Reviewer's report

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

Version: 1
Date: 3 February 2014

Reviewer: Sonja Lumme

Reviewer's report:

This paper introduces an interesting approach to estimate occupational based income in survey studies using Standard Occupational Classification (SOC) of occupations.

[Major Compulsory Revisions]

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.

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.

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.

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.

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

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.

[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).

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

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.

4. (Table 1, 3 and 4) Too many digits shown, especially in table 1.

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.

[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.

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.).

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

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

[Minor issues not for publication]

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

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

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests.