Author's response to reviews

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

Authors:

Alexandar Rusanov (ar2765@columbia.edu)
Nicole G Weiskopf (ngw2105@columbia.edu)
Shuang Wang (sw2206@columbia.edu)
Chunhua Weng (cw2384@columbia.edu)

Version: 3
Date: 2 May 2014

Author's response to reviews: see over
Dear Editor,

In this revision,

1. We included a title page at the front of the manuscript file.

2. We structured our abstract and made sure it is compliant with the Word limit 350.

3. We inserted IRB approval number in the Methods section.

4. We inserted an author contribution statement.

Below are our detailed responses to the previous review comments.

**REVIEWER 1**

Reviewer: Sebastien Haneuse

Our thanks to the reviewer for his comments and suggestions. His critique of the statistical methods used in our manuscript was especially helpful. We have done our best to respond to all his suggestions and made corrections where appropriate, and hope that he is satisfied with our updated manuscript.

**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).

I apologize if we do not make it clear in the paper, but we are interested in the relationship between patient health and data quantity at a given point in time, not whether or not health status predicts how much data will be accumulated going forward. Specifically, we wanted to look at how patient health at the time of data extraction is related to the amount of available data at that same time, which is what will be most relevant in secondary use scenarios. We wanted to know if sick patients have more data, not if they will go on to accrue more data in the future.

We would also argue that ASA status is more representative of health during the preceding year than during the following year for two reasons: 1) ASA is intended to be a retrospective measurement that relies upon information about the patient’s preexisting medical experiences and status, and 2) ASA assessment is frequently performed as a preliminary step for surgical procedures, which can have a major impact on patient health status. Therefore, ASA status may not be a valid assessment of patient health going forward. Moreover, the surgical procedure itself is likely to result in major changes in data quantity, even if a patient is healthy. Patients undergoing surgery are more likely to have more lab tests done over the following year, as follow up to their procedure. In general EHR sufficiency for all patients
will increase following surgery just due to the surgery itself and not, necessarily as a function of the patient’s health status preceding the surgery.

As an example, imagine a patient who is determined to be healthy (ASA 1) and thus, as we hypothesize, has little data. Now this patient undergoes surgery and gets a prescription for pain medication added to their EHR. In the operating room a healthy patient undergoing general anesthesia will almost always receive antibiotic and DVT (heparin) prophylaxis, local anesthetic for the skin incision, an anxiolytic (benzodiazepine), numerous doses of analgesics (opioids and NSAIDS), a sedative hypnotic (propofol, etomidate, dexmedetomidine, etc), a volatile anesthetic (sevoflurane, nitrous oxide, etc), a muscle relaxant (succinylcholine, rocuronium, etc) and then a reversal agent (neostigmine) with an antimuscarinic (atropine, glycopyrrolate, etc), and an antiemetic (ondansetron), many at high risk for nausea will receive another antiemetic adjunct such as a steroid or a dopamine antagonist (i.e. metoclopramide), many will also require a vasopressor such as phenylephrine or ephedrine. As you can see, this will quickly make the medication counts in the EHR go from zero to more than ten for this healthy patient. This number will be even higher if the patient is admitted to the hospital. Perhaps the sufficiency will increase to different degrees for healthy patients than it does for sicker ones (sicker patients will get all the medications listed above and, often, others) but that is not the question we want to answer (perhaps in a future study this can be addressed). The question we want to answer is whether, throughout the EHR, sicker patients have more data. We chose ASA class as a measure of health status (for reasons outlined in the paper) and thus ground our measure of EHR data at the same time. The other “exposure” variables you mention (patient type and emergency status) are also not “exposure” variables in our conceptualization of the study. They are characteristics of the patient just as sex and age are. If the procedure the patient is scheduled to undergo is not viewed as the “exposure” then all this may make more sense – or so we hope. We have added text to the methods section explaining why we looked retrospectively.

2.a. 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.

You make a very valid point that in a paper dealing with the danger of sufficiency exclusions, it is somewhat ironic to exclude records from patients with infrequently occurring diagnoses or procedures. In light of this, we went back and looked more closely at these infrequently occurring ICD9 and CPT categories with the help of a medical doctor. We found that many of these categories could be merged with other medically similar categories. Due to the large size of our dataset (10,000 patients) we also decided to lower the cutoff (in terms of the number of records) for defining infrequent categories from 5% to 2%. These two changes to the way we approached our data analysis allowed us to include all 10,000 patients in the regression model.

Only 0.2% of the patients in our anesthesia database were ASA 5 and only 0.02% were ASA 6. From a medical perspective these patients are also drastically different from ASA1-4 patients. The ASA 1 -4
patients are patients commonly encountered in a variety of healthcare settings (inpatient and outpatient) and are commonly the populations of interest to researchers. Conversely, ASA5-6 patients are exclusively inpatients and are unique in that ASA 5 ones are unlikely to survive and the surgery is a last ditch attempt at salvage, while ASA 6 ones are already dead and are presenting for one procedure only – organ harvest. These two groups of patients are usually of little interest to researchers. Aside from being very rare, patients classified as ASA 5 and 6 are different from other patients in many aspects aside from just health status. ASA classification is somewhat continuous from 1 to 4, but not beyond that. ASA 5 and 6 are truly unique categories. For these reasons we feel that including these patients in the analysis would not add to our understanding of the relationship between health status and sufficiency for the general population.

2.b. 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?

The population is in fact heterogeneous in terms of age, as well as a host of other variables. As discussed above, we no longer exclude any patients from the analysis. To give more meaning to the marginal rate ratios we grouped the patients into smaller bins (by decade) for reporting of marginal rate ratios in Table 2. This has the effect of decreasing the heterogeneity of each bin and should serve to make the rate ratios more meaningful.

3. There is far too little detail about the statistical analysis. In particular, it’s 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.

To account for overdispersion we have switched from using a ZIP model to a negative binomial regression model for our data analysis. We have provided a better explanation of the model in the Methods section and have also included a proper reference.

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.

As discussed in the answer to the previous comment, we now utilize a negative binomial regression model for our data analysis. Our reason for doing this is to better account for overdispersion. The measures of uncertainty (Standard Errors and 95% Confidence intervals) are now reported in Table 3.

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.
We no longer include level specific p-values for our covariates, only for our primary outcome variable. We believe this has resulted in a clearer and more easily interpretable results, which we presume is what you had intended. Thank you for this suggestion.

**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.

This is a good point. We changed the text “continuous variable” to “count variable” in the Methods section. Thank you.

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?

As the data we are working with contains protected health information it is considered to be highly sensitive. To protect patient privacy and institutional security exposure our IRB has placed strict requirements on the use of such data. Namely, they require that the least amount of data to adequately answer the research question be used. We chose to only extract information for 10,000 patients as we felt this sample would be large enough to answer our research question. While we could have extracted information for more patients, we felt that this would not be necessary as it would significantly increase the security risk without much of an effect on our results.

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

We have added the requested figures (Figure 1), which are indeed useful for visualizing the data.

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

As discussed in the response to Comment 2b in the Major Essential Revisions section above, we now use decades for the age groupings in Table 2. This was done in order to make the age groups less heterogeneous and thus increase the meaningfulness of the reported marginal rate ratios. Age groupings are not reported in Table 3, but rather the overall variable effect of age as a covariate.

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.

Our apologies. We have added the appropriate category and level labels to the Table 2, which should make it easier for readers to follow. ICD-9 and CPT levels are no longer reported in Table 3 (see response to Comment 5 in Major Essential Revisions section above).
REVIEWER 2

Reviewer: Jose Alberto Maldonado

We appreciate the kind comments of the reviewer. He raises a number of points that are of importance in the field of EHR data quality assessment, especially in the area of completeness.

Discretionary Revisions

1. 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.

In response to suggestions from the other reviewer, we have changed our model to a negative binomial regression. We have also added further detail to the manuscript to describe the model and explain why it is an appropriate choice. The changes can be found Method Section. We have also added a reference which provides further details on the model.

2. 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).

A precise definition of sufficiency is of central importance to this paper, so we appreciate the reviewer emphasizing this point. We consider sufficiency to be one of multiple ways to conceptualize completeness. We have added further detail and clarification to the manuscript in the Introduction Section, which read as follows: “Sufficiency can be conceptualized as a type of completeness, which is one of several categories of data quality that are relevant to EHR data reuse.[14] When EHR data are complete according to the requirements of a given task, those data can be considered to be sufficient for that task. Required data may be missing for different reasons: a data point was observed but not documented [12] or it was never observed in the first place, either because the observation was not clinically necessary or because it could not be performed. Regardless of the reason, missing data is very common in today’s EHR databases, leading to datasets that may not be sufficient for work relying on the secondary use of EHR data. Although it has been pointed out that the missing data may cause records to be “visually complete but intellectually insufficient,”[15] the causal effect of health status on data sufficiency is not the focus of this study. Instead, we focus on the correlation between the sufficiency of electronic health record data for clinical research and the underlying patient health status.”

3. 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.
Most of these artifacts are used for expected or required data fields that do not have a known value. For example, a demographic field, such as ethnicity or gender, might be denoted as missing or null for different reasons: not asked, asked but not answered, unknown, etc. This is generally not true of data like medication orders or laboratory results, since the data points do not exist unless they have a value. Therefore, while something like the HL7 null flavors could be very helpful in understanding patterns and causes of missingness for certain types of information within the EHR, we don’t feel that this specific study, which focuses on “optional” data, would benefit from such discussion.

**Minor Essential Revisions**

1. 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.).

We changed all occurrences of “data type” when referring to lab test and med orders in the paper (2 occurrences) to “kinds of data”. This should help decrease ambiguity. Thank you for this suggestion.