Reviewer's report

Title: The Impact of Standardizing the Definition of Visits on the Consistency of Multi-Database Observational Health Research

Version: 2
Date: 12 October 2014
Reviewer: Mark G Weiner

Reviewer's report:

The premise of this manuscript, that the data specification used to define an inpatient encounter type can impact the detection and rates of disease in a population, is important. The application of a standard transformation, and the display of results where differences in disease rates between data sets from two different vendors decrease after the transformation is also compelling. The authors rightfully indicate that while they are not comparing to a gold standard source of truth of the patient's inpatient status, they would rather be consistently wrong than inconsistently right -- an important caveat that should be stated more up front as the goal, and not just in the discussion (perhaps alluding to the distinction between a consistent definition and a gold standard truth on line 78).

While I understand the difficulty in comparing to a true gold standard, knowing that the source data that feeds both vendors’ raw data sources (UB92-> Truven, UB92->Optum) is identical, I need more information about the organization and content of relevant data by these two vendors and the transform of the vendor data to OMOP, and some examples of the differences.

The transformation seems more natural for the Truven data since both the Truven and the CDM definition of an inpatient stay is based on a revenue code consistent with an inpatient stay -- like the "room and board" code. However, the intended and applied specification for inpatient status for Optum seems less related to revenue codes, and seems inconsistently described in the manuscript. The manuscript on line 71 and 72 says that Optum contains a field to indicate claims associated with an inpatient confinement,***but the inpatient classification is defined by the place-of-service.*** (presumably not a revenue code, but this is not clear). Later on (lines 152-154), the manuscript says "Any Optum records with an associated confinement identifier were considered as part of an inpatient visit unless they were identified as ER claims by the place of service field." In this latter statement the inpatient definition focuses on the "confinement identifier" where earlier it was the place of service. This discrepancy must be clarified, with quantitative results reanalyzed if appropriate. Is the "confinement identifier" a revenue code, or is it a calculated field derived by Optum based on their own algorithm? What is the revenue code in Optum that is used to anchor the inpatient status assigned in the CDM?

With the close alignment of Truven and CDM definitions of inpatient status, the
stable rates of consistency seem reasonable. However, I cannot envision the nature of the scenario where the CDM labels an encounter as an inpatient stay, but the "raw" Truven data does not. Furthermore, the Optum data has occurrences where the raw data labels a stay as an inpatient encounter, while the CDM does not -- a situation that never occurs with the Truven data. The variable rate of concordance between CDM and Optum recognition of inpatient status may reflect a change in their use of the "confinement identifier," though, again, that is not the intended inpatient marker based on lines 71-72. Something is clearly going on in 2009 with Optum's definition on inpatient status!

Furthermore, I am curious about other differences in the nature of the inpatient admissions picked up by the CDM definition that are not part of the "raw" definition. The manuscript states (lines 265-67) that "CCAE mentions in their documentation that a small percentage of inpatient services fall into their outpatient services table when no charges are found" If these account for the extra inpatient encounters found by the CDM definition, then what is good for normalizing diagnosis rates may not be so good for normalizing charges. The lengths of stay for these "CDM-only" inpatient encounters may also be skewed.

Bottom line is that there needs to be some illustrative examples of the discrepancies so that a reader can make a judgment about the correctness of the CDM version. It needs to be clear what is a revenue code, a place of service code and what is a derived code in the source (raw) data. Analyses of how the CDM definition of inpatient stay affects other characteristics of relevance to inpatient encounters, like charges and length of stay, should be addressed or at least mentioned.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I have no financial conflicts to report, nor have I worked formally with any of the authors. I have had many informal conversations with Dr. Ryan related to our mutual work on OMOP, though not related to the content of this manuscript.