Reviewer's report

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

Version: 2 Date: 5 January 2007

Reviewer: Cashel D'Arcy James Holman

Reviewer's report:

General

The validity of information 'capture' is a topic of increasing importance in health services research, due to the wider use of multiple, linked administrative databases to answer research questions. The article deals with an issue that is relevant to these developments - what to do when a health outcome is discrepantly identified, or not identified, in multiple databases.

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

I am in agreement with the authors' interpretation of their results and their conclusions with respect to estimation of the prevalence of pregnancy-related hypertension. It seems clear that both databases (the MDC and ISC) underascertain the presence of hypertension, resulting in a differential misclassification error (too many false negatives) when overall ascertainment is based on mandatory dual capture in both databases. The error is corrected, at least partially, when overall ascertainment is based on capture in either database - only then do we come closer to identifying all of the true cases of pregnancy-related hypertension.

My interpretation of the second part of the authors' results, being the analysis of risk factors, is quite different to theirs. It may be irrelevant that I disagree, but the authors should at least consider my alternative interpretation and either reconsider their own interpretation or provide a robust justification for why their original interpretation is preferred.

The authors do not comment on the clinical processes or the nature of the clinical observers who recorded the presence of hypertension in the MDC and ISC. It is likely that a midwife completes the MDS form and a medical resident completes the ISC form or at least records information in case notes that is abstracted by medical record coders. Whatever clinical observation and recording processes were involved, it would be reasonable to suspect that if the pregnancy-related hypertension was more severe, it would be more likely to be observed and recorded in both the MDC and ISC; whereas more marginal cases of hypertension, which were of a relatively lesser clinical relevance to the midwives and doctors, would be less likely to be consistently recorded by both databases. The authors do not address this possibility, but it is in my opinion quite a reasonable conjecture. If correct, it means that hypertension identified on the basis of capture in both databases will have been, on average, more severe than hypertension captured in only one database. The selection of a relatively more severe subset of pregnancy-related hypertensive using dual capture would be consistent with the fact, acknowledged by the authors, that mandatory dual capture in both databases identifies only a subset of the true prevalent pool of women affected by the condition.

It follows that the different risk factor analyses presented in Table 2 do not relate to the same disease, but rather to hypertension of different average levels of severity. It does not surprise me, therefore, that known risk factors have their strongest effect (using a relative effect measure such as the OR) when hypertension is more severe. The absolute risk of more severe hypertension is lower and thus even the same absolute increase associated with the presence of a risk factor will result in a higher relative risk. Moreover, it is usually found that known risk factors cause a higher aetiologic fraction of more severe cases of disease - that is how they become 'known', because they have strong effects compared with unknown causes.

I disagree with the authors, therefore, that the weaker ORs based on capture in any database are biased due to misclassification error. Bias towards the null is the only possible result of non-differential misclassification error, but in this instance the nature of the misclassification error is actually differential, and thus the bias may be in either direction. If the aim is to estimate the effect of risk factors on
pregnancy-related hypertension in general (and not a subgroup of more severe hypertension) then the effect measures based on the mandatory dual capture are the biased ones, being affected as they are by differential misclassification error in outcome ascertainment - the very reason why the dual method also yields biased estimates of prevalence as acknowledged by the authors.

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

Nil.

Discretionary Revisions (which the author can choose to ignore)

The authors may wish to strengthen the context in which they place their work by referring to the following closely related issues and research methods:

1. Sensitivity and PV+ve estimates for the reliability of health information in administrative databases, based on an independent validation study, are uncommonly available, albeit that such estimates were available to the authors in this instance. Therefore, the authors may wish to refer to alternatives such as those in the capture-recapture methods literature. A couple of useful references are:


2. The authors deal with the problem where discrepancies arise in health outcome capture between two different administrative databases. A similar issue may arise when discrepancies occur between linked records for an individual within the same administrative database, such as a hospital morbidity data system. The authors may wish to discuss the implications of their work for using multiple records within the same data set to identify, for example, comorbidity.

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

Level of interest: An article whose findings are important to those with closely related research interests

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.