Reviewer's report

Title: Automated, sustainable, broad-based, near real-time public health surveillance using presentations to hospital Emergency Departments in New South Wales, Australia

Version: 1 Date: 7 June 2005

Reviewer: don weiss

Reviewer's report:

General
Well written with no glaring grammatical or spelling errors.
It is not clear what the authors set out to do and thus it is unclear if they have hit their mark.

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

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

Background
The authors cite the 2001 anthrax and 2003 SARS outbreaks as public health threats and as reasons for enhancing surveillance. Their sole stated objective however is to monitor patterns of ED visits. It is unclear how these are related. Later, they suggest that ED surveillance can be used as an early warning system. An early warning system is that which alerts a governmental authority of a potential problem in time to intervene. In subsequent sections such a system is not clearly described.

Methods
It would help to know what proportion of annual visits are seen at the twelve selected EDs.

It is mentioned that a batch system was created in addition to the HL7 messaging system, but it is not clear which system was operating at the time of the Rugby World Cup. If all EDs ran the same EDIS software and HL7 was created for one, then why would it not be available for all?

The physician-assigned provisional diagnosis codes were used to train the category classifier. Was this performed on archival or baseline data? If so, how much data? How was decision made to allow a single ED visit to be coded into multiple categories? How did this affect the results, what problems did multiple categories per visit create or alleviate? As I understand it, the assignment of an ED visit into a category is based on meeting a threshold probability value. The explanation of how the threshold was chosen is unclear. How is it done empirically and automatically? Sensitivity is said to be maximized, heretofore the determination of sensitivity has not been given. What was used as the gold standard?

The daily report made available to ED and public health personnel is described as 34 groupings of the principal physician-assigned provisional diagnosis. Further, that patient counts for the preceding day plus means for the prior 7 days and 3 weeks. Trends were presented by diagnosis group and ED using proportions (stratified by age, triage acuity and admission status) across the same three time periods. What proportion was used? Diagnosis group to all visits? Where trends presented for all hospitals combined? Where any of the cusum results presented? Why not? Werent these the main outcome? Other than the 95% confidence intervals, where any other measures of data aberration given? What did the public health authorities do with this information? Do you know if it was reviewed daily?
I am not familiar with Loess regression, perhaps other readers are not as well. Please provide an explanation of how this was done, what data was used plus a reference. I am familiar with day-of-week and seasonal effects, however what are day-of-the-week seasonal effects? How were they adjusted?

I understand that the preliminary success of automatic text classification was measured in two ways, a Pearson correlation coefficient and sensitivity. It is not clear how sensitivity was determined. Am I to presume that the provisional diagnosis assigned category was the gold standard? How did you handle the multiple classifications possible under your system? That a given ED visit could be assigned to both the provisional diagnosis categories cough and fever? How can the gold standard have two values? You dropped the ED visits that did not have a provisional diagnosis category assigned, how did this affect the analysis?

The description of cusum is not clear. The authors state that they developed a modified cusum to assess the statistical significance and amplitude of daily ED visits by syndrome and category (61 analyses per day?). Exactly how statistical significance is determined and at what level a signal is decided are not given. It is also stated that the expected value used is that of a single day, the preceding same weekday. My understanding of cusum is that a mean is used as the expected value and the standard deviation around that mean used to determine what gets added cumulatively to the score to determine if a threshold is exceeded. The convoluted double standardization that is presented appears to be adjusting for the fact that a single day was used for the expected value and that any excess over this for the seven days is added to compute the cusum. Using a baseline up to 365 days will not allow the cusum to float with the season, such that it will be very easy to signal during the prevalent season and very hard to signal in the off seasons. For example, if the mean for the year is 10 but in the last three weeks the mean 2, then a value of 10 wouldnt signal but in the context of the recent past it would be unusual. Since so many syndromes and categories were used, how was the problem of multiple comparisons handled? With 61 analyses done daily and a p value of 0.05, one would expect three signals daily.

Results

Nine and one half pages are devoted to describing the methods employed. Results are summarized in 1 pages. This seems wholly inadequate. Were there problems with data quality? Missing values? Was data received every day? Any transmission drops, errors, problems, glitches?

The authors report that no major public health issues occurred during the period of the world cup. What do they mean? None detected via the system or none via other surveillance systems? Or both? That no syndrome ever exceeded the cusum threshold during the 43 days? Or that they dismissed the signals either via investigation of their trivial nature? Authors say except for diarrhea all was within the usual ranges but never say how this was determined.

The abstract discusses the 27 categories the free text was classified into, where is this data? Where are the names, definitions and results on the 27 syndrome categories? What was found? How often were their signals? What were the ad hoc analyses done? Were these investigated by public health? How? How often? How was it decided what to investigate? Who got this information? What was learned?

Did the number of ED visits or traumatic injuries increase during the World Cup compared to the period before with the increased volume of spectators? Were there any independent measures of influx of people into the city? Was this looked at? Hotel reservations? Restaurant sales? Beer sales? What about the variable of place of residence? Can you be certain this system measured what you designed it to do during this time period?

Were there specific categories monitored for infectious agents? Which were these? Wasnt this the purpose? What is the point of Table 1? Was the main objective of the paper to compare a classifying
system using the provisional diagnosis and nurse triage free text fields? Why are there 36 categories when 34 are in the methods (p. 12)? The legend is unclear. Are the correlations presented the same as those mentioned in the text on page 13? Comparing the classification of nurse text and provisional diagnosis? Then the table needs to be labeled to clearly communicate this. The computation of sensitivity remains a mystery. In the discussion section the quality of the physician diagnosis is much maligned. Correlations seem particularly poor for infections symptoms such as cough (0.16) and influenza-like (0.09). These need to be commented on.

Title of figure 1 does not match that in the table which further does not match the text. All should be the same, All Respiratory.

The sentence, that the system provided welcome reassurance to public health personnel, is not a result but a commentary and does not belong here.

The last paragraph in the results section speaks of events that occurred outside of the World Cup period. These are not strictly results either, as no data are actually presented. It does not belong in the results section. Figure 2 is presumably a result but it is discussed only in the context of a finding outside of the time period for the study.

Discussion
The system is said to have been rapidly implemented, however, no evidence of the time it took was presented in the paper.

The data may be updated several times a day but if it is not analyzed and interpreted several times a day what is the benefit of this?

While some of the injury categories are relatively specific, most of the categories that relate to infectious diseases are not, furthermore, no evidence was presented to relate any of these categories to actual, culture proven diagnoses. So, exactly how does the system provide NSW Department of Health with the capacity to monitor acute disease in the community (or more appropriately, of the community) and permit mitigating public health responses? They authors appear to suggest that the system provides as much if not more information than does the reportable disease system. An assertion I severely doubt any public health official would support.

The authors on page 17 in one sentence say how wonderful it was to use both the preliminary diagnosis and free text but then really gone on to say that the diagnosis code had numerous problems. Why was this not quantified in the results section?

The final paragraph of this section on p. 20 raises the real quandary with syndromic surveillance systems. What do you do with the output? While the authors mention that the staffing levels and work practices may not be adequate, they fail to mention the invariably low positive predictive value of syndromic signals and the inherent difficulty in investigating them. The diversion of public health resources based on the system herein described has not been justified.

Conclusion
No mention of costs occurs in the text, we are asked to take the authors word that it was modest. The value of the system is overstated based on the evidence presented. Exactly how is the system to save lives, resources and money in the event of a SARS-like outbreak in New South Wales?
What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No

Declaration of competing interests:

The answer to the first four questions above are all no. We operate a public health syndromic surveillance unit and publish articles on our findings.