PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	A Cohort Profile of the UK Biobank; diagnosis and characteristics of cerebrovascular disease
AUTHORS	Hewitt, Jonathan; Walters, Matthew; Padmanabhan, Sandosh; Dawson, Jesse

VERSION 1 - REVIEW

REVIEWER	Steffen E Petersen Queen Mary University of London, United Kingdom
REVIEW RETURNED	21-Jul-2015

GENERAL COMMENTS	Diagnosis of cerebrovascular disease within the UK Biobank; characteristics of 502650 participants The authors characterise the cohort of UK Biobank participants with stroke and TIA. Major points 1. The overall impression I had reading this manuscript was that authors compared self-reported disease with more accurate nurse-led clinical history taking. This is clearly not the major point of the manuscript. 2. It is difficult to find a clear message in the paper. 3. Language and style will require further work 4. Statistical methodology is not described in enough detail. 5. The study is very limited by its cross-sectional study design. Biases are likely to have contributed to unusual findings (e.g. Associations between alcohol consumption, physical activity, oral contraceptive pill) but will also affect the associations seen with an expected direction of effect. 6. Description of pharmacotherapy will require a lot more detail than described in the manuscript.

REVIEWER	Rustam Al-Shahi Salman University of Edinburgh, UK
REVIEW RETURNED	14-Oct-2015

GENERAL COMMENTS	Thanks for a nice paper. I didn't see a STROBE checklist, so cannot comment on adherence to reporting guidelines. I just have a few minor suggestions: 1. Abstract, methods: Document that you compared characteristics of people with stroke or TIA to those without a past history of stroke or TIA (if I understood correctly). 2. Abstract, results: “The nurse led interview identified 7669 (1.53%)

	people…" – do the authors mean "with stroke"? 3. Abstract, page 2: could the authors clarify the statement, "Women were less likely to have taken the OCP". 4. The authors describe hypotheses on page 4, but I am not sure that they tested them in their study. Perhaps best to omit these hypotheses? 5. Page 5, typo: "…for study to study…" 6. Page 6, typo: "…we could accurately be reproduced…" 7. Methods. The authors helpfully describe the participants' touch screen response options. Could they also provide any relevant details from structured rating forms used by the nurses in their interviews, including definitions (if there were any)? 8. Page 6, line 41, "stroke or TIA" rather than "stroke and TIA" 9. The methods indicate that UK Biobank collected data on 500,000 people, but the results mention 502,650, so correct the former? 10. Could the authors be consistent in their terminology "stroke or TIA", "cerebrovascular disease" and "stroke disease" – all subtly different things! I think they mean the first. 11. The authors state that they compared patients who reported stroke or TIA to those who did not, though the Tables provide summary data on the entire Biobank cohort, not the remainder of the cohort without stroke or TIA. Could this be clarified somewhere (it may be too cumbersome to describe the three non-stroke/TIA separate comparison groups in the Tables, so perhaps just explain this in the methods)? 12. Missing data is an important limitation to note. Given that ~25% of data are missing in some touchscreen questions (Table 1), is it appropriate to quote the frequency in the entire group, or is it better to quote among respondents only? Or does missing imply that the characteristic is absent given the way touchscreen questions were completed (in which case say so in the methods)? Perhaps the authors could clarify. 13. Another caveat about the apparently high frequency of secondary prevention medication use is that not only was the Biobank cohort young, but also they may reflect the 'healthy [or perhaps, adherent] volunteer' effect. Worthy of a comment, perhaps? 14. Table 1 – the numbers in each category do not appear to add up to 502,650. Perhaps the authors could check their sums. 15. N=6699 in Table 1, but n=6669 in Table 2. 16. In the Methods, perhaps indicate that each of the three groups was individually compared with the non-stroke/TIA Biobank population?

VERSION 1 – AUTHOR RESPONSE

Reviewer 1. Professor Peterson has a number of major points. In light of the Editors suggestion that this article be amended to a cohort profile we hope that the majority of his concerns will be ameliorated. Can the editor please ensure that the new title is appropriate for this article in the context of a cohort profile.

We have redrafted the manuscript to be in keeping with the cohort profile. By doing so we have been able to address many of his specific concerns.

Please find below a point by point response

1. The overall impression I had reading this manuscript was that authors compared self-reported disease with more accurate nurse-led clinical history taking. This is clearly not the major point of the manuscript
   Response: We hope that "the point" of the manuscript is now self evident as a cohort profile. We have changed the structure to fit with the journals guidelines, including adding Cohort Profile to the Title. The abstract has also been restructured.

2. It is difficult to find a clear message in the paper.
   Response: We hope that this is now corrected in light of our response to point 1 above. We have added text to the introduction to highlight our message.

3. Language and style will require further work.
   Response: We have tried to improve this throughout. Professor Al-Shahi Salmon (reviewer 2) has made some particular comments regarding this which we have included in the updated text

4. Statistical methodology is not described in enough detail.
   Response: On reviewing the guidelines for a cohort profile, there is specific guidance that detailed statistical methods are to be avoided. That said, we have included additional information in the data collection and follow up section (paragraph 2) which has replaced the methods.

5. The study is very limited by its cross-sectional study design. Biases are likely to have contributed to unusual findings (e.g. Associations between alcohol consumption, physical activity, oral contraceptive pill) but will also affect the associations seen with an expected direction of effect.
   Response: We have inserted a paragraph highlighting the cross sectional bias seen with these three outcome measures.

6. Description of pharmacotherapy will require a lot more detail than described in the manuscript.
   Response: We agree that our pharmacotherapy data is simplistic. However, it is extremely challenging to present these data in a cohort profile type scenario without becoming entangled in huge detail beyond the scope of this cohort profile. We have highlighted this deficiency in our paragraph beginning "What we have not done…".

Reviewer 2. Rustam Al-Shahi Salman. We have provided a point by point response to Professor Al-Shahi Salman's' comments

1. Abstract, methods: Document that you compared characteristics of people with stroke or TIA to those without a past history of stroke or TIA (if I understood correctly).
   Response: We have added this
2. Abstract, results: "The nurse led interview identified 7669 (1.53%) people…" – do the authors mean "with stroke"?
   Response: We have added this
3. Abstract, page 2: could the authors clarify the statement, "Women were less likely to have taken the OCP".
   Response: We have clarified the statement
4. The authors describe hypotheses on page 4, but I am not sure that they tested them in their study. Perhaps best to omit these hypotheses?
   Response: We have deleted these hypotheses
5. Page 5, typo: "…for study to study…"
   Response: We have corrected this error.
6. Page 6, typo: "…we could accurately be reproduced…"
   Response: We have corrected this error.
7. Methods. The authors helpfully describe the participants' touch screen response options. Could they also provide any relevant details from structured rating forms used by the nurses in their interviews, including definitions (if there were any)?
   Response: We have added more information and referenced a link to the UK Biobank website. The information available was fairly general, however, it complements the

additional information we have added in relation to changing the focus of the article to a cohort profile.

8. Page 6, line 41, “stroke or TIA” rather than “stroke and TIA”
Response: We have corrected this error.
9. The methods indicate that UK Biobank collected data on 500,000 people, but the results mention 502,650, so correct the former?
Response: We have changed the former
10. Could the authors be consistent in their terminology “stroke or TIA”, “cerebrovascular disease” and “stroke disease” – all subtly different things! I think they mean the first.
Responses: We have changed stroke disease to cerebrovascular disease where appropriate in the manuscript
11. The authors state that they compared patients who reported stroke or TIA to those who did not, though the Tables provide summary data on the entire Biobank cohort, not the remainder of the cohort without stroke or TIA. Could this be clarified somewhere (it may be too cumbersome to describe the three non-stroke/TIA separate comparison groups in the Tables, so perhaps just explain this in the methods)?
Responses: We have added text to the final paragraph of the methods to highlight this.
12. Missing data is an important limitation to note. Given that ~25% of data are missing in some touchscreen questions (Table 1), is it appropriate to quote the frequency in the entire group, or is it better to quote among respondents only? Or does missing imply that the characteristic is absent given the way touchscreen questions were completed (in which case say so in the methods)? Perhaps the authors could clarify.
Response: We agree that the table is unclear. We have corrected it, we had inserted missing data into the touchscreen section in error and this has been removed.
On reflection we agree that it is also difficult to follow the recording of “missing data” in the touchscreen section. These data represent people who were not asked directly about stroke disease having not flagged it in the touchscreen questionnaire, or volunteered the information spontaneously. We have therefore removed these data from the table.
13. Another caveat about the apparently high frequency of secondary prevention medication use is that not only was the Biobank cohort young, but also they may reflect the ‘healthy [or perhaps, adherent] volunteer’ effect. Worthy of a comment, perhaps?
Response: We have added a comment to this effect on page 12, second paragraph.
14. Table 1 – the numbers in each category do not appear to add up to 502,650. Perhaps the authors could check their sums.
Response: We have corrected these data
15. N=6699 in Table 1, but n=6669 in Table 2.
Response: We have corrected this error, which also occurred in table 3.
16. In the Methods, perhaps indicate that each of the three groups was individually compared with the non-stroke/TIA Biobank population?
Response: This comment is in keeping with another above and we have clarified the test to ensure this comes across more clearly.