Assessing the performance of American chief complaint classifiers on Victorian syndromic surveillance data

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

Syndromic surveillance systems aim to support early detection of salient disease outbreaks, and to shed timely light on the size and spread of pandemic outbreaks. They can also be used more generally to monitor disease trends and provide reassurance that an outbreak has not occurred. One commonly used technique for syndromic surveillance is concerned with classifying Emergency Department data, such as chief complaints or triage notes, into a set of pre-defined syndromic groups. This paper reports our findings on the investigation of the utility and effectiveness of two existing North American methods for free-text chief complaint classification on a large data set of Australian Emergency Department triage notes, collected from two hospitals in the state of Victoria. To our knowledge, these methods have never before been analysed and compared against each other for their applicability and effectiveness on free text chief complaint classification at this scale or in the Australian context.

1 Introduction

Syndromic surveillance is a procedure that is widely used for early detection of disease outbreaks. It may shed timely light on the size, spread, and tempo of outbreaks which can also monitor disease trends and provide reassurance that an outbreak has not occurred [8]. Such information can also be used in the context of disaster medicine for effective and timely medical responses during a major disaster [7].

The practice of disaster medicine has evolved rapidly with the introduction of Electronic Medical Records (EMRs) [1]. EMRs, in most cases, serve as a real-time record of patient encounters. Their use allows clinical staff to enter, store, and share data with their colleagues across time and space [2], and further opens up the potential for real-time monitoring of public health.

While EMRs allow data to be entered in various forms, unstructured free text is the data format most preferred by clinicians and medical officers [12]. One type of EMR free text data is the Chief Complaint (CC), or presenting complaint, typically captured by a triage nurse in an Emergency Department (ED) on reception of an arriving patient. CCs represent professional interpretation of the symptoms or condition which brought the patient to the ED to seek emergency care [15]. They can potentially be used to subset patients into cohorts, initiate decision support, and perform research [12]. CCs usually consist of "a mixture of subjective and objective information describing a patient's status on his/her visit to the ED" [11].

A major use of ED CCs is for syndromic surveillance, in combination with other data elements such as the diagnosis codes assigned at the end of the ED visit and the patient's temperature as measured in the ED [19]. For example, during the Rugby World Cup in Sydney in 2003, electronic triage notes were directly analysed for syndromic surveillance purposes [17]. Recent recommendations from the International Society for Disease Surveillance (ISDS) also include triage notes as part of the recommended minimum data set of electronic health record data for surveillance purposes [9].

To be useful for syndromic surveillance, the free-text triage CCs must first be classified into predefined syndromic categories or into some other type of coded representation that can be manipulated or analysed by a computer program [10]. Once CC texts have been classified into syndromic categories, temporal analysis of categorized results may be conducted to detect possible epidemics and outbreaks [14]. Conway et al. [4] have provided a comprehensive review of fifteen operational English language CC-based syndromic surveillance systems in North America that take different approaches in categorization of CCs. The authors found that most

existing systems utilize keyword-based, linguistic, statistical, and/or character-level data from CC texts in the classification process. Ivanov et al. [11] conducted a retrospective study to ascertain the potential of free-text CCs collected in pediatric emergency departments. They used the Bayesian classifier implemented in Complaint Coder (CoCo) [16]. On a population of children less than five years of age, for early detection of respiratory and gastrointestinal outbreaks, they found that: i) time series of automatically coded free text CCs related to pediatric patients correlated with hospital admissions, and ii) the same time series preceded hospital admissions by the mean of 10.3 and 29 days for respiratory and gastrointestinal outbreaks, respectively.

In this paper, we present an evaluation of the applicability and effectiveness of two existing machine learning-based North American CC classifiers, namely Symptom Coder (SyCo) [5] and Complaint Coder (CoCo). These tools are both part of a computer-based public health surveillance system named Real-time Outbreak and Disease Surveillance (RODS) [6].

Our analysis of SyCo and CoCo on free text CC classification is based on a syndromic surveillance data set that includes ED triage notes from two different hospitals in the Australian state of Victoria. We focus on three primary medical conditions of particular interest in public health surveillance: *Flu Like Illness*, *Acute Respiratory*, and *Diarrhoea*. According to [4], and to our knowledge, SyCo and CoCo have never been evaluated against each other on any surveillance data set at the level of complexity and size as the data we used for this study.

The only previous research that compared the two systems is the work in [18] which only focused on 1,122 CC entries on a single disease, i.e., *Influenza Like Illness*. Our analysis goes beyond this by considering the three above-mentioned diseases and a much larger syndromic surveillance data set (with a total of 314,629 CC records) as will be introduced in section 2.2.

2 Methods

2.1 Symptom Coder and Complaint Coder

Symptom Coder (SyCo) [5] and Complaint Coder (CoCo) [16] are two chief complaint classifier systems developed as parts of the Real-time Outbreak and Disease Surveillance (RODS) system [6]. Both CoCo and SyCo implement Naïve Bayes text classification which assumes conditional independence between features. CoCo is a direct complaint-into-syndrome classifier which finds the posterior probabilities for each of its eight syndromic categories (i.e., constitutional, respiratory, gastrointestinal, hemorrhagic, botulism, neurological, respiratory, and other) given the text of a chief complaint. CoCo can classify CCs into one of the eight syndromic categories constitutional, respiratory, gastrointestinal, hemorrhagic, botulism, neurological, respiratory and other.

The Bayesian probabilities that CoCo calculates are determined using a default probability file (developed by RODS) that was derived from 28,990 CC strings each manually coded by a physician into a syndrome category [16]. Although CoCo has the capability to be retrained by using patient data obtained locally, CoCo, in our experiments, was used with the default probability file and no further training was performed.

SyCo differs from CoCo in that it implements a two-layer classification procedure from chief complaints into syndromes. SyCo first finds the posterior probabilities of a number of symptoms (i.e., 17 in-built symptoms) given the text of the chief complaint. Consequently, SyCo calculates the posterior probabilities of each syndromic group given the (posterior) probabilities of the symptoms. The syndrome with the highest posterior probability is then selected as the syndromic group for the chief complaint.

The textual features that CoCo and SyCo extract from CCs are at the word level only, excluding phrases, n-grams, or any biomedical terminology. The study conducted by Connor et al. [3] showed that the RODS CoCo classifier is outperformed by using a Maximum Entropy Model that was able to overcome the word-level approach of CoCo by considering both sub-word and super-word sequences of characters. The MaxEnt classifier also improved over CoCos ignorance of conditional dependence between textual features.

Silva et al. [18] compared the performance of syndromic classification of CCs achieved using SyCo and CoCo (as well as with their Geographic Utilization of Artificial Intelligence in Real-time for Disease Identification and Alert Notification System that is not available for us to test); however, this analysis was only performed on a single disease (i.e., *Influenza Like Illness*) with a small data set of 1,122 CC records from a single urban academic medical centre. They found similar recall (0.3530) for SyCo and CoCo on that data but slightly different precision values (CoCo = 0.9890% and SyCo = 0.9930%).

2.2 Victorian Data Set

The SynSurv data is data collected from the Emergency Departments of the Royal Melbourne Hospital and the Alfred Hospital during the period 1998-2010 (predominantly from 2000 to 2009). The data were collected on behalf of the Victorian Department of Health (Vic Health) for syndromic surveillance during the 2006 Commonwealth Games held in Melbourne. The Vic Health remains the custodian of the data; however, for the purposes of this project, we were granted permission by Vic Health to work with it.

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

Table 1: The distribution of the different syndromic groups in the partial SynSurv data set used for our analysis.

SynSurv consists of naturally occurring data collected at triage, consisting primarily of free text notes written by a triage nurse for each patient visit to the EDs of the selected hospitals. These notes have been augmented with annotations of diagnostic codes using the ICD-10 version of the International Classification of Disease.

The data set contains a total of 918,330 ED visit records. From these records, there are 730,054 visits with a valid ICD-10 code with the primary diagnosis as entered by a nurse upon triage assessment of the patient. In SynSurv, there are 456,213 records with any textual comment at all. The total number of records with both a diagnosis and a textual nurse note is 316,362 entries. From this data set of 316,362 CC records, we used a subset of 314,629 records that were labelled with one of the three syndromic groups *Flu Like Illness*, *Acute Respiratory*, and *Diarrhoea*, or as other. Table 1 summarizes the distribution of the different syndromic groups in the data set.

3 Supervised Chief Complaint Analysis with SyCo and CoCo

3.1 Experimental Set-up

For the machine learning-based classification experiments that follow, the set of records corresponding to a given syndromic group was used as positive examples for that syndrome and all other records in the relevant subset of SynSurv (see Table 1) were considered to be negative instances for the syndrome.

We trained one binary SyCo classifier for each of the syndromic groups with the training portion of the data set for the given syndromic group and then, tested the effectiveness of the classifier using the corresponding test set for the syndromic group.

To evaluate CoCo, which we did not train with our training data sets, we only used the test sets to measure the classification performance of the tool for each syndromic group. For this, we had to find a mapping between our syndromic groups and those eight syndromic categories that CoCo had been trained with. Table 2 shows the mapping that we considered between the two categories of syndromic groups.

To understand how the performances of SyCo and CoCo compare with those of trivial baseline systems, we developed two trivial baseline classifiers:

- **All-Positive (All+) Baseline**, which assigns a positive class label to every given instance of CCs in the test set. A positive class label means that the chief complaint is labelled as the given syndrome under investigation.
- **Random-50-50 Baseline**, which assigns either a positive or a negative class label to every given instance of CCs in the test set. The assignment of class labels is based on a binary random generator. To more accurately account for the randomness of this baseline system, we generated average evaluation results for 10 consecutive independent runs over the same test set related to each syndromic group.

Syndromic group in Victorian data	Syndromic group in CoCo
Flu Like Illness	Constitutional
Acute Respiratory	Respiratory
Diarrhoea	Gastrointestinal

Table 2: Mapping between our syndromic groups and RODS syndromic groups, used to evaluate CoCo's classification performance.

3.2 Results

We evaluated the classification methods using the number of True Positives (TPs), False Positives (FPs), Precision (i.e., specificity), Recall (i.e., sensitivity), and F1-measure. Table 3 summarizes the results achieved using SyCo and CoCo as well as the two baseline systems All+ and Random-50-50 on the above-mentioned data set.

Method	Syndrome	#Positive instances	#TPs	#FPs	Prec.	Rec.	F1
CoCo	Flu Like Illness	5829	910	3998	0.1854	0.1561	0.1695
	Acute Respiratory	3877	1038	1843	0.3603	0.2677	0.3072
	Diarrhoea	2601	970	4219	0.1869	0.3729	0.2490
SyCo	Flu Like Illness	5829	3135	5542	0.3613	0.5378	0.4322
	Acute Respiratory	3877	1932	5845	0.2484	0.4983	0.3316
	Diarrhoea	2601	1084	3665	0.2283	0.4168	0.2950
All+	Flu Like Illness	5829	5829	98941	0.0556	1.0000	0.1054
	Acute Respiratory	3877	3877	100893	0.0370	1.0000	0.0714
	Diarrhoea	2601	2601	102169	0.0248	1.0000	0.0484
Random-50-50	Flu Like Illness	5829	2922.9	49466.5	0.0558	0.5014	0.1004
	Acute Respiratory	3877	1941.7	50464.8	0.0370	0.5008	0.0690
	Diarrhoea	2601	1294.7	50988.2	0.0248	0.4978	0.0472

Note: TPs=True Positives, FPs=False Positives, Prec.=Precision, Rec.=Recall

Table 3: The results of the SyCo and CoCo classification methods with respect to each syndromic group.

4 Discussion

From the results in Table 3, and considering F1-measure, from the two chief complaint classification systems, SyCo has been demonstrated to perform better on all of the three syndromic groups. In terms of precision and recall, SyCo outperforms CoCo again in most cases, except for the Acute Respiratory group where CoCo shows a higher precision (0.3603 vs. 0.2483). This single scenario where CoCo outperforms SyCo corresponds to the situation in which SyCo returns with a large number of FPs (5,845) versus only 1,843 FPs returned by CoCo for the same syndromic group. This could not be compensated for by the relatively small difference between the numbers of TPs that both methods returned (CoCo: 1,038 vs. SyCo 1,931). One major factor that may explain the superior performance of SyCo over that of CoCo in our experiments is the fact that CoCo was not trained with our training data sets whereas each SyCo binary classifier was trained for each of the syndromic groups with the corresponding training set in our data set.

According to the results in Table 3, SyCo and CoCo result in a large difference in the recall values for the *Flu Like Illness* category (CoCo=0.1561 and SyCo=0.5378). These recall values are both relatively different to what Silva et al. [18] found on the similar disease category *Influenza Like Illness* with 1,122 CC entries (SyCo=CoCo=0.3530). The higher recall value of SyCo here compared with what Silva et al. [18] demonstrated may be due to the larger training data set that has been utilised in this work.

The results in Table 3 also demonstrate that: i) neither of the two baseline classification systems perform well on any of the three syndromic groups, and ii) more importantly, both SyCo and CoCo perform relatively higher than the trivial baseline systems.

To understand how SyCo and CoCo classify each instance of the CCs in the SynSurv data set, we conducted a Cohen's Kappa statistical agreement analysis on the output classifications of the two systems. For this, a pair of output classifications for each CC string was created and the entire set of pairs were analysed for their agreement with each other.

Table 4 shows the results of the Cohen's Kappa statistic for the agreement of the classification outputs generated using SyCo and CoCo for each of the syndromic groups in the SynSurv data set under study. According to the Kappa statistic magnitude guidelines in [13], these results suggest that SyCo and CoCo classify CC entries that relate to *Flu Like Illness slightly* similarly. For both of the other two diseases, i.e., *Acute Respiratory* and *Diarrhoea*, the agreement between the classification outputs is considered as *fair*. All of these results are statistically significant as indicated by p-value$<.0001$.

The relatively weak agreement between SyCo and CoCo on our data suggest that the two classifier systems model and interpret the CCs in relatively different fashions. As a result, there is a dissimilar, possibly complementary, coverage over true positives and true negatives. A possible next step would be to bring the two classifiers together using an ensemble approach.

Syndrome	Kappa	p-value
Flu Like Illness	0.1349	$<.0001$
Acute Respiratory	0.2225	$<.0001$
Diarrhoea	0.2862	$<.0001$

Table 4: The agreement analysis between the classification results of SyCo and CoCo per syndromic group.

5 Conclusion

Syndromic surveillance is a widely-utilised procedure for early detection of salient disease outbreaks. One of the most commonly used techniques in the area of syndromic surveillance is based on supervised classification of Chief Complaints into a set of pre-defined syndromic categories. In this paper, we analysed the performance of two well-known CC classifiers from the Real-time Outbreak and Disease Surveillance (RODS), namely Symptom Coder (SyCo) and Complaint Coder (CoCo). While CoCo was used as an off-the-shelf component, with no training in our experiments, SyCo was trained with the training subsets of the data we used in this work. The results of our analysis on a large data set of Australian ED notes labelled with three syndromic groups *Flu Like Illness*, *Acute Respiratory*, and *Diarrhoea* suggest that, in most cases, SyCo outperforms CoCo in terms of the evaluation metrics. Both SyCo and CoCo outperform the two trivial baseline systems that we developed for performance comparison reasons only. We also found that the two classifiers do not agree on the classification outputs, i.e., the classifiers make different mistakes, which may suggest that an effective approach may be required to combine the two classifiers.

6 Acknowledgements

This work was supported by the Bioterrorism Preparedness task of the Land Personnel Protection Branch, Land Division of the Australian Defence Science and Technology Organisation (DSTO). We also thank the Victorian Department of Health and Human Services for their contribution of the SynSurv data set. This research was conducted under Ethics Approval 21/14 by the Department of Health Human Research Ethics Committee.

References

[1] M. Abir, F. Mostashari, P. Atwal, and Lurie N. Electronic health records critical in the aftermath of disasters. *Prehospital and Disaster Medicine*, 27(6):620–622, 2012.

[2] Alexandra T. Bambrick, Dina B. Passman, Rachel M. Torman, Alicia A. Livinski, and Jennifer M. Olsen. Optimizing the use of chief complaint & diagnosis for operational decision making: An EMR case study of the 2010 Haiti earthquake. *PLoS Currents*, 6:1–18, 2014.

[3] Kevin H.O. Connor, Kieran M. Moore, Bronwen Edgar, and Don Mcguinness. Maximum entropy models in chief complaint classification. In *International Society for Disease Surveillance Conference*, page p. 23, Baltimore, 2006.

[4] 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, 2013. ISSN 15320464.

[5] 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.

[6] 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.

[7] Aaron T. Fleischauer, Stacy Young, Joshua Mott, and Raoult Ratard. Disaster surveillance revisited: Passive, active and electronic syndromic surveillance during hurricane katrina, New Orleans, LA 2005. *Advances in Disease Surveillance*, 2:153, 2007.

[8] Kelly J. Henning. Overview of syndromic surveillance: What is syndromic surveillance? *Morbidity and Mortality Weekly Report (MMWR)*, pages 53(suppl): 5–11, 2004.

[9] International Society for Disease Surveillance. Final recommendation: Core processes and EHR requirements for public health syndromic surveillance. Technical report, ISDS, 2011. URL `www.syndromic.org/projects/meaningful-use`.

[10] C.B. Irvin, P.P. Nouhan, and K. Rice. Syndromic analysis of computerized emergency department patients chief complaints: An opportunity for bioterrorism and influenza surveillance. *Annals of Emergency Medicine*, 41(4):447–452, 2003.

[11] Oleg Ivanov, Per H. Gesteland, William Hogan, Michael B. Mundorff, and Michael M. Wagner. Detection of pediatric respiratory and gastrointestinal outbreaks from free-text chief complaints. *AMIA Annual Symposium proceedings*, pages 318–322, 2003. ISSN 1942-597X.

[12] Yacine Jernite, Yoni Halpern, and Steven Horng. Predicting chief complaints at triage time in the emergency department. In *Machine Learning for Clinical Data Analysis and Healthcare NIPS Workshop 2013*, pages 1–5, Lake Tahoe, Nevada, 2013.

[13] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33: 159–174, 1977.

[14] Hsin-Min Lu, Daniel Zeng, Lea Trujillo, Ken Komatsu, and Hsinchun Chen. Ontology-enhanced automatic chief complaint classification for syndromic surveillance. *Journal of biomedical informatics*, 41(2):340–56, April 2008. ISSN 1532-0480. doi: 10.1016/j.jbi.2007.08.009.

[15] Tomi Malmström, Olli Huuskonen, Paulus Torkki, and Raija Malmström. Structured classification for ED presenting complaints - from free text field-based approach to ICPC-2 ED application. *Scandinavian journal of trauma, resuscitation and emergency medicine*, 20(76), January 2012. ISSN 1757-7241. doi: 10.1186/1757-7241-20-76.

[16] C.A. Mikosz, J. Silva, S. Black, G. Gibbs, and I. Cardenas. Comparison of two major emergency department-based free-text chief-complaint coding systems. *Morbidity and Mortality Weekly Report (MMWR)*, pages 53(suppl): 101–105, 2004.

[17] David J. Muscatello, Tim Churches, Jill Kaldor, Wei Zheng, Clayton Chiu, Patricia Correll, and Louisa Jorm. An automated, broad-based, near real-time public health surveillance system using presentations to hospital emergency departments in New South Wales, Australia. *BMC public health*, 5(1):141, January 2005. ISSN 1471-2458. URL `http://www.biomedcentral.com/1471-2458/5/141`.

[18] Julio C. Silva, Shital C. Shah, Dino P. Rumoro, Jamil D. Bayram, Marilyn M. Hallock, Gillian S. Gibbs, and Michael J. Waddell. Comparing the accuracy of syndrome surveillance systems in detecting influenza-like illness: GUARDIAN vs. RODS vs. electronic medical record reports. *Artificial Intelligence in Medicine*, 59:169–174, 2013. ISSN 09333657. doi: 10.1016/j.artmed.2013.09.001.

[19] Debbie Travers, Stephanie W. Haas, Anna E. Waller, Todd A. Schwartz, Javed Mostafa, Nakia C. Best, and John Crouch. Implementation of emergency medical text classifier for syndromic surveillance. *AMIA Annual Symposium proceedings*, 2013:1365–74, January 2013. ISSN 1942-597X.