Reviewer’s report

Title: Validation of de-identified record linkage to ascertain hospital admissions in a cohort study

Version: 1 Date: 16 December 2010

Reviewer: Don E Detmer

Reviewer’s report:

This is a very important issue but from a policy perspective as well as data management issue (cohort databases from differing sources), I’m not convinced that the authors defined the issue sufficiently nor offered up advice for policy makers and regulators on how to get better data. They more or less assumed that you’ll have ‘so-so’ data and went from there. OK but surely there is some interest in making the data more valid from the ‘get-go’.

From strictly a cohort study question, they may have left out one methodological consideration. It appears that they worked until they found a set of records that were valid using a triangulation of approaches including examination of actual medical records and so forth. So far, so good. However, as I read the manuscript, they then checked their algorithms against the entire common cohort database that had resulted from this data validation exercise. I would have been happier if they had used 60% or even 75% of the common cohort database and then tested their final “best” algorithm against the remaining 25-40% to see how it performed. I don’t consider myself a basically a methodologist so other reviewers should weigh in on this methods issue.

Policy Dimension. From the perspective of policy, at least in the USA where we are trying to demonstrate the value of having a Unique Personal Health Identifier (UPHI), their report is only helpful in part. Obviously, part of the utility, or decay of value, of a unique identifier apparently relates to the nature of how Australia assigns unique identifiers, e.g., unique to families and then members within families which can be confusing and, apparently, a source for slippage in validity as well as how well the data are entered in the first place. These issues being the case, I was a bit puzzled why they ‘declared victory’ with their algorithm rather than also suggesting in the discussion how Australians might improve their system of identifiers both from the perspective of precision as well as administration. Another question of importance relates to the impact of anonymization and what methods are used to anonymize the data. This may have been clear to other reviewers but it was not sufficiently clear for me.

Having made the comment relating to assessing the value of having a unique personal health identifier for improved cohort analyses, I’m not positive how I could ‘cure’ this issue for this study or for any other future study for that matter. It appears that their analysis finds that one gets a ‘false sense’ that one has not lost any value through anonymization by not having a totally unique personal health identifier but that this represents ‘false security’ due to a high ‘false
positive’ rate. Writing as just a single reviewer, what I’m most interested in knowing is how much ‘loss of signal’ results from not having 1) a totally UPHI and 2) accurate recording. I gather that this is asking the wrong question at least from the researcher’s perspective.

Where does this leave me? Happily, I’m not the editor. There may be sufficient value to justify publication of the paper based solely upon what they accomplished. Even if this is the case, I’d be much happier to see discussion of the implications for policy and administration and not just how researchers address such slippery slopes encountered in their work.

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

**Quality of written English:** Acceptable

**Statistical review:** Yes, but I do not feel adequately qualified to assess the statistics.

**Declaration of competing interests:**

I declare that I have no competing interests