Reviewer’s report

Title: Adjusting selection bias in German health insurance records for regional prevalence estimation

Version: 0 Date: 18 Sep 2018

Reviewer: Falk Hoffmann

Reviewer’s report:

The manuscript deals with a very relevant problem when conducting research based on health insurance claims data in Germany. The authors call this problem fund-specific selectivity and there has not yet been any approach how to quantify it on a regional basis and how to deal with it. Therefore, the authors provide an innovative approach. However, there are some methodological issues that I do not fully understand (but this might to some extent be due to the fact that I am not a statistician or economist). Anyhow, some methodological points might be described in some more detail and the approach should be described more easily that a broader audience can critically understand what the approach is.

1. The background of the abstract is somewhat too long. Instead, a more comprehensive methods and results section would be desirable in my opinion.

2. The introduction gives a good overview on this problem. On page 4, line 7 the authors should also describe another point why diabetes is a good example to conduct this analysis. There are large regional differences in its prevalence, that have also been shown by the DIAB-CORE Consortium.

3. At the beginning of the methods there should be a section describing the databases used. Some of these information can be found in the introduction and in the results but they have to be the first part of the methods and have to be more comprehensive.

4. Thereafter, please describe the methods first in a few sentences that are easy to understand and then describe details about the statistical framework and the model.

5. There are also several points about the methodological approach that should be more transparently reported, for instance:

   a. Which data were used (e.g. only inpatient data of the AOK)?

   b. How were patients with diabetes selected in both databases? Which diagnoses were used? Were only main diagnoses in the DRG data used?

   c. What was the denominator in the AOK analyses? Were there any selection criteria (e.g. only persons that were insured the whole year)?
d. What was the denominator of the DRG analyses (as there are no information on persons at risk and not whether a person was hospitalized more than once)?

e. What was compared between both analyses?

f. Were the prevalences age- and sex-adjusted?

6. "Stationary prevalence" (page 6) might be the wrong wording. Please use inpatient or hospital instead of stationary.

7. The authors often use the term "prevalence rate". A prevalence is not a rate, as this implies a person-time in the denominator. Please use the term prevalence instead.

8. It would be interesting to present more data on the distribution of the correction factor in the results section. How does this factor vary between regions?

9. There are several comparisons with the literature and references in the results section. This should be shifted to the discussion section.

10. What about Figure 3? This is not described in the results section or in the discussion. Why was this analysis done and what do the authors conclude about the quite different picture compared to figure 2? In my opinion, one possible explanation might be that type 1 diabetes is more often wrongly coded as type 2. For type 1, there might be not that much regional variance because the socioeconomic influence is smaller than for type 2 diabetes.

11. The last paragraph on page 8 describes a comparison with data from the Zi, which is not mentioned anywhere else. Was another data source used (then introduce it in the methods section)? Or were results only compared to other findings (then put this in the discussion section)?

12. The discussion is quite brief and should be enlarged.

a. For which further diseases or health services used are those methods applicable?

b. Which other data sources instead of the DRG data might be used (for instance, when other outcomes are studied)?

c. The small case numbers in some strata might also be a further limitation. How large (or small) were cell sizes?

d. What are further limitations of the methods proposed? One limitation might be that several diseases itself show regional variations in hospitalizations although no regional differences in its prevalence exist. An example might be knee replacement, for which the diagnosis of arthrosis is coded.
13. What is the conclusion for research using claims data of health insurance claims data?

**Level of interest**
Please indicate how interesting you found the manuscript:

An article of importance in its field

**Quality of written English**
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published

**Declaration of competing interests**
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal.