Reviewer's report

Title: Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research

Version: 2
Date: 21 March 2014

Reviewer: Sebastien Haneuse

Reviewer's report:

This paper demonstrates the selection bias that results when one restricts analyses to individuals with “sufficient” data in an EMR. The premise is intuitive (i.e. that sicker patients would have more information in the EMR) but it may be an issue that is nonetheless not taken as seriously as it should. While I believe this is an important and well-motivated issue (the introduction is particularly well written) I found the methods and analysis to be very disappointing.

Major essential revisions

1. The main problem with the analysis is that the “outcomes” occur prior in time to the “exposure” of interest. Patients are identified on the basis of their ASA classification and yet the number of days with medication orders and lab results is computed over the previous year. Since all of this is from an EMR, why not extract the counts for the subsequent year? The same issue goes for several other “exposure” variables such as emergent status and patient type (i.e. both are subsequent to the outcome).

2. Numerous exclusions are imposed without sufficient detail, either in their motivation or in a discussion of the implications. It seems somewhat ironic, for example, that the paper is all about the “danger” of sufficiency exclusions and yet the authors exclude records from patients with “infrequently” occurring ICD-9 codes or CPT categories. Similarly, the authors excluded ASA 5 and 6 cases. It’s fine to say that these groups have small numbers but it seems antithetical to the point of the paper. Along these lines, the authors impose no restrictions on the basis of age with the result that the samples consists of patients aged 1 to 102 years. This is an incredibly heterogeneous population. Does a (marginal) rate ratio of approximately 4.0 have much meaning when the population is so heterogeneous?

3. There is far too little detail about the statistical analysis. In particular, its not obvious to me that readers will be familiar with the details of the ZIP model and the authors should provide a description of the model, ideally with actual notation, along with a generic interpretation of its components. Referring to SAS is insufficient and a proper reference should be given. A key component of the ZIP model, which distinguishes it from a straightforward Poisson model, is the additional term that permits more ‘zeros’ than would be dictated by a Poisson distribution. This term (usually a proportion) should also be reported as part of
the results.

4. The methods also don’t describe how uncertainty is evaluated. In particular, overdispersion is a common statistical challenge with count data. Did the authors account for this? Are the standard errors based on a sandwich estimator or a model-based estimator? I suspect that it won’t make much difference either way but they should at least be described. Along these lines, nowhere in the paper are measures of uncertainty reported such as confidence intervals. These should be reported at least in the tables, results section and abstract. A reliance on p-values is insufficient.

5. There are far too many p-values in the paper. In particular, it’s fine to report omnibus tests for covariates that have multiple levels but level-specific p-values are essentially meaningless.

Minor essential revisions

1. As a minor editorial point, just prior to the data analysis section, the authors refer to having conceptualized data sufficiency as a “continuous” variable … arguably a more precise description of what they did in the paper was to treat it as a count variable.

2. Why did the authors only sample 10,000 of the 24,073 cases? If data is being abstracted from the EMR, why not extract the information for all of them?

3. It would be helpful to see the actual marginal distributions of the two outcome variables, stratified by the 4 ASA classes.

4. How did the authors choose the age groupings in Tables 2 and 3?

5. The labeling of the ICD-9 and CPT levels in Tables 2 and 3 is very poor and the authors shouldn’t require the reader to either know the codes or have to look elsewhere to find out their meaning.

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.