Syndromic surveillance on the Victorian chief complaint data set using a hybrid statistical and machine learning technique

Hafsah Aamer[1], Bahadorreza Ofoghi[1], and Karin Verspoor[1,2]

[1]Department of Computing and Information Systems
[2]Health and Biomedical Informatics Centre
The University of Melbourne
Melbourne, Victoria, Australia

Abstract

Emergency Department Chief Complaints have been used to detect the size and the spread of disease outbreaks in the past. Chief complaints are readily available in digital formats and provide a good data source for syndromic surveillance. This paper reports our findings on the identification of the distribution of a few syndromes over time using the Victorian Syndromic Surveillance (SynSurv) data set. We utilized a machine learning-based Naïve Bayes classifier to predict the syndromic group of unseen chief complaints. Then, we analyzed the patterns of the distributions of three syndromes in the SynSurv data, specifically the Flu-like Illness, Acute Respiratory, and Diarrhoea syndromes, over sliding windows of time using the EARS C1, C2, and C3 aberrancy detection algorithms. The results of our analyses demonstrate that applying aberrancy algorithms over the variance data between two consecutive weeks reduces the large number of possible disease outbreaks detected using raw frequencies of the syndromic groups in the same time period, resulting in a more feasible approach for practical syndromic surveillance.

1 Keywords

Syndromic surveillance, Disease outbreak, Aberrancy detection, Chief complaints.

2 Introduction

The risk of bio-terrorist attacks and pathogenic diseases such as SARS have resulted in an increased need for ongoing syndromic surveillance of life-threatening diseases. During the past few decades, Natural Language Processing techniques have been widely used in biosurveillance for detecting disease outbreaks from health-related data sets [2]. Chief complaints and triage notes at an Emergency Department provide a good data source for the detection of such disease outbreaks. A chief complaint is a short summary of symptoms experienced by the patient at the time of arrival at an Emergency Department. The confirmation of an infectious disease or pathogen-borne illness usually requires extensive laboratory testing and may take days. This is a time consuming process which may result in the late-identification of a significant disease outbreak; however, with constant monitoring of chief complaints the usual delays in the process can be significantly reduced.

Syndromic surveillance has a focus on following the size, spread, and tempo of outbreaks with the aim of monitoring disease trends and providing reassurance that an outbreak has not occurred [11]. If chief complaints can be classified into syndromic groups, then abnormally high visit rates with a specific syndrome can easily send an early signal of a disease outbreak in real-time. After the anthrax letter threat in 2001, a range of projects were initiated to classify chief complaints into syndromic categories and detect disease outbreaks using supervised machine learning methods [6, 5, 3, 19]. Supervised algorithms use existing chief complaints and develop a classification system to predict the syndromic group of new unseen chief complaints.

In this work, we aim to find disease outbreaks through the analysis of chief complaint texts in the emergency department. We approach this task using a combination of supervised machine learning methods and statistical aberrancy detection algorithms over shifting windows of time. We analyse the performance of the approach using the Victorian Syndromic Surveillance (SynSurv) data set, and consider the effectiveness of the algorithm

to retrospectively identify outbreaks. We compare the raw frequencies over consecutive time periods with the use of the differences between total frequencies of the positive classified syndromes in detecting aberrancies.

3 Methods

3.1 SynSurv Data Set

The SynSurv syndromic surveillance data was collected from the Emergency Departments of two of the main hospitals in the city of Melbourne, Australia, the Royal Melbourne Hospital and the Alfred Hospital. The data was collected on behalf of the Victorian Department of Health, initially to enable monitoring during the 2006 Commonwealth Games held in Melbourne. The data covers the period July 2005 to August 2009. The syndromes that we focused on in the data set were Flu-like Illness, Diarrhoea and Acute Respiratory. The SynSurv data set contained a total number of 314,630 chief complaints already labeled with one of the syndromic groups. In the SynSurv data set, a chief complaint can be a part of more than one syndromic group at a time. For example, a chief complaint can belong to the Flu-like Illness category as well as Acute Respiratory at the same time since these two diseases are very similar. Table 1 summarizes the distribution of the different syndromic groups in the SynSurv data set. The data was split into training and testing portions by using 2/3 of the total data set for each syndrome for training and the remainder for testing.

Syndromic group	#Training records	#Testing records
Flu-like Illness	11,398	5,829
Acute Respiratory	7,431	3,877
Diarrhoea	5,066	2,601
Other	185,965	92,462
Total:	*209,860*	*104,769*

Table 1: The distribution of chief complaints over various syndromes in the training and testing data sets

3.2 Data Pre-processing

Chief complaints in the SynSurv data set are in the free text format; therefore, they require some pre-processing before they can be used as training data for the classification process. First of all, the ID field was removed from the daily chief complaints. The ID field is unique for each chief complaint and does not represent a meaningful and informative feature for the Vector Space Model used for classification. Then, the stop-words were removed from the texts of the chief complaints using the set of the English stop-words presented by van Rijsbergen [20] since the stop-words do not play any role in distinguishing between syndromic groups. The chief complaints also had already been expanded with terminological codes from the two biomedical vocabularies; SNOMED and ICD-10. To keep the chief complaints in their original format, the already assigned ICD-10 and SNOMED codes were removed from the chief complaints. The texts of the chief complaints were then lowercased and tokenized to be utilized in a bag-of-words model.

3.3 Chief Complaint Classification into Syndromic Groups

To analyze the distribution of syndromes over a period of time, the first step is to classify the chief complaints reported in that period of time into specific syndromic groups. This requires either an off-the-shelf classifier to be applied on the data or a machine learning-based classifier to be trained with some pre-labelled chief complaints. In a previous work [16], two existing North American chief complaint classifiers, Symptom Coder (SyCo) [7] and Complaint Coder (CoCo), were tested on the SynSurv data set. These machine learning-based classifiers are parts of the Real-time Outbreak and Disease Surveillance system (RODS) [8]. The results of those experiments showed moderate performances by the two classifiers on the SynSurv data set; SyCo outperformed CoCo achieving a highest F-measure of 0.432 on the Flu-like Illness syndromic group.

In this study, a new classifier was trained on the SynSurv data set for chief complaint classification. Once the chief complaints were pre-processed, the Naïve Bayes classifier in MALLET [15] was utilized for this experiment. MALLET is a package of statistical machine learning and natural language processing algorithms developed in Java. MALLET provides a pipeline of necessary processes to prepare the data for classification purposes. Built-in tokenization and conversion of the token space to a vector space model were among the processes we included in the MALLET pipeline. In this case, a bag-of-words model (tokens only) was used with no additional features from the chief complaints. The Naïve Bayes classifier was trained on the set of chief complaints in the training portion of the SynSurv data set and its performance was evaluated for each of the three syndromic groups with the testing portion of the SynSurv data set. Table 2 summarizes the performances achieved on the SynSurv data using the MALLET Naïve Bayes classifier.

As shown in Table 2, the results of the new classifier in this study show an improvement over those reported on the same data set in [16] for the three syndromic groups.

Syndromic group	Precision	Recall	F1-measure	Accuracy	F1-measure(prev.) [16]
Flu-like Illness	0.534	0.754	0.625	0.950	0.432
Diarrhoea	0.446	0.512	0.477	0.972	0.295
Acute Respiratory	0.433	0.658	0.522	0.955	0.332

Table 2: The results of the Naïve Bayes classification method for the three syndromic groups Flu-like Illness, Diarrhoea, and Acute Respiratory. The last column shows the results of the best-performing system (i.e., SyCo) in the previous work on the same data set.

3.4 Aberrancy Detection Algorithms

Statistical aberrancy detection algorithms have been a vital method for syndromic surveillance [4]. These algorithms can detect large sudden deviations of occurrences of specific events that significantly depart from the norm over time. We applied Early Aberration Reporting System's C1, C2, and C3 aberrancy detection algorithms [12] to the syndromic group data extracted for shifting windows of time over the SynSurv data set. The details of this procedure are given in the next section.

The formulas to calculate the EARS C1, C2, and C3 algorithms can be found in [4, 10]. Briefly, they are based on a calculation of the occurrence of specific diseases, relative to an expected value for that occurrence. The C1 algorithm requires a 7-day baseline data starting from $t-7$ to $t-1$, where $t = \textit{present day}$, to calculate the mean and standard deviation over a sample. It then calculates how much the value at day t varies from the expected value. If the variance exceeds a pre-set threshold, then an aberrancy in the data is detected. The C2 algorithm adds a 2-day lag to the baseline, starting from $t-9$ to $t-7$, while C3 uses the current and previous two values of C2 to detect possible aberrancies. The thresholds for C1 and C2 were set to $[(\textit{sample mean} + (3 \times \textit{sample standard deviation})]$; $C1(t) > 3$ and $C2(t) > 3$; while for the C3 Algorithm, any $C3(t) > 2$ would signal an aberrancy at t.

The C2 algorithm is known to perform better on serially correlated data [21]. For comparison purposes, we implemented all of the three aberrancy detection algorithms on the SynSurv data set.

4 Experiments

4.1 Aberrancy Detection Setup on SynSurv

The trained classification system (MALLET's Naïve Bayes) was used to predict the syndromic group of the chief complaints in the test SynSurv data set. Then, the positive classified cases for each syndrome and the dates they occurred on were tracked. Once we knew how many cases of Flu-like Illness, Diarrhoea, and Acute Respiratory syndromic groups occurred per day in the SynSurv test data set, we calculated the differences between the frequencies of cases for each syndromic group over shifting windows of time. The time windows were seven days long, and each subsequent window was shifted by one day. Therefore, the first window of time was from day 1 to day 7, the second from day 2 to day 8, and so on. This procedure formed the data that we refer to as the *predicted* data set. A similar procedure was utilized on the same SynSurv test data set, but using the actual gold standard labels of the chief complaints rather than text-based predictions to derive syndrome frequencies; we refer to this as the *actual* data set. Finally, the aberrancy detection algorithms C1, C2, and C3 were applied to both Predicted and Actual data sets to find any outbreaks of the three syndromic groups in the SynSurv data set.

Since C1, C2, and C3 algorithms do not require prior training, their application introduces a hybrid approach combining unsupervised statistical methods with supervised classification techniques. This hybrid method will enable the analysis of large volumes of data collected at emergency departments and will draw health practitioners' attention to any statistical aberrancies in the data that could indicate significant outbreaks.

The aberrancy detection algorithms were also applied to raw frequencies of each syndromic group per window of time, in addition to the differences between the consecutive time windows. Therefore, we discuss two types of methods here: i) the *raw frequency method*, that focuses on the raw frequencies of positive cases of each syndromic groups over a period of time, and ii) the *variance method*, that considers the differences between the frequencies in consecutive time windows for positive cases of each syndrome. We treated the raw frequency methodology as the baseline method for comparison with new variance method.

4.2 Results and Discussion

Before applying the aberrancy detection algorithms using the raw and variance methods, we wanted to understand how the distribution of the positive cases of each syndromic group compare using both the classification system and the gold standard labels. The positive instances were counted and the summary result is shown in Figure 1. As shown in this figure, during the syndromic class prediction process, the chief complaint classifier produced a number of false positives as indicated by the larger numbers of the positive instances compared

with those of the actual cases for all of the three syndromic groups. However, the trends in the predicted cases followed those of the actual cases, i.e., the distribution of false positives is uniform.

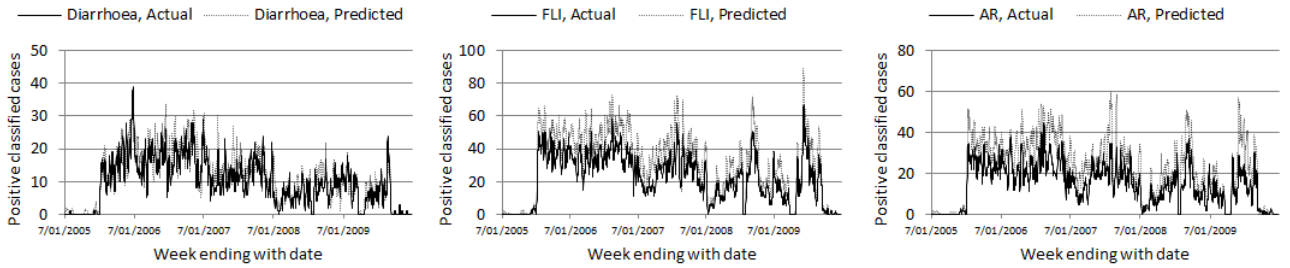

Figure 1: The raw frequency of actual and predicted positive classified cases of the three syndromic groups starting from the week ending 7/01/2005. Note: FLI=Flu-like Illness, AR=Acute Respiratory.

Since we are using aberrancy detection algorithms, a uniform distribution of false positives will not negatively affect the detection process. The aberrancy detection algorithms consider the mean and the standard deviation of time windows and set a threshold accordingly before indicating an abnormal deviation in the number of positive cases of the syndromes. The false positives are therefore adjusted by the sample means and sample standard deviations. In other words, the aberrancy detection algorithms can play a pivotal role in situations where the underlying data are noisy and the classification system produces a reasonable amount of false positives.

Another advantage of using the EARS aberrancy detection algorithms is that only a 7-day baseline is required. This helps to quickly identify any health-related outbreaks within a short period of time without the need for a longer-term data background. Moreover, due to varying seasonal trends and varying numbers of patients visiting medical centres over weekends and weekdays, a more systematic way of detecting disease outbreaks is required. Aberrancy detection algorithms provide inherent measures to control such false alarm rates and the different algorithms are categorized according to their sensitivity in finding abnormal deviations; C1 having *mild*, C2 *medium*, and C3 *ultra* sensitivity[13] for detecting aberrancies.

We applied the aberrancy detection algorithms in the way discussed in the last section to the SynSurv data set. C3 detected approximately 400 aberrancies for each syndrome using the raw data. Analysing each of these 400 alerts is an intensive process for a health practitioner. Although this large number of aberrancies were cut to nearly half for the variance data, the number of aberrancies is still reasonably high. Therefore, we used the C2 algorithm as it has medium sensitivity, resulting in more manageable numbers of aberrancies. Figure 2 depicts the results for the C2 algorithm on the predicted data only for space limitations.

From Figure 2, it can be seen that there are a large number of peak lines above the threshold value 3 on the left, while the curves on the right have a smaller number of peaks crossing the threshold value detecting possible outbreaks. Therefore, finding the differences in the number of diseases between two consecutive weeks, i.e., the variance method, results in a less noisy output and is a more feasible method for finding possible outbreaks compared with using raw frequencies between consecutive time periods.

It is difficult to assess whether the detected aberrancies represent real disease outbreaks. To address this, we compared the onset dates detected by the C2 algorithm with the predicted data (the variance method) with the Australian National Influenza Surveillance Scheme reports, which we consider to be the best available source of information pertaining to influenza outbreaks.

In 2005, our system predicted aberrancies in the SynSurv data in early July, which falls in the flu season in Australia (i.e., from June to August each year). "Influenza infections are seasonal in temperate climates (June to September in the Southern Hemisphere and December to April in the Northern Hemisphere)" [9]. Another detected date was in the week ending October 2, 2005. "The 2006 Australian influenza season was mild in comparison to previous years and was predominantly due to influenza A infections" as reported in the Influenza Annual Report 2007 [18]. In 2006, our system (even with an increased threshold above 3) did not find any outbreak. In 2007, the dates detected were in September only. The 2008 influenza also followed the traditional flu season pattern [14]; however, there was a gradual increase in notifications above non-seasonal levels from much earlier in the year. The out of season dates our system detected in 2008 also started from mid January to the end of February. For 2009, we detected aberrancies in most of April and mid August, again within the flu season. According to [1], Flu-like Illness presentations to emergency departments remained steady and slightly above background levels in 2009.

It should be noted that the Australian Annual Influenza reports make use of various surveillance methods including reports from emergency departments, general practitioners, and laboratory confirmed cases all over the country. The SynSurv data, however, includes Victorian emergency department data only which may not be a comprehensive representation of the national data. We interpret our results to be reliable as long as the detected aberrancies fall within the flu season of Australia. No official reports exist for the Acute Respiratory and Diarrhoea groups and therefore we cannot directly assess performance of the method for these diseases.

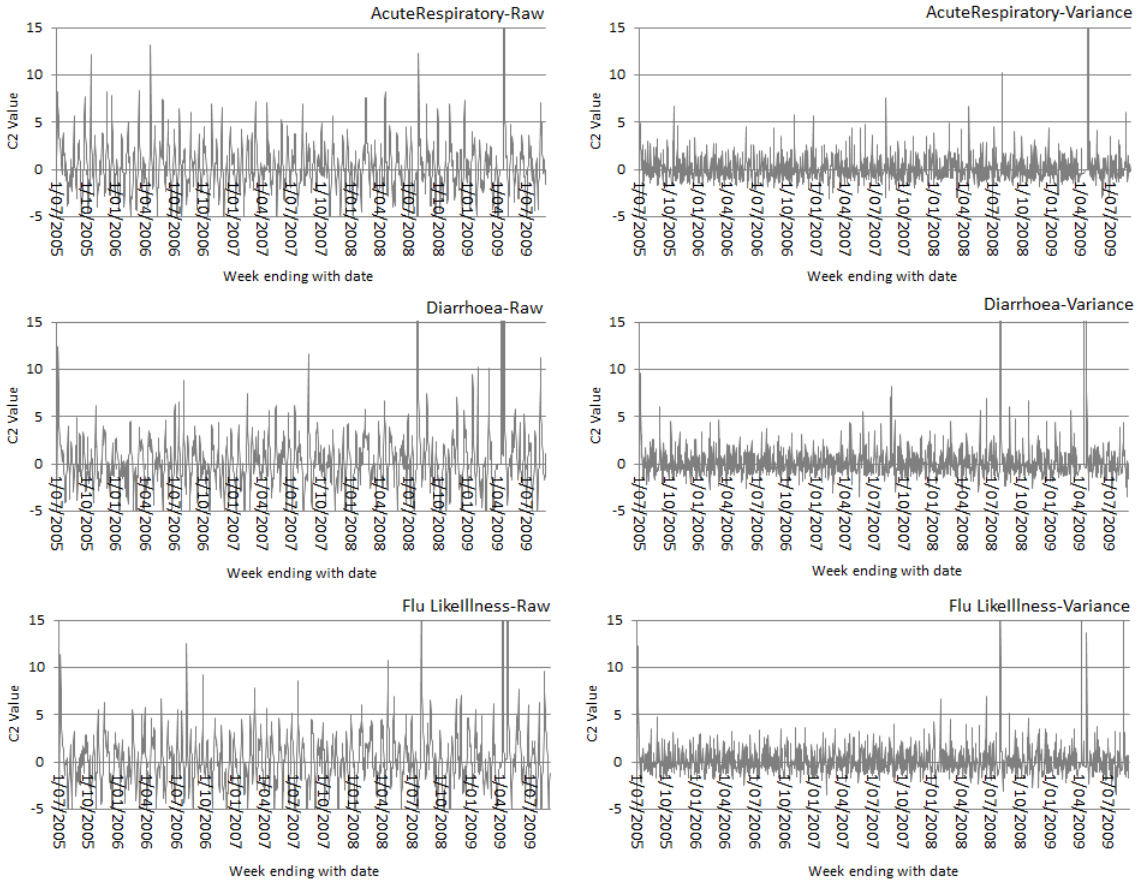

Figure 2: Left: C2 algorithm applied to the weekly predicted frequencies of a syndrome (i.e., the raw frequency method). Right: C2 algorithm applied to the variance of the predicted cases of a syndrome over consecutive weeks (i.e., the variance method). For the C2 aberrancy detection algorithm, the threshold was set to 3 by default. The y-axis has been trimmed to [-5,15] for improved visual interpretability.

5 Conclusion and Future Work

We employed a new technique to perform retrospective syndromic surveillance of three specific syndromic groups, i.e., Flu-like Illness, Acute Respiratory, and Diarrhoea over the Victorian Syndromic Surveillance (SynSurv) data set consisting of a large number of emergency department chief complaints. The process started with the analysis of the unstructured text of the chief complaints and the classification of these into the three syndromic groups using supervised machine learning methods. Then, aberrancy detection algorithms were utilized on both the raw frequencies of positive disease cases and the variances between the number of positive cases of each syndromic group over shifting windows of time (each window comprising 7 days). While the machine learning classifier produced a number of false positives for each syndromic group, the aberrancy detection procedure was insensitive to those (uniformly distributed) false positives, due to its consideration of the mean and standard deviation of frequency differences over time. More importantly, we found that the detection of possible disease outbreaks using our new variance method, which considers the differences between disease frequencies as the inputs to the aberrancy detection process, will result in a more effective outbreak detection compared with the standard method that uses the raw frequencies of the positive cases.

Our study has limitations based on the effectiveness of the SynSurv data set for outbreak detection. The SynSurv data set only contains data from Emergency Departments of hospitals. Real-time syndromic surveillance may require incorporation of other sources of data such as those from medical practitioners for more effective surveillance.

In future work, we are planning to further investigate the textual features of chief complaints using more in-depth natural language processing techniques to find other (complementary) methods for identifying disease outbreaks from chief complaints. Similar to our related work on Ebola [17], we would like to understand whether any specific lexical properties of chief complaints, such as the distribution of linguistic structures, are associated with any significant deviations in the number of positive cases of syndromic groups.

References

[1] Australian influenza report 2009-12-18 december 2009. *Communicable Diseases Information*, 2009.

[2] Overview of biological agents that could be used in a terrorist act, 2015. URL http://www.health.gov.au/internet/main/publishing.nsf/content/health-pubhlth-strateg-bio-agents.html.

[3] Colleen A. Bradley, H. Rolka, D. Walker, and J. Loonsk. Biosense: implementation of a national early event detection and situational awareness system. *MMWR Morb Mortal Wkly Rep*, 54(Suppl):11–19, 2005.

[4] David L. Buckeridge, Anna Okhmatovskaia, Samson Tu, Martin O'Connor, Csongor Nyulas, and Mark A. Musen. Understanding detection performance in public health surveillance: modeling aberrancy-detection algorithms. *Journal of the American Medical Informatics Association*, 15(6):760–769, 2008.

[5] Wendy W. Chapman, Lee M. Christensen, Michael M. Wagner, Peter J. Haug, Oleg Ivanov, John N. Dowling, and Robert T. Olszewski. Classifying free-text triage chief complaints into syndromic categories with natural language processing. *Artificial intelligence in medicine*, 33(1):31–40, 2005.

[6] Mike Conway, John N. Dowling, and Wendy W. Chapman. Using chief complaints for syndromic surveillance: A review of chief complaint based classifiers in north america. *Journal of Biomedical Informatics*, 46(4):734 – 743, 2013. ISSN 1532-0464. doi: http://dx.doi.org/10.1016/j.jbi.2013.04.003. URL http://www.sciencedirect.com/science/article/pii/S1532046413000464.

[7] Jeremy U. Espino, John Dowling, John Levander, Peter Sutovsky, Michael M. Wagner, and Gregory F. Cooper. SyCo: A probabilistic machine learning method for classifying chief complaints into symptom and syndrome categories. *Advances in Disease Surveillance*, 2(5), 2007.

[8] J.U. Espino, M.M. Wagner, F.C. Tsui, H.D. Su, R.T. Olszewski, Z. Lie, W. Chapman, X. Zeng, L. Ma, Z.W. Lu, and J. Dara. The RODS open source project: Removing a barrier to syndromic surveillance. *Medinfo*, 11(Pt 2):1192–1196, 2004.

[9] Simon Firestone, Ian Barr, Paul Roche, and John Walker. Annual report of the national influenza surveillance scheme, 2005. *Communicable Diseases Intelligence*, 30(2), 2006.

[10] Ronald D. Fricker Jr, Benjamin L. Hegler, and David A. Dunfee. Comparing syndromic surveillance detection methods: Ears'versus a cusum-based methodology. 2008.

[11] Kelly J. Henning. What is syndromic surveillance? *Morbidity and Mortality Weekly Report*, 53:7–11, 2004. ISSN 01492195, 1545861X. URL http://www.jstor.org/stable/23315680.

[12] Lori Hutwagner, William Thompson, G. Matthew Seeman, and Tracee Treadwell. The bioterrorism preparedness and response early aberration reporting system (ears). *Journal of Urban Health*, 80(1):i89–i96, 2003.

[13] Lori Hutwagner, Timothy Browne, G. Matthew Seeman, and Aaron T. Fleischauer. Comparing aberration detection methods with simulated data. *Emerg Infect Dis*, 11(2):314–316, 2005.

[14] Ian G. Barr Marlena Kaczmarek, Rhonda Owen. Annual report of the national influenza surveillance scheme, 2008. *Communicable Diseases Intelligence*, 34(1), 2010.

[15] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[16] Bahadorreza Ofoghi and Karin Verspoor. Assessing the performance of American chief complaint classifiers on Victorian syndromic surveillance data. In *Proceedings of Australia's Big Data in Biomedicine & Healthcare Conference*, Sydney, Australia, 2015.

[17] Bahadorreza Ofoghi, Meghan Mann, and Karin Verspoor. Towards early discovery of salient health threats: A social media emotion classification technique. In *Pacific Symposium on Biocomputing*, pages 504–515, 2016.

[18] Rhonda Owen, Ian G. Barr, Andrew Pengilley, Conan Liu, Bev Paterson, and Marlena Kaczmare. Annual report of the national influenza surveillance scheme, 2007. *Communicable Diseases Intelligence*, 32(2), 2008.

[19] Fu-Chiang Tsui, Jeremy U. Espino, Virginia M. Dato, Per H. Gesteland, Judith Hutman, and Michael M. Wagner. Technical description of rods: a real-time public health surveillance system. *Journal of the American Medical Informatics Association*, 10(5):399–408, 2003.

[20] C.J. van Rijsbergen. *Information Retrieval. 1979.* Butterworth, 1979.

[21] Yiliang Zhu, W. Wang, D. Atrubin, and Y. Wu. Initial evaluation of the early aberration reporting system–florida. *Morbidity and Mortality Weekly Report*, 54(123):1, 2005.