Author's response to reviews

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

Authors:

- David J Muscatello (dmusc@doh.health.nsw.gov.au)
- Tim Churches (tchur@doh.health.nsw.gov.au)
- Jill Kaldor (jkald@doh.health.nsw.gov.au)
- Wei Zheng (wzhen@doh.health.nsw.gov.au)
- Clayton Chiu (cchiu@doh.health.nsw.gov.au)
- Patricia Correll (patriciac@woolcock.org.au)
- Louisa Jorm (ljorm@doh.health.nsw.gov.au)

Version: 3 Date: 19 August 2005

Author's response to reviews: see over
Dear Sir/Madam

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

I am writing to provide a revised manuscript in response to reviewer comments. We are grateful to the reviewers for providing thorough and valuable feedback.

Our responses are shown in **bold** following the reviewers' comments:

**REVIEWER 1**

Discretionary Revisions (which the author can choose to ignore)
1. The paper appears to focus exclusively on temporal aggregation and the issue of spatial aggregation is not discussed. It is not clear whether each ED is regarded as a distinct geographic unit or whether all regions are amalgamated. In my experience, finer scale geographic localization is a practical challenge (eg does the post or zip code of residence serve as a reasonable proxy of the spatial component of the real event?) and may well be a very "noisy" covariate - however, it may be useful when there are relatively fine scale geographic events? Although this is probably another paper, the authors might point out that their system is purely temporal in contrast to some other more complex approaches.

We have inserted a paragraph in the Discussion on P25 discussing this point:

"Unlike other ED-based systems, we have not yet put a strong emphasis on geographically based statistical aberration techniques [refs]. For simplicity, we have elected to focus on the ED as a marker of geographic location rather than the patient's residential address, although we will be trying alternatives as further development of the system progresses. While the address of the ED attended is a rather coarse measure of the relevant geographic location, home address may also be of limited value because exposure to a biological agent could occur at work, at school or elsewhere in the community."

2. Although already present in the MS, I wonder if the authors might want to give even more emphasis to the simplicity, low infrastructure and opportunity costs, and low impact of the system on data providers, since in my experience, these are not widely understood as issues in the design and implementation of large scale, sustainable surveillance systems.

We have modified our conclusion to reflect this:

"Through secondary use of routinely collected ED information system data, a high degree of automation, and thoughtful use of readily available web-based reporting
technology, we developed a public health surveillance system that provides a much more timely, broad-based surveillance capability than was previously available in NSW.

REVIEWER 2

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. The Background has been substantially revised and reorganised to state more clearly the aims of the surveillance system, the aims of the manuscript, and how these aims are related to the demand for early warning systems in public health. The remainder of the paper has also been revised with this in mind to ensure the aims are met:

"We here describe the methods and review the early results from the new ED-based syndromic surveillance system implemented for the Rugby World Cup period. Our aims were to implement comprehensive, sustainable, and rapid population health surveillance that relied on data collected in ED information systems as part of their routine work flow. We also present the early results from automatic syndrome categorisation that uses presenting problem and triage nurse assessment text to assign syndrome categories."

Methods
It would help to know what proportion of annual visits are seen at the twelve selected EDs. This information is now included at the bottom of page 6 and the top of page 7:

"They account for almost one-third of the approximately 1.2 million ED visits captured annually in electronic ED information systems in NSW."

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? This is now clearly stated at the beginning of the results, page 18/19, with an explanation of why we did not implement HL7 more extensively early on:

"At the start of the Cup tournament on 10th October 2003, of the twelve hospitals selected for participation, four were providing data. One of these was providing real-time HL7 messaging and the remainder provided four to six hourly FTP batch feeds.
Given the greater complexity of installing HL7 messaging, we elected to install only the simpler FTP batch module at subsequent hospitals during the Cup period, to ensure that we could quickly complete our planned ED participation timetable.

There is also discussion of this point in the Discussion:
"...barriers to greater implementation of HL7 messaging include the need for a strong level of local information system support at participating sites, the variety of EDIS software versions from the same vendor that are implemented at different EDs (only recent versions have HL7 capacity), and the need for HL7 message routing infrastructure within the regional WAN. The FTP batch method, because it only relies on a standalone software module that communicates directly with our data centre, is much less dependent on the currency of EDIS software versions in use in each ED."

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? These are now clearly stated on the bottom of page 13. "Training of the classifier for each syndrome is performed using all available baseline records in the real-time surveillance database that meet the diagnosis-based syndrome criteria at the time the training procedure executes."

How was decision made to allow a single ED visit to be coded into multiple categories? This justification is now clearly stated on page 12:

"To exploit fully the often detailed and multi-faceted triage text that is available, a separate classifier is trained for each category. This means a single ED visit may be assigned multiple syndrome categories. This approach potentially allows monitoring of derived syndrome groupings that represent multiple symptoms occurring together; for example 'fever' plus 'cough'."

How did this affect the results, what problems did multiple categories per visit create or alleviate? This will have to be the subject of formal evaluations of the classification currently underway. The aim of this paper is to report early methods and results for our system.

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? Clarified on page 12:

"These values are chosen by our software tools by cycling through permutations of the number of features and probability thresholds and selecting the combination that maximises the Pearson correlation between the daily counts of the baseline computer-assigned syndromes and the diagnosis-based syndromes."

Sensitivity is said to be maximized, heretofore the determination of sensitivity has not been given. What was used as the gold standard? Added to page 13:

"The ED diagnosis-based syndrome was used as a gold standard for the sensitivity calculation."

"To exploit fully the often detailed and multi-faceted triage text that is available, a separate classifier is trained for each category. This means a single ED visit may be assigned multiple syndrome categories. This approach potentially allows monitoring of derived syndrome groupings that represent multiple symptoms occurring together; for example 'fever' plus 'cough'."
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? Answered on page 16:

"These summaries showed the proportion of ED visits in each time period for that syndrome that fell into: each of four age categories; a category that combined the most urgent triage acuity categories; and a category representing patients admitted to hospital for further treatment."

Were trends presented for all hospitals combined? Answered on page 16:

"In both types of reports, two levels of ED aggregation were presented for all statistical results. To evaluate geographically localised trends, statistics were available for each ED. To evaluate large scale trends, the same statistics were presented for the aggregated data for all EDs combined. Web page 'hyperlinks' were included in the web reports to navigate among the various reporting levels."

Were any of the cusum results presented? Why not? Werent these the main outcome? These were presented and this is stated on page 16. The new Figure 1 shows an example of these summary results:

"The provisional diagnosis summaries automatically prepared each day included counts of patient presentations for the previous day, the daily average count for the seven days up to and including the previous day, the daily average count for the three weeks prior to that, and the cusum-based index of increase (Figure 1)."

Other than the 95% confidence intervals, where any other measures of data aberration given? The statistics given are now clearly explained and an example given as per the above comments.

What did the public health authorities do with this information? Do you know if it was reviewed daily? The Public Health Response section of the Methods on page 17 now describes the way the reports were monitored and interpreted:

"During the Cup tournament, we monitored the automatic reports daily and performed ad hoc analyses to investigate signals. First, any signal was compared against a non-real-time, historical database of ED visits collected by the NSW Department of Health to determine whether a seasonal increase was expected. If an increase was expected then the signal was considered a false alarm. If the surveillance epidemiologist considered that seasonal influences did not adequately explain the increase, then a number of activities were undertaken to evaluate the alarm, including: evaluating the size of the counts that contributed to the signal; the magnitude of the index of increase; the proportions of patients falling within various subgroups based on age, admission status or triage category to determine whether unusual epidemiological features, including changes in the distribution of the severity of cases, were present; descriptive epidemiological analysis to compare the demographic characteristics of the period of increase with a control period, such as a recent period or the same period
in previous years from the historical ED database. Finally, if the epidemiologist judged that a concern remained after these analyses, the Rugby World Cup Public Health Committee or other relevant authority, such as the NSW Department of Health Centre for Drug and Alcohol, would be informed."

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. The description of the use of loess regression was removed because, as the reviewer points out, readers may not be familiar with the technique and we don't believe that maintaining its presence in the paper contributes anything substantive to the aims of the paper or interpretation of our results.

I am familiar with day-of-week and seasonal effects, however what are day-of-the-week seasonal effects? How were they adjusted? Day-of-week effects are an example of a seasonal effect that has a periodicity of 7 days. There is no requirement that the authors are aware of that states that seasonal effects have to exhibit annual periodicity. Nevertheless, we have rephrased the description of this seasonality to avoid confusion:

"Using the same weekday for comparing counts avoids the problem of comparing two successive days that may be subject to different day-of-week effects. These effects are strongly evident in ED visits and reflect lack of availability of other primary care services after hours and on weekends and varying social risk behaviours, such as alcohol consumption on weekends."

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? Yes, this is now described on page 13:

"The ED diagnosis-based syndrome was used as a gold standard for the sensitivity calculation."

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? Ideally, at least one of the automatically classified syndromes should match the diagnosis-based syndrome. We have expanded the description of the methods of this preliminary assessment as follows:

"While multiple text-based syndromes could be assigned to a visit, one of those syndromes would ideally match the diagnosis-based syndrome. In this preliminary assessment, we did not attempt to assess the accuracy of other text-based syndromes assigned to a visit. However, there should have been sufficient examples of all syndromes and diagnoses throughout our database to allow a reasonable assessment of the success of all the text-based syndrome classifiers."

You dropped the ED visits that did not have a provisional diagnosis category assigned, how did this affect the analysis? We discuss the limitations of the provisional diagnoses in the Discussion. On page 24, we have included a
statement that a better gold standard is required for more formal evaluation of the success of syndrome classification:

"More formal evaluation is required, one that uses a better gold standard than the provisional diagnoses, such as systematic classification by humans who are blinded to the final diagnosis."

The description of cusum is not clear. The description of the cusum has been substantially revised and expanded to make it clearer (page 14).

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?). Hopefully the revised description on pages 14 and 15 and the example of our report page as described above makes the volume of results clearer.

Exactly how statistical significance is determined and at what level a signal is decided are not given. The level of statistical significance was determined empirically in the early stages of our project (page 15):

"Based on empirical testing with many syndrome time series we heuristically selected an index value of ten as a signalling threshold of both statistical and practical significance."

We believe the technique again needs more formal testing, but we also believe that reporting these early methods will be valuable to others working in this field and it may provide useful ideas for others working in this area.

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 authors' understanding is that cusum techniques can be implemented in several ways, and that the original method assumes a normal distribution of the underlying process, but several papers report using alternative implementations. Examples of alternative implementations are included in the references we provide in this manuscript, such as reference 22 which describes 'Counted data cusums'.

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. We believe we have now more clearly explained and justified the 'convoluted' adjustment process on page 15:

"To make the cusum independent of the level and variability of the background syndrome incidence, two forms of standardisations are applied successively: the first for level and the second for variability. This allows the same interpretation of the degree of increase for different syndromes. The first standardisation is achieved by dividing the accumulated cusum value by the mean syndrome count for the available baseline up to a maximum 365 days. The second standardisation step was introduced because we found the response of this 'mean-standardised' cusum to be sensitive to the
variance of the baseline time series; series with high variance were more likely to signal inappropriately than those with low variance. We also found that one-day 'differencing' of the strongly autocorrelated mean-standardised cusum produced an approximately normally distributed series. Differencing is commonly used for detrending time series to allow appropriate parametric analysis of their behaviour. This final standardisation was therefore achieved dividing the mean-standardised cusum by the standard deviation of its differenced values, again using whatever baseline is available up to a maximum of 365 previous days. We called this fully standardised cusum an 'index of increase' as it offered a measure of growth in incidence that was comparable across different syndromes, independent of the size and variance of the original counts. Based on empirical testing with many syndrome time series we heuristically selected an index value of ten as a signalling threshold of both statistical and practical significance.

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 wouldn't signal but in the context of the recent past it would be unusual. Whether such a circumstance should result in a signal or not is a matter for argument that does not belong in this paper. This cusum is designed to signal during a period of low incidence if a sustained rise in incidence occurs, although in that situation it would not be quite as sensitive, as the reviewer argues. Our experience is that an overly sensitive signalling technique such as the simple Poisson count method we used will create false-alarm fatigue in those required to monitor the syndromes. The cusum method we designed grows as incidence grows, and will remain static if incidence remains static. We believe this is useful property of the index of increase because a rising value will continue to demand the attention of surveillance personnel. These benefits and limitations are now discussed in more detail in the Discussion, page 25:

"The cusum-based index of increase we developed appears to provide a more useful and easily interpretable assessment of a sustained increase in incidence than does the simple Poisson count method we used. We find that a rising value of the index will continue to demand the attention of surveillance personnel because this means incidence is continuing to rise. The Poisson method was oversensitive and appeared to trigger false alarms simply because of day-of-week effects and because of chance increases that are inevitable when making so many multiple comparisons. Because we are monitoring many health categories, it was important to have a measure that did not require a different interpretation for each category. There are other advantages to the cusum approach: its use of the previous weekday count instead of a longer term average for the expected value means it only needs a short baseline to become useful, as long as the baseline time series is operating under non-epidemic conditions. Shorter baseline methods have been shown to be as effective as those requiring several years of baseline. The missed diarrhoeal and pertussis outbreaks before and during the Cup highlight the risk of relying on too short a baseline in the early life of a system, however. Unlike the traditional cusum, our modified cusum quickly returns to baseline levels when the underlying process does, without needing to be reset. It's value provides both a threshold for triggering, but after it triggers it can be monitored to assess the relative and ongoing scale of continued increase in incidence. Its primary
disadvantage is that it does not adjust for annual seasonal trends. We therefore have to manually examine historical trends to assess whether seasonal increases are progressing beyond seasonal levels. In some cases this is not a disadvantage. We believe it useful to trigger an alert at the start of the annual influenza season, for example, even though it occurs every year. The impact of the circulating influenza on ED visits as marked by the index of increase may provide a useful assessment of the virulence or degree of transmission of the circulating influenza strain, although this needs evaluation. A related disadvantage is that a sustained seasonal increase will affect the 365-day mean used in the cusum's adjustment, making it less sensitive to subsequent increase in syndrome incidence. However, the moving 365-day mean changes more slowly than the short-term increase that the cusum is designed to detect, making this less of a problem. The choice of a maximum 365-day baseline was a balance between preventing the standardised cusum from being unduly influenced by long-term secular trends while not being too sensitive to short term trends. The cusum continues to provide a useful signalling tool while we explore more sophisticated statistical methods."

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. We believe that because the cusum responds to sustained rises in incidence this problem is reduced and that it can be controlled through the choice of alert threshold, although further discussion of this point beyond what is already stated above is beyond the scope of this paper. We do discuss it in relation to the Poisson count method we used:

"The Poisson method was oversensitive and appeared to trigger false alarms simply because of day-of-week effects and because of chance increases that are inevitable when making so many multiple comparisons."

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? This is a valid point and we have included more results that discuss these points and provided more examples in Figures 1 and 2 to demonstrate the results:

"At the start of the Cup tournament on 10th October 2003, of the twelve hospitals selected for participation, four were providing data. One of these was providing real-time HL7 messaging and the remainder provided four to six hourly FTP batch feeds. Given the greater complexity of installing HL7 messaging, we elected to install only the simpler FTP batch module at subsequent hospitals during the Cup period, to ensure that we could quickly complete our planned ED participation timetable. Nevertheless, the HL7 site provided data continuously throughout the Cup. By 13th November, all 12 EDs were providing data.

We have records of data completeness for 37 of 44 days (84%) in the Cup period. A median of 93% (range: 58% - 100%) of the ED visits actually occurred were captured in the system by the time the automated report was generated and checked each day. Median completeness of the provisional ICD-coded diagnosis variable was 76%
While 12 EDs were participating, the median number of ED visits reported per day was 997 (range: 941-1077).

Reports were available for checking on every day during the Cup tournament. Figure 1 shows an example of a report page automatically generated for the final day of the Rugby World Cup (22 November 2003), based on coded diagnosis-based syndromes. Small increases in incidence for that date are evident using the Poisson-based confidence intervals for the 'Head injury', 'Open wounds', 'Alcohol-related provisional diagnosis', and 'Skin problems' syndromes. These were not considered sufficient to cause concern, particularly because the Poisson method we used did not adjust for day-of-week effects. The cusum-based index of increase, which does take into account day-of-week influence, supported this conclusion. An example from the automated text summary web page for that day was:

"The index of increase was incorporated into our automatically generated reports on 29 October 2003. From then to the end of the Cup, we observed 21 signals based on a threshold of ten. This represented an average of 0.8 signals per day, range 0-3 per day. After evaluation, as described in the Methods section above, none of these signals were considered by the surveillance epidemiologist to warrant further action."

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 results shown above address these points. Plus we have added additional information that provides support for this comment, but also provide information that highlights a limitation of the system in having a baseline that was too short to allow detection of problems that were already in progress before monitoring started (pages 20-21):

"Subsequent review of the non-real-time ED database held by the Department revealed a larger-than-usual seasonal increase in diarrhoea-related ED visits in children that had peaked prior to the beginning of the Cup. This coincided with a seasonal increase in rotavirus identification by a major public health laboratory. Examination of the cusum algorithm revealed that the index of increase did not reach a statistically significant value because the diarrhoea-related syndrome counts were already well above background levels when real-time monitoring started, and the available baseline (starting 1 September 2003) was too short to cause the cusum to signal.

The only known outbreak of a reportable disease in progress in urban areas during the months of the Cup was a state-wide outbreak of pertussis. Again, this had commenced prior to the start of the baseline data for the real-time system and as a result the system did not detect an increase in the provisional diagnosis-based pertussis syndrome."

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? The syndromes are listed in detail in Table 1. We do not understand what the reviewer means, although we have modified the column
headings and titles for Table 1 to hopefully make it more relevant to the text. We have also revised the list of syndromes to reflect what was current at the time of the Rugby World Cup; we have since added more syndromes.

How often were their signals? **This is presented in the results for the cusum as shown above. We did not present these results for the Poisson count method, because as the reviewer correctly asserts, the cusum was our primary method of alerting.**

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? **These points are now address in the Public Health Action section of the Methods, as already described above.**

What was learned? **We believe the Discussion in the current version of the manuscript adequately describes 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? **We had no evidence of this. Having experienced the Olympics in Sydney in 2000, we anecdotally had the impression that the Rugby Cup was dramatically smaller in scale in terms of its impact on the city. Nevertheless, because we do not have any evidence of such an impact we have not presented any results that suggest such an impact.**

Were there any independent measures of influx of people into the city? Was this looked at? Hotel reservations? Restaurant sales? Beer sales? **We report in the Introduction the available measure of visitor influx from the Australian Bureau of Statistics:**

"During the months preceding and during the Cup, the number of short-term overseas visitors to Australia increased by approximately 60,000 compared with the same period in the previous year."

What about the variable of place of residence? **We did not rely on this in the early life of the system, but we are exploring its value as time goes on. We discuss geographic issues in the Discussion, page 26:**

"Unlike other ED-based systems, we have not yet put a strong emphasis on geographically based statistical aberration techniques. For simplicity, we have elected to focus on the ED as a marker of geographic location rather than the patient's residential address, although we will be trying alternatives as further development of the system progresses. While the address of the ED attended is a rather coarse measure of the relevant geographic location, home address may also be of limited value because exposure to a biological agent could occur at work, at school or elsewhere in the community."

Can you be certain this system measured what you designed it to do during this time period? **The extended results hopefully provide more information to help answer this question. The system requires further evaluation, but we believe it is of**
sufficient value to warrant reporting our early methods and results to add a 
contribution to this field.

Were there specific categories monitored for infectious agents? Which were these? 
These are included in Table 1. Wasnt this the purpose? The primary aim of the system was bioterrorism that can manifest as a chemical, biological or radiation incident. However, for mass gatherings, we traditionally focus on a broader range of issues, hence the broad scope of our system. During the Olympics in 2000 in Sydney, there were a number of injury problems identified by ED-based surveillance, for example, so we included these broader syndrome topics in our newer system.

What is the point of Table 1? The point of Table 1 is hopefully clearer now - it has several uses in the paper, both to show the syndrome definitions used and the results of the evaluation of the text-based classification:

"Table 1 - Provisional diagnosis and triage text-based syndromes used for statistical reporting and for training the automatic triage text classification system, with correlation between daily counts from the two categorisation methods and the sensitivity of the automatic text classifier in assigning a category that matched the provisional diagnosis category."

Was the main objective of the paper to compare a classifying system using the provisional diagnosis and nurse triage free text fields? As discussed above, the aims are now more clearly stated in the Introduction.

Why are there 36 categories when 34 are in the methods (p. 12)? We have reviewed the number of categories that we report and now a consistent set of numbers is reported throughout the paper.

The legend is unclear. Are the correlations presented the same as those mentioned in the text on page 13? Yes.

Comparing the classification of nurse text and provisional diagnosis? Then the table needs to be labeled to clearly communicate this. See comments above.

The computation of sensitivity remains a mystery. This point was addressed above.

In the discussion section the quality of the physician diagnosis is much maligned. The Discussion now provides a much more balanced discussion of the relative value of diagnosis and text-based syndromes (pages 23-24):

"Diagnosis and free text-based syndromes each have advantages and disadvantages for use in syndromic surveillance. While the diagnoses can be more specific than automatically classified syndromes, they are entered by busy ED staff who are not trained in medical coding and therefore their accuracy can vary. Secondly, diagnoses are not entered until the end of the visit, or some time thereafter, and are sometimes not entered at all, as evidenced by the one quarter of ED visits that did not have a diagnosis at the time our daily reports were checked during the Rugby World Cup."
The triage text is routinely entered in the ED information system shortly after the patient's arrival at the ED and is therefore available for syndrome classification as soon as it arrives in our database. Thirdly, depending on the level of diagnostic certainty available at the end of the ED visit, the diagnosis entered may reflect a symptom or a specific diagnosis, but not both. For example, a patient presenting with chest pain, who is later diagnosed with a cardiac event, may be assigned a principal diagnosis code that represents only chest pain, or a diagnosis code that represents the cardiac event. This is an additional, unwanted source of variability. Fourthly, a mean of only 1.1 diagnosis codes are entered in the ED information system for each patient presentation to an ED. Thus, only a limited part, if any, of the presenting syndrome may be reflected in the ICD codes that are recorded. In contrast, the free text classification algorithm can assign as many categories as are warranted by the textual description., thereby potentially offering a fuller picture of the presenting syndrome. Finally, new text-based syndrome categories can be introduced into the system and retrospectively assigned to ED visits as required, simply by training a classifier and reprocessing historical text information through it."

Correlations seem particularly poor for infections symptoms such as cough (0.16) and influenza-like (0.09). These need to be commented on. See Discussion page 24. We have removed any reporting of the cough syndrome because this was not available at the time of the Rugby World Cup:

"Based on this preliminary evaluation, automatic classification of triage nurse text into syndromes worked well for many symptoms and syndromes, particularly those with large counts and broad meaning, such as injury and respiratory disease. Specific diagnoses such as pertussis and pneumonia showed poor results, probably because it is unrealistic to expect the classifier to identify specific diagnoses given only the information available at triage. This confirms that automated text classification is best used for syndromes that are symptom rather than diagnosis-based. Other categories, particularly those with small counts such as “respiratory distress”, were not classified well."

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. Done.

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

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. Agreed. These results were removed.

Discussion
The system is said to have been rapidly implemented, however, no evidence of the time it took was presented in the paper. This can be discerned from dates provided in the Introduction and Results, but is now explicitly stated at the beginning of the Discussion:
"In just over three months from the commencement of funding, we implemented an automated syndromic surveillance system of broad scope that took maximal advantage of existing information systems and communications infrastructure."

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 we did not take full advantage of this in the initial system, it does offer greater real-time monitoring potential, as now stated on page 22:**

"While current reporting is up to midnight on the previous day, near instantaneous HL7 messaging and/or frequent automated batch feeds give us the future flexibility to report on trends within recent hours rather than the most recent day."

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. **As described above, pertussis is now mentioned in the results. Further detailed evaluation of this question will need to be the subject of future papers. For example, we are currently undertaking an evaluation of the sensitivity and specificity of ED diagnoses for culture-proven meningococcal diagnoses, but this is just one disease and including these results would add further size to an already long paper and would further delay its publication.**

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? **More correctly it provides NSW Health with the capacity to monitor a greater range of syndrome-based health trends than before. We now provide a much more measured Discussion of the advantages and disadvantages of this kind of surveillance system (page 26-27):**

"While the system has since proven useful for a range of public health activities and problems as diverse as influenza surveillance, short-term increases in acute asthma incidence in children and in recreational drug misuse, more formal, objective evaluation of its likely performance in recognising the disasters that it is primarily designed to prevent is required.

... Public health services are also challenged to develop appropriate responses to the information generated by new, non-traditional public health surveillance systems such as ours. As we have demonstrated, it is relatively easy to establish a system which can collect large volumes of public health surveillance data with very little delay. It is much more difficult to determine, on a continuous basis, whether that data contains information which demands public health action, or more likely, warrants further investigation. Near real-time information demands a near real-time response, which may be difficult to achieve with existing staffing levels and work practices. Such systems also potentially generate information on issues that may not have previously been the subject of rapid, organised public health responses, such as clusters of presentations related to illicit drugs, specific causes of injuries or asthma. Further, identifying and ascribing syndrome trends to specific causes or organisms remains a
challenge that is the subject of ongoing evaluation and development. It is difficult to mount a public health response in response to an increase in the incidence of a symptom that in most cases represents a naturally occurring, self-limiting illness. Identifying that the observed increase relates to the early activity of an introduced bioterrorism agent requires extremely careful judgement on the part of surveillance system staff and extensive knowledge of the epidemiology of the non-specific syndromes they are monitoring. For this reason, in judging the importance of any unusual syndrome activity, we rely heavily on variables from the ED information system that mark the severity of the presenting syndrome, as well as the scale of an apparent increase in syndrome activity."

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. **Agreed. The revised discussion shown above should limit this impression. We also now explicitly state that the system complements existing communicable disease surveillance (Conclusion page 28):**

"This does not replace notifiable communicable disease surveillance, but complements its greater specificity but narrower scope."

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. **This issue was addressed above. A much more balanced assessment of the advantages and disadvantages of each method is now discussed.**

Why was this not quantified in the results section? **Quantification is now included in the additional results described above.**

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. **Agreed. The Discussion is now much more measured in its assessment of the value of this system, as mentioned above in response to the reviewer's point about how the system increases NSW Health's capacity to monitor acute diseases. For example, we now say:**

"It is difficult to mount a public health response in response to an increase in the incidence of a symptom that in most cases represents a naturally occurring, self-limiting illness. Identifying that the observed increase relates to the early activity of an introduced bioterrorism agent requires extremely careful judgement on the part of surveillance system staff and extensive knowledge of the epidemiology of the non-specific syndromes they are monitoring. For this reason, in judging the importance of any unusual syndrome activity, we rely heavily on variables from the ED information system that mark the severity of the presenting syndrome, as well as the scale of an apparent increase in syndrome activity."
No mention of costs occurs in the text, we are asked to take the authors word that it was modest. **Discussion of cost was removed.**

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? **See points above about our more measured Discussion of the merits of the system.**

I hope these changes are satisfactory and look forward to your response.

Yours sincerely

David Muscatello
Senior Epidemiologist
Centre for Epidemiology and Research