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: 31 March 2014

Reviewer: Jose Alberto A Maldonado

Reviewer’s report:

The main topic of the paper is quite interesting and it is aligned with two current hot topics in (bio)medical informatics: the reuse of EHR data for research and data quality. The paper focused on sufficiency of EHR data, and explores how sufficiency may bias towards sick patients. The paper is well written and structured and the limitations are clearly stated. My concerns are:

Discretionary Revisions

Although, my field of expertise is not statistics, I miss a clearer justification of the statistical method (ZIP) used for analysis and the reason for not using other methods.

The paper hypothesis seems sensible to me and it is something I would expect: sicker patients have more data recorded in their EHRs. The EHR is a record of health and care provision to a subject of care, thus sicker patients have more data: more medication orders, more lab test results, etc. A different issue is whether the data recorded in the EHR accurately reflect the health and care provision; in this regard a desirable property is completeness. I think that a clear distinction between completeness and sufficiency in the text would help readers (completeness is the most commonly assessed dimension of data quality) to clearly focus the scope of the paper (sufficiency). In this regard the authors could give some details about the completeness of their data sets (lab results and medication orders).

EHR information models have some artifact to deal with missing data such as null flavors in HL7 world. It could improve the paper if the authors could discuss about how these kinds of artifact may help.

Minor Essential Revisions

The authors categorize “laboratory results” and “medication orders” as data types. I recommend using other term such as “domain concept” or even “data set” since the term “data type” has a particular meaning in computing (integer, string, etc.) and even in EHR information architectures (timestamp, ratio, coded value, etc.).

Major Compulsory Revisions

none
Level of interest: An article of importance in its field

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