.26)	(1.16)	(0.68)
α_1: cons	-1.12	0.03	-4.37**	-1.57*
	(1.09)	(0.32)	(1.41)	(0.81)
α_2: educ	0.43	0.46*	2.66**	1.17**
	(0.83)	(0.24)	(1.10)	(0.58)
α_2: cons	-1.64	-0.65**	-3.61**	-1.37*
	(1.02)	(0.30)	(1.34)	(0.70)
α_3: educ	0.38	-0.05	2.57**	0.97*
	(0.85)	(0.26)	(1.10)	(0.58)
α_3: cons	-1.91*	-0.04	-3.80**	-1.15
	(1.05)	(0.34)	(1.34)	(0.71)
α_4: educ	-0.02	0.05	2.49**	1.12*
	(1.03)	(0.37)	(1.18)	(0.64)
α_4: cons	-1.70	-0.27	-4.09**	-1.49*
	(1.29)	(0.48)	(1.45)	(0.78)
N	12,908	12,908	7,340	7,340

	Interaction w. log HH income			
	Dead	*diff.*	Cond.	*diff.*
lnehi	-0.16	0.05	-0.19	-0.07
	(0.18)	(0.04)	(0.24)	(0.11)
α_1: lnehi	0.14	0.05	0.07	0.08
	(0.21)	(0.07)	(0.27)	(0.13)
α_1: cons	-1.18**	-0.19	-0.35	-0.22
	(0.58)	(0.18)	(0.79)	(0.36)
α_2: lnehi	0.01	-0.06	0.01	0.09
	(0.20)	(0.06)	(0.26)	(0.12)
α_2: cons	-1.15**	0.09	-0.43	-0.24
	(0.56)	(0.17)	(0.75)	(0.34)
α_3: lnehi	0.03	-0.03	0.01	0.08
	(0.21)	(0.08)	(0.26)	(0.12)
α_3: cons	-1.51**	-0.02	-0.74	-0.24
	(0.60)	(0.23)	(0.76)	(0.34)
α_4: lnehi	0.18	-0.09	-0.05	0.08
	(0.28)	(0.12)	(0.29)	(0.13)
α_4: cons	-2.31**	0.04	-0.95	-0.39
	(0.87)	(0.38)	(0.88)	(0.40)
N	12,908	12,908	7,340	7,340

	Interaction w. male			
	Dead	*diff.*	Cond.	*diff.*
male	0.55**	-0.04	-0.17	-0.04
	(0.24)	(0.05)	(0.30)	(0.12)
α_1: male	-0.10	-0.03	0.18	0.07
	(0.28)	(0.09)	(0.34)	(0.16)
α_1: cons	-0.74**	-0.04	-0.24	-0.03
	(0.21)	(0.06)	(0.23)	(0.12)
α_2: male	0.26	0.17**	-0.20	0.01
	(0.27)	(0.09)	(0.32)	(0.13)
α_2: cons	-1.27**	-0.19**	-0.34	0.01
	(0.22)	(0.07)	(0.23)	(0.11)
α_3: male	0.03	-0.05	0.11	0.08
	(0.29)	(0.11)	(0.32)	(0.13)
α_3: cons	-1.46**	-0.08	-0.77**	-0.05
	(0.23)	(0.09)	(0.23)	(0.11)
α_4: male	-0.63*	-0.20	0.55	0.25
	(0.38)	(0.16)	(0.37)	(0.16)
α_4: cons	-1.43**	-0.13	-1.37**	-0.29**
	(0.28)	(0.12)	(0.27)	(0.13)
N	12,908	12,908	7,340	7,340

Standard errors in parentheses

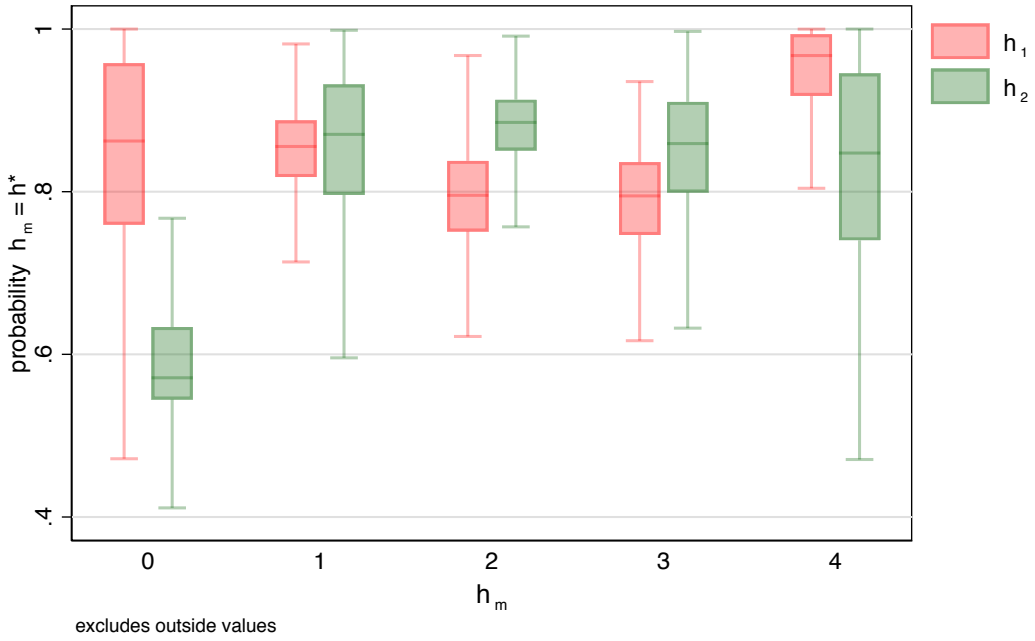
* $p < 0.10$, ** $p < 0.05$

	Interaction w. age			
	Dead	*diff.*	Cond.	*diff.*
age	-3.18	-4.19	9.04	1.76
	(6.44)	(3.30)	(7.38)	(3.96)
agesq	13.17**	3.48	-8.47	-1.92
	(5.70)	(2.66)	(7.61)	(3.78)
α_1: age	0.86	4.90	-3.62	-1.56
	(7.33)	(4.52)	(7.96)	(5.22)
α_1: agesq	-1.47	-4.12	4.72	1.56
	(6.43)	(3.67)	(8.22)	(5.07)
α_1: cons	-0.76	-1.46	0.38	0.35
	(2.06)	(1.36)	(1.88)	(1.30)
α_2: age	1.83	3.00	-2.47	-2.73
	(6.99)	(3.57)	(7.71)	(5.50)
α_2: agesq	-2.18	-2.44	5.19	3.01
	(6.19)	(2.96)	(7.96)	(5.34)
α_2: cons	-1.38	-0.97	-0.52	0.60
	(1.95)	(1.06)	(1.82)	(1.36)
α_3: age	-4.55	3.51	-3.01	-1.55
	(6.95)	(3.61)	(7.73)	(6.82)
α_3: agesq	3.63	-2.92	6.65	1.81
	(6.25)	(3.05)	(8.00)	(6.66)
α_3: cons	-0.10	-1.10	-0.91	0.31
	(1.90)	(1.05)	(1.82)	(1.68)
α_4: age	-0.43	7.64*	-6.67	-3.73
	(8.05)	(4.08)	(8.19)	(10.13)
α_4: agesq	-0.13	-6.61*	9.83	3.76
	(7.31)	(3.51)	(8.53)	(10.14)
α_4: cons	-1.44	-2.28*	-0.37	0.70
	(2.17)	(1.17)	(1.91)	(2.39)
N	12,908	12,908	7,340	7,340

Standard errors in parentheses

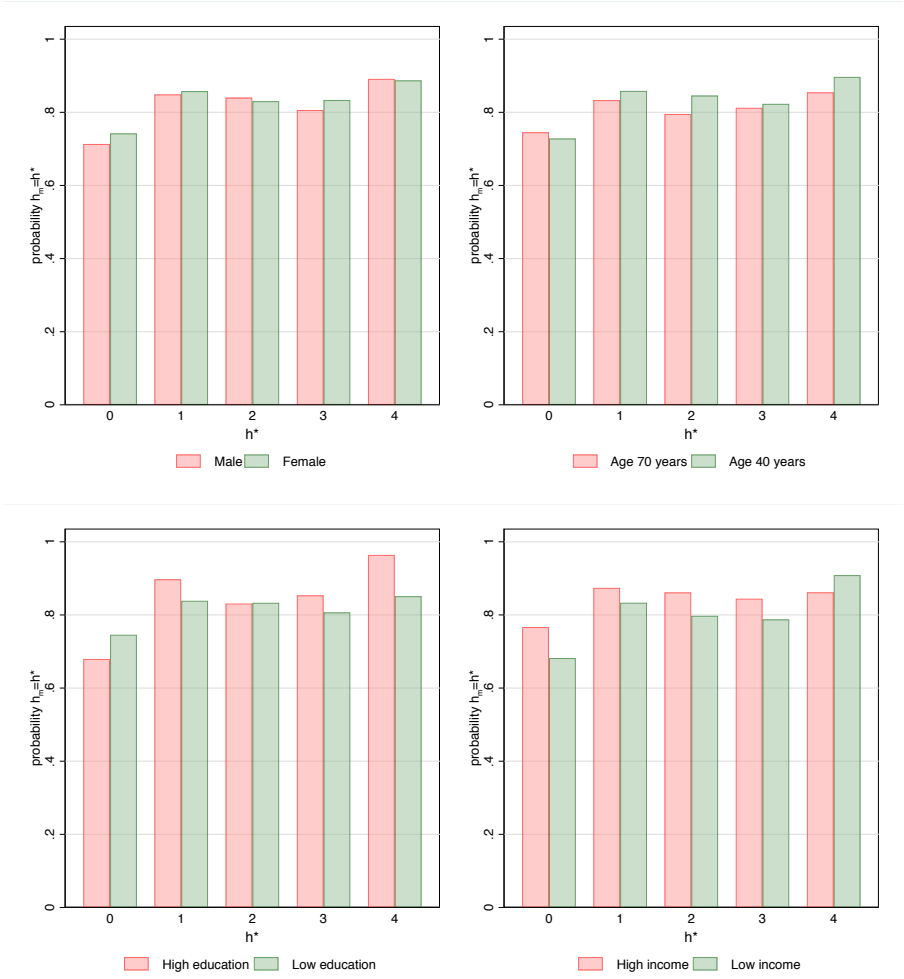
* $p < 0.10$, ** $p < 0.05$

Figure A1: Heterogeneity in the probability of reporting true health status, $N = 12,908$



Notes: Data from HILDA waves 1 and 16 for individuals who responded to SRH questions in wave 1.

Figure A2: Marginal effects of reporting true health status, $N = 12,908$



Notes: Data from HILDA waves 1 and 13 for individuals who responded to SRH questions in wave 1.