Author’s response to reviews

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

Authors:

Jan Burgard (burgardj@uni-trier.de)
Joscha Krause (s4jokrau@uni-trier.de)
Ralf Thomas Münnich (muennich@uni-trier.de)

Version: 1 Date: 23 Oct 2018

Author’s response to reviews:

Adjusting selection bias in German health insurance records for regional prevalence estimation
Revision 1: Answers to the reviewers
Jan Pablo Burgard, Joscha Krause, Ralf Münnich

23/10/2018

Dear Editor,

We greatly appreciate the very constructive comments made by you and the reviewers. Our answers are given directly to each comment.

Concerning the language: we could not give this paper to a language correction while avoiding a delay in resubmission due to the relatively short revision time. The language service of our institution could not provide us with a corrected version in time. However, in a next round, we certainly would include this step.

Best regards

Jan Pablo Burgard, Joscha Krause, and Ralf Münnich

COMMENT_REV1_1: The background of the abstract is somewhat too long. Instead, a more comprehensive methods and results section would be desirable in my opinion.
ANSWER_REV1_1: We agree with your assessment. The background section of the abstract was shortened and the methods section of the abstract was extended by further explanations on how the fund-specific selectivity is adjusted within the model.

COMMENT_REV1_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.

ANSWER_REV1_2: Thank you very much for this suggestion. We included that aspect and the corresponding reference in the last paragraph of the background section.

COMMENT_REV1_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.

ANSWER_REV1_3: We agree that including additional information on the data bases used makes the methodological aspects more transparent. Therefore, we followed your advice and provided a distinct data subsection prior to the statistical framework in the methods section. As a consequence, some details on the data bases were deleted in the results section to avoid repetitions.

COMMENT_REV1_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.

ANSWER_REV1_4: We added an additional paragraph (Data usage) at the end of the data subsection as you suggested. It briefly describes how the data bases are used jointly within our approach to adjust fund-specific selectivity. We hope that this marks an intuitive summary of the methodology and allows for a better understanding of the subsequent statistical framework.

COMMENT_REV1_5: There are also several points about the methodological approach that should be more transparently reported […].

ANSWER_REV1_5: Thank you very much for these important questions. Apparently, we didn’t state these points with sufficient transparency in the last version. Hereafter, we answer your questions one after another.

COMMENT_REV1_5a: Which data were used?
ANSWER_REV1_5a: We used AOK member records, inpatient diagnosis frequencies, and population statistics. More details on data usage are now provided in the data subsection you suggested. Hopefully, this ensures more transparency.

COMMENT_REV1_5b: How were patients with diabetes selected in both databases? Which diagnoses were used? Were only main diagnoses in the DRG data used?

ANSWER_REV1_5b: The methodology only requires the identification of patients with diabetes within AOK member records. For this, an intersectoral disease profile was used. It is now included in appendix A for more transparency. However, note that the term ‘identification’ only refers to obtaining the number of AOK members with diabetes per aggregate. No actual identification of individuals is necessary. Further, our methodology does not require to identify any patients within the DRG Statistics at all. Using the aggregated records on inpatient diagnosis frequencies is sufficient when assuming equality in conditional expectations (equation 2 in the methods section). For the inpatient diagnosis frequencies, both main and secondary diagnoses were considered. We made these points more transparent by elaborating more on data usage in the data subsection.

COMMENT_REV1_5c: What was the denominator in the AOK analyses? Were there any selection criteria (e.g. only persons that were insured the whole year)?

ANSWER_REV1_5c: AOK members were defined as every citizen that has been enrolled to the company for at least one day in the report year. A short corresponding explanation is now included in the data subsection.

COMMENT_REV1_5d: 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)?

ANSWER_REV1_5d: You are right. Due to the lack of information on persons at risk, it is not possible to know how many people were hospitalized or whether a person was hospitalized more than once. However, within the proposed methodology, such knowledge is not required. It is sufficient to have aggregated records of inpatient diagnosis frequencies as we only need auxiliary information on the aggregation level we seek prevalence estimates for. Regarding the required demographic information on the national population (how many citizens are associated with a given cell), we retrieved corresponding figures from the population statistics of the German Federal Statistical Office. These points are now included in the data subsection you suggested.

COMMENT_REV1_5e: What was compared between both analyses?

ANSWER_REV1_5e: We retrieved aggregated inpatient diagnosis frequencies of diagnoses closely related to diabetes. The identification of these diagnoses was performed by means of
variable selection in the health insurance records of the AOK member population. These aspects are now stated in the data subsection.

COMMENT_REV1_5f: Were the prevalences age- and sex-adjusted?

ANSWER_REV1_5f: We presented crude prevalence figures only. However, as our prevalence estimates are on cell level (cell = cross combination of district, age and sex), they allow for an age- and sex-specific depiction of the regional prevalence. This point is now stated in the results section more clearly.

COMMENT_REV1_6: "Stationary prevalence" (page 6) might be the wrong wording. Please use inpatient or hospital instead of stationary.

ANSWER_REV1_6: Thank you very much for this suggestion. The expression “stationary prevalence” was substituted by “inpatient diagnosis frequency” where appropriate.

COMMENT_REV1_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.

ANSWER_REV1_7: Thank you very much for this correction. The term “prevalence rate” was substituted by “prevalence”.

COMMENT_REV1_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?

ANSWER_REV1_8: We agree that analysing the regional distribution of the correction factor would be an interesting addition to the paper. Unfortunately, by confidentiality restrictions, we are not allowed to provide such information. The AOK is a union of multiple regional health insurance companies whose individual morbidity structures must not be published. Therefore, we cannot display any geographical insights on the prevalence adjustment. The focus of the paper thus must be on the regional diabetes prevalence of the national population and not of the AOK member population. We kindly ask for your understanding. However, we are allowed to display the age-specific correction factor distribution. This provides more insights on the adjustment behaviour of the model compared to the old version of the paper. We now included a corresponding plot in the results section (Figure 2).

COMMENT_REV1_9: There are several comparisons with the literature and references in the results section. This should be shifted to the discussion section.

ANSWER_REV1_9: We agree with your assessment that comparisons to the literature should not be included directly next to the prevalence estimates. However, we think that the discussion
section is primarily suitable for elaborating on methodological aspects, such as advantages and limitations of the approach. Therefore, we provided a new subsection (Validation) to the results section and relocated the comparisons there. We hope you find this appropriate.

COMMENT_REV1_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.

ANSWER_REV1_10: Thank you very much for this observation. By unknown reasons, this paragraph was finally lost in the last version of the paper. It is now restated in the results section. Your thought regarding the wrong coding of diabetes type 1 seems plausible and marks an interesting aspect for future research. However, note that Figure 3 (now Figure 4) doesn’t correspond to diabetes type 1, but to diabetes type 2 only. Our argumentation is that these prevalence differences between regions regarding diabetes type 2 can be partially explained by regional demography. Figure 3 (now Figure 4) shows the prevalence for citizens that are 34 or younger. Compared to Figure 2 (now Figure 3), which depicts the unadjusted total prevalence of all age groups, Figure 3 (now Figure 4) shows much less regional patterns. We hope that this aspect now becomes more transparent as we included the missing paragraph.

COMMENT_REV1_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)?

ANSWER_REV1_11: Thanks for the comment. Indeed, we didn’t state this point sufficiently in the last version of the paper. Several data sources were used for validation. We compared both overall prevalence and regional patterns using the following sources:


We now also considered the results of the DIAB-CORE Consortium you suggested. In terms of the location within the main text, we provided an additional subsection Validation within the results section.

COMMENT_REV1_12: The discussion is quite brief and should be enlarged.

ANSWER_REV1_12: Thank you very much for this suggestion. We now elaborated on your points within the discussion section of the paper. Subsequently, we comment on them directly.

COMMENT_REV1_12a: For which further diseases or health services used are those methods applicable

ANSWER_REV1_12a: We currently have a joint research project together with the AOK where the methodology is applied to many different diseases. First results show that the method works very well for common diseases on the one hand, and diseases with strong predictors within the DRG-Statistics on the other hand. The latter are usually diseases with comorbidities that are recorded in hospitals. However, rare diseases without closely related inpatient diagnoses, such as multiple sclerosis, cannot be estimated with the DRG-Statistics as auxiliary data source. For corresponding diseases, other data sets for prevalence adjustment need to be explored. We now included these points in the discussion.

COMMENT_REV1_12b: Which other data sources instead of the DRG data might be used (for instance, when other outcomes are studied)?

ANSWER_REV1_12b: We primarily used the DRG-Statistics due to its high geographic and demographic detail, as well as the records corresponding to both statutory and private HI members. However, using other data sources for prevalence adjustment is generally possible. It basically comes down to the question for which population prevalence estimates are required for, and on which level of aggregation prevalence adjustment is suitable. With the DRG-Statistics, we are not only able to provide prevalence estimates on the cell level, but also account for systematic differences between the member population and the national population on the cell level. We now added some of these aspects to the discussion section.
COMMENT_REV1_12c: The small case numbers in some strata might also be a further limitation. How large (or small) were cell sizes?

ANSWER_REV1_12c: You are right, this could be a major issue in a classical estimation framework. If the number of AOK members within a cell is small relative to the overall cell size, then a classical estimation would be subject to unacceptably high variances and, then, corresponding prevalence estimates suffer from high uncertainty. However, in our approach, we use small area estimation techniques (SAE). In order to produce prevalence estimates for a given cell, we not only use information from the corresponding cell, but from other cells as well. This is usually referred to as borrowing strength within the SAE literature and reduces the classical sampling error considerably. The basic idea is to exploit the functional relation between the cell prevalence and auxiliary information by calculating the corresponding conditional expectation over all cells in a region. This conditional expectation is then extrapolated in order to obtain prevalence estimates. Accordingly, the size of the cell or the number of AOK members within a cell is much less relevant as long as the auxiliary information has sufficient explanatory power for the cell prevalence. Nevertheless, a cell size-related issue remains. If the number of AOK members in a region is so small that the corresponding AOK member prevalence is zero for a large fraction of the cells within a region, then zero-inflation may play a role. In that case, the probability distribution of the AOK member prevalence has a considerable mass point on zero, which can lead to severe bias within linear models. In such a situation, different model types have to be considered. We now included these points in the discussion section.

COMMENT_REV1_12d: 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.

ANSWER_REV1_12d: Yes, this is a potential counter example. It basically comes down to which inpatient diagnoses are selected for prevalence adjustment and whether they include the true regional patterns of the disease of interest. We made variable selection via correlation analysis. Hence, we chose diagnoses for prevalence adjustment only based on statistical arguments. However, this does not replace an informed epidemiologic researcher who is aware of potential issues like the one you mentioned. We provided a corresponding comment in the discussion section.

COMMENT_REV1_13: What is the conclusion for research using claims data of health insurance claims data?

ANSWER_REV1_13: The presented approach was developed to allow for regional prevalence estimation from health insurance records despite fund-specific selectivity. Accordingly, we see this as the main contribution of the paper. Nevertheless, it would be interesting to investigate how essential ideas of the approach could be applied to other medical research as well. An example could be to use routine data for sophisticated benchmarking within propensity score
matching. Although being subject to future research, we provided a short comment on this in the discussion section.

COMMENT_REV2 MAJOR1: 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.

ANSWER_REV2 MAJOR1: We agree with your assessment that providing more information on data usage improves the clarity of the model. Therefore, we provided an additional data subsection in the methods section. It gives more insights on the data sets and how they were used for regional prevalence estimation. Further, we provided a list of selected inpatient diagnoses used for prevalence adjustment in the results section, as you suggested. We hope that this allows for more transparency.

COMMENT_REV2 MAJOR2: 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.

ANSWER_REV2 MAJOR2: The purpose of our paper is to improve regional prevalence estimation. We understand that given this purpose it would be better to validate our methodology on regional levels, and not nationwide, if accurate data is available. However, this is problematic for reasons that motivated the study in the first place: there is no reliable information on the regional diabetes prevalence, at least not in terms of actual prevalence figures. There are a few national health surveys with 10,000 to 20,000 participants. These surveys include only very few observations per region, which ultimately leads to very high standard errors of corresponding regional prevalence estimates. There are even some regional health surveys, but the majority of them is outdated and only includes a few hundred observations. In fact, some of the regional reference figures actually contradict each other. Thus, corresponding estimates suffer from very high uncertainty as well. As this marks the fundamental problem for which small area estimation methods were developed, we think that a numerical validation of our estimates with these results is misleading since the comparison would be against inaccurate figures. We validated the prevalence estimates on a level for which reliable figures are available, which is unfortunately only nationwide. In small area estimation, it is important that (accurate) aggregates on higher levels using classical sampling-based inference methods are met. Nevertheless, in order to follow your suggestion and provide at least some comparison on regional levels, we made an additional subsection (Validation) in the results section. There, we located all our comparisons to past studies, including a now conducted comparison to survey-based regional confidence intervals.
obtained from a national health survey. However, we explicitly pointed out the limited validity of this regional comparison due to the high uncertainty of the reference figures.

COMMENT_REV2_MAJOR3: 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.

ANSWER_REV2_MAJOR3: You are right in your assessment that the approach works especially well for diagnoses that have typical comorbidities which can be retrieved from medical routine data. Nevertheless, it does also work for a variety of other diseases. We currently have a joint research project with the AOK in which the methodology is applied to over 30 different diseases (e.g. dementia, coxarthrosis, cancer, dorsalgia). Our research showed that the approach works well for diseases that are either common, or have strong predictors within the DRG-Statistics. This is of course related to your point regarding the comorbidities. Rare diseases without strong predictors, such as multiple sclerosis, cannot be estimated in the presented manner. We elaborated more on the limitations of the approach in the discussion section, as you suggested. Generally, note that due to disclosure restrictions we cannot display detailed results for multiple diseases. We primarily chose diabetes because it is known to be asymmetrically distributed between HI member populations and regions, as well as being an important cost factor in health care systems all around the globe.

COMMENT_REV2_MINOR1: Suggest changing term "gender" to "sex" throughout text.

ANSWER_REV2_MINOR1: Thank you very much. We changed the terms as you suggested.

COMMENT_REV2_MINOR2: Instead of using the term "total population", use a more specific descriptor such as "national population" or "regional population".

ANSWER_REV2_MINOR2: We agree that this wording improves the paper in terms of transparency. The term “total population” was substituted by a more accurate expressions whereever suitable.
COMMENT_REV2_MINOR3: 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.

ANSWER_REV2_MINOR3: We agree with your observation. We reformulated corresponding phrases.

COMMENT_REV2_MINOR4: 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.

ANSWER_REV2_MINOR4: We agree that providing more information on the German health insurance system at the beginning of the paper makes the study more accessible to the reader. However, reviewer #1 pointed out that the background section of the abstract is too long and needs to be shortened. Therefore, in order to find an adequate compromise between the reviewers, we provided a comprehensive overview of the German health insurance market in the introduction (page 3-4), but left it out in the abstract.

COMMENT_REV2_MINOR5: 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.

ANSWER_REV2_MINOR5: Thank you very much for your comments. In the new version of the paper, we elaborated more on essential aspects like benchmarking and the data used for validation for more clarity. We further replaced phrases like “stationary prevalence” to “inpatient diagnosis frequency”. Hopefully, these provide a better understanding. However, we would like to keep the official expression “statutory HI” as the presented methodology is applied to regional prevalence estimation in Germany and thus used to account for fund-specific selectivity in on the German HI market. Since we enlarged the description of the German system according to your above suggestions, we hope it reads more appropriate now.