Reviewer’s report

Title: A comparison of estimators from self-controlled case series, case-crossover design, and sequence symmetry analysis for pharmacoepidemiological studies

Version: 0 Date: 08 Aug 2017

Reviewer: Aksel Jensen

Reviewer’s report:

Review: A comparison of estimators from self-controlled case series, case-crossover design, and sequence symmetry analysis for pharmacoepidemiological studies.

The current manuscript provides a brief overview and detailed simulation studies investigating three case-only designs and their robustness to violation of some of the fundamental assumptions underlying the three designs. Time-invariant and time-varying confounders for the association between an exposure and a recurrent outcome (all binary) are studied in the simulations. In addition, misspecification of the "risk period" duration and trends in the underlying exposure and event risks are also investigated in the simulations. Finally, the authors apply the three designs on real data based on electronic medical records.

As stated in the title, the manuscript compares three designs. Under the scenarios investigated the sequence symmetry analysis seems to perform somewhat better compared with the other two designs. Further, the analysis based on real data indicates that result can differ substantially depending on the employed case-only design.

Overall, I find the manuscript interesting, well written and well structured. It focuses on a highly relevant topic: identifying and assessing the importance of underlying assumptions and choices made when conducting a case-only analysis. Such considerations seem to be somewhat overlooked when case-only designs are applied in practice.

I have the following more specific comments to the authors:

Major comments:

Several underlying assumptions for the three case-only designs are investigated in the simulations and discussed in the text. However, the key assumption of independence between an event and future exposure is barely mentioned in the text (only once I think, in connection with the introduction of the case series methods (l. 112-113)). The importance of this assumption is stressed several times by Whitaker et al. in their 2005 tutorial on the case series method, e.g., "The most important, and possibly restrictive, assumption of the case series method is that occurrence of an event does not alter the probability of subsequent exposure." (Whitaker et al., Statist. Med. 2005; 0:1-31). The current manuscript and the concrete simulations consider a setting where this assumption is fulfilled. However in practice this is a very restrictive
assumption. Further, only two of the three case-only designs studied in the simulations rely on this assumption, namely the self-controlled case series and the sequence symmetry analysis whereas the case-crossover design does not rely on this assumption. Any comparison of the three designs should be very explicit about this assumption, especially if the comparison leads to more general recommendations favoring one specific design such as the sequence symmetry analysis in the current manuscript. Moreover my intuition is that the different sources of bias often tend to amplify each other when working with case-only methods. If this holds true even minor deviations from the assumption of independence between an event and future exposure could alter the magnitude of the bias caused by e.g., time-varying confounders when using the sequence symmetry analysis.

Concluding, I think the authors should refrain from general recommendations unless they are much more explicit about the restrictive setting that the recommendations are based upon i.e. a recurrent event setting where occurrence of an event does not alter the probability of a future event or future exposure.

Minor comments:

Throughout the text, more references linking to the relevant tables could improve the readability.

(l. 73-74) Strictly speaking CCO does not estimate a relative risk but an odds ratio as the authors also state explicitly later, so why write that all three methods estimate relative risks?

(l. 105) Shouldn't the baseline incidence rate be exp(Φ) instead of Φ?

(l. 153-154) I would prefer the wording "fixed" instead of "stable" and in addition explicitly write that different durations are investigated (or even write the specific durations i.e. 5-day up till 30-day). Further, what if a new exposure is initiated before the stable period has ended? I assume the stable period is just extended? Or is the new "exposure-initiation" ignored?

(l. 156 and l. 164) The equations defining the rate parameters in connection with X(t) and Y(t) respectively are key and should be presented in separate lines with equation numbers.

(l.172-180) I think it will improve the readability if the authors for each scenario presented explicitly mention the relevant terms from the rate parameter equations for X(t) and Y(t). More generally, now that the parameters have been introduced, why not use them more throughout the text to be more specific about assumptions and scenarios? See comments relating to line 230-232 below.

(l. 198+ l.207) It will improve the readability if sentences start with words instead of parameters.

(l.216-220) I'm a little confused here. Do the authors considered calculation of NSR a natural component in the sequence symmetry analysis design? Or is it an extension of the design aimed at dealing with time trends similar to the case-time-control design extension of the case-crossover design? When one of the aims of the simulations is to investigate robustness towards
time trends in exposure and events it's hardly surprising that the symmetry analysis performs well with respect to time trends given that the NSR component should address exactly such time trends. However, the NSR component/approach relies on additional assumptions I guess? One would be that the exposure time trends are similar among the (future) cases and the non-cases i.e. the way a drug is prescribed and/or used among cases and non-cases are similar. Otherwise the correction could potentially bias the estimate even more. In conclusion, even though I find the underlying assumptions for the symmetry analysis somewhat unclear, I suspect the construction of the simulations "fit" these assumptions very well and that should be recognized and more explicitly stated in the text.

(l. 230-232) Here the simulation scenario is nicely and very reader friendly presented with specific parameter values relating to the assumptions and specific durations explored.

(l- 234-258) I think it's highly relevant to include a case study based on real data, especially to illustrate the differences between designs and within designs depending on choices such as period length. But is the earlier discussed assumption of independence between an event and future exposure really fulfilled in the case study?

(l.272) "…15-, 30-, or 45-day period." Before or after first prescription right?

(l. 273) Does "each dataset" refer to the different datasets created by the different definitions of risk period length?

(l.273-275) "For the calculation of NSR, we used information from the patients who experienced either an exposure or event in each observation period." I find this unclear. Do the authors mean information from each period where the patient experiences either an event or an exposure, possible both?

(l. 303-305) "when there were either exposure or event time trend, there were no substantial biases observed for all methods." I find this somewhat surprising, but is that not because the outcome is recurrent and not affected by previous outcome?

(l.309-315) I find this part of the simulations and discussion especially interesting because misspecification of the risk period duration is a consequence of a researcher decision in opposition to e.g., lacking knowledge of unmeasured time-varying confounders.

(l.334-338) "The extent of differences in log-scale point estimates between the three methods ranged from 0.52 to 0.68 (1.48 to 1.90 on ratio scale)." I would prefer that the authors just reports the actual RR and OR estimates in the text. It could e.g., be done as follows: "For all settings of risk period durations (15,30 and 45 days), the highest point estimate of the effect measure was obtained from CCO design (xx,xx,xx), followed by SCCS (xx,xx,xx) and SSA (xx,xx,xx)."

(l.352-354) "If large SEs in SSA estimates are not a major concern, this method may be a suitable choice for applications in pharmacoepidemiological studies" I don't think the simulations support such a general formulation and I don't think focus should be on finding "the best" case-only design. Each design has its advantages but the key point is how well the
underlying assumptions for the designs are fulfilled with right to the specific data. That should guide the choice of design.

(1.400) For readability, instead of 0.0000002, I would write 2*10^{-7} or 2 per xx person days at risk or something similar.

(1.404-407) "Although previous studies showed that exposure time trends introduced bias into CCO design estimates [15, 16], the magnitude of bias would not be considerable when the effect of time trend is moderate (as set in this study)." Aren't this because the outcome in the simulations is not terminal?

(1.408-417) Very nice summary/explanation

(1 417-419) Which scenarios are referred to here?

Table 2+3+4: I would prefer to also have the estimated \(\exp(\beta X)\) values included in the tables to supplement the Bias (log scale) values.

Table 7: As in Table 2-4, I would prefer to have the RR and OR with 95% CI included in the table instead of less interpretable log scale values.

Figure 1: Very nice illustration.

**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.

No

**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/). 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