Methodological Review

Using chief complaints for syndromic surveillance: A review of chief complaint based classifiers in North America

Mike Conway [a,*], John N. Dowling [b], Wendy W. Chapman [a]

[a] Division of Biomedical Informatics, University of California, San Diego, 9500 Gilman Dr. MC 0505 La Jolla, California 92093, United States
[b] Department of Biomedical Informatics, University of Pittsburgh, 5607 Baum Boulevard, Pittsburgh, PA 15260, United States

ABSTRACT

A major goal of Natural Language Processing in the public health informatics domain is the automatic extraction and encoding of data stored in free text patient records. This extracted data can then be utilized by computerized systems to perform syndromic surveillance. In particular, the *chief complaint*—a short string that describes a patient's symptoms—has come to be a vital resource for syndromic surveillance in the North American context due to its near ubiquity. This paper reviews fifteen systems in North America—at the city, county, state and federal level—that use chief complaints for syndromic surveillance.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

A major goal of Natural Language Processing (NLP) in the medical domain is the automatic extraction and encoding of data stored in free text patient records. This extracted data can then be utilized by Information Technology systems to perform syndromic surveillance. In particular, the *chief complaint*—a short string that describes a patient's symptoms before even a preliminary diagnosis has been made—has come to be a vital resource for syndromic surveillance in the North American context. Despite the acknowledged importance of chief complaints for syndromic surveillance, considerable variation exists in how they are used in practice by system builders, both in terms of algorithms chosen to map chief complaint strings to syndromes, and the syndrome definitions themselves.

This paper reviews fifteen currently operational English language[1] chief complaint based syndromic surveillance systems in North America. These systems were developed and deployed at the local, provincial or national level, by government agencies, universities, national public health bodies, military organizations, or local hospitals.

The paper is divided into three sections. Firstly, we briefly outline the advantages (and disadvantages) of using chief complaints for syndromic surveillance. We then go onto describe each of the fifteen systems in turn. Finally we present a discussion section, including an outline of emerging syndrome definition standards for the field.

2. Syndromic surveillance and chief complaints

Conventional public health disease surveillance relies on the routine manual or electronic filing (by clinicians and laboratories) of reportable and unusual diseases that alert public health officials to disease outbreak clusters of interest [2–4]. That is, conventional surveillance mechanisms depend on confirmatory laboratory testing *after* preliminary diagnosis by a physician. In many cases, the confirmation of an infectious disease or a bioterrorism agent takes days of testing and epidemiological analysis before an outbreak is identified. Traditional disease surveillance methods that rely on the passive and voluntary reporting of cases of specific diseases by practitioners and laboratories may not be timely enough to provide the information needed to detect and monitor a rapidly evolving outbreak [2,3,5,6].

Timely outbreak detection requires the identification of suspicious patterns that occur *early* in the course of an illness [2,7,8]. Syndromic surveillance focuses on the early symptom (prodromal) period before clinical or laboratory confirmation of a particular disease [7,9,10] and may utilize both clinical and alternative data sources that reflect measurable alterations in personal behaviors that may precede a clinical diagnosis.[2] Syndromic surveillance

---

[2] Note that the term *syndromic surveillance* can be used in a more restrictive sense to refer only to the analysis of medical reports (of various kinds), with alternative non-medical data sources specifically excluded [11,12]. In this paper, we use the term *syndromic surveillance* in its wider sense, to include over-the-counter medication sales and absenteeism data.

systems often utilize data sources that already exist but have not been designed specifically for public health surveillance purposes. These sources could include patients' chief complaints in medical facilities, prescriptions filled, retail drug and product sales, and school or work absenteeism [3,6,7,13].

Many automated syndromic surveillance systems use triage chief complaints to classify patients into syndromic case definitions [8,14–19]. A chief complaint is a short phrase entered by a triage nurse or admission clerk describing the reason for a patient's visit to a medical facility[3] [20]. It has the advantages of being nearly ubiquitously available in the United States, routinely generated during normal hospital operations [21], and available electronically during or shortly after a patient's visit, thus providing a basis for real-time surveillance [22] (for a random sample of chief complaints, see Table 1). The chief complaint is routinely collected during a patient's encounter, because it represents a central piece of information that directs care. Various clinical, research, and administrative objectives all rely on the presence of an easily identifiable and unambiguous chief complaint. However, to be useful for syndromic surveillance, the free-text triage chief complaints must first be classified into syndromic categories or into some other type of coded representation that can be manipulated by a computer [14]. Hand-coding data into syndrome categories, whether performed onsite in the medical facility or offsite, requires considerable time and labor [22]. To make chief complaint data more realistically usable for ongoing surveillance, automated syndromic categorization applications have been developed. However, automated chief complaint categorization still suffers from the challenging nature of the data (that is, prevalence of abbreviations and misspellings, context-sensitive vocabulary, inter-hospital variation) and usability considerations (for example, providing a means for refining syndrome criteria) both of which must be overcome to classify chief complaints efficiently and effectively [22]. Furthermore, chief complaints vary in accuracy because they are recorded prior to physician involvement in care and can therefore lack the diagnostic precision of physician generated reports [23].

While capturing surveillance data using chief complaints does not incur human effort, delay, or drastic reduction in information obtained by coding, the use of chief complaint data does have associated problems—identified by Shapiro [24]—caused by linguistic variation:

1. A single symptom can be described in multiple ways by using synonyms and paraphrases.
2. Medical concepts are often recorded using abbreviations and acronyms that are idiosyncratic to individual hospitals.
3. The same concept can be indicated with different parts of speech.
4. Words are frequently misspelled or mistyped in busy medical settings, causing the continual appearance of new, previously unseen errors.

The use of free text in chief complaint based syndromic surveillance systems requires managing the substantial variation that results from the use of synonyms, abbreviations, acronyms, truncations, misspellings and typographic errors (examples of the surface variation found in chief complaints can be seen in Table 2). Failure to detect these linguistic variations could result in missed cases, and traditional methods for capturing this variation require ongoing labor intensive maintenance. One way of addressing this problem is to *preprocess* the string before it is handed over to the syndrome classifier (see Fig. 1, for a generic preprocessing module). Alternatively, supervised learning methods, if provided with enough labeled training data, are able to process chief complaints without

---

[3] *Chief complaints* are known as *presenting complaints* in some anglophone countries like the United Kingdom, Australia and the Republic of Ireland.

Table 1
Randomly sampled chief complaints.

injury, toe	migraine	fell off bus
confused	weakness	psychiatric evaluation
detox from heroin	vomiting up blood	right knee pain
crying/vomiting	rash on face	injured finger
right shoulder injury	slurred speech	head injury
stomach cramps	cold	tired/dizzy
medical	diff. swallowing	followup
l. hip pain	dental filling	labial swelling
body ache	optical exam	throat swelling
visual disturbance	earache	nausea
sprained ankle	grion pain	eye injuery
trouble urinating	palpitations	diabetic
injured leg	sores on back	foreign body, throat
ruq pain	inj lt ear	seizure
epistaxis	left hip pain	chest pain

Table 2
Chief complaint example.

Chief complaint	Expansion
nausae	Nausea
vomiting and headache	Vomiting, headache
sore throat headache 4 days	Sore throat, headache
abd pain	Abdominal pain
sob	Shortness of breath
n/v	Nausea, vomiting
headache	Headache
dirreach	Diarrhea
rlq pain	Right lower quadrant pain
chi	Chill
body ache	Body ache
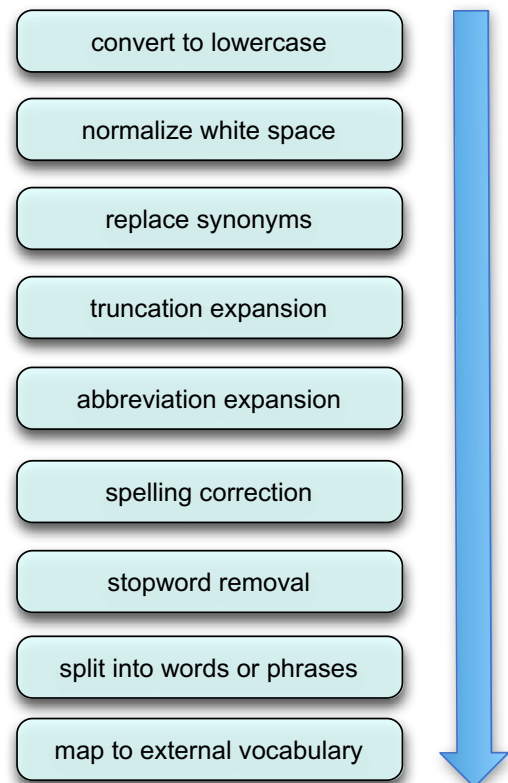
Fig. 1. Generic preprocessing steps.