Reviewer's report

Title: Novel use of three administrative datasets to establish a cohort for environmental health research

Version: 2
Date: 11 December 2014

Reviewer: Rosa Gini

Reviewer's report:

This manuscript contains very valuable information on cohorts available for public health research in New Zealand, and there is no doubt that it deserves publication.

However, its content is very badly organized, so that the manuscript needs to be rewritten. There seems to be a misconception in the methods section, that 'methods' is the way the analysis dataset was obtained only. Most of the design choices and of the statistical methods are either implicit or described in the Results section. On the other hand, the Results section contains a lot of comments that should be shaped in a more complete and agile Discussion section, where limitations of the study should be grouped but recommendations arising from the study should be summarized as well.

More detailed comments follow.

Abstract

What is nationwide housing data is not clear from the abstract. Please introduce the acronym QV and refer to it throughout the abstract.

"We then compared population distributions and hospitalisation rates across the Census, NHI and matched cohort populations": 'compare population distribution' is not clear. Indeed, distribution of several variables were compared in the different population.

"Nationwide housing data match rates were 68%." This sentence is not clear, and overall the results subsection is difficult to understand. The matched cohort is the list of subjects in NHI whose address is available and can be matched with a QV record, is this correct? The 68% of the subjects had their address matched to QV, is this correct? Main results from comparison of variable distributions in the different populations should be summarized here, as well as results from comparison of hospitalization rates.

The sentence "The NHI has some advantages over the Census as a denominator, and the ability to access individual-level data through address matching provides many opportunities for use in environmental health studies" is more for an introduction, as it is a general statement and is not supported nor rejected by the study results. In the conclusion there should be a quick discussion of the main
results (why is the matching between NHI and QV not complete? why is the number of subjects in NHI overall higher than the number of subjects in the census?) before recommendations are made.

Background

"However, there has been little description of how such data have been used": this is not factual. descriptions ahve been made in different countries: US, canada, several european countries, australia, several asian countries...; and for different objectives: pharmacoepidemiology, public health, health services research...). please put this sentence in context. see below some reference

Tricco AC, Pham B, Rawson NSB. Manitoba and Saskatchewan administrative health care utilization databases are used differently to answer epidemiologic research questions. Journal of Clinical Epidemiology. 2008;61(2):192–7.e12.

Methods

the NZDep index of deprivation should be introduced in the methods section ("We used the census area unit-based NZDep index of deprivation to assign socio economic status[14]" is now in the results)

the subsection on statistical analysis "Population and hospitalisation comparisons" should be expanded, and made more clear: for instance, the term "match rate" used in the results should be introduced here. in generale all the analysis leading to the results contained in tables and figures should be
introduced here, with a precise description of methods and rationale. for instance instead of "the matched cohort population distribution" analysis should be described as the distribution of age, gender, deprivation index, ethnicity... in the matched cohort population.

"As all three populations were so large, even small size differences between them were statistically significant, so decisions on which differences mattered and which did not could only be made subjectively": would it be feasible to fit a multivariate model, that estimates within the NHI matching probability with QV according to several variables at a time? that would help summarizing the results

Results

figures and tables should all be introduced in the main text, as explained in the methods. all the main results should be described in this section. results relevant for discussion should be highlighted, but only discussed in the Discussion section.

"After exclusions and filtering": exclusion and filtering methods should be described more extensively in the Methods section, and the dataset before exclusion and filtering should be briefly described

"NHI population was approximately 6% larger than the comparable 7 March 2006 Census population" this is a different figure wrt the abstract "The NHI provided a raw population 22% larger than the Census", why? "The main reason the NHI was larger was that it included some duplicates and some New Zealanders who were living overseas": how is this supported? is this factual or just an hypothesis? "As noted below, some of these additional NHI records could be excluded": all exclusions should be described in the methods, and their effect only should be described in the Results.

subsection "NHI vs 2006 Census population comparisons" is not clear: what is a 'group'? and what are the results contained in this subsection? it sound more like a discussion

"Consequently, results for this age group should be treated with caution": this sentence belongs to discussion "and were excluded from hospitalisation rate calculations for the total population": this sentence belongs to methods. in general, the subsections on age and on age and sex could be merged

in all the following subsections, again results should be described but not commented. all decision about data analysis should be moved to the methods section and all comments to the Discussion section.

subsection "NHI Data limitations" should be moved altogether to the Discussion section, as well as the following subsections.

Discussion

this section should include all comments to the results that are now in the Results
section, and should be better organized.

Conclusion

the main findings and recommendations should be repeated here

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests