Reviewer's report

Title: Identification of Pneumonia and Influenza Deaths using the Death Certificate Pipeline

Version: 2 Date: 8 January 2012

Reviewer: David Hanauer

Reviewer's report:

This paper is much improved and easier to follow. I have a few further comments and multiple grammatical corrections and suggestions:

DISCRETIONARY REVISIONS:
----------------------------------

Minor editorial suggestions:

Page 8: "had a processing rate of 95-97%"
--> What does this mean? An accuracy rate? It is not clear what a "processing rate" means in this context.

Figures generally have a brief caption. I assume they should have captions here as well? For example, in Figure 3 you could use the caption to state whether the electronic death certificates have any coded section or if all of the boxes are for free text entry.

Page 16: Is precision, recall, and F-measure calculated in various ways? If not (if it is standard) then these definitions can probably be removed from the paper to make it a bit shorter.

Table 2 is helpful, but takes up too much space. I think you can -snip- a large part of the text out and just show the most relevant sections of the XML output for the reader.

----------------------------------

Grammatical suggestions:

I have seen the term "key word" (with a space) and "keyword" without a space. Please be consistent.

Similarly I saw "P&I" as well as "P-I" being used.

Page 2, "However to this this type of data. . ."
-->Change to, "However, to use these types of data. . ."
Page 3, "... run in real time form with modest ..."
--> remove the word "form"

Page 4, "Change "from the 1635 to 1830s" to "from 1635 to the 1830s"

Page 4, in general people say "morbidity and mortality". On one part of this page it is reversed. I would suggested changing "mortality and morbidity" to "morbidity and mortality".

Page 4, "3 million persons acquire pneumonia and, depending on the severity of the influenza season, 15 to 61 million ..."
--> add two commas as shown above

Page 4, "Moreover, pneumonia and influenza together cost ..."
--> switch the location of "together" as shown above.

Page 5, "varies by states and, as a result, there is variability"
--> add the word "and" as shown above.

Page 8, "processed through the remaining components"
--> remove "by" as show above

Page 9, "capability to automatically code death certificates for outbreak and disease"
--> remove "of" as shown above

Page 10, "encodes death certificates using an NLP tool"
--> change "and" to "an" as shown above

Page 10, change "and or weekly counts to "and/or weekly counts"

Page 12, I would suggest breaking one sentence into two:
"The pipeline consisted of two components. The first component ..."

Page 14, Change "Table 3 and 4" to "Tables 3 and 4"

Page 15, "a sub-query was run" (change from ran to run).

Page 20, "thus yielding two codes" (remove the word "in")

Page 21, "needed to avoid overestimating" (remove the word "the")

Page 21, change sentence to "extracting causes of death from free-form"

Page 21, change to "present in the UMLS which resulted in good recall"

Page 21, change "system weakest point" to "system's weakest point"
**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Needs some language corrections before being published

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

**Declaration of competing interests:**

I declare that I have no competing interests