Reviewer’s report

Title: The impact of the lookback period and definition of confirmatory events on the identification of incident cancer cases in administrative data

Version: 0 Date: 13 May 2017

Reviewer: David Goldsbury

Reviewer's report:

Thank you for the opportunity to review this article. This is a useful paper about methods for identifying incident cancer cases in administrative data that will support further research using these data. It is a good additional use of information that is already routinely collected and the authors have tested many scenarios for the best use of these data. The paper is an important resource for highlighting that caution is required in using these data to estimate cancer incidence. The manuscript is of good quality and is clear and well written. There are some minor corrections and/or suggestions but overall the authors are to be congratulated for the quality of the paper.

What are the results for an algorithm without a confirmation record? It is very plausible that someone might only have one record and therefore not have the chance to have a confirmation record. Given the difference in results for confirmation periods of 1 quarter and 4 quarters, this would form another useful component of sensitivity analyses.

Indicate earlier in the paper that it is a form of ecological study, as there is no actual linkage between cancer registry and the other administrative data for validation. This wasn't clear until towards the end of the paper (line 284).

Indicate in the abstract that the data relate to Germany. e.g. line 32 "using German cancer registry data as a benchmark".

Please explain the meaning of "status post" (lines 102, 121).

The relevance of the exact outpatient diagnosis dates mentioned in line 105. It wasn't clear if these were actually used. What were the proportions of diagnosis dates assigned using exact dates and quarterly dates? This could impact upon the temporality of estimates.

In the limitations (lines 293-295) the authors seem to be suggesting that the benchmark cancer registry data are not valid. I suggest re-phrasing this to say the cancer registry data are not perfect but still provide a valid benchmark, if that is the case.

The keywords (line 49) include "pancreatic neoplasms", this should be "prostate neoplasms".

Grammatical issue line 225 "twice as much the number of".
Grammatical issue line 233 "or is treated non-operative palliative or curative (colorectal)."

If looking to tighten up the tables, the denominators in Tables 1, A1 and A2 could be listed in a footnote rather than taking up 3 separate columns or repeated values.

The word "respective" is used in several places where it isn't needed (e.g. lines 124, 212). This might just be a stylistic issue and if everyone else is OK with these then I am comfortable with it.

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

Yes

**Does the work include the necessary controls?**
If not, please specify which controls are required in your comments to the authors.

Yes

**Are the conclusions drawn adequately supported by the data shown?**
If not, please explain in your comments to the authors.

Yes

**Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?**
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

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

Acceptable

**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.