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

November 30, 2020

Abstract

Self-assessed health (SAH) is often used in health econometric models as the key explanatory variable or as a control variable. However, there is evidence questioning its test-retest reliability, with up to 30 percent of individuals changing their response. Building on recent advances in the econometrics of misclassification, we develop a way to consistently estimate and account for misclassification in reported SAH by using data from a large representative longitudinal survey where SAH was elicited twice. From this we gain new insights into the nature of SAH misclassification and its potential for biasing health econometric estimates. The results from applying our approach to nonlinear models of long-term mortality and chronic morbidities reveal that there is substantial heterogeneity in misclassification patterns. We find that adjusting for misclassification is important for estimating the impact of SAH. For other explanatory variables of interest, we find significant but generally small changes to their estimates when SAH misclassification is ignored.

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, Mark Harris, Joe Hirschberg, Maarten Lindeboom, Jenny Lye, Frank Windmeijer, Rainer Winkelmann, Eugenio Zuccheli, the participants of the European Workshop on Econometrics and Health Economics (Groningen), 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. Petrie acknowledges support from the Australian Research Council through grant DE150100309. Staub acknowledges support from the Australian Research Council through grant DE170100644. Alex Ballantyne and Edwin Chan provided excellent research assistance. The names of the authors are listed in alphabetical order.

[†]Melbourne School of Population and Global Health, 207 Bouverie Street, 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@ndph.ox.ac.uk

[§]Centre for Health Economics, Monash Business School, 900 Dandenong Road, Caulfield East, Victoria 3145, Australia. E-mail: dennis.petrie@monash.edu

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

1 Introduction

Self-assessed health (SAH) is a ubiquitous measure in the health economics literature and, more broadly, in social science research (Au & Johnston, 2014). It is often asked as a simple question, "in general, how would you rate your health?", where respondents select from categories such as excellent, very good, good, fair or poor. SAH is used variously in econometric models as the outcome variable, as the key explanatory variable or as a control variable to prevent health from confounding the effect of interest. However, there is a large literature calling into question the reliability of reported SAH, as up to 30% of individuals change their response when re-asked about their SAH (Crossley & Kennedy, 2002; Clarke & Ryan, 2006; Black *et al.*, 2017a). This paper takes advantage of recent econometric developments on misclassification and information from a prominent longitudinal survey—the Household, Income and Labour Dynamics in Australia (HILDA) survey—to gain new insights into the nature of misclassification in reported SAH and its potential for biasing estimates of the effects of SAH and other explanatory variables in health econometrics models. In particular, our analysis uses data from the 2001 wave of HILDA, which records the same individual's SAH responses in two different but similar questionnaires in the same wave (face-to-face or over-the-phone interview, and on a self-completion questionnaire), and combines this information with longitudinal data on mortality and the development of chronic health conditions 15 years later. We develop a new likelihood-based nonlinear estimator which uses this information to jointly estimate the misclassification in both reported SAH measures as well as the effects of SAH on mortality and morbidity.

Two independent misclassified measures of a categorical variable, such as SAH, supplemented with data on an outcome, such as mortality, can identify all the misclassification probabilities and the effect of the variable on the outcome (Hu, 2008, 2017). This can be done without imposing virtually any restrictions on the misclassification patterns, such as assuming that the probabilities of certain forms of misclassification are zero, that certain misclassification probabilities are larger than others, or that the misclassification probabilities are the same for both SAH measures. In our case, the flexibility of allowing that each measure may have differing levels of misclassification is important because the mode with which the question was asked is different.[1] While infinitely many misclassification patterns are compatible with the observed data on only two reported SAH measures,[2] adding information about an outcome affected by SAH such as mortality allows us to pin down the misclassification probabilities. The reason is that each possible misclassification pattern implies a unique distribution of SAH within each reporting group, so that the average outcome within each group provides the missing information needed to reveal the misclassification pattern present in the reported SAH data. For instance, consider (a) the group of individuals reporting "excellent" health according to the first measure and "very good" according to the second, versus (b) the group responding "excellent" in both. If the individuals in

[1] And, more generally, the misclassification probabilities may vary across the measures due to the nature of priming questions. For example, if the respondent had first been asked others questions about their health conditions, this may reduce the level of misclassification; or, if the question came late in the survey when respondents may be fatiguing and losing concentration, this might increase the likelihood of misreporting SAH.

[2] For example, it is not possible to distinguish the case of the first reported SAH being severely misclassified and the second being almost error-free from the case of misclassification being equally severe for both measures.

both groups are mainly in "excellent" health, then they should have similar mortality. But if the individuals in (a) are mainly in "very good" health whereas those in (b) are mainly in "excellent" health, then the mortality of group (a) is likely to be different to the mortality of (b). Thus, looking at these three variables jointly (the two reported SAH plus mortality) can identify all underlying misclassification probabilities. And, conversely, because identifying misclassification is tantamount to knowing the underlying distribution of SAH within each group, it also makes it possible to back out the true impact of SAH on the outcome. In the next section (Section 2), we present a more detailed example of this identification strategy, and in Section 3 we show how this formally generalises to a full econometric model which can be richly parametrised in terms of covariates. However, estimating such a model is not straightforward.

While Hu (2008) discusses a nonparametric estimator for this setting, the implementation of that estimator is non-trivial and its computation is prohibitive when, as in our case, the model has many covariates, the potentially misclassified variable (SAH) has many categories, and the sample size is large.[3] Therefore, we develop a more easily implementable parametric likelihood-based estimator. An important advantage of our estimator is that the effects of categorical SAH are specified by including dummy variables for each category of SAH in the outcome model, the standard way SAH is included as a categorical regressors in the health economics literature. Our approach also lends itself easily to specifications with interaction effects where the impact of unobserved SAH differs depending on other individual characteristics. Such specifications, common in applied work to investigate the heterogeneity of the effect of SAH, have received little attention in the misclassification literature so far. Another advantage is that, because the model has a finite-mixture representation, our estimation approach is a flexible parametric specification estimated via a standard expectation-maximisation (EM) algorithm, which offers fast and reliable computation. The flexibility and richness of our model, where we allow unrestricted patterns of misclassification that depend on all covariates, means that the likelihood is complex and difficult to maximise. The EM algorithm provides the key to a simpler and more direct path to the solution. By holding misclassification constant in the maximisation step, the resulting log likelihood is substantially simpler: it becomes additively separable, so that components can quickly be maximised separately. Moreover, because it is likelihood-based, our estimator can be easily adapted to encompass several outcomes jointly (such as, in our case, mortality and chronic morbidity) and be further extended to consider the pensalisation of misclassification parameters to improve stability and efficiency. Section 4 provides simulation evidence on our estimators finite sample performance.

The focus of this paper is the application of our proposed estimator to the HILDA data with the aim of making two key contributions to the health economics literature. First, we go beyond the current literature which only documents observed differences in multiple reported measures of SAH, typically by regressing an indicator of conflicting SAH answers on a set of explanatory variables (Black *et al.*, 2017a). As an example of the difficulties associated with interpreting some of the estimates produced

[3] To the best of our knowledge, to date there is no paper that uses this estimator in a setting comparable to ours, with many covariates and a misclassified variable that has many categories. The illustrative application in Hu (2008) is an order of magnitude smaller than ours in terms of sample size (N=1,688), has a only five covariates, and the specification of the outcome model is restricted by assuming that the key categorical variable has a linear effect on the outcome.

with this method, consider for instance the finding that individuals with lower education are more prone to giving conflicting SAH answers when asked twice. It is generally not possible to conclude from such a finding which of the two reported SAH questions is answered more accurately, and which types of specific mistakes are made with which frequency. It is not even possible to conclude that individuals with lower education tend to have generally higher rates of misclassification than higher education individuals, since it could be, for instance, that face-to-face SAH from low education individuals is much less reliable, while self-completed SAH from low education individuals is somewhat *more* reliable. In contrast to these reduced-form approaches, our new framework provides estimates of the complete set of probabilities of misreporting each category of each measure by covariates such as education, income, etc. By linking differences in SAH to underlying misclassification probabilities, it makes it possible to address behavioural questions about the extent, patterns and heterogeneity of individuals' responses. It also makes it possible to assess questions pertaining to survey methodology, such as the type and incidence of response errors associated with each of the two survey instruments—face-to-face interview and self-completion questionnaire.

Second, our results make it possible to assess how biased conventional estimates of the effects of reported SAH are by misclassification. In our approach, the outcome model takes the form of a standard nonlinear model, such as a logit model, and can be specified not just in terms of SAH but also by including a vector of covariates. This makes it straightforward to compare our estimates of the outcome model to naïve estimates which ignore misclassification—that is, simple logit models of mortality and morbidity that include either the first or second reported SAH measure, as widely encountered in the health economics literature. Our estimator provides a way to adjust for misclassification in reported SAH when estimating the effects of SAH on mortality and morbidity in such models. As mentioned above, once the misclassification probabilities are identified, the effect of SAH on, say, mortality can be backed out because, for each reported SAH group, the group's distribution of SAH can be inferred and linked to the group's mortality. Similarly, our approach also makes it possible to assess how biases stemming from misclassification of reported SAH affect the estimates of *other* regressors of interest. Such spillover of the bias in reported SAH to other regressors can occur if the latter are correlated to SAH. Intuitively, one can understand the use of reported SAH as introducing a type of omitted variable problem: part of SAH is missing in the reported measure. If covariates are correlated to SAH (and thus also to the omitted part of SAH), this will bias the coefficients on these other variables. And due to the bias on the effect of SAH itself, even the non-omitted part is not being adjusted for appropriately, which will also further spill over to these correlated variables. Bago d'Uva *et al.* (2011) also look at such spillovers, albeit for a different outcome and with an approach based on vignettes.

Understanding and dealing with measurement errors in SAH has been and still is an active area of research within health econometrics. While Greene *et al.* (2018) and Brown *et al.* (2018) adjust for untruthful reporting in discrete dependent variable models, our focus lies in the case of discrete SAH taking the role of a regressor. In models seeking to explain labour supply decisions a key focus has been on individuals misclassifying (under-reporting) their SAH to justify not working (Bound, 1991; Currie & Madrian, 1999; Lindeboom & Kerkhofs, 2009; Black *et al.*, 2017b).[4] This can be

[4] Black *et al.* (2017b) find, using HILDA data, that, compared to workers, individuals not working were more likely to

problematic in these models because it can upwardly bias the estimated coefficient on SAH, but what is less commonly noted is that other reasons why SAH is misclassified will cause bias in the other direction (Bound, 1991). In the current paper, we consider 'to justify not working' as one of many reasons which may explain an individual's propensity (or probability) to misclassify their SAH. Our estimator can fully account for misclassification related to work status or any other factor, as long as the respondents are not misclassifying SAH to justify the outcome we use for identification (in our case mortality or the onset of chronic condition 15 years in the future). Our paper also contributes to the literature which investigates the association between "objective" and self-assessed health measures (Bound, 1991; Mossey & Shapiro, 1982; Butler *et al.*, 1987; Baker *et al.*, 2004; Doiron *et al.*, 2015). We consider substantially longer-term associations between SAH and mortality (and morbidity) than in these studies (15 years vs 3-6 years), and we adjust the association by accounting for misclassification in reported SAH. The most closely related studies to ours are Crossley & Kennedy (2002), Clarke & Ryan (2006) and Black *et al.* (2017a), which also consider the change in an individual's response when SAH is asked twice; however, none of these papers estimates the impact of misclassification when reported SAH is used as a regressor, nor do they study the underlying misclassification probabilities.

Our main results are discussed in Sections 5 and 6. In Section 5, we document the empirical salience of the problem of differing answers to repeated SAH questions throughout the HILDA survey, which motivates our research, and we replicate the previous literature's reduced-form results by regressing these differences in SAH responses on covariates. As discussed, it is difficult to link such results to the underlying misclassification. Section 6 presents the results using our estimator on the HILDA data, which overcomes these problems. We find strong evidence for the presence of misclassification and for heterogeneity in misreporting behaviour across different population subgroups, such as male vs females and low vs high income earners. For instance, we find that men who are in excellent health almost never fail to report this in interviews, but not all men who report being in excellent health are truthfully reporting their SAH. We also document that there is less measurement error in the SAH question elicited by face-to-face interviews than in the one from the self-completion questionnaire. The results indicate that misclassification leads to statistically significant biases in the parameters of the mortality and morbidity models. While the bias is similar in absolute size across the models, this translates to relative biases in the coefficients of SAH ranging mostly from 10 to 20 percent in the mortality model, and as high as 100 percent for the morbidity model. For the coefficients of other covariates, the biases, while statistically significant, are more moderate and around 10 percent. Finally, we use our approach to estimate potential heterogeneity in the effect of SAH by specifying models with interactions of SAH with sex, age, education and income. With the exception of gender differences in mortality, the results indicate that the long-term effects of SAH on mortality and new chronic conditions are quite homogenous.

We conclude the paper in Section 7. Our findings suggest that when specifying models where SAH is

reclassify themselves as having a disability after they were asked their work status. This suggests that the wider context in which people are asked questions about their self-assessed health may also impact on their propensity to misclassify. But even with the results from Black *et al.* (2017b), we do not know the extent to which the prior question about work decreased the likelihood of the respondents concealing their disability, or increased the respondents' likelihood of reporting a non-existent disability—our framework can reveal these underlying patterns.

the regressor of interest, it is important to adjust for misclassification. In case this is not possible, SAH measures from face-to-face interviews should be strongly preferred over self-completed SAH measures. On the other hand, our findings also indicate that when specifying models where SAH is used as a key control variable, there is likely to be little contamination of the variables of interest from the misclassification in SAH.

2 Identifying misclassification in SAH: An intuitive example

To fix ideas and give the intuition behind the identification, we discuss in this section a simple example where we have two binary misclassified SAH measures with potentially different misclassification probabilities and where SAH influences the probability of being dead at some point in the future.

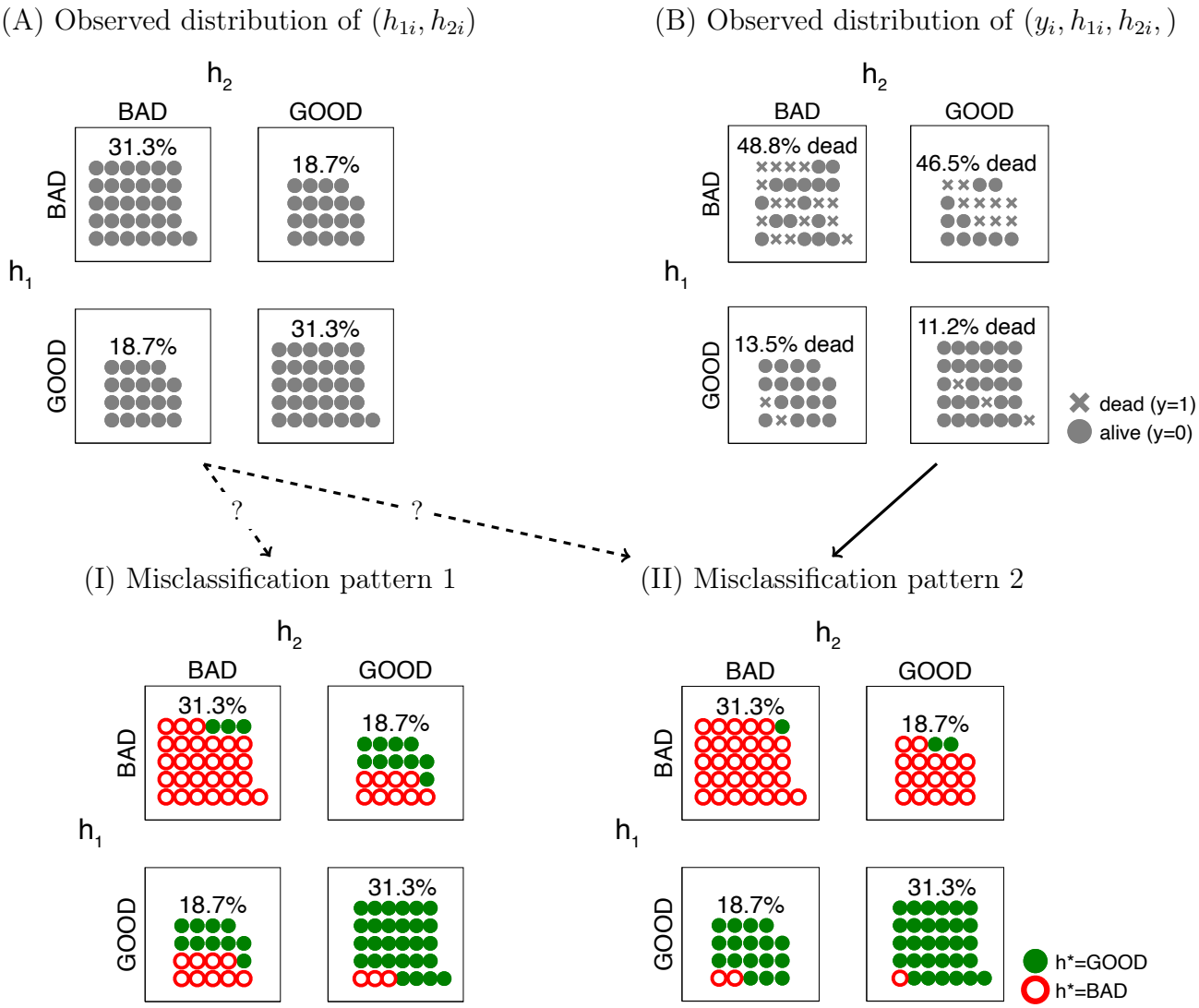
Consider a simple hypothetical setting, where we assume that individuals' self-assessed health, h^*, is either good or bad, and each of these two health groups has a fixed probability of being alive or dead in some future year. However, we do not observe individuals' SAH, h^*; we only observe two potentially misreported measures of it for each individual—h_1 and h_2—, and whether or not, by some time in the future, they are dead (y). We use this example to give the intuition on how the observed data (y, h_1, h_2) and assumptions about the nature of misclassification provide information to identify both the rates of misclassification in SAH and the relationship between SAH and mortality (Figure 1).

Panel (A) depicts the observed population distribution for the two potentially misreported measures of SAH, which results in four distinct subpopulations (the four cells of the 2×2 matrix). In this example, we use a symmetrical distribution for convenience; however, the intuition is the same for asymmetrical distributions. In the top-left cell, 31.3 percent of the population report bad health for both measures; and 31.3 percent also report good health for both measures in the bottom-right cell. However, there is some misclassification in at least one of the reported measures, as 18.7 percent of the individuals report bad health for h_1 but report good health for h_2; and, in addition, 18.7 percent report good health for h_1 but bad health for h_2. In each of these four groups there is likely to be some individuals with good and bad health (h^*).

Assume that misclassification in h_1 and h_2 are independent; that is, misreporting in h_1 does not change the probability of misreporting in h_2 or vice versa. This implies that five factors determine Panel (A), the joint distribution of (h_1, h_2): the true proportion of the population who are in good health, plus four misclassification probabilities, two for each measure (the conditional probability of reporting good health when SAH is bad, and the conditional probability of reporting bad health when SAH is good). Now given independence in misclassification and the portion of the population in the off-diagonals (18.7 percent in each), this narrows the possible set of misclassifications probabilities that could result in this observed population distribution. But, unfortunately, there is still an infinite number of possibilities.

Two such possible underlying misclassification patterns that would result in the distribution observed in Panel (A) are shown in Panels (I) and (II). The green solid dots represent those whose SAH h^* is good, while the red hollow dots represent those whose h^* is bad. Panel (I) represents a case where

Figure 1: Example: Identification of misclassification patterns from observed data



Notes: Graphical representation of identification of unobserved misclassification patterns from observed data using a fictional example of a binary unobserved health variable h^* (=1 if "GOOD",=0 if "BAD") with $P(h^* = 1) = 0.5$, two misclassified measures of health, h_1 and h_2, and an outcome y representing mortality (=1 if "dead", =0 if "alive"). Results rounded to one decimal place (numbers) or whole dots (graphs). Panels (A) and (B) contain the joint distribution of h_1 and h_2. Panel (B) superimposes the outcome y and indicates death rate $P(y = 1|h_1, h_2)$ in each cell. Panels (I) and (II) indicate two potential underlying misclassification patterns compatible with the observed distribution of h_1 and h_2. Panel (II) is the only misclassification pattern compatible with the observed distribution of h_1, h_2 and y. Details of the data generating processes: For both panels (I) and (II), $P(y = 1|h^* = 1) = 0.1$ and $P(y = 1|h^* = 0) = 0.5$. For panel (I), $P(h_m = j|h^* = k) \equiv \delta^m_{j|k} = 0.25$ for m=1,2 and all $j \neq k$. For panel (II), $\delta^1_{j|k} = 0.05$ and $\delta^2_{j|k} = 0.36$ for all $j \neq k$. See Tables A1 and A2 in the Appendix for the general equations giving the cells of these matrices in terms of the parameters $\pi \equiv P(h^* = 1)$ and $\delta^m_{j|k}$ $(m = 1, 2, j \neq k)$.

both measures h_1 and h_2 have similar rates of misclassification, whereas in Panel (II), h_1 has low rates of misclassification while h_2 has high rates of misclassification. With only data on (h_1, h_2) it is impossible to distinguish between the patterns of (I) and (II), but we now show how using the additional information about the mortality of each of the four groups (Panel (B)) can be used to identify which is the true misclassification pattern. The crosses in Panel (B) represent those in each group that had died by some future year. We can see that those who report bad health for both measures have a mortality rate of 48.8 percent, while those who report bad health for h_1 and good health for h_2 have a mortality rate of 46.5 percent. This suggests that the composition of these two groups in terms of the underlying proportion with good and bad health (h^*) is very similar. This is

compared with those who report good health for h_1 and bad health for h_2, whose mortality rate is much lower (13.5 percent), which suggests that this group has a very different underlying composition of h^* and is more similar to the group that reports good health for both measures (mortality 11.2 percent). This indicates that the misclassification pattern in Panel (I) is incompatible with the data, and that the actual misclassification in this case is that of Panel (II).

Once the misclassification is known, it is straightforward to identify the effect of health h^* on mortality, that is, the difference in mortality rates by good and bad h^*. In each of the four groups the proportion of those with good and bad health is known and so is each group's mortality rate. Thus, to reveal the two unknown mortality probabilities for those in good and bad health we can use the mortality of any two cells of Panel (B), for instance from the leading diagonal groups, since this simplifies to two equations with two unknowns,

$$\begin{aligned} 0.971\ \bar{Y}_B + 0.029\ \bar{Y}_G &= 0.488, \\ 0.029\ \bar{Y}_B + 0.971\ \bar{Y}_G &= 0.112, \end{aligned}$$

where $\bar{Y}_B \equiv P(y = 1|h* = \text{BAD})$ is the mortality rate of the group in bad health, and $\bar{Y}_G$ the analogue of those in good health. The first equation represents the top diagonal cell reporting bad health for both measures; and the second, the bottom diagonal cell. The mortality in each cell is a weighted average of the unknown mortality for those in good and bad SAH (e.g., for those reporting bad health for both measures, 97.1 percent are in bad health and 2.9 percent are in good health). Solving the equations reveals that, in our example, $\bar{Y}_G = 0.10$ and $\bar{Y}_B = 0.50$, so that the effect of SAH on mortality is -0.40.

Appendix A.7 provides the formulas for the proportion of the population in each group and the average mortality of each group which shows that the general solution for identifying the misclassification pattern and the effect of health on mortality equates to solving seven equations with seven unknowns (the five parameters of the misclassification, and the two of the mortality). Next, we consider how to set up and solve such a case in a regression framework.

3 Econometric Methods

In this section, we translate and genrealise the previous example of misclassification in SAH to a formal regression framework that can be easily applied to commonly estimated health economic models and accommodates covariates, interaction effects with unobserved SAH, multinomial SAH with more than two categories, and heterogeneous misclassification probabilities (Section 3.1). We then present an expectation-maximisation (EM) algorithm to estimate this model, and discuss two ways of potentially improving the estimation in finite samples: penalisation and system estimation (Section 3.2).[5]

[5]Econometric approaches related to ours include Gosling & Saloniki (2014) and Kane *et al.* (1999); but these papers do not include regressors and are limited to linear models, respectively. Battistin *et al.* (2014) develop a Bayesian approach. In all these papers the misclassified regressor is binary. Hu (2008) provides a nonparametric estimator for the same

3.1 Model specification

Consider a logit model for mortality, an outcome we will use in our application in Sections 5 and 6. The outcome y_i equals 1 if individual i is dead 15 years after the initial survey, and 0 otherwise. We are interested in how SAH, h_i^*, at the time of the initial survey, is related to mortality y_i. SAH is an (ordinal) categorical variable with five outcomes, $h_i^* \in \{0, 1, 2, 3, 4\}$, where $h_i^* = 0$ indicates poor health and $h^* = 4$ excellent health. The key feature of the models we consider is that SAH, h_i^*, is unknown; what is known instead is an individual's reported SAH, and this might be misclassified. Each individual reports his or her SAH twice, thus providing two potentially misclassified measures. SAH is related to the probability of dying as follows:

$$P(y_i = 1|h_i^*, \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{d}_i^{*'}\boldsymbol{\alpha} + \boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{d}_i^{*'}\boldsymbol{\alpha} + \boldsymbol{x}_i'\boldsymbol{\beta})} \equiv \Lambda(\boldsymbol{d}_i^{*'}\boldsymbol{\alpha} + \boldsymbol{x}_i'\boldsymbol{\beta}). \tag{1}$$

where represents a $\boldsymbol{x}_i'\boldsymbol{\beta}$ a linear index in $\boldsymbol{x}_i$, a $K \times 1$ vector of covariates (including a constant term) with conforming coefficient vector $\boldsymbol{\beta}$. The main interest lies in the linear index $\boldsymbol{d}_i^{*'}\boldsymbol{\alpha}$ which captures the impact of SAH. The vector $\boldsymbol{d}_i^* = (d_{1i}^*, d_{2i}^*, d_{3i}^*, d_{4i}^*)'$ consists of a set of indicator variables of a particular health status, $d_{ji}^* = \mathbb{1}(h_i^* = j)$ for $j = 1, 2, 3, 4$, where $\mathbb{1}(\cdot)$ denotes the indicator function; and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)'$ is the corresponding vector of coefficients.

If h_i^* were observed, (1) 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 $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by using two potentially misclassified SAH measures 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 SAH—as

$$\delta_{k|j}^m = P(h_{mi} = k|h_i^* = j, \boldsymbol{x}_i) \quad \forall j, k = 0, 1 \ldots, 4, \text{ and } j \neq k, \tag{2}$$

for the two reported SAH measures $m = 1, 2$. To complete the description of the system, we define the distribution of SAH as $P(h_i^* = j|\boldsymbol{x}_i) \equiv \pi_{ji}$. The marginal distributions of the two reported SAH measures can then be expressed as $P(h_{mi} = k|\boldsymbol{x}_i) = \sum_j \pi_{ji}\delta_{k|j}^m$.

The parameters of the outcome equation (1) can be estimated using the outcome model, the joint conditional distribution of the data $(y_i, h_{1i}, h_{2i}|\boldsymbol{x}_i)$ and two assumptions:

> **CIA** (Conditional Independence Assumption): Conditional on SAH 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.
>
> **NMA** (No Mirror Assumption): $\delta_{j|j}^m > \delta_{k|j}^m, \quad \forall j, k,$

The first assumption, which relates to the relationship between the misclassified health measures and the outcome, is used to simplify the structure of the joint distribution of $(y_i, h_{1i}, h_{2i}|\boldsymbol{x}_i)$, see Appendix

setting that we consider. See Schennach (2016) and Hu (2017) for overviews of the recent econometric measurement error literature.

A.1 and A.2 for details. Intuitively, one can see the identifying value of this assumption by considering the opposite degenerate case where $h_{1i} = h_{2i}$ (perfect positive dependence); then, our setting effectively reduces to the naïve case where all the misclassification is still "hidden" in one measure. As we move away from the case of perfect dependence towards independence, more and more of the misclassification is revealed through the off-diagonals of the joint distribution of h_{1i} and h_{2i} which indicate conflicting answers. Thus, even under moderate violations of CIA the estimator is likely to be more informative than the naïve approach treating h_{1i} or h_{2i} as if it was h_i^*. We discuss some empirical simulation evidence of this robustness in the simulations in the next section. The second assumption is used to ensure a unique solution for the estimation. By assuming that the probability of truthfully reporting a health level j is larger than any probability of misreporting it, NMA rules out the "mirror solutions" in which probabilities of misreporting and correctly reporting are switched along with the impact of each health level on the outcome y_i is also switched.[6]

As we show in Appendix A.2, if the regressors $\boldsymbol{x}_i$ are discrete, all objects of the system (the parameters of the outcome model, as well as all misclassification probabilities and the distribution of SAH) are identified and directly estimable. When some of the regressors are continuous, the system remains nonparametrically identified and nonparametric estimation (Hu, 2008) is possible in principle. But in practice the flexibility offered by such an approach might come at the cost of high small sample bias and a computational intensity that might quickly prove prohibitive if there are numerous regressors over which the level of misclassifications may vary. Instead, we proceed by using a more standard parametric approach for the misclassification, but which has the advantage of reducing the impact of small sample bias and of easily being able to incorporate many regressors in the misclassification equations.[7] As a basic specification, we assume that the misclassification probabilities and SAH are known multinomial-logit-based functions of the regressors:

$$\delta_{k|j}^m = \frac{\exp(-\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{k|j}^m))}{1+\sum_{k:k\neq j}\exp(-\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{k|j}^m))}, \quad \pi_{j,i} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\eta}_j)}{1+\sum_{j=1}^4\exp(\boldsymbol{x}_i'\boldsymbol{\eta}_j)}. \tag{3}$$

The added exponential function in the argument of the multinomial logit form for $\delta_{k|j}^m$ directly implements the constraint implied by NMA that $\delta_{j|j}^m > \delta_{k|j}^m$.

Finally, the proposed approach can be extended to nonlinear outcome models other than logit, such as Poisson count models or Weibull duration models (see Appendix A.5). Another potentially useful extension of model (1) is to the case where the impact of unobserved SAH on the outcome may differ based on an individual's characteristics. Our flexible approach allows us to go further and also accommodate interaction terms between all or some of the regressors $\boldsymbol{x}_i$ and unobserved health:

$$y_i = \mathbb{1}\left(\sum_{j=1}^J d_{ji}^*\alpha_j + \sum_{j=1}^J d_{ij}^* x_{ki}\alpha_{j,x} + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0\right), \quad i = 1,\ldots,N, \tag{4}$$

[6] See Hu (2008) for a discussion of an alternative identifying mirror assumption: that the estimated direction of the impact of each health level on the outcome is known; that is, in this case, that $\hat{\alpha}$ is negative.

[7] With our approach it is also straightforward to increase the flexibility of the parametric form in the misclassification equations to test the sensitivity of the results to a particular functional form. More generally, because of the underlying nonparametric identification of the misclassification, our approach can also serve as the basis of a nonparametric estimation via a series estimation approach (such as by including polynomials or splines of the linear indices, cf. Newey, 1994).

for some variable of interest x_{ki} such as, say, education (cf. Appendix A.2).

3.2 Estimation: A penalised finite mixture (PFM) approach

The full likelihood of the model takes the form of a finite mixture or latent class model (see eq. (16) in Appendix A.3). We propose to estimate the model via the Expectation-Maximisation (EM) algorithm (Dempster *et al.*, 1977) which we found to be substantially faster and more stable than competitor approaches such as standard maximum likelihood or GMM (see Appendix A.4), making it our only viable estimator. The EM algorithm iterates between the maximisation or M-step, and the expectation or E-step. The nth iteration of the M-step is

$$\hat{\boldsymbol{\theta}}^n = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \tilde{\ell}_i(\boldsymbol{\theta}; y_i, h_{1i}, h_{2i}, \boldsymbol{x}_i, \hat{w}_i^n), \tag{5}$$

where $\boldsymbol{\theta}$ collects all the parameters of the system—$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\gamma}_{k|j}^m$ for $m = 1, 2$ and $j \neq k$— and

$$\tilde{\ell}_i(\cdot) = \sum_{j=0}^{4} \hat{w}_{ji}^n \left(\ln F(y_i|h_i^*=j, \boldsymbol{x}_i) + \ln F(h_{1i}|h_i^*=j, \boldsymbol{x}_i) + \ln F(h_{2i}|h_i^*=j, \boldsymbol{x}_i) + \ln \pi_{ji} - \ln \hat{w}_{ji}^n \right) - \frac{t}{N} \boldsymbol{\gamma}_j^{m\prime} \boldsymbol{\gamma}_j^m, \tag{6}$$

where all $F(\cdot|\cdot)$ denote likelihood functions of their arguments (see Appendix A.3 for details). Equation (6) is a penalised estimate of the likelihood obtained by conditioning on and summing over estimates of the posterior probabilities $\hat{w}_{ji}^n = \hat{P}^n(h^* = j|y_i, h_{1i}, h_{2i}, \boldsymbol{x}_i)$. The term $\frac{t}{N} \boldsymbol{\gamma}_j^{m\prime} \boldsymbol{\gamma}_j^m$ is a ridge penalty term for the parameters of the misclassification probabilities ($\boldsymbol{\gamma}_j^m$) with scalar tuning parameter t. We discuss the advantages and disadvantages of penalisation below; for $t = 0$, equation (6) is a standard M-step. In the $(n+1)$th iteration of the E-step, we update the posterior probabilities as follows:

$$\hat{w}_{ji}^{n+1} = \frac{\hat{\pi}_{ji}^n \, \hat{F}^n(y_i|h_i^*=j, \boldsymbol{x}_i) \, \hat{F}^n(h_{1i}|h_i^*=j, \boldsymbol{x}_i) \, \hat{F}^n(h_{2i}|h_i^*=j, \boldsymbol{x}_i)}{\sum_{j=0}^{4} \hat{\pi}_{ji}^n \, \hat{F}^n(y_i|h_i^*=j, \boldsymbol{x}_i) \, \hat{F}^n(h_{1i}|h_i^*=j, \boldsymbol{x}_i) \, \hat{F}^n(h_{2i}|h_i^*=j, \boldsymbol{x}_i)}, \tag{7}$$

where all $\hat{F}^n(\cdot|\cdot)$ correspond to terms evaluated at $\hat{\boldsymbol{\theta}}^n$.

The increased stability and speed of EM comes from the fact that, first, as opposed to the full likelihood $\ell_i(\cdot)$ (cf. A.3), in $\tilde{\ell}_i(\cdot)$ of the M-step, the logarithm goes through the sum of the finite mixture components of the joint distribution $F(y_i, h_{1i}, h_{2i}|\boldsymbol{x}_i)$; and, second, these components depend on separate sets of parameters (since the w_{ji} are fixed in the M-step), meaning that each can be maximised 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 $\boldsymbol{\gamma}_{k|j}^1$ and $\boldsymbol{\gamma}_{k|j}^2$ vectors (with $j \neq k$), respectively; and $\pi_i = \pi(\boldsymbol{x}_i)$ is a function only of $\boldsymbol{\eta}$.

The penalisation of misclassification parameters in our estimation addresses a potential finite sample issue of low statistical power given that misclassification probabilities (i) may depend on many parameters (if the dimension of $\boldsymbol{x}_i$ is large) and (ii) may be small. This implies that in practice the misclassification probabilities may be identified from potentially low frequency cells of the joint distribution of $(y_i, h_{1i}, h_{2i}|\boldsymbol{x}_i)$. In the most extreme case, the sample likelihood function may be maximised

for a value of a misclassification probability equal to zero, which may manifest itself as a convergence problem in the maximum likelihood procedure since parameters will tend to infinity. But even in less extreme cases, where estimates are finite, they might be biased. These issues, while originating in the estimates of the misclassification probabilities, may spill over to the parameter estimates of the outcome equation. To overcome such convergence issues and reduce the small sample/low statistical power bias we suggest implementing a penalised estimation by setting $t > 0$ in (6). This rules out infinite estimates and reduces extreme misclassification probabilities. The tuning parameter determines the weight given to the penalty. However, as with all penalised likelihood estimations, it can introduce bias: if the penalty is too harsh (that is, if t is too large), the overall bias of the estimator may increase. Since our primary objective with the penalisation is to ensure finiteness of all estimates, we choose a small $t > 0$ in our application and check for sensitivity of the results to changes of the chosen value. With increasing N, the weight of the penalisation with a fixed tuning parameter decreases and the penalised estimator converges towards the unpenalised one.

In addition to penalisation, a second possible avenue for reducing low power issues in the estimation of the misclassification parameters is using more than one outcome variable, say $\boldsymbol{y}_i = (y_{1i}, y_{2i})$. If more than one possible outcome is available which is dependent on SAH and conditionally independent of misclassification, then the 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 both as conditionally independent of misclassification but potentially correlated with each other. The connection between the two (or more) models is that the unobserved SAH is obviously the same for each observation across both outcome 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.[8] Adapting the EM algorithm is straightforward. In equations (5)-(7), the terms $F(y_i|h_i^* = j, \boldsymbol{x}_i)$ are simply replaced by $F(y_{1i}|h_i^* = j, \boldsymbol{x}_i)F(y_{2i}|h_i^* = j, \boldsymbol{x}_i)$.

4 Monte Carlo experiments

Next, to benchmark the performance of our proposed finite mixture (FM) and penalised FM (PFM) estimators, we compare their performance to the ideal estimator that uses the unobserved SAH status which is infeasible in practice, and, on the other end of the spectrum, to the naïve estimator that just uses the first observed reported SAH measure, treating it as if it was the SAH status. As further points of comparison, we also examine four potential competitor estimators, which address misclassification in ad-hoc ways sometimes encountered in the literature. We examine the estimators' finite sample performance in a Monte Carlo simulation study.

The baseline design we use is a simple data generating process (DGP) with a single regressor x_i and a binary SAH indicator h_i^*. Details of the parameter specification choices and the drawing procedure

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

Table 1: Simulation results: Baseline DGP, N=10,000

		h^*	h_1	$\bar{h}$	$\bar{\bar{h}}$	$\hat{h}_1$	$\hat{e}_1$	FM	PFM
Parameters of the outcome model									
$\hat{\alpha}$	Bias	0.002	-0.457	-0.259	-0.245	0.602	0.643	0.008	0.005
	RMSE	0.050	0.460	0.268	0.253	0.647	0.686	0.085	0.083
$\hat{\beta}$ const	Bias	-0.002	0.260	0.165	0.158	-0.245	-0.258	-0.010	-0.000
	RMSE	0.049	0.265	0.172	0.171	0.272	0.284	0.118	0.098
$\hat{\beta}$ slope	Bias	0.003	0.156	0.141	0.057	-0.144	-0.140	0.012	0.021
	RMSE	0.084	0.177	0.164	0.120	0.179	0.176	0.154	0.124
Parameters of the misclassification probabilities									
$\hat{\eta}$ const	Bias							0.036	0.003
	RMSE							0.393	0.289
$\hat{\eta}$ slope	Bias							-0.064	-0.120
	RMSE							0.517	0.363
$\hat{\gamma}^1_{1\|0}$ const	Bias							0.039	0.034
	RMSE							0.373	0.252
$\hat{\gamma}^1_{1\|0}$ slope	Bias							0.010	-0.199
	RMSE							0.612	0.424
$\hat{\gamma}^2_{1\|0}$ const	Bias							-0.012	0.048
	RMSE							0.323	0.227
$\hat{\gamma}^2_{1\|0}$ slope	Bias							0.022	-0.156
	RMSE							0.547	0.380
$\hat{\gamma}^1_{0\|1}$ const	Bias							-0.010	0.007
	RMSE							0.249	0.195
$\hat{\gamma}^1_{0\|1}$ slope	Bias							0.028	0.031
	RMSE							0.337	0.246
$\hat{\gamma}^2_{0\|1}$ const	Bias							-0.029	0.044
	RMSE							0.235	0.194
$\hat{\gamma}^2_{0\|1}$ slope	Bias							0.049	-0.016
	RMSE							0.281	0.221

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), the average of h_1 and h_2 ($\bar{h}$), h_1 in the sample restricted to i with $h_{1i} = h_{2i}$, predicted h_1 ($\hat{h}_1$), the residual from a prediction of h_1 ($\hat{e}_1$), and the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. Observations are drawn from $y_i = \mathbb{1}(\alpha\, h_i^* + \beta_{\text{const}} + \beta_{\text{slope}} x + \varepsilon_i > 0)$, with $\alpha = 1$, β const=0, β slope=1; with the distribution of h_i^* given by $\pi_i = \Lambda(\eta_{\text{const}} + \eta_{\text{slope}} x_i)$, with η_1=1.5 and η_0=-0.1342; and the misreporting probabilities by $\delta^m_{k|j} = \Lambda(-\exp(\gamma^m_{k|j}\text{const} + \gamma^m_{k|j}\text{slope}\; x_i)), m = 1, 2, j \neq k = 0, 1$, where $\gamma^m_{k|j}\text{slope} = 1$ for all m, k, and $\gamma^1_{0|1}$const=-0.25, $\gamma^2_{0|1}$const=-0.75 , $\gamma^1_{1|0}$const=0, and $\gamma^2_{1|0}$const=-0.5. See Appendix B.1 for more details on the DGP.

are given in the table notes and, more fully, in Appendix B.1. Similar to the survey data used in our application, the reported SAH measures in our chosen simulation DGP have distributions which are similar to each other while at the same time there is a substantial share of conflicting answers. We use a sample size of $N = 10,000$ and replicate the estimations 500 times. The results in Table 1 show that the infeasible estimator that uses the unobserved SAH status (in columns "h^*") is, as expected, virtually unbiased. The naïve estimator which uses the misreported SAH measure h_1 (depicted in columns "h_1"), is severely biased: the average estimate of α is about 45 percent below its true value of 1, illustrating the pernicious effects of misclassification.

The next four columns depict estimates of ad-hoc approaches to dealing with misclassification. In column "$\bar{h}$", the average of the two SAH measures is used as the regressor in the models. In column "$\bar{\bar{h}}$", all observations of individuals whose second response to the SAH question is different from the first were dropped from the estimation sample, leaving a sample of individuals with what sometimes is called "consistent responses" (although this includes individuals who misreport SAH twice). The next two columns contain estimates from approaches that mimic two-stage least squares in linear models. They consist of 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 these estimators then includes the first-stage predictions as the regressor in the outcome model (column "$\hat{h}_1$"). 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 with the mismeasured SAH response in the outcome model (column "$\hat{e}_1$"). This is a version of the control function approach and is inconsistent, in general, when the endogenous variable (here, SAH) is discrete.[9] The results in the table show that for all ad-hoc approaches all estimated parameters, including the slope of x_i, are very distorted. Thus, such approaches, while well-suited to classical measurement error in linear models, cannot be recommended as solutions to the measurement error problem at hand.

The remaining columns present estimates from the proposed finite mixture ("FM") estimator and the penalised finite mixture ("PFM") estimator. Both are able to greatly reduce the bias in the key parameter α from h_1 from 46 to less than 1 percent. The other parameters of the outcome model, β_0 and β_1, are estimated similarly well. The lower part of the table contains the parameters of the misclassification probabilities, which are only estimable with FM and PFM. These parameters are more difficult to estimate, as evidenced by their larger root-mean squared error (RMSE). The PFM estimator has uniformly better RMSE than FM, sometimes by as much as 50 percent. However, these gains in RMSE come at the cost of introducing some bias.[10]

In Table 2, we present results for the parameters of the outcome model for two further DGPs (estimates of the misclassification probabilities are omitted for brevity). In Panel (A), we extend the specification of the outcome model by including an interaction term between SAH and the regressor ($h_i^* \times x_i$) so that the impact of SAH on the outcome varies with x_i. That this is a more challenging DGP to estimate can be seen by observing the RMSE for the infeasible estimator, which almost doubles for the constant in α (and more than triples for the slope in α, i.e. the interaction coefficient) relative to the baseline case from Table 1. Both FM and PFM perform well, with PFM tending to have a marginally lower bias and RMSE than FM. The next panel (B) extends the SAH status from being a binary variable to being a multinomial variable with five categories, like the one in the HILDA data. However, the

[9] There are very 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). 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). 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).

[10] However, even the largest biases in the misclassification parameters, such as -0.199, only translate into biases in average marginal effects that are unlikely to be meaningful in practice; for instance, for the mentioned case, the estimate of a one standard deviation change in x_i on the misclassification probability is 0.048 where the true effect is 0.056.

Table 2: Simulation results: Further DGP

(A) *Interaction effect*, $N = 10,000$		h^*	h_1	FM	PFM
$\hat{\alpha}$ const	Bias	0.008	-0.596	0.020	0.002
	RMSE	0.090	0.602	0.186	0.177
$\hat{\alpha}$ slope	Bias	-0.011	-0.409	-0.018	0.012
	RMSE	0.178	0.442	0.295	0.286
$\hat{\beta}$ const	Bias	-0.003	0.324	-0.015	-0.009
	RMSE	0.063	0.330	0.149	0.131
$\hat{\beta}$ slope	Bias	0.006	0.514	0.018	0.024
	RMSE	0.129	0.528	0.222	0.203
(B) *Categorical SAH*, $N = 10,000$		h^*	h_1	FM	PFM
$\hat{\alpha}_1$	Bias	0.017	-0.293	0.039	0.013
	RMSE	0.094	0.302	0.175	0.169
$\hat{\alpha}_2$	Bias	0.005	-0.534	0.028	0.012
	RMSE	0.093	0.539	0.157	0.149
$\hat{\alpha}_3$	Bias	0.003	-0.754	0.019	0.001
	RMSE	0.082	0.758	0.155	0.145
$\hat{\alpha}_4$	Bias	0.003	-0.937	0.027	0.010
	RMSE	0.087	0.940	0.130	0.131
$\hat{\beta}$ const	Bias	-0.011	0.660	-0.030	-0.009
	RMSE	0.077	0.662	0.128	0.124
$\hat{\beta}$ slope	Bias	0.005	0.165	0.005	0.019
	RMSE	0.071	0.180	0.073	0.077

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), and the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. In panel (A), the DGP of the outcome model is $y_i = \mathbb{1}(\alpha_{\text{const}} h_i^* + \alpha_{\text{slope}} h_i^* x_i + \beta_{\text{const}} + \beta_{\text{slope}} x + \varepsilon_i > 0)$, with $\alpha_{\text{const}} = 1$, $\alpha_{\text{slope}} = 1$, $\beta_{\text{const}} = -0.375$ and $\beta_{\text{slope}} = 1$. Remaining parameters are as in the baseline DGP from Table 1. See Appendix B.1 for more details. In panel (B), the DGP of the outcome model is $y_i = \mathbb{1}(\alpha_1 h_{1i}^* + \alpha_2 h_{2i}^* + \alpha_3 h_{3i}^* + \alpha_4 h_{4i}^* + \beta_0 + \beta_1 x + \varepsilon_i > 0)$, with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)' = (0.5, 1.0, 1.5, 2.0)'$, $\beta_{\text{const}} = 0$ and $\beta_{\text{slope}} = 1$. See Appendix B.4 for more details on this DGP.

performance for FM and PFM remains similar to the one in the baseline DGP.

Appendix B contains extensive additional results, which shed light on a number of further issues. For instance, simulations with N=1,000 show that the advantage of PFM over FM is more pronounced in such cases of less statistical information (B.2, B.3, B.4). The estimators are also applied to other outcome models, such as a Poisson count data model and a Weibull duration data model (B.5). They are applied to DGPs with multiple outcome models, showing that combining them and estimating them jointly can further enhance the quality of the estimates (B.6). The performance of FM and PFM is also examined and compared to the results of Hu's (2008) nonparametric estimator results in DGPs where the misclassification probabilities follow a different functional form than assumed (B.7). The fact that FM and PFM outperform a nonparametric estimator in this setting illustrates how the potential disadvantage of a misspecified parametric estimator can be compensated by low bias and RMSE in

finite samples. Finally, simulations where the conditional independence assumption (CIA) is violated by an additional regressor in the DGP which is omitted from the estimation show our approach's sensitivity to one of the key identifying assumptions (B.8). The results show that the approach is robust to mild and moderate departures from CIA. Where violations induce biases, they act similarly to classical omitted variable bias, and its effect is not worse than would be on the infeasible estimator. We use the lessons from the simulations in this section to inform our application of these methods to real world data. In the next section, we estimate a joint logit-logit model for mortality and morbidity using two five-category reported SAH measures and a sample of close to 13,000 individuals.

5 SAH misclassification in the HILDA data

In this section we first outline the HILDA data and describe the repeated reported measures of SAH in some HILDA waves. We then use these measures to replicate the descriptive reduced-form approach from the literature which involve regressing indicators of differences in the SAH measures on a vector of socioeconomic variables, and discuss what we can and cannot learn about misclassification from such estimates.

HILDA is an annual Australian household-based longitudinal survey that began in 2001 (Summerfield *et al.*, 2014). The survey covers social and economic topics such as household structure, income, work and health. Individuals aged 15 or over are asked to respond and reasons for non-response are recorded where known. Wave 1 (2001) covered 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. Overall, there are roughly 13,000 respondents in each wave of the HILDA Survey from 2001 to 2016. While non-response due to death is recorded annually where known, the survey sample was also linked in 2014 to the National Death Index so that details of individuals' year and age of death are available for all those originally in the survey, including the subsequent 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.[11] The SAH question is the first question in the health section, and is followed by a number of other health-related questions such as long-term conditions and disabilities. We designate this SAH variable as h_1. Respondents are asked to choose their health on a 5-option scale "Poor", "Fair", "Good", "Very Good", and "Excellent" which we label as $0, \ldots, 4$. Respondents are then issued with the self-completion Questionnaire, which is to be filled in by themselves and collected by the interviewer that day or posted back after completion. In this questionnaire, the same SAH question is asked again at the beginning. We designate this SAH variable as h_2, and label it in the same way as h_1. The dates of completing both questionnaires are only available for waves 9 and 13. The median time between completion of the two questionnaires is 1 day in both waves and on average, the questionnaires were completed only 4.8

[11] Some of these interviews were conducted over the phone, but, for convenience, we refer to the SAH question from the Person Questionnaire as face-to-face in the remainder of the text.

Table 3: Joint distribution of reported SAH 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

Notes: Source: HILDA waves 1, 9 and 13. Cell entries show relative frequencies in per cent for joint and marginal distribution of the reported SAH measures h_1 and h_2. Labels: 0="poor"; 1="fair"; 2="good"; 3="very good"; 4="excellent".

days and 4.6 days apart in 2009 and 2013, respectively. Since the surveys were taken close together, the likelihood of an actual meaningful change in health is fairly low. As a result, we believe that the majority of differential responses to the SAH question are random and unlikely to be related to changes in an individual's underlying health status.

The top panel of Table 3 reports the joint-distribution (in percent of respondents) from the two SAH questions in wave 1. About 27.8 percent of respondents changed their health status between h_{1i} and h_{2i}, similar to that reported by Clarke & Ryan (2006) and similar to Crossley & Kennedy (2002) where SAH was asked twice in a different survey. 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 (middle and bottom panels of Table 3), and so is 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,

Table 4: 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 out of labour force)	12,908	0.344	0.475
unemployed (=1, if unemployed)	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

*N*otes: Source: HILDA waves 1 and 16.

this was not driven by the same individuals switching in each wave. 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 a large amount of randomness.

Given the two questionnaires were completed around the same time, 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 3 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 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.

From here on, we focus on h_{1i} and h_{2i} for the first wave in 2001 because we want to study the implications of SAH for long-term (15-year) mortality and morbidity. There are 12,908 individuals with responses on h_{1i} and h_{2i}. Descriptive statistics for selected demographic and socio-economic characteristics of these individuals are given in Table 4. We begin our empirical investigation by applying a reduced form strategy used in previous literature to characterise individual misclassification behaviour (Black *et al.*, 2017a). In Table 5, we present estimates of logit models where the dependent

Table 5: Estimation results: Logit models for changes in SAH response

Dep. var.	$\mathbb{1}(h_1 \neq h_2)$ (1)	$\mathbb{1}(h_1 > h_2)$ (2)	$\mathbb{1}(h_1 < h_2)$ (3)
age/100	-0.17	0.88	-0.79
	(0.65)	(0.94)	(0.76)
age^2/100	1.01	-0.83	1.91**
	(0.67)	(0.98)	(0.78)
male	0.05	0.23**	-0.08
	(0.04)	(0.06)	(0.05)
education/10	-0.59**	-0.44**	-0.55**
	(0.11)	(0.16)	(0.13)
log HH income	-0.13**	-0.16**	-0.07*
	(0.03)	(0.05)	(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

*N*otes: Source: HILDA wave 1, own calculations. Cells represent estimated coefficients, and robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$

variable is an indicator that an individual gave two conflicting reports of SAH, $\mathbb{1}(h_{1i} \neq h_{2i})$ (Column 1), an indicator that they gave a higher SAH in the face-to-face questionnaire, $\mathbb{1}(h_{1i} > h_{2i})$ (Column 2), and that they gave a higher SAH in the self-completion questionnaire, $\mathbb{1}(h_{1i} < h_{2i})$ (Column 3). Explanatory factors include age, sex, education, income, 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. However, insignificant estimates are harder to interpret. There could be different types of misclassification which 'average out', resulting in small insignificant effects on a change in reported SAH ($\mathbb{1}(h_{1i} \neq h_{2i})$). For instance, the regressor *male*, has an effect on the dependent variable '$\mathbb{1}(h_{1i} > h_{2i})$ despite not having a significant effect on $\mathbb{1}(h_{1i} \neq h_{2i})$. In addition,

more complex misclassification patterns can be completely undetectable with these dependent variables and they also tell us nothing about which of the two measures are more likely to be misclassified. 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.

6 Estimating SAH misclassification and the effects of SAH on mortality and morbidity

In this section we present our estimates of our joint model of SAH misclassification and of the association between SAH and two outcomes measured 15 years after the initial survey: mortality (whether the individual is deceased) and, if the individual is still alive, whether they developed any chronic conditions in the 15-year period. We examine the estimated misclassification patterns in Section 6.1 and discuss the estimates from the outcome equations in Section 6.2, where we also compare our results to naïve estimates obtained ignoring misclassification. In our empirical model, the misclassification probabilities, mortality and chronic conditions further depend on the same covariates $\boldsymbol{x}_i$ as the reduced-form models in the previous section. There were 12,908 individuals in the 2001 survey of which 10.9% were deceased by 2016 and we can obtain information on the possible development of new chronic conditions in 2016 for 7,340 individuals. Means and standard deviations for these outcome variables are reported in Table 4, along with those of the covariates. We estimate the two outcomes jointly using the penalised finite mixture estimator (with the tuning parameter set to a value of t=0.5).[12]

6.1 SAH and misclassification

Figure 2(a) shows the average predicted probabilities of reporting behaviour by gender. Each of the five panels illustrates how males (M) and females (F) in each health state ($h^* = j$) are likely to respond when answering h_1 and h_2. We find there is substantial misclassification in both reported SAH measures but more so for the self-completion questionnaire (h_2) with this being concentrated in those with excellent and poor health. This leads to a pattern consistent with a tendency towards ticking middle boxes in the self-completion questionnaire. As expected we see that most misclassification is by only one category with other larger misclassifications rare.

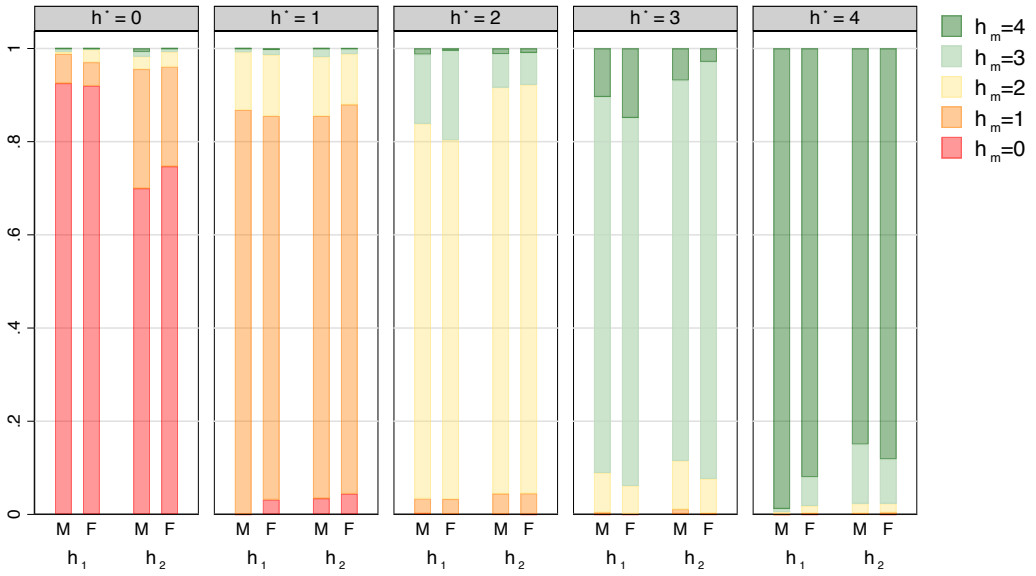
As already noted there was a discrepancy between the two SAH measures in those reporting being in "excellent" health, with "excellent" health being reported more often in face-to-face interviews (h_1) than when filling out the questionnaire privately in the self-completion questionnaire (h_2). The results seen in (a) suggest that this is because males in excellent health ($h^* = 4$) are more likely to under-report their health in the self-completion questionnaire (very few males in excellent health under-report their health in the face-to-face case[13]) and males and females in very good health ($h^* = 3$) are more

[12] Tables C1–C2 in the Appendix report robustness tests from using a range of different values for t that show that our results are not sensitive to changes of t around the chosen value of 0.5.

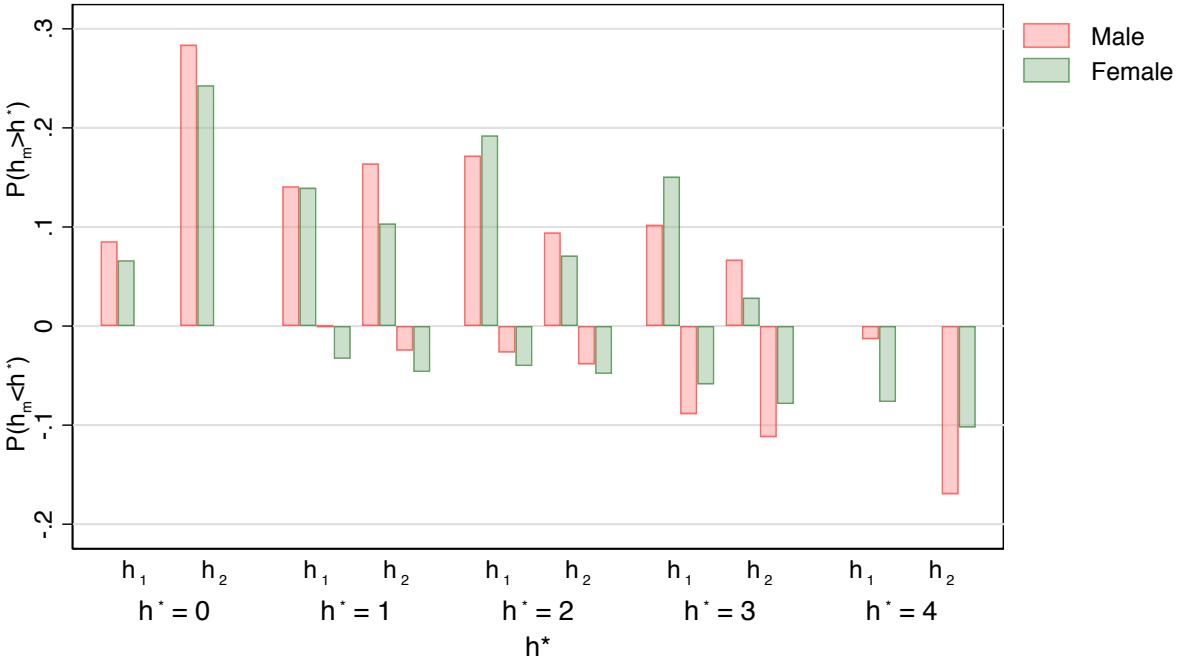
[13] Indeed, males in excellent health reporting being in excellent health in the face-to-face questionnaire has the highest

Figure 2: Misclassification in SAH for males and females

(a) Reporting for males (M) and females (F)



(b) Average predicted upward and downward misreporting conditional on covariates



Notes: Estimates from HILDA data waves 1 and 16 for individuals who responded to SRH questions in wave 1 (N=12,908). In Panel (a), weighted average predicted probabilities are presented for the separate samples of males and females. In Panel (b), weighted average predicted probabilities are presented when probabilities are evaluated at male=0 and then male=1. While in (a) differences in reporting may also be due to other charateristics which differ across males and females, (b) attempts to isolate the role of gender itself on reporting behaviour such that the differences give the average marginal effects of gender on misreporting.

likely to over-report their health in the face-to-face questionnaire. For males and females in good health ($h^* = 2$) we also observe similar over-reporting patterns with nearly 20 percent over-reporting

share of truthful reporting of all categories shown in the figure: it would seem men like to say that they are in excellent health.

their health in the face-to-face case compared to just under 10 percent over-reporting in the privately answered questionnaire. Conversely, for males and females in poor health ($h^* = 0$) the share truthfully reporting their health status is much higher in the face-to-face interviews with over 20 percent over-reporting their health in the self-completion questionnaire. Perhaps poor health individuals are more honest in the face-to-face case because their poor health is evident or they can verbally justify claiming they have poor health. Recency effects (in HILDA the interviewer reads out the category "poor" last) might additionally reinforce truthful reporting of this category in interviews. For those in fair and good health we see very little evidence of under-reporting.

Can the gender-specific reporting patterns seen in Figure 2a be explained by potential differences in the observed covariates such as age, education, income, etc.? In Figure 2(b), we show a graph that relies on gender differences that have been adjusted for differences in the covariates. The figure shows the average predicted posterior misclassification probabilities grouped into upward and downward misclassification when those in each SAH status is assumed to be male and female respectively (i.e., the difference between the two gives the average marginal effect of gender on misclassification). We see that the patterns are largely similar to those observed when we look at the misclassification for each subgroup in the population. That is, the answer to the question posed at the start of this paragraph is that the role of gender seen in the subgroup analysis (a) is not being masked by the role of other individual characteristics on misclassification.

While Figure 2 considers misclassification probabilities for males and females there is also considerable heterogeneity in the individual probabilities of reporting health status truthfully by other individual characteristics. In the appendix we present the equivalent figures but where we consider the role of income, age and education on misclassification (Figures C1, C2 and C3). Of note, we see that, after controlling for other characteristics, low income individuals are more likely to over-report their health compared to high income individuals, older individuals are more likely to misclassify in general compared to younger individuals, and there are limited differences in misclassification related to education apart from the extreme categories in the self-completion questionnaire, where those with low education are more likely to over-report their health as excellent, while high education individuals are more likely to under-report their health as poor.

6.2 Impact of SAH misclassification on the outcome model

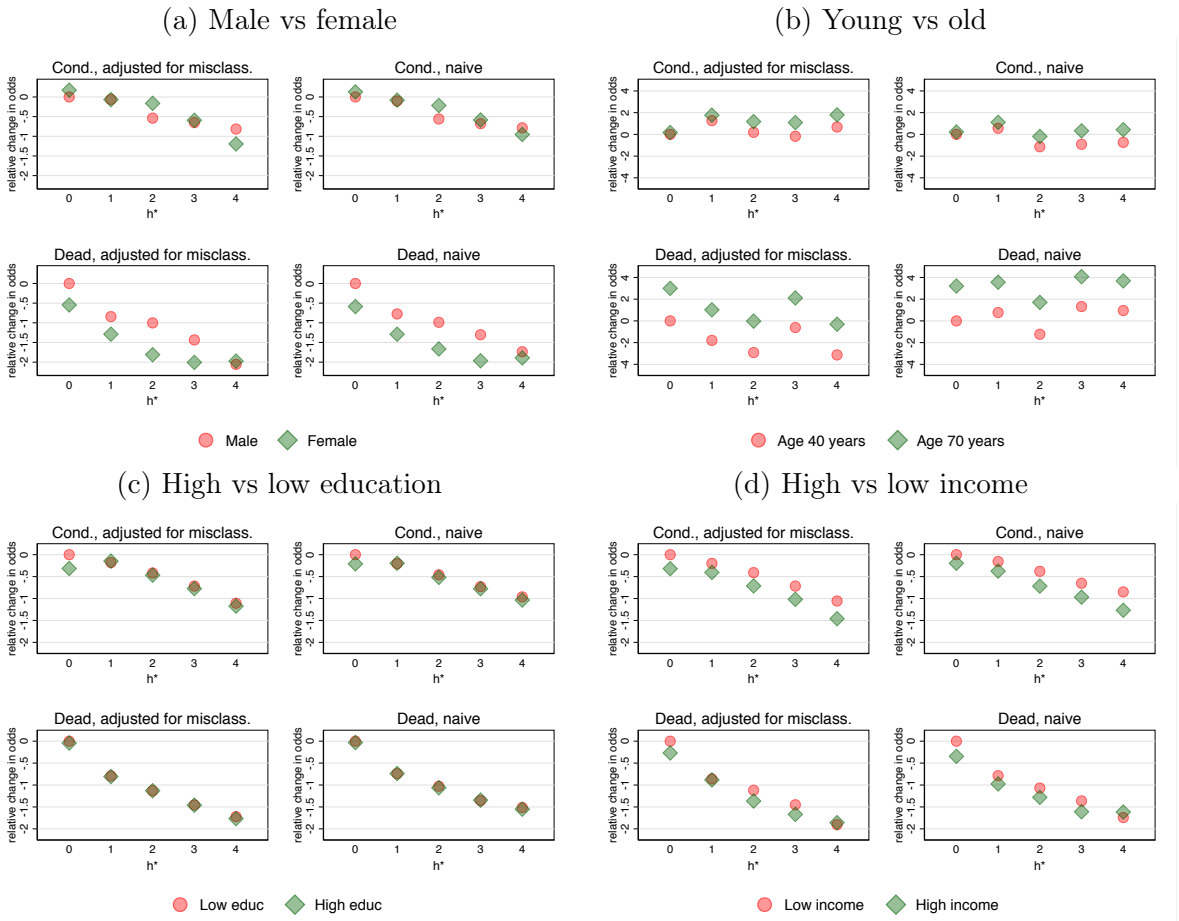
Table 6 contains the estimated parameters of our penalised finite mixture models for mortality and morbidity (Columns 1 and 4) along with the difference compared to what the estimates would have been using naïve estimators h_1 and h_2 (Columns 2–3, and 5–6). For the key parameters of interest, the health coefficients α_j, the differences to the naïve approaches are often significant and range from small to large. While they are of a similar absolute magnitude across outcomes (that is, the differences in Column 2 are about the same as those in Column 5; and those in Column 3, as those in Column 6), they are quite different in relative terms: For mortality, the PFM estimates differ by about 10 percent to 20 percent; but some PFM estimates of α_j are more than twice as big in the chronic condition equation. In most cases, we find that by using our PFM approach there are larger difference in health

Table 6: Estimation results: System Penalised Finite Mixture (PFM) models for mortality (dead: yes/no) and morbidity (chronic condition: yes/no)

Dep. var.	Dead			Chronic cond.		
	PFM	Diff. to naïve		PFM	Diff. to naïve	
		h_1	h_2		h_1	h_2
	(1)	(2)	(3)	(4)	(5)	(6)
α_1	-0.80**	-0.06	-0.17	-0.15	0.01	-0.23*
	(0.14)	(0.04)	(0.11)	(0.17)	(0.08)	(0.12)
α_2	-1.13**	-0.10**	-0.23**	-0.41**	0.03	-0.21**
	(0.15)	(0.05)	(0.09)	(0.17)	(0.08)	(0.10)
α_3	-1.46**	-0.12*	-0.21**	-0.71**	-0.02	-0.25**
	(0.16)	(0.06)	(0.09)	(0.18)	(0.08)	(0.10)
α_4	-1.77**	-0.24**	-0.26**	-1.11**	-0.16*	-0.37**
	(0.20)	(0.09)	(0.12)	(0.20)	(0.09)	(0.11)
age/100	-3.90**	-0.11	-0.50**	6.28**	-0.08	-0.28**
	(1.56)	(0.08)	(0.12)	(1.39)	(0.09)	(0.09)
age^2/100	13.20**	0.04	0.56**	-3.08**	0.00	0.29**
	(1.44)	(0.08)	(0.14)	(1.49)	(0.10)	(0.11)
male	0.58**	-0.00	-0.02**	-0.12*	0.01	-0.01
	(0.08)	(0.00)	(0.01)	(0.07)	(0.00)	(0.00)
education/10	-0.12	0.03**	-0.01	-0.52**	0.01	-0.00
	(0.22)	(0.01)	(0.02)	(0.18)	(0.01)	(0.01)
log HH. income	-0.10*	0.01**	0.02**	-0.17**	0.01**	0.02**
	(0.06)	(0.00)	(0.01)	(0.06)	(0.00)	(0.00)
chronic condition	0.27**	-0.03*	-0.07**	0.39**	-0.00	-0.05**
	(0.09)	(0.02)	(0.02)	(0.09)	(0.02)	(0.02)
married	-0.38**	-0.01**	0.01	-0.14*	0.00	0.01*
	(0.08)	(0.00)	(0.01)	(0.08)	(0.00)	(0.00)
overseas	-0.25**	0.01**	0.02**	-0.05	0.01**	0.01**
	(0.09)	(0.00)	(0.01)	(0.08)	(0.00)	(0.00)
not in labour force	0.07	-0.02**	-0.05**	0.13	-0.00	-0.02**
	(0.11)	(0.01)	(0.01)	(0.09)	(0.01)	(0.01)
unemployed	0.07	0.01	0.03**	0.31*	0.01	0.01*
	(0.25)	(0.01)	(0.01)	(0.17)	(0.01)	(0.01)
smoker	0.60**	0.00	0.02**	0.29**	0.00	0.01
	(0.08)	(0.01)	(0.01)	(0.07)	(0.00)	(0.00)
N	12,908	12,908	12,908	7,340	7,340	7,340

*N*otes: Source: HILDA waves 1 and 16, own calculations. Bootstrap standard errors in parentheses. Columns (1) and (4) present the coefficients of the joint PFM model estimated for the binary dependent variables 'Dead' and 'Chronic cond.'. Columns (2), (3) and (5), (6) show the PFM coefficient estimate minus the coefficient estimate from models using h_1 or h_2 (and the corresponding bootstrapped standard error for this difference). The independent variable 'chronic condition' refers to whether the individual had a chronic condition in 2001, while the dependent variable 'Chronic cond.' refers to whether they had developed an additional chronic condition by 2016. * $p < 0.10$, ** $p < 0.05$

Figure 3: Heterogeneity in the effect of health on morbidity (cond) and mortality (dead): Relative change in odds



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

outcomes between SAH categories than when misclassification is ignored. The biases in estimates of α_j using the face-to-face responses (h_1) are smaller than when using the self-completion responses (h_2). This is consistent with the findings of the misclassification probabilities presented in Figure 2, Panel (a), which showed that, on average, h_1 was less affected by misclassification than h_2.

There are also significant biases in the estimated coefficients on other explanatory variables, but in most cases these biases are small (in relative terms for those factors that are highly associated with health outcomes, or in absolute terms of those that are not strongly associated with the outcomes) and in most cases less than 10 percent in relative terms. For example, in our PFM model for both mortality and future chronic conditions we now find that income has a significantly smaller effect after we take into account that low income individuals are more likely to over-report their SAH; i.e., some of their poorer future health outcomes are actually due to their current poorer SAH that they do not always disclose. Some of the largest effects in absolute terms of ignoring misclassification in the outcome equation are for chronic conditions in 2001 and whether they are in the labour force.

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 C3 (and additional descriptive statistics for the discretised variables in Table C4).

We find broadly similar results to the ones in our baseline specification with continuous regressors, although differences tend to be somewhat larger.

Finally, we examined potential heterogeneity of the effect of SAH on mortality and morbidity by estimating specifications with interaction effects in SAH. We ran four separate specifications where we interacted health with education, household income, sex, and age. To facilitate interpretation, the results are presented graphically in Figure 3, which gives differences in the relative odds for each category and outcome, each evaluated at two points: male vs female, young (40 years old) vs old (70 years), low education (12 years of schooling) vs high education (16 years), and low income (25th percentile) vs high income (75th percentile).[14] The results are presented both for the PFM estimator which adjusts for misclassification as well as for the naïve estimator which does not.

The extent of effect heterogeneity in Figure 3 is reflected by how parallel the dots of the two groups (red circles vs green diamonds) remain along the x-axis of health categories. While the trajectories in the graphs are not perfectly parallel, the results point to a substantial homogeneity in the effect of SAH for these outcomes, especially considering that the groups represent rather large differences in the interacting variable (e.g. four years of education, 30 years of age, or a 50 percentile change in the income distribution). Of the presented graphs, the one showing the most amount of heterogeneity is the one giving the effect of SAH on mortality by gender: mortality differences tend to be larger for the middle SAH categories, and almost nonexistent for "excellent" health. Given the absence of striking patterns of heterogeneity in the PFM estimates, it may be unsurprising that the differences to the heterogeneity in the naïve estimates are also minor. While there are some statistically significant differences (namely in sex for mortality and education for both outcomes; see Table C5), these differences do not substantially change the patterns of the odds ratios of mortality and development of future chronic conditions for those in each SAH category.

7 Conclusions

While previous literature has documented that a large share of individuals report different SAH when asked twice, several important questions raised by this issue have so far remained unanswered. Because many forms of misclassification are compatible with observed differences in reported SAH, questions such as whether reported SAH is inherently unreliable or whether observed differences stem from one particular deficient measure could not be addressed. Similarly, it was not possible to know what patterns of misclassification occur nor how these vary with individual characteristics. Given that SAH is arguably one of the most widely used variables to measure health status or health capital in health economics, a key question is also how this misclassification in reported SAH translates to biases in estimates of models where reported SAH is used as a regressor. The question is not only if the effects of SAH are biased by misclassification, but also whether these biases spill over to the effects of other regressors.

[14] The corresponding regression results in table form are in the Appendix, Table C5.

Making use of recent advances in the econometrics of misclassification, we provide answers to these questions with data from a prominent household survey where SAH was reported twice in the same wave. Thanks to the setup of two measurements and at least one outcome, it is possible to identify the entire system of misclassification probabilities and the effects of SAH on the outcome without the need for additional arbitrary instruments or exclusion restrictions (Hu, 2008). Another advantage of the approach presented in this paper is that it specifies the effects of the categorical SAH variable directly by including dummy variables for each category of SAH in the outcome model, which is the standard way in which the applied literature includes categorical SAH variables as regressors in models. This avoids the additional modelling step of linking categorical SAH to some latent underlying continuous health, which would require additional assumptions.

Our results showed that there is substantial misclassification in both reported SAH measures, face-to-face interviews and self-completion questionnaires, and that misclassification patterns further vary by individual characteristics. When considering the role of current SAH in predicting future mortality and chronic conditions, comparisons between our proposed PFM approach and the naïve approaches using the misclassified responses showed that the naïve approaches were affected by statistically significant biases that ranged in magnitude from small to large. One result worth noting is that there were smaller biases in the outcome equations when the face-to-face questionnaire responses were used compared to the self-completion responses. We found significant but small biases for the role of explanatory variables other than SAH when misclassification is ignored. The small magnitude of these differences suggest that the use of SAH as a control variable might not be badly compromised despite the large shares of inconsistent answers in SAH. This is a result which might be useful to researchers who do not have multiple measure of SAH available but rely on including SAH in their empirical analysis as a useful way of addressing omitted variable bias from health status. However, when trying to estimate the role of SAH on outcomes it is likely to be more important to account for misclassification, especially if SAH is obtained through a self-completion questionnaire.

In this paper, we focussed on explaining the large share of different responses by the same individuals when asked to report SAH twice. This type of misclassification, which we found to be virtually uncorrelated over time, fits our assumption of conditional independence and our modelling of misclassification as conditionally random. However, if there is additional misclassification that is not conditionally independent then this remaining misclassification remains hidden. An example are further variables that are not accounted for in our empirical specification but that potentially influence misclassification. Our simulations show that while the proposed method might not fully account for all misclassification, such as systematic misclassification tied to omitted variables, adjusting for conditionally random misclassification will generally lead to visibly improved estimates.

The proposed method for nonlinear regression models where the key regressor is a categorical health variable and two misclassified measures of the regressor are available, can naturally be applied to contexts other than SAH. Our simulation results on its finite sample performance 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. Finite sample bias is also expected to be smaller with dependent variables with more possible outcomes, such as counts or durations, compared to binary dependent variables. Using a penalised estimator can visibly improve the performance of the estimator, especially when there is limited statistical power. Using several dependent variables jointly can also help reduce finite sample bias. Thus, our parametric approach with penalisation may have advantages over nonparametric approaches in finite samples where there is limited statistical power (e.g small sample size or many explanatory factors over which the misclassification probabilities may vary). 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 is constant across both studies (see Appendix A.6). This might be especially useful for exploring the sensitivity to misclassification in studies which only have one mismeasured SAH variable or have small sample sizes available.

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] d'Uva, Teresa Bago, Maarten Lindeboom, Owen O'Donnell, & Eddy Van Doorslaer, 'Education-related inequity in healthcare with heterogeneous reporting of health.' *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174** (3), pp. 639–664, *2011*.

[3] 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*.

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

[5] 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*.

[6] Black, Nicole, David W Johnston, Michael A Shields, & Agne Suziedelyte, 'Who provides inconsistent reports of their health status? The importance of age, cognitive ability and socioeconomic status.' *Social Science & Medicine*, **191**, pp. 9–18, *2017a*.

[7] ——— , ——— , & Agne Suziedelyte, 'Justification bias in self-reported disability: New evidence from panel data.' *Journal of Health Economics*, **54**, pp. 124–134, *2017b*.

[8] Bound, John, 'Self-Reported Versus Objective Measures of Health in Retirement Models.' *Journal of Human Resources*, **26** (1), pp. 106–138, *1991*, URL: http://www.jstor.org/stable/145718.

[9] Brown, Sarah, Mark N Harris, Preety Srivastava, & Xiaohui Zhang, 'Modelling illegal drug participation.' *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **181** (1), pp. 133–154, *2018*.

[10] 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*.

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

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

[13] Currie, Janet & Brigitte C Madrian, 'Health, health insurance and the labor market.' *Handbook of labor economics*, **3**, pp. 3309–3416, *1999*.

[14] Dempster, Arthur P, Nan M Laird, & Donald B Rubin, 'Maximum likelihood from incomplete data via the EM algorithm.' *Journal of the royal statistical society. Series B*, pp. 1–38, *1977*.

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

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

[17] Greene, William, Mark N Harris, Preety Srivastava, & Xueyan Zhao, 'Misreporting and econometric modelling of zeros in survey data on social bads: An application to cannabis consumption.' *Health economics*, **27** (2), pp. 372–389, *2018*.

[18] 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*.

[19] ——— , 'The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics.' *Journal of Econometrics*, **200** (2), pp. 154–168, *2017*.

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

[21] 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*.

[22] 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*.

[23] Newey, Whitney K, 'Series estimation of regression functionals.' *Econometric Theory*, **10** (1), pp. 1–28, *1994*.

[24] Schennach, Susanne M, 'Recent advances in the measurement error literature.' *Annual Review of Economics*, **8**, pp. 341–377, *2016*.

[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 Econometrics Methods: Details and additional results

A.1 Identification in a logit model with binary SAH and without covariates

Consider a simple logit model for mortality y_i (=1 if individual i is dead) with binary SAH; $h_i^* = 1$ indicates that individual i is in good health; and $h_i^* = 0$ that i is in bad health:

$$y_i = \mathbb{1}(\alpha\, h_i^* + \beta_0 + \varepsilon_i > 0), \quad i = 1, \ldots, N, \tag{8}$$

where $\mathbb{1}(\cdot)$ represents the indicator function, α and β_0 unknown scalars and ε_i an IID logistically-distributed idiosyncratic error. Thus, the probability of mortality as a function of SAH is

$$P(y_i = 1|h_i^*) = \frac{\exp(\alpha h_i^* + \beta_0)}{1 + \exp(\alpha h_i^* + \beta_0)} \equiv \Lambda(\alpha h_i^* + \beta_0), \tag{9}$$

a special case of the general model (1). As before, h_i^* is unobserved, but two potentially misclassified SAH measures h_{1i} and h_{2i} are available to the econometrician. In this minimal example, the corresponding four misclassification probabilities are

$$\delta^m_{0|1} = P(h_{mi} = 0|h_i^* = 1) \quad \text{and} \quad \delta^m_{1|0} = P(h_{mi} = 1|h_i^* = 0), \quad \text{for } m = 1, 2, \tag{10}$$

an the distribution of SAH is determined by the single parameter

$$P(h_i^* = 1) \equiv \pi. \tag{11}$$

The observed marginal distributions of the reported SAH measures can then be expressed as functions of the parameters defined in equations (10) and (11):

$$P(h_{mi} = 1) = \pi_i(1 - \delta^m_{0|1}) + (1 - \pi_i)\delta^m_{1|0}. \tag{12}$$

All parameters, $\theta = (\alpha, \beta_0, \pi, \delta^1_{0|1}, \delta^1_{1|0}, \delta^2_{0|1}, \delta^2_{1|0})$, are identified from the joint distribution of the data, (y_i, h_{1i}, h_{2i}) by using the structure of the outcome equation (9) and the two assumptions CIA and NMA. In the context of the minimal model without regressors, the conditional independence assumption requires a stronger formulation: Conditional on h_i^*, the reported h_{1i} and h_{2i} are independent of each other and of y_i. That is, independence is assumed marginal of any regressors $\boldsymbol{x}_i$. And in the context of binary h_i^*, the no-mirror assumption amounts to $\delta^m_{0|1}, \delta^m_{1|0} < 0.5$.

The joint distribution of the 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,

$$\begin{aligned} F(r_0, r_1, r_2) &= \pi\, F(r_0, r_1, r_2|h_i^*{=}1) \quad + \quad (1 - \pi)\, F(r_0, r_1, r_2|h_i^*{=}0) \\ &= \pi\, F(r_0|h_i^*{=}1)\, F(r_1|h_i^*{=}1)\, F(r_2|h_i^*{=}1) \\ &\quad + (1 - \pi)\, F(r_0|h_i^*{=}0)\, F(r_1|h_i^*{=}0)\, F(r_2|h_i^*{=}0), \end{aligned} \tag{13}$$

where

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

The second equality in (13) follows from CIA. To see an example of one of the expressions in (13), consider $F(1,1,1)$:

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

The model fulfils a necessary condition for identification since the data provides seven linearly independent quantities $F(r_0, r_1, r_2)$, which map to the seven parameters of the model: α, β_0, π, $\delta^1_{0|1}$, $\delta^1_{1|0}$, $\delta^2_{0|1}$, $\delta^2_{1|0}$.[15] However, there are two solutions to this problem. NMA obtains a unique solution and identifies the parameters by selecting the solution where the probabilities of reporting truthfully are greater than the probabilities of misreporting. That is, NMA rules out the "mirror solution" in which probabilities of misreporting and correctly reporting are switched and the impact of each health level on the outcome y_i is also switched: $\tilde{\alpha} = -\alpha$ and $\tilde{\beta}_0 = \beta_0 + \alpha$.[16]

Thus, under CIA and NMA, 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).

A.2 Identification in the general logit model

The model discussed so far is quite minimal. Not only does it not include any other regressors apart from the health indicator, but the misclassification probabilities are the same across all individuals. The identification results translate easily to the case of covariates in the outcome equation and heterogeneous misclassification probabilities which depend on these covariates as well. First, consider the

[15] Note that for Eq. (13) to provide seven linearly independent quantities we need the regularity condition that the outcome be informative of the SAH status; that is, we need to assume that $\alpha \neq 0$. Thus, while one can identify if α is equal to 0 or not, it is not possible in the case of $\alpha = 0$ to further estimate the misclassification probabilities because the outcome does not inform us about which SAH group each person falls into.

[16] See Hu (2008) for a discussion of an alternative identifying mirror assumption: that the estimated direction of the impact of each health level on the outcome is known; that is, in this case, that $\hat{\alpha}$ is negative.

case of modifying the minimal model by only adding covariates to the outcome equation. 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 (13) and the corresponding expressions are then simply to be taken conditional on $\boldsymbol{x}_i$. The number of parameters to be estimated is now $6 + K$ (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.

With covariates, it is also possible to weaken the stronger version of the conditional independence assumption invoked in above in Section A.1.[17] Violation of the stronger version of the CIA can occur, for example, if men and women have different misreporting probabilities. In such a case, the assumption does not hold because the two misreported measures will be dependent through the impact of gender. Thus, by explicitly making the misclassification probabilities dependent on x_i and only requiring CIA to hold conditional on some x_i identification can be achieved under weaker conditions.

Is the model with covariates still identified under CIA and NMA? 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 under CIA follows immediately from it being equivalent to the identification of each subsample under the constraint that α and $\boldsymbol{\beta}$ are the same across subsamples.

This reasoning also shows identification of the more general outcome model with interactions between SAH and all or some of the regressors $\boldsymbol{x}_i$, as presented in equation (4). Suppose 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 for each subsample (where each is already identified as before) without imposing the restriction that the slopes on $\boldsymbol{x}_i$ be the same across subsamples.

When some of the other regressors are continuous, we cannot directly resort to this simple approach of identification in each subsample. However, the model remains identified: Intuitively, in an infinitely large sample, we could discretize the continuous regressors ever more finely and then apply identification in each subsample. A formal proof of the model's nonparametric identification has been given by Hu (2008). Because of the underlying nonparametric identification of the misclassification, our parametric approach (which consists in specifying multinomial logit based forms for the misclassification probabilities and SAH—cf. equations (3)) can be implemented flexibly. For instance, it is straightforward to increase the flexibility of the parametric form in the misclassification equations to test the sensitivity of the results to a particular functional form. More generally, our parametric approach can also serve as the basis of a nonparametric estimation via a series-estimation approach (such as by including polynomials or splines of the linear indices, cf. Newey, 1994).

To conclude, we discuss identification for the general case of a categorical SAH that has five outcomes,

[17] The strong version of CIA was used, for instance, by Gosling & Saloniki (2014) in an application to misreported binary disability status.

$h_i^* \in \{0,1,2,3,4\}$; that is, the full model defined in equations (1)-(3). The outcome equation (1) has $4+K$ parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, and there are now twenty misreporting probabilities (equation (2)) per measure h_{mi}, $m = 1, 2$. When parametrised in terms of $\boldsymbol{x}_i$ as in (3), these are $20K$ parameters per measure. In addition, there are four probabilities of the distribution of SAH, $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)'$, where $\pi_j \equiv P(h_i^* = j)$, which parametrised as in (3) adds $4K$ parameters. Thus, the grand total is $45K + 4$ 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 $2 \times 5 \times 5 = 50$ 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 same arguments as before can be made. NMA ensures the uniqueness of the solution with categorical or ordinal SAH by discarding the 119 "mirror solutions" that violate the condition that $\delta^m_{j|j} > \delta^m_{k|j}$ (there are 120 sets of solutions here; i.e., 120 ways to order the 5 groups that correspond to each health level). The counterpart to equation (13) in the case of an ordinal regressor h_i^* is

$$\begin{aligned} F(r_0, r_1, r_2) &= \sum_{j=0}^{4} p_j^*\, F(r_0, r_1, r_2 | h_i^* = j) \\ &= \sum_{j=0}^{4} p_j^*\, F(r_0 | h_i^* = j) F(r_1 | h_i^* = j) F(r_2 | h_i^* = j), \end{aligned} \tag{14}$$

where, naturally, $\pi_0 = 1 - \sum_{j=1}^{4} \pi_j$, and conditioning on $\boldsymbol{x}_i$ is omitted for notational simplicity. Again, the second equality follows from CIA.

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 $F(r_m | h^*)$:

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

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

$$\delta^m_{j|j} = P(h_{mi} = j | h_i^* = j) = 1 - \sum_{k \neq j} \delta^m_{k|j}. \tag{15}$$

A.3 Additional EM estimation details

The model can be estimated in a number of ways based on the joint distribution function (13), which is in the form of a finite mixture (FM) or latent class model. One option is GMM estimation, which is presented below in A.4. Another option is maximum likelihood estimation. The maximum likelihood estimator of the parameters of the model is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}; y_i, h_{1i}, h_{2i}, \boldsymbol{x}_i) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \sum_{r_0} \sum_{r_1} \sum_{r_2} I_i^{r_0 r_1 r_2} \ln(F(r_0, r_1, r_2)), \tag{16}$$

where $\boldsymbol{\theta}$ collects all the parameters: $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\gamma}_{k|j}^m$ for $m = 1, 2$ and $j \neq k$. $I_i^{r_0 r_1 r_2}$ is an indicator variable equal to one if $y_i = r_0$, $h_{1i} = r_1$ and $h_{2i} = r_2$. Maximisation can be implemented, in principle, directly via a standard Newton-Raphson procedure based on (16). However, we found that, especially in models with categorical health and several regressors, maximum likelihood estimation via the Expectation-Maximisation (EM) algorithm (Dempster *et al.*, 1977) was substantially faster and more stable, than either GMM or standard maximum likelihood, making it our only viable estimator.

The EM algorithm iterates between the maximisation or M-step, and the expectation or E-step. The nth iteration of the M-step is

$$\hat{\boldsymbol{\theta}}^n = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \tilde{\ell}_i(\boldsymbol{\theta}; y_i, h_{1i}, h_{2i}, \boldsymbol{x}_i, \hat{w}_i^n), \tag{17}$$

where

$$\tilde{\ell}_i(\cdot) = \sum_{j=0}^{4} \hat{w}_{ji}^n \Big(\ln F(y_i | h_i^* = j, \boldsymbol{x}_i) + \ln F(h_{1i} | h_i^* = j, \boldsymbol{x}_i) + \ln F(h_{2i} | h_i^* = j, \boldsymbol{x}_i) + \ln \pi_{ji} - \ln \hat{w}_{ji}^n \Big), \tag{18}$$

and all $F(\cdot|\cdot)$ denote corresponding terms defined in (13), and the $\hat{w}_{ji}^n$ are estimates of the posterior probabilities $P(h^* = j | y_i, h_{1i}, h_{2i}, \boldsymbol{x}_i)$ defined in equation (7), and which constitute the E-step of the algorithm. Note that equation (18) is the default M-step and its use leads to what we have called the FM estimator, while using the penalised counterpart, equation (6), leads to what we have called the PFM estimator. Both estimators rely on the same E-step, eq. (7).

A.4 GMM estimation

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

$$E\big(\left[I_i^{r_0 r_1 r_2} - F_i(r_0, r_1, r_2) \right] \boldsymbol{x}_i \big) = 0, \tag{19}$$

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

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \boldsymbol{Q}_i(\boldsymbol{\theta})' \boldsymbol{W}_N \boldsymbol{Q}_i(\boldsymbol{\theta}), \tag{20}$$

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.5 Counts and durations: Poisson and Weibull models

The proposed approach, which we have presented for a logit binary outcome, can be extended to many common nonlinear models that follow the form

$$f(y_i | h_i^*, \boldsymbol{x}_i) = g(\alpha\, h_i^* + \boldsymbol{x}_i' \boldsymbol{\beta};\ \omega), \tag{21}$$

where $f(\cdot|\cdot)$ is a functional of the conditional distribution of y_i given 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

$$P(y_i | h_i^*, \boldsymbol{x}_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \qquad \lambda_i = \exp(\alpha\, h_i^* + \boldsymbol{x}_i' \boldsymbol{\beta}), \tag{22}$$

where the left-hand-side of (22) corresponds to $f(\cdot|\cdot)$ and the right-hand-side to $g(\cdot)$. We can use the EM algorithm described in (5)-(7) 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 (22). 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

$$E\Big((y_i - \lambda_i)\boldsymbol{x}_i\Big) = 0, \tag{23}$$

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

$$f(y_i|h_i^*, \boldsymbol{x}_i) = \lambda_i \omega y_i^{\omega-1} \exp(-\lambda y_i^{\omega}), \qquad \lambda_i = \exp(\alpha\, h_i^* + \boldsymbol{x}_i'\boldsymbol{\beta}). \tag{24}$$

Results from a Monte Carlo simulation on the performance of the FM and PFM estimators for Poisson and Weibull models are presented in Section B.5.

A.6 Using estimates from one dataset to adjust for misclassification in other datasets

Estimates of the misclassification system obtained from one sample might be used to adjust for misclassification in some other sample. We discuss two possible cases in which this is possible here.

Case 1: Same covariate specification, different data

In this case, we want to use the PFM estimates to obtain estimates for another outcome and another dataset but where a second measure of SAH is unavailable. However, the single available SAH as well as the covariates are the same as in the original dataset. Same here means that they are measured/coded in the same way (for instance, same number and labelling of categories of SAH).

Suppose Dataset A consists of $i = 1, \ldots, N^A$ observations drawn from $(y_i^A, h_{1i}, h_{2i}, \boldsymbol{x}_i)$, and that both CIA and NMA are satisfied. We can then obtain estimates of the entire vector of model parameters θ via the PFM estimator. In particular, there are estimates of the misclassification parameters γ at hand, say $\hat{\gamma}^A$, and of the parameters of SAH, say $\hat{\delta}^A$.

Suppose Dataset B consists of $j = 1, \ldots, N^B$ observations drawn from $(y_j^B, h_{1j}, \boldsymbol{x}_j)$.

We can obtain estimates of the parameters of the outcome equation for y_j^B, say $(\alpha^B, \boldsymbol{\beta}^B)$, by estimating the following modified version of the EM algorithm presented in equations (5)-(7). The nth iteration of the M-step is

$$(\hat{\alpha}^{B,n}, \hat{\boldsymbol{\beta}}^{B,n}) = \arg\max_{\alpha^B, \boldsymbol{\beta}^B} \sum_{j=1}^{N^B} \tilde{\ell}_j(\alpha^B, \boldsymbol{\beta}^B; y_j^B, h_{1j}, \boldsymbol{x}_j, \hat{w}_j^n), \tag{25}$$

with

$$\tilde{\ell}_j(\cdot) = \sum_{k=0}^{4} \hat{w}_{kj}^n \left(\ln F(y_j^B|h_j^*{=}k, \boldsymbol{x}_j) + \ln \hat{F}^A(h_{1j}|h_j^*{=}k, \boldsymbol{x}_j) + \ln \hat{\pi}_{kj}^A - \ln \hat{w}_{kj}^n \right), \tag{26}$$

where $\hat{F}^A(h_{1j}|h_j^*{=}k, \boldsymbol{x}_j)$ is evaluated at estimates of $\hat{\gamma}^A$; and $\hat{\pi}_{kj}^A$, at estimates of $\hat{\delta}^A$.

In the $(n{+}1)$th iteration of the E-step, the posterior probabilities are updated according to

$$\hat{w}_{kj}^{n+1} = \frac{\hat{\pi}_{kj}^A\ \hat{F}^n(y_j^B|h_j^*{=}k, \boldsymbol{x}_j)\ \hat{F}^A(h_{1j}|h_j^*{=}k, \boldsymbol{x}_j)}{\sum_{k=0}^{4} \hat{\pi}_{kj}^A\ \hat{F}^n(y_j^B|h_j^*{=}k, \boldsymbol{x}_j)\ \hat{F}^A(h_{1j}|h_j^*{=}k, \boldsymbol{x}_j)}. \tag{27}$$

Thus, compared to (5)-(7), here there is only one reported measure of SAH, h_1. Moreover, both the terms $\hat{F}^A(h_{1j}|h_j^*=k,\boldsymbol{x}_j)$ and $\hat{\pi}_{kj}^A$ are constructed from the PFM estimates obtained from Dataset A and are not updated during the estimation using Dataset B.

Case 2: Different specifications, same data

In this case, we want to use the PFM estimates to obtain estimates for another outcome but for the same individuals and where a second measure of SAH is unavailable. An example would be a panel survey where two measures of SAH are available only in some waves and some outcomes of interest are not available in the waves with two SAH measures (such as is the case across different waves of HILDA). In this case, the covariates included in the models of the two samples can but need not be the same.

Suppose Dataset A consists of $i=1,\ldots,N$ observations drawn from $(y_i^A, h_{1i}, h_{2i}, \boldsymbol{x}_i^A)$, and that both CIA and NMA are satisfied. We can then obtain estimates of the entire vector of model parameters θ via the PFM estimator. In particular, there are estimates of the misclassification parameters γ at hand, say $\hat{\gamma}^A$, and of the parameters of SAH, say $\hat{\delta}^A$.

Suppose Dataset B consists of $i=1,\ldots,N$ observations drawn from $(y_i^B, h_{1i}, \boldsymbol{x}_i^B)$.

We can obtain estimates of the parameters of the outcome equation for y_i^B, say $(\alpha^B, \boldsymbol{\beta}^B)$, by estimating the following modified version of the EM algorithm presented in equations (5)-(7). The nth iteration of the M-step is

$$(\hat{\alpha}^{B,n}, \hat{\boldsymbol{\beta}}^{B,n}) = \arg\max_{\alpha^B,\boldsymbol{\beta}^B} \sum_{i=1}^{N} \tilde{\ell}_i(\alpha^B, \boldsymbol{\beta}^B; y_i^B, h_{1i}, \boldsymbol{x}_i^A, \boldsymbol{x}_i^B, \hat{w}_i^n), \tag{28}$$

with

$$\tilde{\ell}_i(\cdot) = \sum_{j=0}^{4} \hat{w}_{ji}^n \left(\ln F(y_i^B|h_i^*=j, \boldsymbol{x}_i^B) + \ln \hat{F}^A(h_{1i}|h_i^*=j, \boldsymbol{x}_i^A) + \ln \hat{\pi}_{ji}^A - \ln \hat{w}_{ji}^n \right), \tag{29}$$

where $\hat{F}^A(h_{1i}|h_i^*=j,\boldsymbol{x}_i^A)$ is evaluated at estimates of $\hat{\gamma}^A$; and $\hat{\pi}_{ji}^A$, at estimates of $\hat{\delta}^A$.

In the $(n+1)$th iteration of the E-step, the posterior probabilities are updated according to

$$\hat{w}_{ji}^{n+1} = \frac{\hat{\pi}_{ji}^A \, \hat{F}^n(y_i^B|h_i^*=j,\boldsymbol{x}_i^B) \, \hat{F}^A(h_{1j}|h_i^*=j,\boldsymbol{x}_i^A)}{\sum_{j=0}^{4} \hat{\pi}_{ji}^A \, \hat{F}^n(y_i^B|h_i^*=j,\boldsymbol{x}_i^B) \, \hat{F}^A(h_{1i}|h_i^*=j\boldsymbol{x}_i^A)}. \tag{30}$$

Similar to Case 1, there is again only one reported measure of SAH, h_1, and both $\hat{F}^A(h_{1i}|h_i^*=j,\boldsymbol{x}_i^A)$ and $\hat{\pi}_{ji}^A$ are constructed from the PFM estimates obtained from Dataset A and are not updated during the estimation using Dataset B. Note that there is no restriction on the relationship between $\boldsymbol{x}_i^A$ and $\boldsymbol{x}_i^B$. They can be identical, disjoint, or overlapping. The key, however, is that CIA holds conditional on $\boldsymbol{x}_i^A$. In practice, therefore, for many applications it might be sensible that $\boldsymbol{x}_i^A$ be a subset of $\boldsymbol{x}_i^B$.

A.7 Formulas for the matrices in Panels (A) and (B) of Figure 1

Table A1: Formulas for Panel (A) of Figure 1: Joint distribution of (h_{1i}, h_{2i})

	h_{2i}	
h_{1i}	BAD ($h_{2i}=0$)	GOOD ($h_{2i}=1$)
BAD ($h_{1i}=0$)	$F_{h_1h_2}(0,0) =$ $\pi\delta^1_{0\|1}\delta^2_{0\|1}+$ $(1-\pi)(1-\delta^1_{1\|0})(1-\delta^2_{1\|0})$	$F_{h_1h_2}(0,1) =$ $\pi\delta^1_{0\|1}(1-\delta^2_{0\|1})+$ $(1-\pi)(1-\delta^1_{1\|0})\delta^2_{1\|0}$
GOOD ($h_{1i}=1$)	$F_{h_1h_2}(1,0) =$ $\pi(1-\delta^1_{0\|1})\delta^2_{0\|1}+$ $(1-\pi)\delta^1_{1\|0}(1-\delta^2_{0\|1})$	$F_{h_1h_2}(1,1) =$ $\pi(1-\delta^1_{1\|0})(1-\delta^2_{1\|0})+$ $(1-\pi)\delta^1_{1\|0}\delta^2_{1\|0}$

Notes: The table shows all the outcomes of the joint distribution of h_{1i} and h_{2i}, $F_{h_1h_2}(r_1,r_2) \equiv P(h_{1i}=r_1, h_{2i}=r_2)$, where $r_1 \in \{0,1\}$ and $r_2 \in \{0,1\}$. The right-hand side of the equations in the cells follow from the fact that $F_{h_1h_2}(r_1,r_2) = F_{h_1h_2|h^*=1}(r_1,r_2) + F_{h_1h_2|h^*=0}(r_1,r_2)$, and, by independence, $F_{h_1h_2|h^*=j}(r_1,r_2) = F_{h_1|h^*=j}(r_1)F_{h_2|h^*=j}(r_2)$, where $F_{h_1h_2|h^*=j} \equiv P(h_1=r_1, h_2=r_2|h^*=j)$, $F_{h_m|h^*=j}(r_m) \equiv P(h_m=r_m|h^*=j)$, $j \in \{0,1\}$ and $m=1,2$.

Table A2: Formulas for Panel (B) of Figure 1: Joint distribution of (y_i, h_{1i}, h_{2i})

	h_{2i}	
h_{1i}	BAD ($h_{2i}=0$)	GOOD ($h_{2i}=1$)
BAD ($h_{1i}=0$)	DEAD ($y_i=1$): $F_{y\,h_1h_2}(1,0,0) = \pi\bar{Y}_G\delta^1_{0\|1}\delta^2_{0\|1}$ $+(1-\pi)\bar{Y}_B(1-\delta^1_{1\|0})(1-\delta^2_{1\|0})$ ALIVE ($y_i=0$): $F_{y\,h_1h_2}(0,0,0) = \pi(1-\bar{Y}_G)\delta^1_{0\|1}\delta^2_{0\|1}$ $+(1-\pi)(1-\bar{Y}_B)(1-\delta^1_{1\|0})(1-\delta^2_{1\|0})$	DEAD ($y_i=1$): $F_{y\,h_1h_2}(1,0,1) = \pi\bar{Y}_G\delta^1_{0\|1}(1-\delta^2_{0\|1})+$ $(1-\pi)\bar{Y}_B(1-\delta^1_{1\|0})\delta^2_{1\|0}$ ALIVE ($y_i=0$: $F_{y\,h_1h_2}(0,0,1) = \pi(1-\bar{Y}_G)\delta^1_{0\|1}(1-\delta^2_{0\|1})+$ $(1-\pi)(1-\bar{Y}_B)(1-\delta^1_{1\|0})\delta^2_{1\|0}$
GOOD ($h_{1i}=1$)	DEAD ($y_i=1$): $F_{y\,h_1h_2}(1,1,0) = \pi\bar{Y}_G(1-\delta^1_{0\|1})\delta^2_{0\|1}+$ $(1-\pi)\bar{Y}_B\delta^1_{1\|0}(1-\delta^2_{0\|1})$ ALIVE ($y_i=0$): $F_{y\,h_1h_2}(0,1,0) = \pi(1-\bar{Y}_G)(1-\delta^1_{0\|1})\delta^2_{0\|1}+$ $(1-\pi)(1-\bar{Y}_B)\delta^1_{1\|0}(1-\delta^2_{0\|1})$	DEAD ($y_i=1$): $F_{y\,h_1h_2}(1,1,1) = \pi\bar{Y}_G(1-\delta^1_{1\|0})(1-\delta^2_{1\|0})+$ $(1-\pi)\bar{Y}_B\delta^1_{1\|0}\delta^2_{1\|0}$ ALIVE ($y_i=0$): $F_{y\,h_1h_2}(0,1,1) = \pi(1-\bar{Y}_G)(1-\delta^1_{1\|0})(1-\delta^2_{1\|0})+$ $(1-\pi)(1-\bar{Y}_B)\delta^1_{1\|0}\delta^2_{1\|0}$

Notes: The table shows all the outcomes of the joint distribution of y_i, h_{1i} and h_{2i}, $F_{y\,h_1h_2}(r_0,r_1,r_2) \equiv P(y_i=r_0, h_{1i}=r_1, h_{2i}=r_2)$, where $r_0 \in \{0,1\}$, $r_1 \in \{0,1\}$ and $r_2 \in \{0,1\}$. The right-hand side of the equations in the cells follow from the fact that $F_{y\,h_1h_2}(r_0,r_1,r_2) = F_{y\,h_1h_2|h^*=1}(r_0,r_1,r_2) + F_{y\,h_1h_2|h^*=0}(r_0,r_1,r_2)$, and, by independence, $F_{y\,h_1h_2|h^*=j}(r_0,r_1,r_2) = F_{y|h^*=j}(r_0)F_{h_1|h^*=j}(r_1)F_{h_2|h^*=j}(r_2)$, where $F_{y\,h_1h_2|h^*=j} \equiv P(y=r_0, h_1=r_1, h_2=r_2|h^*=j)$, $F_{y|h^*=j}(r_0) = P(y=r_0|h^*=j)$, $F_{h_m|h^*=j}(r_m) \equiv P(h_m=r_m|h^*=j)$, $r_m \in \{0,1\}$ and $m=1,2$. Further, for compactness, in the cells the shorthand notation $\bar{Y}_G \equiv F_{y|h^*=1}(1)$ and $\bar{Y}_B \equiv F_{y|h^*=0}(1)$ is used.

B Monte Carlo Simulation: Details and additional results

B.1 Baseline simulation DGP

In the baseline design, $\boldsymbol{x}_i = (1, x_i)$, where $x_i \sim U(0,1)$; health status h_i^* is drawn from a Bernoulli distribution with probability π_i; ε_i, from a logistic distribution. Survival status y_i (=1 if alive) is generated as

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

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 unobserved 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 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 $P(h^* = 1) = 0.7$. The reported health measures have marginal distributions $P(h_1 = 1) = 0.61$ and $P(h_2 = 1) = 0.57$. The share of conflicting answers is $P(h_1 \neq h_2)$=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 ($\delta^2_{1|0}$)).

Sample sizes are $N = \{1000; 10000\}$ and the number of replications is 500.

B.2 Baseline simulation results and comparison to ad-hoc approaches to adjust for misclassification

Table B1 contains the results for the infeasible estimator using the unobserved h_i^* as a regressor (column "h^*"), the naïve estimator using the misreported h_{1i} as a regressor ("h_i") as well as the FM and PFM estimators, for the two sample sizes of 1,000 and 10,000 observations. The results for 10,000 observations are the ones presented in Table 1 in the paper.

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

Table B1: 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

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), and the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. The true values of the parameters in the DGP are $\alpha = 1$, β const=0, β slope=1, η const=-0.1342, η slope=1.5, $\gamma^m_{k|j}$slope $= 1$ for all m, k, and $\gamma^1_{0|1}$const=-0.25, $\gamma^2_{0|1}$const=-0.75 , $\gamma^1_{1|0}$const=0, and $\gamma^2_{1|0}$const=-0.5. See Appendix B.1 for more details on 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, for the parameters of the misclassification system, at N=1,000, there are larger biases, ranging up to about 20 percent; 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 in the RMSE of these parameters that range from 50 to almost 90 percent. This improvement in the estimation of the misclassification parameters 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.

Table B2: Simulation results: Ad-hoc misclassification approaches

		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

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), the average of h_1 and h_2 ($\bar{h}$), h_1 in the sample restricted to i with $h_{1i} = h_{2i}$, predicted h_1 ($\hat{h}_1$) and the residual from a prediction of h_1 ($\hat{e}_1$). The true values of the parameters in the DGP are α=1, β const=0, β slope=1. See Appendix B.1 for more details on the DGP.

Table B2 contains the results for four potential competitor estimators, which address the misclassification in an ad-hoc way and are sometimes encountered in the literature. The top panel corresponds to results for a sample size of 1,000; the bottom panel, to results for a sample size of 10,000. The latter panel is also included in Table 1 in the paper. Below we give some more detail on these ad-hoc approaches.

First, we experiment by using the average of the two SAH measures as the regressor in the models (in tables, we denote this estimator as"$\bar{h}$"). 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 may 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 ad-hoc estimators included in the simulation correspond to approaches that mimic two-stage least squares in linear models. They consist of 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 these estimators then includes the first-stage predictions as the regressor in the outcome model ("$\hat{h}_1$"), an inconsistent approach for nonlinear models, but unfortunately often encountered in the literature. 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, but not in general when the endogenous regressor (here, SAH) is discrete.

The results in Table B2 show that for 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 -25 percent for both estimators. Thus, these procedures improve over the estimation using a single reported measure, but the bias is still very large.

The columns "$\hat{h}_1$" and "$\hat{e}_1$" report the results for the possible *ad hoc* methods related to IV estimation. The control function approach "$\hat{e}_1$" has been advocated as a potentially useful remedy that might not cure the problem but reduce it in some circumstances even if its assumptions are violated (Basu & Coe, 2015; Wooldridge, 2014). However, 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 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 approaches the bias as variances shrink with increasing N.

B.3 Simulation DGP with interaction effect in unobserved health

The ability to easily specify interaction effects is a hallmark of our approach, and in this section we simulate from a DGP where the impact of SAH on the outcome varies with x:

$$y_i = \mathbb{1}\left(\alpha\, h_i^* + \alpha_x h_i^* x_i + \beta_0 + \beta_1 x + \varepsilon_i > 0\right), \tag{31}$$

where α_x is the coefficient on the new interaction between health and x. Table B3 shows the results from this DGP. The results for N=10,000 (right panel) correspond to Panel (A) in Table 2 in the paper. Compared to the baseline, this DGP is more difficult to estimate. For instance, compared to the RMSE of α in the baseline case from Table B1 the RMSE at N=1,000 for the infeasible estimator almost doubles for α_{const} (and quadruples for α_{slope}, i.e. the interaction coefficient). 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 of about 25 percent with N=1,000. However, the PFM estimator is able to obtain improved estimates, 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, the FM estimator works well and the advantages of PFM over FM in this DGP are only marginal.

Table B3: 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

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), and the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. The true values of the parameters in the DGP are α const=1, α slope=1, β const =-0.375, β slope=1; all misclassification parameters are kept at their baseline values (see notes of Table 1); see Appendix B.1 for more details on the DGP.

B.4 Simulation DGP for multinomial health with five categories

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

$$y_i = \mathbb{1}\left(\alpha_1 h^*_{1i} + \alpha_2 h^*_{2i} + \alpha_3 h^*_{3i} + \alpha_4 h^*_{4i} + \beta_0 + \beta_1 x + \varepsilon_i > 0\right), \tag{32}$$

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 unobserved 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 B4 for N=1,000 (left panel) and N=10,000 (right panel). Results for the right panel correspond to those presented in Table 2 in the paper. Again, that this is a more challenging DGP can be seen in the

Table B4: 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

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), and the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. The true values of the parameters in the DGP are $\alpha = 0.5$, $\alpha = 1$, $\alpha = 1.5$, $\alpha = 2$, β const=-1, β slope=1. See Appendix B.4 for more details on the DGP.

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.

B.5 Simulations for count data and duration data DGPs

Table B5 explores different nonlinear outcome models. Keeping the same values for the parameters as in the baseline, we now change the outcome model to one for counts and one for durations. The count model is a Poisson regression model where y_i is drawn from a Poisson distribution with mean λ_i:

$$y_i \sim Poisson(\lambda_i^{\text{Pois}}), \qquad \lambda_i^{\text{Pois}} = \exp(\alpha\, h_i^* + \beta_0 + \beta_1 x). \tag{33}$$

For the duration model ($y_i > 0$) we use a Weibull regression model with hazard function

$$\lambda_i^{\text{Weib}} = \exp(\alpha\, h_i^* + \beta_0 + \beta_1 x) y_i^{\omega - 1}. \tag{34}$$

The (ancillary) parameter ω determines the form of the duration dependence. We set $\omega = 1.5$, which, being larger than 1, results in positive duration dependence (i.e., all else equal, the hazard increases with the duration of the spell). The Poisson results are in the left panel, the Weibull results in the right panel. We only report the estimates for the smaller sample of $N = 1,000$. With these models, the proposed estimators perform already very well at the smaller sample size. For instance, biases are below 1 per cent for all outcome parameters for FM, and below 2 per cent for PFM. Clearly, the estimators benefit from the added information in these outcomes relative to the sparser case of a binary outcome. For both Poisson and Weibull the naíve estimates of α relying on the observed SAH are biased downwards by roughly 47 per cent.

Table B5: 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

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), and the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. The true values of the parameters in the DGP are $\alpha = 1$, β const=0, β slope=1, η const=-0.1342, η slope=1.5, $\gamma^m_{k|j}$slope $= 1$ for all m, k, and $\gamma^1_{0|1}$const=-0.25, $\gamma^2_{0|1}$const=-0.75 , $\gamma^1_{1|0}$const=0, and $\gamma^2_{1|0}$const=-0.5. For the Weibull DGP, the true value of ω=1.5. Poisson and Weibull models were parametrised as described in Appendix **??**. See Appendix B.1 for more details on the simulation DGP.

B.6 Simulation DGP with two outcomes

Next, we consider the case of a multivariate outcome. In Table B6 we present results from estimations with two outcomes, simulated 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 increasing the information (statistical power) available to estimate these parameters. The extent to which pooling both outcomes adds information depends on the degree of the dependence between the two errors, ε_1 and ε_2 (though this dependence is not estimated with our method). In the worst case, $\varepsilon_1 = \varepsilon_2$ and joint estimation will bring no advantage. 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.

Table B6: 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

Notes: Cell entries show bias and root mean square error for parameters estimated over 500 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), and the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators that adjust for misclassification. For the PFM, the tuning parameter is set to $t = 0.5$. See Appendices B.1 and B.6 for more details on the DGP.

The case ρ=1 is the same as the baseline, and indeed we get very similar results. For both FM and

Table B7: 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

Notes: Cell entries show bias and root mean square error for parameters estimated over 200 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), the Nonparametric IV estimator from Hu (2008) (NPIV), the Finite Mixture (FM) and Penalised Finite Mixture (PFM) estimators. For the PFM, the tuning parameter is set to $t = 0.5$. For NPIV, the results are taken from Hu (2008). The DGP is that from Hu (2008, Table 1, p.45) and given in Appendix B.7. The true values of the parameters in the DGP are α=1, β const=0.5, β slope=1. In Scenario 1, the misclassification probabilities depend negatively on x; in Scenario 2, positively.

PFM, as the correlation decreases, the estimators in general become progressively more successful at reducing the biases, 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 somewhat larger reductions in RMSE are achieved for the parameters of the misclassification system.

B.7 Performance of FM and PFM in a misspecified DGP

So far we have evaluated the performance of the FM and PFM estimators in DGPs where they correctly specify the misclassification system. Here, the proposed parametric estimators are evaluated in a DGP where the misclassification probabilities are misspecified. We use the same DGP 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 unobserved health distribution. First, by using flexible specifications of the linear indices $\boldsymbol{x}_i'\boldsymbol{\gamma}^m_{k|j}$, many functional forms may be approximated well. Second, compared to more nonparametric approaches, even if FM/PFM might be inconsistent due to misspecified functional forms, they might still be preferable in terms of RMSE for finite samples. This simulation gives some evidence of the second point. That is, we do not explore potential further improvements by specifying polynomials of $\boldsymbol{x}_i$ in the linear indices.

In the DGP from Hu (2008), misclassification does not follow our logit-based functional forms. Rather, some misclassification probabilities, for instance, are partially linear functions with kinks. Table B7 shows our results for FM and PFM from this DGP, with N=500 and 200 replications as in the original Hu (2008) paper, next to the h_i^*, h_1 and Hu (2008) NPIV results from their paper. Scenarios 1 and 2 depicted in the table correspond to two variants of the DGP in which the probabilities $\delta^m_{0|1}$ 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. It could be that the good results of FM and PFM were achieved by chance: small sample bias and misspecification bias could be offsetting each other, yielding the low biases observed in the table. To check whether this was the case, we repeated the simulations for N=10,000 for PFM in Scenario 2. This resulted in biases of -0.001, 0.001, and -0.016 for $\hat{\alpha}$, '$\hat{\beta}$ const' and '$\hat{\beta}$ slope', thus dispelling the concern that an equal and opposite small sample bias might be concealing what potentially could be large misspecification biases. It also highlights that the bias due to incorrect functional form of the misclassification in this case is already small.

Details on the DGP

We use the setup of Hu (2008) as reported in Hu (2008, Table 1, p.45). The DGP is for a probit outcome y_i, a binary misclassified regressor h_i^*, and a normally distributed covariate x_i. We adjust our outcome model to be a probit, but leave the misclassification probabilities and π_i as logistic, while in the DGP they are not. Specifically, the DGP is

$$P(y_i = 1|h_i^*, x_i) = \Phi(\alpha h_i^* + \beta_0 + \beta_1 x_i),$$

where $\Phi(\cdot)$ denotes the standard normal CDF, α=1, β_0=0.5 and $\beta_1 = 1$, and $x_i \sim N(0, 0.25)$. Health status is defined as $h_i^* = \mathbb{1}(\epsilon < 0.6)$, where $\epsilon \sim Uniform(0, 1)$. The reported health measures h_1 and h_2 are defined as follows: $h_{2i} = \mathbb{1}(\epsilon + \delta < 0.6)$, where $\delta \sim N(0.0.04)$. For h_{1i},

$$P(h_{1i} = 0|h_i^* = 1, x_i) = \min(1, \max(0, p_i)), \quad P(h_{1i} = 1|h_i^* = 0, x_i) = \min(1, \max(0, q_i)).$$

In Table B7, for results in Panel "*Scenario 1*",

$$p_i = 0.3 - 0.1x_i,$$
$$q_i = 0.2 + 0.1x_i;$$

and for results in Panel "*Scenario 2*",

$$p_i = 0.3 + 0.1x_i,$$
$$q_i = 0.2 + 0.1x_i.$$

B.8 Sensitivity of PFM to the Conditional Independence Assumption

To explore the sensitivity of the proposed estimator to violations of the conditional independence assumption (CIA), we produce the following experiment. We generate the data from a process where the population consists of two groups with different levels misclassification, but that group membership is ignored by the econometrician. This induces dependence between the reported health measures through the unobserved group membership—thus violating CIA—and corresponds to the omission of a dummy variable from the covariate vector of the misclassification probabilities.

Table B8: Sensitivity to conditional independence assumption (CIA), N=1,000

		No omitted variable			Violation of CIA through omitted variable in: (h_1, h_2)			(h_1, h_2, y)			(h_1, h_2, h^*, y)		
		h^* (1)	h_1 (2)	PFM (3)	h^* (4)	h_1 (5)	PFM (6)	h^* (7)	h_1 (8)	PFM (9)	h^* (10)	h_1 (11)	PFM (12)
Parameters of the outcome model													
$\hat{\alpha}$	Bias	0.006	-0.464	0.035	0.010	-0.460	-0.002	-0.008	-0.462	0.005	-0.063	-0.490	-0.058
	RMSE	0.164	0.490	0.256	0.151	0.486	0.258	0.144	0.487	0.254	0.157	0.515	0.257
$\hat{\beta}$ const	Bias	0.007	0.277	0.056	-0.029	0.239	0.065	-0.037	0.224	0.081	-0.005	0.248	0.120
	RMSE	0.152	0.314	0.242	0.158	0.280	0.210	0.156	0.265	0.214	0.153	0.287	0.222
$\hat{\beta}$ slope	Bias	0.016	0.167	-0.018	0.045	0.201	0.026	0.053	0.204	-0.001	0.063	0.201	0.008
	RMSE	0.283	0.326	0.321	0.275	0.323	0.293	0.276	0.323	0.291	0.275	0.321	0.289
Parameters of the misclassification probabilities													
$\hat{\eta}$ const	Bias			-0.253			-0.287			-0.384			-0.424
	RMSE			0.551			0.531			0.584			0.609
$\hat{\eta}$ slope	Bias			-0.039			-0.041			-0.022			-0.010
	RMSE			0.538			0.499			0.491			0.491
$\hat{\gamma}11\|0$ const	Bias			-0.017			-0.028			-0.081			-0.066
	RMSE			0.367			0.333			0.330			0.331
$\hat{\gamma}11\|0$ slope	Bias			-0.607			-0.574			-0.618			-0.631
	RMSE			0.729			0.721			0.758			0.761
$\hat{\gamma}21\|0$ const	Bias			0.147			0.170			0.114			0.126
	RMSE			0.370			0.360			0.337			0.334
$\hat{\gamma}21\|0$ slope	Bias			-0.516			-0.554			-0.588			-0.561
	RMSE			0.636			0.678			0.708			0.677
$\hat{\gamma}10\|1$ const	Bias			0.206			0.283			0.339			0.336
	RMSE			0.397			0.441			0.482			0.480
$\hat{\gamma}10\|1$ slope	Bias			-0.136			-0.208			-0.222			-0.220
	RMSE			0.357			0.413			0.417			0.427
$\hat{\gamma}20\|1$ const	Bias			0.385			0.407			0.446			0.451
	RMSE			0.478			0.493			0.527			0.531
$\hat{\gamma}20\|1$ slope	Bias			-0.379			-0.343			-0.360			-0.365
	RMSE			0.504			0.494			0.509			0.509

Notes: Cell entries show bias and root mean square error for parameters estimated over 200 Monte Carlo replications for the estimators using actual SAH (h^*), reported SAH (h_1), and Penalised Finite Mixture (PFM) estimators. For the PFM, the tuning parameter is set to $t = 0.5$. The DGPs used in columns (4)-(12) include as a covariate an indicator variable which is omitted from estimation. See Appendices B.8 and B.1 for details.

To implement this we modify the baseline DGP by adding a second regressor to the specification of the misreporting probabilities

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

where d_i is the binary regressor omitted in estimation. Both outcomes of the regressor are equally likely and we set all slope coefficients of d_i equal to 0.5. All remaining parameters of the model are left at baseline values (see B.1 for details). The neglected regressor thus has a substantial effect on the misclassification; the variance of the neglected regressor ($\gamma^m_{k|j,d} d_i$) is 0.0625, which is close to the variance of the included regressor ($\gamma^m_{k|j,x} x_i$), 0.0833. Finally, in order to make the DGP comparable to the baseline DGP, we recentre $\gamma^m_{k|j,d} d_i$ around zero. In this way, average levels of misclassification marginal of d_i (i.e., omitting d_i) are the same as in the baseline. Otherwise, any differences we find between this DGP and the baseline could be due to variation in the total amount of misclassification between the two DGPs.

The results of the Monte Carlo simulation with $N = 1,000$ are shown in Table B8. Columns (1)-(3) give results for the baseline case (which is obtained by setting $\gamma^m_{k|j,d} = 0$), for the infeasible, naive and PFM estimators. As is to be expected, these results are virtually identical to the ones in Table B1 (N=1,000). The next three columns (4-6) depict bias and RMSE for the DGP with a common omitted

variable in h_1 and h_2, thus violating CIA. The results show that despite the omission of the unobserved regressor, the performance is not meaningfully different to the baseline case.

Next, we aggravate the problem of omitted variables, by adding the unobserved regressor d_i to the specification of the outcome equation,

$$y_i = \mathbb{1}\left(\alpha\, h_i^* + \beta_0 + \beta_1 x_i + \beta_2 d_i + \varepsilon_i > 0\right),$$

where, as before, we set the slope coefficient of d_i to $\beta_2 = 0.5$ and recentre the neglected part of the linear index to zero. The results for this DGP, where the econometrician has failed to include d_i in estimation, are depicted in Columns (7)-(9). Again, the performance of PFM (and the other estimators) is not materially different from the baseline case (1)-(3) where CIA holds.

A final specification includes the regressor d_i also in the distribution for unobserved health, h_i^*, whose distribution is now

$$\pi_i = \Lambda(\eta_0 + \eta_1 x_i + \eta_2 d_i),$$

with $\eta_2 = 0.5$ and $\eta_2 d_i$ centred on zero. The results (Columns 10-12) show that there is a modest increase in the absolute value of the bias of PFM for the key parameter α from 3.5% to 5.8%. Because d_i is now directly related to the unobserved health h^*, neglecting d_i leads to a classical omitted-variables-bias situation. This can be seen in the fact that the infeasible regressor, which includes h^* but also omits d_i, is equally biased, with a bias (in absolute value) of 6.3%.

To summarise, the simulations show that the PFM approach is reasonably robust to some forms of violations to the conditional independence assumption (CIA). In particular, just omitting an important regressor that influences only the reporting of health but is unrelated to other regressors, health and the outcome, is unlikely to bias key parameters. An important class of practical examples of such regressors are the many types of interviewer effects and survey effects. The approach is also reasonably robust to the case in which, apart from reporting, an unobserved regressor also affects the outcome. If, however, the neglected regressor also impacts health, this will bias the coefficients of the outcome equation, but biases are not substantially worse than those that an infeasible estimator would suffer.

From a practical perspective, this is an important lesson, since it suggests that such violations to CIA result in similar biases than would be obtained if health was observed without potential misclassification. Thus the differences in estimates can be thought of in terms of short versus long regressions (Angrist and Pischke, 2009), that is, in terms of partial effects of regressors in conditional versus marginal models, rather than in terms of biases.

We expect similar results if the omitted regressors are correlated to the included regressors, x_i. In that case, differences in estimates would extend to the slope parameters of the misclassification probabilities and of h^*, but, again, they would largely remain interpretable as (approximate) relationships of the marginal or short model.

C Additional estimation results

Table C1: Sensitivity to tuning parameter. Outcome:*Dead*

Tuning parameter $t =$	(1) 0.25	(2) 0.50	(3) 1.00	(4) 1.50	(5) 2.00	(6) 3.00
α_1	-0.79**	-0.80**	-0.80**	-0.80**	-0.81**	-0.82**
	(0.14)	(0.14)	(0.14)	(0.14)	(0.14)	(0.14)
α_2	-1.13**	-1.13**	-1.13**	-1.14**	-1.14**	-1.15**
	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)
α_3	-1.45**	-1.46**	-1.46**	-1.46**	-1.46**	-1.47**
	(0.16)	(0.16)	(0.16)	(0.16)	(0.16)	(0.16)
α_4	-1.76**	-1.77**	-1.77**	-1.78**	-1.79**	-1.81**
	(0.20)	(0.20)	(0.21)	(0.21)	(0.21)	(0.21)
age	-3.90**	-3.90**	-3.90**	-3.90**	-3.90**	-3.90**
	(1.56)	(1.56)	(1.56)	(1.56)	(1.56)	(1.56)
agesq	13.20**	13.20**	13.20**	13.20**	13.21**	13.22**
	(1.44)	(1.44)	(1.44)	(1.44)	(1.44)	(1.44)
male	0.58**	0.58**	0.58**	0.58**	0.58**	0.58**
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
educ	-0.12	-0.12	-0.12	-0.12	-0.12	-0.11
	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)
lnehi	-0.10*	-0.10*	-0.10*	-0.10*	-0.10*	-0.10*
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
condi	0.27**	0.27**	0.27**	0.27**	0.27**	0.27**
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
married	-0.38**	-0.38**	-0.38**	-0.38**	-0.38**	-0.38**
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
overseas	-0.24**	-0.24**	-0.24**	-0.24**	-0.24**	-0.25**
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
nlf	0.07	0.07	0.06	0.06	0.06	0.06
	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)
unemp	0.07	0.07	0.07	0.07	0.07	0.07
	(0.25)	(0.25)	(0.25)	(0.25)	(0.25)	(0.25)
smoker	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
N	12,908	12,908	12,908	12,908	12,908	12,908

Source: HILDA waves 1 and 16, own calculations. See notes in Table 6 for more information.
* $p < 0.10$, ** $p < 0.05$

Table C2: Sensitivity to tuning parameter. Outcome: *Chronic cond.*

	(1)	(2)	(3)	(4)	(5)	(6)
Tuning parameter t =	0.25	0.50	1.00	1.50	2.00	3.00
α_1	-0.16	-0.15	-0.15	-0.14	-0.14	-0.12
	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)
α_2	-0.41**	-0.41**	-0.40**	-0.40**	-0.40**	-0.38**
	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)
α_3	-0.72**	-0.71**	-0.71**	-0.71**	-0.70**	-0.69**
	(0.17)	(0.18)	(0.18)	(0.18)	(0.18)	(0.18)
α_4	-1.12**	-1.11**	-1.11**	-1.10**	-1.10**	-1.08**
	(0.20)	(0.20)	(0.20)	(0.20)	(0.20)	(0.20)
age	6.28**	6.28**	6.29**	6.29**	6.29**	6.30**
	(1.39)	(1.39)	(1.39)	(1.39)	(1.39)	(1.39)
agesq	-3.08**	-3.08**	-3.08**	-3.09**	-3.09**	-3.10**
	(1.49)	(1.49)	(1.49)	(1.49)	(1.49)	(1.49)
male	-0.12*	-0.12*	-0.12*	-0.12*	-0.12*	-0.12*
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
educ	-0.52**	-0.52**	-0.52**	-0.52**	-0.52**	-0.52**
	(0.18)	(0.18)	(0.18)	(0.18)	(0.18)	(0.18)
lnehi	-0.17**	-0.17**	-0.17**	-0.17**	-0.17**	-0.17**
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
condi	0.39**	0.39**	0.39**	0.39**	0.39**	0.39**
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
married	-0.14*	-0.14*	-0.14*	-0.14*	-0.14*	-0.14*
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
overseas	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
nlf	0.13	0.13	0.13	0.13	0.13	0.13
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
unemp	0.31*	0.31*	0.31*	0.31*	0.31*	0.30*
	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)
smoker	0.29**	0.29**	0.29**	0.29**	0.29**	0.29**
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
N	7,340	7,340	7,340	7,340	7,340	7,340

Source: HILDA waves 1 and 16, own calculations. See notes in Table 6 for more information.

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

Table C3: Estimation results: Specification with discretised continuous variables

Dep. var.	Dead			Chronic cond.		
	PFM	Diff. to naïve		PFM	Diff. to naïve	
		h_1	h_2		h_1	h_2
	(1)	(2)	(3)	(4)	(5)	(6)
α_1	-0.78**	-0.08	-0.17*	-0.16	-0.00	-0.23**
	(0.13)	(0.06)	(0.09)	(0.17)	(0.10)	(0.11)
α_2	-1.12**	-0.14**	-0.22**	-0.44**	0.00	-0.23**
	(0.14)	(0.06)	(0.08)	(0.17)	(0.09)	(0.10)
α_3	-1.45**	-0.16**	-0.20**	-0.74**	-0.03	-0.26**
	(0.15)	(0.07)	(0.08)	(0.17)	(0.09)	(0.10)
α_4	-1.72**	-0.25**	-0.25**	-1.14**	-0.17*	-0.37**
	(0.20)	(0.11)	(0.10)	(0.20)	(0.10)	(0.11)
age: 30s	1.27**	-0.00	0.01	0.56**	0.01	0.01**
	(0.26)	(0.01)	(0.01)	(0.13)	(0.00)	(0.01)
age: 40s	1.66**	-0.01*	-0.02*	0.86**	-0.01	-0.01
	(0.25)	(0.01)	(0.01)	(0.12)	(0.01)	(0.01)
age: 50s	2.41**	-0.03**	-0.02	1.08**	-0.00	0.01
	(0.24)	(0.01)	(0.01)	(0.13)	(0.01)	(0.01)
age: 60s	3.61**	-0.01	0.01	1.43**	-0.02**	0.01
	(0.24)	(0.01)	(0.01)	(0.14)	(0.01)	(0.01)
age: 70 plus	5.16**	-0.04**	0.03**	1.72**	-0.05**	0.00
	(0.24)	(0.01)	(0.01)	(0.17)	(0.01)	(0.01)
male	0.58**	0.00	-0.02**	-0.15**	0.00	-0.01**
	(0.08)	(0.01)	(0.01)	(0.07)	(0.00)	(0.00)
education: year 12	0.14	0.04**	0.04**	-0.10	0.02**	0.02**
	(0.14)	(0.01)	(0.01)	(0.11)	(0.01)	(0.01)
education: certificate	-0.13	0.00	-0.00	-0.03	-0.01	-0.01
	(0.09)	(0.01)	(0.01)	(0.08)	(0.01)	(0.00)
education: bachelor	-0.07	0.01	-0.01	-0.31**	0.00	-0.01
	(0.13)	(0.01)	(0.01)	(0.11)	(0.01)	(0.01)
HH income, 2nd quint.	-0.06	-0.01	-0.00	-0.25**	-0.01*	-0.00
	(0.10)	(0.01)	(0.01)	(0.11)	(0.01)	(0.01)
HH income, 3rd quint.	-0.13	-0.00	0.01	-0.22**	-0.00	0.01*
	(0.12)	(0.01)	(0.01)	(0.11)	(0.01)	(0.01)
HH income, 4th quint.	-0.03	0.02**	0.03**	-0.23**	0.00	0.01**
	(0.12)	(0.01)	(0.01)	(0.11)	(0.01)	(0.01)
HH income, 5th quint.	-0.26*	0.04**	0.04**	-0.29**	0.03**	0.03**
	(0.14)	(0.01)	(0.01)	(0.12)	(0.01)	(0.01)
chronic condition	0.30**	-0.06**	-0.08**	0.40**	-0.02	-0.06**
	(0.09)	(0.02)	(0.02)	(0.09)	(0.02)	(0.01)
married	-0.53**	-0.01*	0.00	-0.11	-0.01*	-0.00
	(0.08)	(0.01)	(0.01)	(0.08)	(0.00)	(0.00)
overseas	-0.24**	0.00	0.02**	-0.02	0.01*	0.01**
	(0.08)	(0.01)	(0.01)	(0.08)	(0.01)	(0.00)
not in labour force	0.20*	-0.03**	-0.05**	0.11	-0.00	-0.02**
	(0.11)	(0.01)	(0.01)	(0.09)	(0.01)	(0.01)
unemployed	0.05	-0.03*	-0.01	0.27	-0.02	-0.01
	(0.26)	(0.02)	(0.01)	(0.17)	(0.01)	(0.01)
smoker	0.49**	-0.00	0.01	0.30**	-0.00	0.00
	(0.08)	(0.01)	(0.01)	(0.07)	(0.00)	(0.00)
N	12,908	12,908	12,908	7,340	7,340	7,340

*N*otes: Source: HILDA waves 1 and 16, own calculations. See notes in Table 6 for more information.
* $p < 0.10$, ** $p < 0.05$

Table C4: 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

*N*otes: Source: HILDA waves 1.

Table C5: Estimation results: System PFM specifications with interactions in health (and difference to naïve estimator using h_1)

	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

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

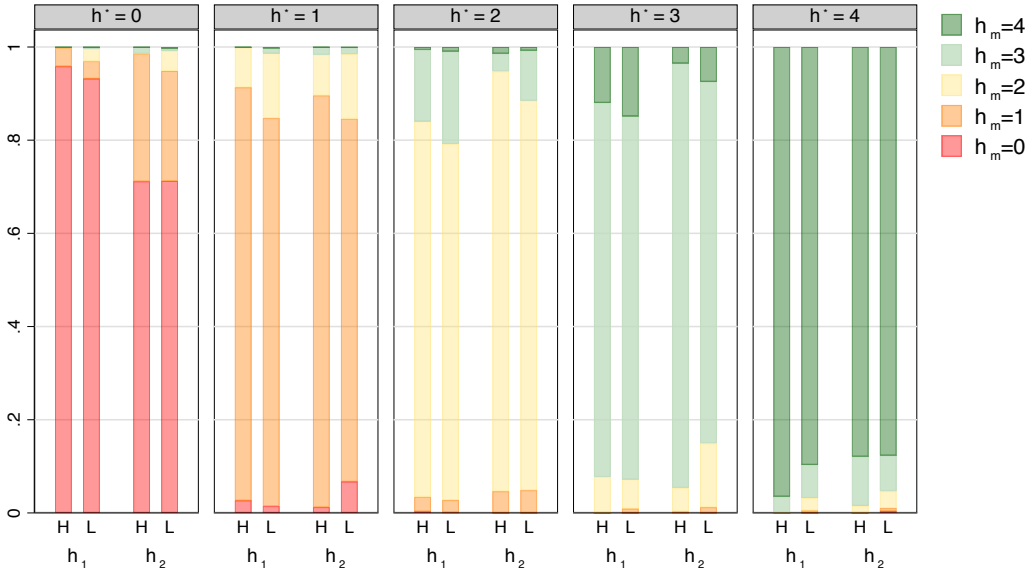
	Interaction w. age			
	Dead	*diff.*	Cond.	*diff.*
age	-3.18	-4.19	9.04	1.76
	(6.44)	(3.30)	(7.38)	(3.96)
agesq	13.17**	3.48	-8.47	-1.92
	(5.70)	(2.66)	(7.61)	(3.78)
α_1: age	0.86	4.90	-3.62	-1.56
	(7.33)	(4.52)	(7.96)	(5.22)
α_1: agesq	-1.47	-4.12	4.72	1.56
	(6.43)	(3.67)	(8.22)	(5.07)
α_1: cons	-0.76	-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

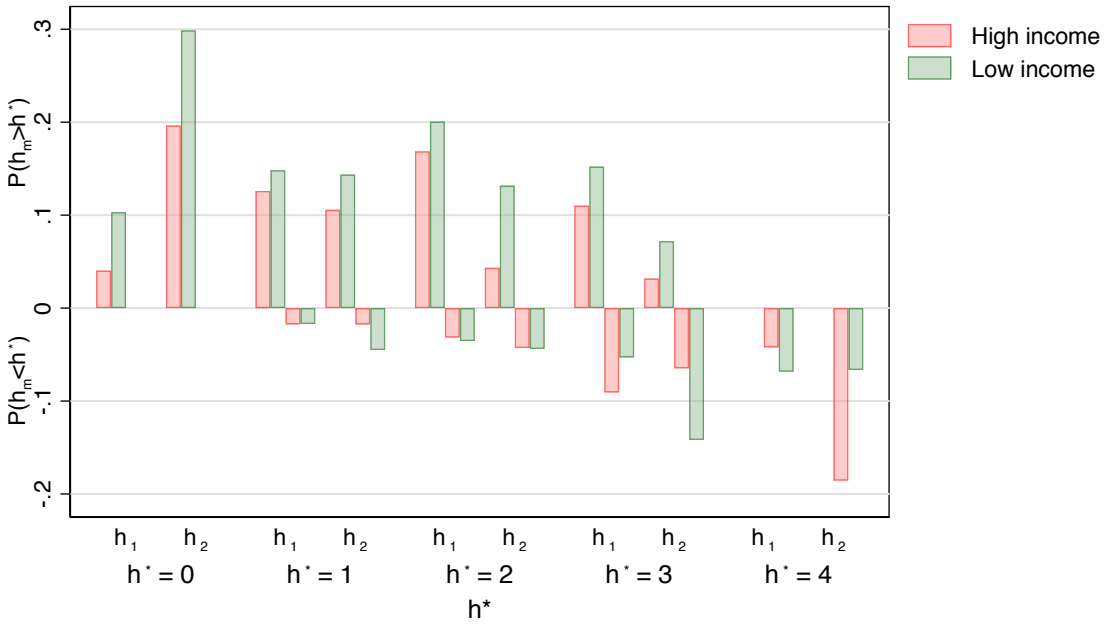
* $p < 0.10$, ** $p < 0.05$

Figure C1: Misclassification in SAH for low and high income individuals

(a) Reporting for high (H) and (L) income individuals



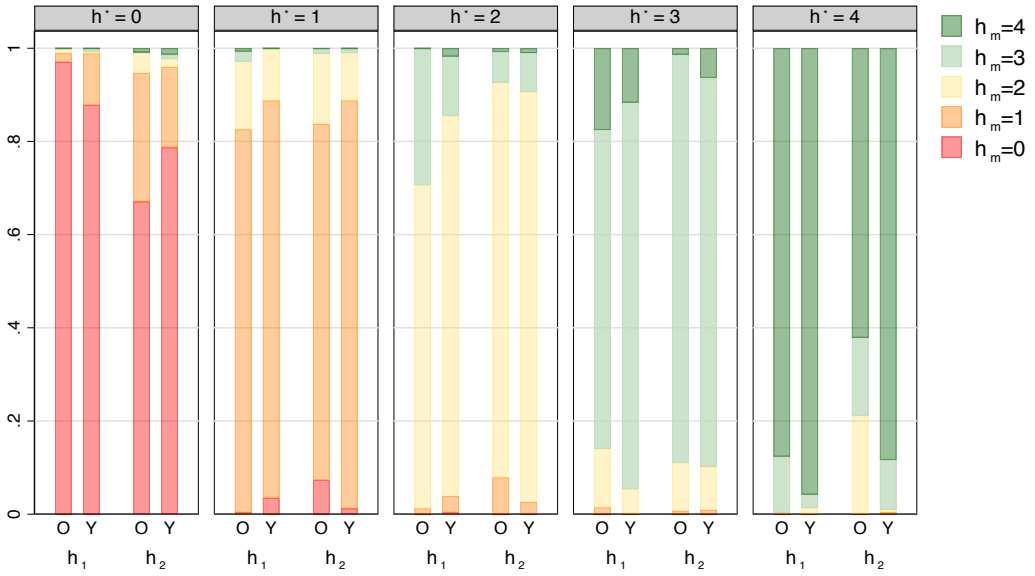
(b) Average predicted upward and downward misreporting



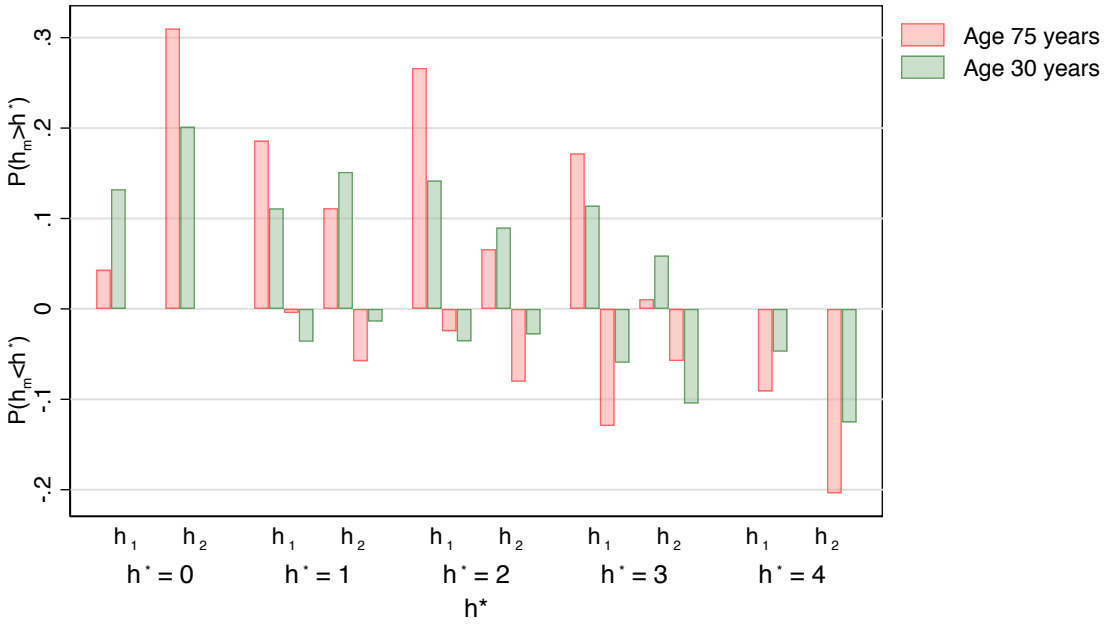
Notes: Estimates from HILDA data waves 1 and 16 for individuals who responded to SAH questions in wave 1. In Panel (a), high income individuals are those in the top quintile and low income individuals those in the bottom quintile of the distribution of equivalised yearly household income. In Panel (b), predicted probabilities are averaged over the whole sample and evaluated at the mean income of the top quintile (High income) and the mean income of the bottom income (Low income).

Figure C2: Misclassification in SAH for old and young individuals

(a) Reporting for old (O) and young (Y) individuals



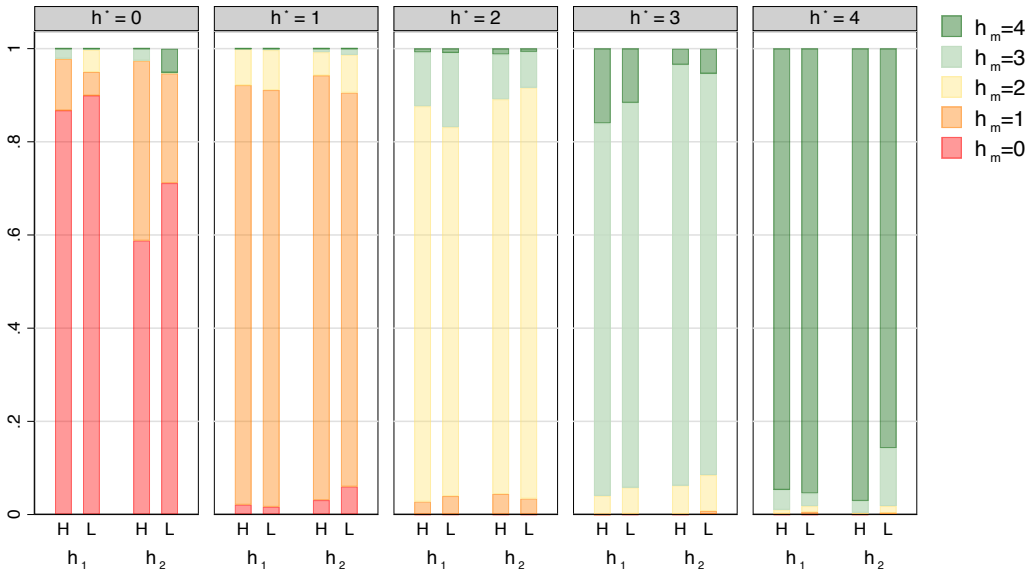
(b) Average predicted upward and downward misreporting



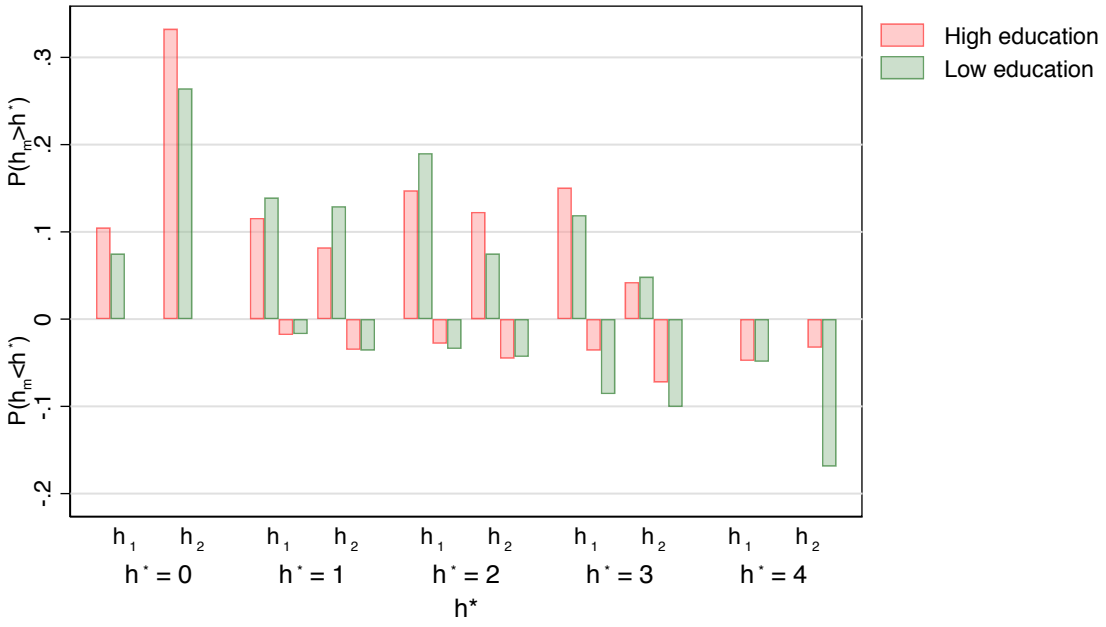
Notes: Estimates from HILDA data waves 1 and 16 for individuals who responded to SAH questions in wave 1. In Panel (a), old individuals are those over the age of 70 years and young individuals those 40 years of age or younger. In Panel (b), predicted probabilities are averaged over the whole sample and evaluated at age 75 years and 30 years.

Figure C3: Misclassification in SAH for individuals with high and low education

(a) Reporting for high (H) and low (L) education individuals



(b) Average predicted upward and downward misreporting



Notes: Estimates from HILDA data waves 1 and 16 for individuals who responded to SAH questions in wave 1. In Panel (a), high education individuals are those whose highest education degree is a bachelor (education=16) and low education individuals those whose highest degree is Year 12 (education=12). In Panel (b), predicted probabilities are averaged over the whole sample and evaluated at education=16 (High education) and education=12 (Low education).