Author's response to reviews

Title: Methods for dealing with discrepant records in linked population health datasets: a cross-sectional study

Authors:

Christine L Roberts (c Roberts@med.usyd.edu.au)
Charles S Algert (calgert@george.org.au)
Jane B Ford (j ford@med.usyd.edu.au)

Version: 2 Date: 28 November 2006

Author's response to reviews: see over
28 November 2006

Deborah Saltman MD, PhD
Editor-in-Chief
BMC Health Services Research

Dear Dr Saltman

Thank you for considering the manuscript entitled ‘Methods for dealing with discrepant records in linked population health datasets: a cross-sectional study’ for publication in BMC Health Services Research.

We have used two linked population health datasets (a legislated perinatal data collection and hospital discharge data on over 250,000 women) to examine how discrepancies in case reports can affect the results of analysing the data. We have examined four different methods of dealing with the discrepancies and conclude that how the outcome of interest is classified will affect the estimates of prevalence and of risk for specific factors. We also use reporting characteristics from prior validation studies to calculate “corrected” prevalences, as an alternative method of assessing what the best estimate of prevalence is. Briefly, we recommend making use of all reports of a condition for prevalence estimates but only perfect matches for risk factor analyses.

We believe this article should be published in BMC Health Services Research for the following reasons:

- The use of population health data to examine health outcomes and health services is increasing and data linkage has been recommended as a way of reducing the under-ascertainment associated with many population health datasets. However, this raises the possibility of having discrepant case reports from different datasets.
- We could not identify any published methods for dealing with discrepant reports.
- We used data from two well-validated population health datasets to examine four different methods of dealing with discrepant reports and compared these with using either dataset alone.
- The method of using linked population data to minimise the under-reporting by accepting a reported condition from either dataset may result in misclassifying non-cases as cases and under-estimate the impact of risk factors. This is particularly true for rare conditions where the number of false positive reports may equal or outnumber the true positive reports.
Our analyses illustrate the issue that misclassification error in population data can be a larger problem than sampling error, and that the narrow confidence intervals for analyses using large population datasets can be misleading.

Although the study is based on perinatal data, the findings are likely to be generalisable to the record linkage of any specialised dataset (e.g. cancer, cardiovascular) to hospital discharge data.

The data for the study were obtained from the de-identified linked birth and hospital discharge files of the NSW Department of Health. Neither the manuscript nor any part of it has been published nor submitted for publication elsewhere. There are no financial or other relationships that might lead to a conflict of interest with the study. All the authors have approved submission of the manuscript. If you have any queries regarding the manuscript please do not hesitate to contact me.

Yours sincerely

Christine Roberts

Address for Correspondence:
Dr Christine Roberts
Clinical and Population Perinatal Health Research
The Kolling Institute of Medical Research
Level 4, Wallace Freeborn Building (E25)
Royal North Shore Hospital 2065
AUSTRALIA
Phone: International + 61 2 9926 7013
Fax: International + 61 2 9906 6742
Email: clroberts@med.usyd.edu.au