Author's response to reviews

Title: Are standardised occupational classifications a better estimator of income than self-reported income? Validation of a synthetic weekly wage measure estimated from occupational descriptions for medical research.

Authors:

Tom Clemens (tc245@st-andrews.ac.uk)  
Chris Dibben (chris.dibben@ed.ac.uk)

Version: 2  
Date: 4 March 2014

Author's response to reviews: see over
1. The terms wage and income are used interchangeably throughout the text, but they can have two very different meanings. Wages mean earnings from work performed, while income generally includes both wages and other sources (such as a pension). For the most part, I would replace "income" with either "wage" or "salary" throughout the paper except when self-reported income is used.

   We have amended the text with the word wage where appropriate i.e. where we are talking about wage rather than income but have retained use of the word income in certain places where it is clear that we are talking about income (ie where others are talking about income).

2. It should be made clearer throughout the manuscript that this strategy only applies to survey responses from current workers and their households. One would not be able to apply this strategy to survey respondents that are often the focus of medical researchers (i.e. disabled adults living alone, elderly retired, and the unemployed). While it may seem obvious to the authors that SOC data would only be available for the subset of respondents with a current occupation, it will not be immediately apparent to others. Nor is the text very clear on this until well into the manuscript. I would strongly suggest adding the qualifier "among workers" or "among households with a wage-earner" throughout the text.

   We have added a number of statements in the text section to make this clearer.
Reviewer 2: Lumme

1. (Title) I suggest that the authors revise the title. In the paper, they do not evaluate occupation as an estimator of income compared to self-reported income, but occupation-based synthetic income. The synthetic wage is an estimation of income by occupation groups. Thus, I would recommend changing or defining the word occupation in the title.

   We have amended the title to include the term standardised occupational classifications to clarify.

2. (Abstract: Results: First sentence and Discussion) To state that the synthetic estimates provided independent explanatory power over models containing other measures of SEP requires showing results proving this. The paper does not include univariate models using only social class or small area based measures or models comparing synthetic estimates with these other measures of SEP. In Table 3, social class and SIMD are included in the same model as the wage.

   We have changed the word “over” to the word “within” when talking about the explanatory power of the synthetic estimates which we think clarifies this point.

3. (Abstract: Conclusions and Discussion) The statement that the direct survey measurement of income may not be necessary might oversimplify the need of income data, since anyway some estimation on income is needed (the prediction model i.e. the synthetic measure of wage needs up to date information on income distribution). So this is true assuming that there is available updated general information on income by occupation groups from large survey data or register data. In discussion the authors say that this methodology can be applied to a wide range of studies. However, it should be noted that the reference data is needed anyway (from which the synthetic model is drawn) and it needs to be timely and each country needs to have own data due to varying income distributions between occupations and different occupational hierarchies. This paper did not evaluate the appropriateness of the synthetic wage estimate in time. Income changes as well as distribution between occupations all the time. Thus the model should likely be updated from time to time. One solution would be to link (in countries where such registries available) surveys to register data including income information. Register data has the advantage of large coverage and reliable information. This could allow omitting collecting income in surveys as suggested by the authors.

   This is a valid point though we would assert that we do not advocate in the paper that the collection of income in all surveys could be replaced. Clearly, the technique relies on there existing at least one source of individual level income data (such as labour force surveys like the one we have used here) from which to estimate the prediction model. However, we argue, based on published evidence that often the level of non-response in surveys depends on the subject matter of the survey as well as other factors. So, for example, in health and medical survey data sources it might make more sense to measure occupation rather than income given the evidence that exists showing that income questions result in not just missing cases for the income question but also for the survey as a whole. However, we do agree that this point could be made more clearly in the abstract and have made amendments to this sentence. However, we believe the other sentences in the discussion and conclusions are sufficiently clear that we are talking about health survey data.
4. (Data: Master data) The data should be described in more detailed. Was the information on income derived from registers (i.e. was the survey linked into registers) or was it based on self-reported income? For readers outside the UK, the UK Labour Force Survey (LFS) is not necessarily familiar. Thus, I suggest inserting also information on the study population, age groups, does it include students, part-time workers. etc.

   *We have added a fuller description of the LFS in the master data section.*

5. Results are not easy to interpret without any accompanying text. They should be interpreted easily alone as well.

   *We have added various annotations to the tables to help with interpretation and added more detailed text to the table titles.*

6. (Discussion: first paragraph: third sentence) The use of other indicators of SEP when income is missing cannot be used to approximate income, since (although overlapping) they describe different aspects of SEP. Thus, I recommend changing the last word of the sentence as SEP (instead of income) or modifying the sentence somehow other way.

   *We have amended this sentence to reflect this.*

[Minor Essential Revisions]

1. (Data: Validating data) It would improve the paper if the definition of the health outcome variables (health questions in the two surveys) was described here. What was the question asked? Also the classification of the variables should be stated more clearly somewhere (good health = reference and fair, bad and very bad = other group).

   *We have added more information about the health variables in the SHS and the UKHLS to the “External validation of the wage estimates” section of the text.*

2. (Table 1) I feel that it is unnecessary to show both residual variance and standard deviation.

   *We have removed the residual standard deviation figures from table one.*

3. (Table 1 vrs. Results section: first paragraph) The connection between these is rather loose. The table 1 is quite difficult to interpret without the text, but on the other hand the values are not described in the text.

   *We have added annotation to table 1 which we think has improved the interpretation of it when read in isolation of the text.*
4. (Table 1, 3 and 4) Too many digits shown, especially in table 1.

   We have changed the values to 2 sig figures.

5. (Results: Internal validation of the prediction equations) It would make reading easier if it was described how the values (£5, £356, £65) were drawn.

   We have added a sentence which details the mean wage value and also how the other figures were calculated.

[Discretionary Revisions]

1. (Introduction, third paragraph) It might improve the Introduction if the relationship between income and occupation is expanded. For example the different dimensions of these indicators of SEP could be briefly described and in addition some results on the correlation between these variables from previous studies.

   We have decided to remove the part of the sentence that talks about “the close relationship between occupation and income level”

2. (Data: Validating data) The sample sizes of the SHS and UKHLS would be informative to give already here. Also the comparability between these Surveys and LFS (study population, ages, etc.).

   We have added additional text to the relevant section.

3. (Table 4) P-values for the correlation coefficients?

   We have added these.

4. (Table 4) For the lowest income group (reference group) odds ratio (1) could be inserted.

   We have added this.

[Minor issues not for publication]

1. (Table 3 and 4) Please, mark confidence intervals (CIs) systematically (, or -)

   We have amended this.

2. (Table 3) *(p<.10) is unnecessary, since it is not generally set as a level of significance.

   We argue that some studies use this convention to indicate significance at the 90% level and that it adds a little more information for the reader.