Author's response to reviews

Title: Automatic Classification of Diseases from Free-text Death Certificates for Real-time Surveillance

Authors:

Bevan Koopman (bevan.koopman@csiro.au)
Sarvnaz Karimi (Sarvnaz.Karimi@csiro.au)
Anthony Nguyen (Anthony.Nguyen@csiro.au)
Rhydwyn McGuire (rmcgu@doh.health.nsw.gov.au)
David Muscatello (dmusc@doh.health.nsw.gov.au)
Madonna Kemp (adonna.Kemp@csiro.au)
Donna Truran (Donna.Truran@csiro.au)
Ming Zang (Ming.Zhang2@csiro.au)
Sarah Thackway (sthac@doh.health.nsw.gov.au)

Version: 5 Date: 26 March 2015

Author's response to reviews: see over
Response to Reviews – “Automatic Classification of Diseases from Free-text Death Certificates for Real-time Surveillance”
Bevan Koopman, Sarvnaz Karimi, Anthony Nguyen, Rhydwyn McGuire, David Muscatello, Madonna Kemp, Donna Truran, Ming Zhang, Sarah Thackway

Dear BMC Medical Informatics and Decision Making editor and reviewers,

Thank you for your detailed comments on our manuscript. A number of important issues were raised and have been addressed in the revised submission.

Our response to each reviewer’s comments is provided below in the following format:

• Original review comments are included verbatim in italic typesetting.
• The authors’ response is provided below in normal typesetting.

Reviewer 1

1. p. 7, line 39, the authors describe the "class confusion" error, in which the more specific codes were confused. Can they explain the source of the confusion? For example, was it based in [sic] insufficient information provided in the text, or lack of distinction in the code names?

We have added an explanation of the class confusion in the same section (pg. 8 line 16):

3. I am curious why the researchers did not include a spelling correction step. Or list of common misspellings in their system. As they note on p. 7, such a step would have eliminated many classification errors.

An accurate spell checker could potentially prevent some of the errors. However, spelling errors were partly fixed in the rule-based method and some were trained in the supervised method. One downfall of using spell checkers is that they may not be very accurate themselves and introduce their own errors, as well as not covering valid word variations which are not errors, but just different choice of typography. Finally, given the specialist medical vocabulary, a specialist medical spellchecker would be required; thus introducing other possible errors or changes.

4. In thinking about applications of such a classification system, is there any reason to treat recall and precision as equally important, or would there be some advantage in maximizing one at the expense of the other? For example, suppose false positives caused more work on the part of public health officials, or lead to faulty decisions – that would suggest that minimizing them has real value.

Importance of precision and recall is indeed relative to the application. We questioned the epidemiologists on a number of occasions about the important of precision vs. recall. For rarer causes-of-death (e.g., Influenza), they said recall was more important (it is important to not miss any cases). In contrast, for common causes-of-death (e.g., Pneumonia), they said precision was more important (as the review points out, false positives caused more work). Overall, given the range of causes-of-death, they felt that considering both recall and precision equally met their requirement (hence, the use of F-measure as the primary measure).

Some of the above explanation is provided on page 6 from line 30 onwards.
As a further comment, one way to favor precision or recall is by adjusting the decision threshold of the SVM. Increasing the decision threshold makes the classifier more conservative in assigning the positive class and therefore reduces false positives. In contrast, a lower decision threshold means more false positives but less false negatives. However, because the death certificates are so short, features tend to be binary rather than weighted and hence the likelihood of the SVM is not discriminative (i.e., the likelihood is either very close to 0.0 or very close to 1.0).

5. **Given the relative strengths and weaknesses of the machine learning and rule-based systems, is there a need to combine the two approaches in some way? Or is the plan to focus on improving just one?**

We are currently devising a methodology to effectively fuse results from the rules and SVM classifiers (we are investigating a number of fusion strategies in this regards).

In general, there is a tendency towards the machine learning based system because they are easier to adopt to other diseases, while for the rule-based method, the process of creating rules has to be started from the scratch for each disease that may be added. Also, rule-based methods can only contribute to identification of the disease of interest and lower level distinction (individual ICD10 classes) are best to be done with the supervised classifiers.

**Minor Essential Revisions**

6. p. 3, 34-36. The meaning of “lower timeliness” isn’t clear; the sentence itself is awkward. Please clarify.

   Sentence re-written on pg. 3 line 35 as:

   “While death is the ultimate outcome of disease and its occurrence lowers urgency of the case to be investigated, it is nevertheless still a high priority for syndromic surveillance”

7. p 6, 30 Typo: “Data did not”

   Fixed.

8. p. 7, 24-32. Paragraph is ungrammatical, please revise

   This has been revised on page 8, line 1.


   Fixed.

**Reviewer 2**

1. **Major Compulsory Revisions**

1.1 What is unclear to me is whether the methods are being employed for automatic classification of just the underlying cause of death or causes of death within the sequence leading to death (both the
underlying, and as I understand it, alternative causes of death). I suspect the latter, but it is unclear.

The reviewer’s assumption is correct: the methods are employed to identify ANY of the causes-of-death mentioned in the death certificate. For syndromic surveillance, it is desirable to consider any cause-of-death; the fact that it is underlying or not is less relevant. Page 2 line 40 has been updated to clarify this.

1.2 Table 4. Text defining Macro-average and Micro-average should be added as footnotes.

Both definitions are added to Table 4, page 17.

1.3 The discussion section should compare the results obtained in the present study with some of the relevant, related work cited in the background.

Paragraph beginning pg. 7, line 25 added comparing results to previous studies.

1.4 Table 4. The ‘ground truth’ numbers for influenza and pneumonia do not match between the rule-based and machine learning sections (for example for influenza in the rule based section there are 36 true cases while in the machine learning section there are 38 cases) shouldn’t these match?

Thank you for the astute observation. The error occurred when multiple rules fired for a single death certificate (e.g., a cause of death for Pneumonia and Influenza). This has been corrected and the evaluation measures updated. Note, there is a slight reduction in the f-measure for Influenza and Pneumonia.

2. Minor Essential Revisions

2.1 Page 4, line 12 "detailed features were extracted .." Could you be more specific on what was extracted? For example (based on the standard US death certificate), in addition to the immediate and sequential causes leading to death, were items such as location of death (in-hospital, DOA, ER), whether or not an autopsy was performed, whether or not tobacco contributed to death, etc extracted?

The features were terms, term n-grams and SNOMED CT concepts as explained in the “Feature Extraction Methods” immediately after this sentence. We only had access to the cause-of-death text so other features such as location of death were not available. We’ve added a comment on this in “Limitation and Future Work” (pg. 9 line 23).

2.2 Page 4, line 15 "The model was then be used to classify...") suggest changing this to "The model was then used to classify...".

Changed “be used” to “used”.

2.3 Page 4, line 36 "The vector comprised of binary values indicating..." suggest changing to "The vector comprised binary values indicating...".

Removed “of” as suggested.

2.4 Page 6, line 30 "Data did no contain variables..." suggest changing to "Data did not contain variables...".


2.5 Page 6, line 39. "The rule machine rule-based approach..." suggest changing to "The rule-based approach...".

Removed extra words “rule machine” as suggested.

2.5 Results and discussion, page 6 lines 45-46, the fact that recall is higher than precision may be more reflective of disease prevalence than false positives or false negatives that arise from classification. While the prevalence of diabetes and pneumonia are somewhat high (in terms of cause of death) prevalence of HIV and influenza are very low which accentuates the effect of false positives for any given value of sensitivity. A strength of the article that should be noted by the authors is that it does include causes of death that are fairly high in terms of prevalence and very low in terms of prevalence.

Indeed, the choice of diseases of interest (diabetes, influenza, pneumonia and HIV) was specifically made to cover both high prevalence and low prevalence cases. The Data & Experimental Setup section (Page 6 line 12) has been updated to highlight this. In addition, the discussion now highlights that the methods are effective across both high and low prevalence diseases. (Page 7, line 7).

2.6 Results and discussion. The authors should make note of any limitations in their study.

A “Limitation and Future” section has been added to pg. 9.

2.7 Table 5. Why are there no results for individual HIV causes of death? Table 3 suggested that all 5 HIV causes were suitable for individual classifiers.

In consultation with our epidemiologist collaborators, they advised that differentiating between specific HIV variants was not a requirement and that an overall HIV classification was what was required.

3. Discretionary Revisions

3.1 Page 5 line 38, would like to see the “Free-text cause of death ” sentence expanded to include that both the immediate and conditions leading to death were used as inputs to make it clear to the reader what is being done.

First bullet point on page 5, line 38 updated to clarity that both immediate and conditions leading to death were considered.

Associate Editor's Comment

All those from the two reviewers. In addition, the editor recommends the authors to better present the results in terms of classifier performance; following the concerns of reviewer #1, it’d be useful adding graphical representation of performance through precision/recall curves or ROCs, comparing with the other methodologies as suggested. In addition, confidence intervals from the extra-sample error estimations will be useful.

We believe we have address each item identified by the two reviewers. We have added a graphical representation in Figure 1 that provides a succinct overview of performance of the
different methods. (The full details of Table 4 are retained for completeness.)

For the ROC: Death certificates are very short so most feature weights tend to be a binary value (i.e., the feature appears in the death certificate or not). As a result, the SVM likelihood is not discriminative – you basically get a value very close to 0.0 or very close to 1.0. Hence the decision threshold makes little difference and ROC does not show anything.

On behalf of the authors,

Dr. Bevan Koopman.