Reviewer’s report

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

Version: 0 Date: 24 Sep 2018

Reviewer: Spencer James

Reviewer’s report:

In general, this article describes an interesting solution for an important problem in population health research, which is using medical record data for population health measurement while addressing selection biases in the composition of the medical records data. The application of this method to estimate type 2 diabetes mellitus (T2DM) is somewhat secondary to the description of the methods.

The premise of the study, to my understanding, is essentially as follows. A given health insurance fund covers only a select population within or across geographical regions of Germany. The fund records have accurate medical diagnostic information, but do not capture a representative population within the region for purposes of population health estimation. The research team cites one solution where prevalence estimates for a disease are estimated within an insured group and then adjusted to match demographic distributions of the actual population of interest, but indicate how this leads to biased estimates due to selection of the insured group. Therefore, the research team identified a dataset (DRG-statistics) that was representative of the total population and contained detailed medical information, most pertinently primary and secondary diagnoses for clinic and hospital encounters. They then develop a statistical model that measures the relationship between different predictors in the clinical records and the prevalence of the disease of interest, and they then apply the coefficients from this model to the same predictors in the insured group to estimate an adjusted prevalence of the disease of interest. They apply Monte Carlo simulations on this process to estimate uncertainty. Their model form is a linear mixed effects model which is fit for each region. They apply this framework using the AOK insured group as the test case while DRG-Statistics 2014 are used to derive the model coefficients that then adjust the AOK prevalence estimates for T2DM. They then calculate a prevalence rate in the population after adjusting the AOK-specific prevalence rate based on the model fit, and validate the adjusted prevalence values with administrative data.

Overall, the premise of the method is logical, and their solution could be useful for other epidemiological tasks where representative data are not available. As a reviewer and as a reader, I would find additional information on their process useful, however, and therefore request the following revisions.

Major revisions
1. While the statistical framework is clear, the actual data inputs and outputs are not obvious. The actual predictors/covariates that were used for the T2DM model should be included as a table in the main text. My understanding is that the fixed effects are diagnoses, age, and sex, but it's not clear how these are actually used - specifically, are the diagnosis variables coded as prevalence for that cell for every other comorbid condition? More details on the actual covariates used for the T2DM model would help improve the clarity.

2. I don't quite understand why they report the overall prevalence estimate to highlight their model validity when the premise for the study was to improve regional estimation. It would better validate the method to compare the adjusted prevalence estimate for a given region with a separate process/study that was used in a given region. I understand that the source used for validation is nationally-representative, but it seems likely that there have been regional assessments of T2DM that are regionally-representative that could be used to support the model estimates.

3. Compared with many other diseases, T2DM is much more predictable in general using comorbidities and other medical history. Therefore, the coefficients between T2DM prevalence and the prevalence of an array of other conditions are likely to be more reliably predictive than they would be for other conditions of interest that have heterogeneous patterns. The authors cite checking myocardial infarction as additional validation, and it's not surprising that it also performed accurately. That the study uses T2DM as a test case doesn't really concern me, as it is still an interesting condition to study that is globally prevalent, but I think this should be better highlighted as a limitation of the method. I would recommend that this is either more emphasized in the Introduction and Discussion and Conclusion, or that the authors evaluate the model performance for a condition which is not necessarily highly prevalent in older ages (e.g. dementia, which they also cite) and which is not typically associated with a select group of diagnoses. For example, validating with a condition such as asthma would be interesting, as asthma has fewer reliable comorbid conditions that would be coded in claims data.

Minor revisions

1. Suggest changing term "gender" to "sex" throughout text.

2. Instead of using the term "total population", use a more specific descriptor such as "national population" or "regional population".

3. Some methodological commentary in methods section - e.g. "trade-off between local heterogeneity and model parameter variance..." and "AOK as initial data source is interesting..." sound more like discussion items than methods items.

4. In the abstract and in the introduction, it would be useful to briefly describe the health insurance structure in Germany to familiarize the reader with the context of the problem.

5. In general, some of the terminology used in the manuscript is a bit difficult to follow, at least to this reviewer. Specifically, the terms "benchmark" and "reference" are used frequently and
generally imply a validation dataset or outcome, but it is a bit difficult to keep track of what exactly is the data requiring validation versus being used for validation. There are some other unique terms that are used, too, that I haven't encountered elsewhere in epidemiological literature. For example, I understood what "stationary prevalence" meant, but I don't think it's a common epidemiological term. Similarly, I think it'd be more clear to use the term "national HI" instead of "statutory HI" as "statutory" seems to be a term specific to the German health insurance system.

**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:

Not suitable for publication unless extensively edited

**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/](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