## PEER REVIEW HISTORY



## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | **Assessing generalizability through the use of disease registers: findings from a diabetes cohort study** |
| **AUTHORS** | David, Michael; Ware, Robert; Donald, Maria; Alati, Rosa |

## VERSION 1 - REVIEW

| | |
|---|---|
| **REVIEWER** | ***Nick Taub***<br>Research Fellow in Medical Statistics<br>Dept. Health Sciences<br>University of Leicester<br>22-28 Princess Rd West<br>Leicester. LE1 6TP<br><br>I have no competing interests with the work presented in the paper that I have reviewed. |
| **REVIEW RETURNED** | 04-Apr-2011 |

| | |
|---|---|
| **GENERAL COMMENTS** | This papers reports a study of the degree to which research carried out using disease registers can be considered to be representative. The study has been well planned and analysed and provides useful warnings to future researchers especially in the fields of public health and epidemiology, on the implications of recent and developing moves to require the consent of an individual before their personal data can be analysed in research. This study is a difficult one to present clearly – the authors have done well, but some improvements could be made.<br><br>The flow-chart, Figure 1, documents the status of the participants in the study, and is clearly needed. It is however not easy to read, so that there should, if possible, also be included a Venn diagram, or something similar - showing how the set of participants is enclosed in the set of invitees, which is in turn enclosed in the set of … , and so on – for each possible status, giving the number of individuals.<br><br>Page 7, text from “A non-invitee was defined … ”. The words ‘ … these two latter groups were combined … ’ appear to define ‘non-invitees’ and ‘non-study registrants’ to be the same group of individuals – needs clarification.<br><br>Page 8, ‘Data analysis’. It needs to be made clearer what importance, if any, the study’s own sampling scheme and corresponding survey weighting have to the analysis of representativeness – and especially whether the adjustment for certain factors in the logistic regression analyses makes allowance for the sampling scheme.<br><br>Table 1. It would be useful to include column totals at the head of the first three columns - Participants, Non-participants, and Non- |

| | |
|---|---|
| | study registrants. Footnotes should remind the reader i) for the which variables the comparison of Participants vs Non-participants was adjusted, and ii) for which variables, if any, the comparison of Participants vs Non-study Registrants was adjusted. |

| | |
|---|---|
| **REVIEWER** | ***Sonia Napravnik***<br>Assistant Professor<br>The University of North Carolina at Chapel Hill<br>United States |
| **REVIEW RETURNED** | 13-Apr-2011 |

| | |
|---|---|
| **THE STUDY** | This is a well written paper on a topic of interest to many areas of clinical and public health research. In its current form there are substantial questions about the methods used to compare participants and non-registrants. It is stated that only aggregate data was available on non-registrants however methods appropriate for statistical analyses using aggregate and individual level data are not presented. It is unclear whether or how multivariable analyses were conducted for assessing these associations. It is difficult to evaluate the results without this information. Also more information should be presented on how models accounted for sampling probabilities. |

## VERSION 1 – AUTHOR RESPONSE

RESPONSES TO REVIEWER 1: Mr. Nick Taub
Paragraph 1
"Research ethics have not been mentioned."

Ethical approval to undertake the study was gained from The University of Queensland's Behavioural and Social Sciences Ethical Review Committee. A statement has been included on page 8.

Paragraph 2
No response as no clarifications or changes recommended.

Paragraph 3
"The flow-chart, Figure 1, documents the status of the participants in the study, and is clearly needed. It is however not easy to read, so that there should, if possible, also be included a Venn diagram, or something similar - showing how the set of participants is enclosed in the set of invitees, which is in turn enclosed in the set of … , and so on – for each possible status, giving the number of individuals."

We appreciate the different groupings of registrants are complicated. In an attempt to clarify the groups we have included a Venn diagram (this becomes figure 1 and the original figure becomes figure 2 in the revised paper) as suggested to clarify the various subgroups and the various relationships that exist between them for the purposes of our comparative analyses. From the diagram it should be clearer to see that participants are a subset of invitees which is a subset of consenting registrants which is a subset of registrants. Additionally, it should be clearer to see that in the final analysis that the reference group of non-study registrants consists of: non-participants, non-sampled consenting registrants and non-consenting registrants. The frequencies of all the groups have been included in the the diagram and in the footnotes.

Paragraph 4

"Page 7, text from “A non-invitee was defined … ”. The words ‘ … these two latter groups were combined … ’ appear to define ‘non-invitees’ and ‘non-study registrants’ to be the same group of individuals – needs clarification."

The description of the mutually exclusive subgroups has been modified in order to eliminate any ambiguity surrounding study 'non-invitees' and 'non-study registrants' (second reference group) such that it is clear that:
non-study registrants=non-participants+non-invitees where
non-invitees=non-sampled consenting registrants+non-consenting registrants
The inclusion of a schematic diagram (new Figure 1) should also make the distinctions between these subgroups clearer (especially non-invitees and non-study registrants).

Paragraph 5
"Page 8, ‘Data analysis’. It needs to be made clearer what importance, if any, the study’s own sampling scheme and corresponding survey weighting have to the analysis of representativeness – and especially whether the adjustment for certain factors in the logistic regression analyses makes allowance for the sampling scheme."

Addtitional comments have been included in Methods section (page 7) to indicate that oversampling was undertaken in the sampling process in three of the four locations or strata, thus resulting in the use of probability weights in the analyses. In our original submission unweighted results where presented as in the main they differed minimally from the weighted results e.g. due to increased standard errors from weighting, the associtions between sex, registrtion year and attrition changed from being slightly significant to being slightly non-significant at the 5% level of significance. These associational changes have not affected the main finding of this study as concerning the influence of research consent on generalizability.

Paragraph 6
"Table 1. It would be useful to include column totals at the head of the first three columns - Participants, Non-participants, and Non-study registrants. Footnotes should remind the reader i) for the which variables the comparison of Participants vs Non-participants was adjusted, and ii) for which variables, if any, the comparison of Participants vs Non-study Registrants was adjusted."

As suggested column totals have been included for participants, non-participants and non-study registrants in Table 1. A description of the controlled variables in the comparative analysis (participants vs non-participants) has also been included in the footnotes. No similar descriptions have been provided in the footnotes for the other comparative analyses, as no variables were controlled or adjusted for. As such, both analyses provide crude or unadjusted odd ratios.

RESPONSES TO REVIEWER 2: Ms. Sonia Napravnik
Paragraph 1
"This is a well written paper on a topic of interest to many areas of clinical and public health research. In its current form there are substantial questions about the methods used to compare participants and non-registrants. It is stated that only aggregate data was available on non-registrants however methods appropriate for statistical analyses using aggregate and individual level data are not presented. It is unclear whether or how multivariable analyses were conducted for assessing these associations. It is difficult to evaluate the results without this information."

The description of the Data Analysis on page 9 has been expanded to clarify the comparative analyses used in this study with the results shown in Table 1. Specifically, that it was not possible to compare participants with non-study registrants (reference group) by a multivariable analysis because

individual level data did not exist for those in the reference group. Hence, the results provided in Table 1 (last column) are crude odds ratios (with 95% CIs) and not adjusted odds ratios (which would have been preferable if individual data had been available and is therefore a limitation of the study and is mentioned in the Disscusion on page). As a consequence, our paper comprises of three comparative analyses. They are in order:
1) participants vs non-participants (univariable analysis)
2) participants vs non-participants (multivariable analysis)
3) participants vs non-study registrants (univariable analysis)

Paragraph 1
"Also more information should be presented on how models accounted for sampling probabilities."

More information has been presented regarding the usage of probability weights in the analysis resulting from our sampling design which utilized oversampling. This need for more clarification and information on this point was also mentioned by the co-reviewer. Rather than repeat my response(s), please see above - Paragraph 5: Weighting in the analysis.

Thank-you for you comments.

**VERSION 2 - REVIEW**

| | |
|---|---|
| **REVIEWER** | ***Nick Taub*** |
| **REVIEW RETURNED** | 17-May-2011 |

| | |
|---|---|
| **RESULTS & CONCLUSIONS** | The authors have generally responded well to the previous comments.<br><br>I have only one concern. I suggested that the authors insert a Venn diagram to clarify the structure of the data. They have instead given a display which clearly divides the 133851 registrants into 4 mutally exclusive groups, and gives the number of people in each.<br><br>This is fine - EXCEPT that the display has the structure of a Venn diagram, as if these mutually exclusive groups were containing each other - that needs to be corrected. |

**VERSION 2 – AUTHOR RESPONSE**

RESPONSE TO REVIEWER: Mr. Nick Taub
"I have only one concern. I suggested that the authors insert a Venn diagram to clarify the structure of the data. They have instead given a display which clearly divides the 133851 registrants into 4 mutally exclusive groups, and gives the number of people in each. This is fine - EXCEPT that the display has the structure of a Venn diagram, as if these mutually exclusive groups were containing each other - that needs to be corrected."

Figure 1 has been modified so that the 4 groups: participants; non-participants; non-sampled consenting registrants and non-consenting registrants are definitely seen to be mutually exclusive from each other and that no group is a subset of the one or more of the remaining 3 groups. In addition, Figure 1 now shows clearly the three groups that consitute non-study registrants.