Author's response to reviews

Title: Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing

Authors:
Qiu-Yue Zhong (qyzhong@mail.harvard.edu)
Elizabeth Karlson (EKARLSON@partners.org)
Bizu Gelaye (bgelaye@hsph.harvard.edu)
Sean Finan (Sean.Finan@childrens.harvard.edu)
Paul Avillach (Paul_Avillach@hms.harvard.edu)
Jordan Smoller (jsmoller@hms.harvard.edu)
Tianxi Cai (tcai@hsph.harvard.edu)
Michelle Williams (mawilliams@hsph.harvard.edu)

Version: 2 **Date:** 27 Mar 2018

Author's response to reviews:

Dr. Mike Conway

BMC Medical Informatics and Decision Making Editorial Office

Re: MIDM-D-17-00368R1

Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing

Dear Dr. Conway,

Thank you very much for the review of our manuscript.

We appreciate the insightful comments and suggestions provided by you and the expert reviewers. We have revised the manuscript taking into account the reviewers' comments. We have indicated the changes by highlighting the changes.

Below, we provide point-by-point response to the reviewers' comments.

Thank you for providing us the opportunity to revise and improve our manuscript.

Sincerely,

Qiu-Yue Zhong, MD, ScM, ScD Candidate

Department of Epidemiology, Harvard T.H. Chan School of Public Health

677 Huntington Avenue, Room Kresge 502A

Boston, MA 02115

TEL: 617-432-1108

FAX: 617-566-7805

E-mail: qyzhong@mail.harvard.edu

Reviewer reports:

Sumithra Velupillai (Reviewer 1):

Overview:

The authors present a comprehensive study on screening pregnant women for suicidal behavior in EHRs using diagnostic codes as well as NLP, and provide an analysis of coverage, PPV and other aspects with these methods. The main contributions could be summarized as a) a large-scale analysis of prevalence of suicidal behaviour in a particularly vulnerable clinical cohort, and b) a comprehensive analysis of advantages and disadvantages of using structured codes and/or NLP to screen for suicidal behaviour in EHR data. The authors conclude that the use of NLP improves sensitivity/recall considerably, but at a cost of PPV/precision.

This is a very interesting study of great importance to the research community. Overall, the manuscript is very well-written and -structured, and provides interesting findings that have been analyzed in depth.

Thank you for noting the merits of our study.

There are, however, some concerns with this paper, that need clarifications and/or elaborations.

* Reference group and statistical analysis: the addition of a reference/control group is great, but there is very little to no actual analysis of this in the manuscript - this is not mentioned as the main focus of the study, of course - but if it is to be included, it would need some further discussion with respect to the reported results. Also, the statistical analysis that is provided is quite shallow (proportions mainly) - the study would be strengthened by including at least some statistical significance numbers.

Thank you for this comment. We would like to clarify that the reference group was included for the purposes of comparting the performance of different methods in identifying suicidal behavior among pregnant women. As indicated by the reviewer, we provided a detailed description (please see Page 13-14) and discussion (please see Page 18-19) of the descriptive analysis. In addition, as suggested by the reviewer, we provided the statistical significance in Table 1 (please see Page 12 and Table 1). However, considering a woman could have multiple encounters with

suicidal behavior during pregnancy and multiple comorbidities (i.e., categories in Table 2 and 3 were not mutually exclusive), we were not able to provide statistical significant numbers for Table 2 and Table 3.

* Sampling procedure and descriptive statistics: it is not quite clear on what basis the samples were drawn, e.g. the reasons behind the different matching ratios (1:30 in Sect. 2.3, 1:100 in sect. 2.4), why exclusion of Partner HealthCare employees was done, and why this was done after the sampling instead of before, and on what basis the samples for the chart review samples was performed?

These are all great points. We identified a random sample for NLP analysis because we were not able to process the entire set of clinical notes for the 273,410 women without any diagnostic codes related to suicidal behavior at a very short time. Our rationale for choosing a 1:30 ratio for subsequent NLP was twofold: (i) to provide a sample size that was large enough allowing a general view of distributions of CUIs and (ii) to minimize the NLP processing time.

Our decision to choose 1:100 ratio for reference group is also for the aforementioned pragmatic reasons. Since we did not need to process the clinical notes for reference group, we included a relatively larger sample size.

The exclusion of Partner HealthCare employees was done to comply with requirement of the Partners IRB. This exclusion is a standard procedure performed by the database administrators after we submitted our dataset request and before we receive the dataset (please see Page 9).

* Chart review: the chart review process and subsequent analysis is a bit difficult to understand and relate to the other parts of the study, e.g. was this done "blindly" or was the NLP output provided to the reviewer in the analysis? In the description of the chart review (Sect. 2.5), a total of 350 patients is said to have been analyzed (50+100+100+100), with a total of at least 1316

clinical notes (196+486+634, no number given for the group who had neither diagnostic codes nor term mentions) - then in the Results section, it is reported that 682 pregnant women were identified as having been screened positive for suicidal behaviour - on what data is this number based?

In general, it would be very useful if all these steps and numbers were reported in an extended version of Figure 1, so that it becomes clear for the reader what the underlying data and methodological steps are for the entire study setup - including also descriptive statistics not only on the number of patients, but the number of documents too.

We apologize for the confusion.

The chart review was done without checking the NLP results. We were interested in these four groups for chart review because they provided the PPV for diagnostic codes, PPV for term mention & NLP positive, NPV for the NLP step, and NPV for the term mention step, respectively.

Among the 5,282 women who did not have any diagnostic codes related to suicidal behavior, the number of women who had neither diagnostic codes nor term mentions was 4,162 (5,282-1,120); the number of women who had term mention but not diagnostic codes was 1,120. Among the 1,120 women who had term mentions related to suicidal behavior, 486 were screened positive by NLP and 634 were screened negative.

Pregnant women were identified as having been screened positive for suicidal behavior (N=682=196+486) by either diagnostic codes (N=196) or by NLP (N=486).

We modified our Figure 1 extensively to provide more details as you suggested.

Some additional, more minor comments:

* Which version of cTAKES was used? Was there a specific reason for it to be used "off-the-shelf" as opposed to analyzing its performance on a small subset to identify potential additional keywords for negation or other attributes? It would also be interesting if the distributions of all parts of the NLP output (affirmed, temporally relevant, subject relevant) on the entire sample was provided, to get a better idea of the number of cases that were filtered out by each of the attributes. Did you consider trying other existing NLP tools and compare the performance of more than one system?

We used the cTAKES 3.2.3 (please see Page 9). We used “off-the shelf” tool because we aimed to develop a screening procedure that can be easily generalized to other healthcare systems. We added the distributions of all parts of the NLP output on the entire sample (please see Page 10 and Supplemental Table 4). We agreed with the reviewer that understanding the performance of existing NLP tools and comparing their performance is an important topic. However, we respectfully argue that such comparison is beyond the scope of this manuscript.

* It would also be very interesting if a more in-depth discussion of the appropriateness of the selected suicide-related keywords and CUIs was provided - since there was an extensive chart review performed, would you say that these keywords/CUIs are sufficient for an NLP system to accurately initially identify relevant cases, and that therefore, the main remaining challenge for NLP to work well on this type of problem is to better identify semantic attributes? Or is there also a need to further develop the NLP system to identify other ways of expressing suicidal behaviour?

We agree with the reviewer that this is an interesting point to discuss. We are confident that these suicide-related keywords/CUIS/terms were able to capture the majority of relevant cases. However, there are definitely some linguistic variations in ways to document suicidal behavior in clinical notes. For example, the following two sentences did not explicitly mention the word “suicide” but expressed the meaning of suicide. Further development of the NLP system capturing the meaning of texts (e.g., word embedding [1–3]), in addition to better identification of semantic attributes, might solve this problem.

1. “She reports she wants to die, and wants to just crawl in a ball and disappear and be invisible, but denies any intent or plan.”

2. “This is a 34-year-old female who is currently at thirty-two weeks gestation who states that this evening she ingested 12 grams of Tylenol at 11 p.m. She was feeling desperate and says she doesn’t know what she was thinking.”

We amended our Discussion section to incorporate this suggestion (please see Page 20).

* 42 day cut-off: is this a clinically motivated cut-off time typically used for analysis of pregnant women?

We used the 42 day cutoff to comply with the WHO definition for maternal death: “the death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes” [4]. We provided the reference in our manuscript (please see Page 8&9).

* Psychiatric comorbidities: just out of curiosity, are these diagnostic codes more reliable than coding related to suicidal behaviour, or could these also be underestimates?

We thank the reviewer for this comment. The diagnostic codes for psychiatric comorbidities could also be underestimated [5,6]. However, to our best knowledge, we did not find any study that provides any estimate regarding whether diagnostic codes for other psychiatric disorders are more reliable as compared with diagnostic codes for suicidal behavior.

* References: one reference mentioned in the manuscript text is missing in the reference list (Raymond Francis Sarmiento, pg. 5), and the two Ford et. al 2016 papers mentioned on pg. 16 should be marked with a and b.

We thank the reviewer for noting this. We fixed our reference list.

To conclude, this manuscript presents a very interesting study of great relevance to the audience of this journal, but the methodological setup and study design needs to be clarified.

Thank you again for noting the merits of our study.

Angus Roberts, PhD (Reviewer 2):

The paper reports on an analysis of suicidal behaviour in pregnancy, as recorded in both the structured and unstructured portions of EHRs from a large US health provider. A comparison is given between EHRs recording suicidal behaviour through diagnostic codes, and those recording suicidal behaviour in the text of the EHR. Text was processed using string matching and the cTAKES NLP system. The authors show improved sensitivity by including NLP of the EHR, but PPV of NLP was lower than of diagnostic codes.

The paper is well written and easy to read. It is well motivated, and previous work is appropriately cited. The methods are sound for a paper of this type.

Thank you for noting the merits of our study.

Detailed comments below.

P4L51: methods for suicidal behavior is -> methods ... are

We apologize for this grammatical errors and we have fixed it as you suggested (please see Page 4).

P6L4: you might like to look at this two papers, either for completeness or to check your assertion in the following lines (I am not an author):

Detection of Suicidality in Adolescents with Autism Spectrum Disorders :

Developing a Natural Language Processing Approach for Use in Electronic Health

Records. / Downs, Jonathan Muir; Velupillai, Sumithra; Gkotsis, Georgios;

Holden, Rachel ; Kikoler, Maxim ; Dean, Harry; Fernandes, Andrea Carmen;

Dutta, Rina. In: Proceedings / AMIA, 05.07.2017.

Thank you for providing this. We included this in our manuscript and updated our references (please see Page 5).

P7L13 two, large -> two large

Thank you and we have fixed it as you suggested (please see Page 7).

P9L12 - I have some doubts as to the validity of the screening. Consider the case of a woman where the search term occurs later in time than the text that cTAKES identifies as suicidal behavior. In your experiments, this case would be included in your sample. In use of such a tool for real time screening however, it could be that the EHR does not yet contain the search term. There are two possibilities:

1. the screening tool consists of both the search terms and cTAKES, in which case the woman would be missed (she does not yet have the search term) and your results are not relevant to this case

2. the screening tool consists of just cTAKES, in which case she would possibly be picked up, but your results are not relevant as you only consider cases that also have search terms

Could you please clarify?

Thank you for raising this point. The terms included in the term search were broader than that of cTAKE (contain all cTAKES terms). In this way, we used cTAKES as a more refined tool to identify suicidal behavior that we are interested in (i.e., non-negated terms related to suicidal behavior during pregnancy, which happened to the pregnant women themselves).

In terms of a real time screening tool, we would use the entire screening tool (term search + cTAKES) every time a patient had encounter with the hospital. If a search term occurs later in time (T2) than the text that cTAKES (C1) identifies as suicidal behavior, the term (T2) would be captured by the cTAKES performed after that specific term (C2) (T1↦C1⇢T2↦C2).

P11L44 - at what point in time were they assessed, and how did this relate to the point in time of the positive screening?

Thank you for noting this. The chart review was performed after all screening processes described previously were done. We clarified this in our manuscript (please see Page 11).

P15L20: We found that women in the diagnostic codes group had more risk factors for suicidal behavior (Turecki & Brent 2016), including low socioeconomic status, being single, and psychiatric comorbidities as compared with those women in the NLP diagnostic group.

Does this mean that women with more risk factors are being coded, whereas those without the risk factors are not being coded - even when mentioned in the text? Some discussion of recording differences would be useful, as it throws light on what the differences are between structured and unstructured, and why both are important. Is there a coding bias?

We agree with the reviewer that coding bias could be one explanation. We have amended this in our manuscript (please see Page 19).

P16L7: how and why are the lists different? did you consider using their list, and do you know what your results would be if you had used their list?

The lists are different given the different purposes of studies. In the study of Haerian et al.[7], they had a more specific focus: suicide by overdose. They included many CUIs for overdose by different medications (e.g., C0573289 diphenhydramine overdose). In our study, we are interested in suicidal behavior, irrespective of methods used or considered. Nevertheless, there is a considerable proportion of overlapped CUIs in these two studies. Had we used the list from Haerian et al., we speculated the results in our study could go either way: we might find more women with suicidal behavior as we had a wider interest in terms of methods for suicidal behavior; we might find fewer women as we did not include more specific CUIs related to drug overdose.

P16L41: structured data only -> structured data alone

Thank you and we have fixed the sentence in question (please see Page 16).

P17L23: it would be useful to have some figures here, and some description of how you carried out your error analysis

Thank you for this comment. We added the description of how we carried out the error analysis (please see Page 17). We respectfully note that Table might provide a more structured way to illustrate our results in this case. As suggested by the reviewer, we added a table (Supplemental Table 6).

P17L28 "almost 50% of the clinical concepts are negated" where? in a sample studied by Chapman? this needs stating in some way, e.g. preface it with "in one study, ..."

We have clarified this in our manuscript (please see Page 17).

P17L47 trained in the Intensive Care Unit discharge summaries -> trained using Intensive Care Unit discharge summaries

Thank you and we have fixed it as suggested (please see Page 17).

P17L52: a considerable amount of suicidal behavior that were negated -> a considerable number of suicidal behavior terms that were negated

Thank you and we have fixed it as suggested (please see Page 18).

P18L4 onwards - figures for these error categories would be very helpful

We added a table (Supplemental Table 6) as suggested by the reviewer.

References:

1. Jagannatha AN, Yu H. Bidirectional RNN for Medical Event Detection in Electronic Health Records. Proc Conf. 2016;2016:473–82.
2. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in Neural Information Processing Systems 26. Curran Associates, Inc.; 2013. p. 3111–9.
3. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space [Internet]. arXiv [cs.CL]. 2013. Available from: http://arxiv.org/abs/1301.3781
4. World Health Organization. International Statistical Classification of Diseases and Related Health Problems. World Health Organization; 2004.

5. Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med. Cambridge Univ Press; 2012;42:41–50.

6. Fischer LR, Rush WA, Kluznik JC, O'Connor PJ, Hanson AM. Abstract C-C1-06: Identifying Depression Among Diabetes Patients Using Natural Language Processing of Office Notes. Clin Med Res. 2008;6:125–6.

7. Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRs. AMIA Annu Symp Proc. 2012;2012:1244–53.