

Taming Variability in Free Text: Application to Health Surveillance

Alan R. Shapiro

New York University School of Medicine, New York, New York

Corresponding author: Alan R. Shapiro, Department of Medicine, New York University School of Medicine, 5 Pheasant Run, Pleasantville, NY 10570. Telephone: 914-747-1804; E-mail: alan.shapiro@med.nyu.edu.

Abstract

Introduction: *Use of free text in syndromic surveillance requires managing the substantial word variation that results from use of synonyms, abbreviations, acronyms, truncations, concatenations, misspellings, and typographic errors. Failure to detect these variations results in missed cases, and traditional methods for capturing these variations require ongoing, labor-intensive maintenance.*

Objectives: *This paper examines the problem of word variation in chief-complaint data and explores three semi-automated approaches for addressing it.*

Methods: *Approximately 6 million chief complaints from patients reporting to emergency departments at 54 hospitals were analyzed. A method of text normalization that models the similarities between words was developed to manage the linguistic variability in chief complaints. Three approaches based on this method were investigated: 1) automated correction of spelling and typographical errors; 2) use of International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes to select chief complaints to mine for overlooked vocabulary; and 3) identification of overlooked vocabulary by matching words that appeared in similar contexts.*

Results: *The prevalence of word errors was high. For example, such words as diarrhea, nausea, and vomiting were misspelled 11.0%–18.8% of the time. Approximately 20% of all words were abbreviations or acronyms whose use varied substantially by site. Two methods, use of ICD-9-CM codes to focus searches and the automated pairing of words by context, both retrieved relevant but previously unexpected words. Text normalization simultaneously reduced the number of false positives and false negatives in syndrome classification, compared with commonly used methods based on word stems. In approximately 25% of instances, using text normalization to detect lower respiratory syndrome would have improved the sensitivity of current word-stem approaches by approximately 10%–20%.*

Conclusions: *Incomplete vocabulary and word errors can have a substantial impact on the retrieval performance of free-text syndromic surveillance systems. The text normalization methods described in this paper can reduce the effects of these problems.*

Introduction

Syndromic surveillance using existing free-text sources (e.g., electronic medical records or emergency department [ED] chief complaints) offers potential advantages in the timeliness and richness of the information that can be provided (1). In particular, capturing surveillance information as free text does not incur the human effort, delay, or drastic reduction in information incurred by coding. However, using free text to track symptom occurrence incurs four particular challenges 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; and 4) words are frequently misspelled or mistyped in busy medical settings, causing the continual appearance of

new, previously unseen errors. This paper discusses new approaches to address these four challenges.

Failure to detect linguistic variations results in missed cases. This problem is potentially severe enough to motivate efforts to develop surveillance systems based on apparently unambiguous numerical codes or standardized vocabularies. One goal of the current study is to analyze ED chief complaints empirically to explore the extent of variation present.

Certain efforts to manage linguistic variations and to increase system sensitivity can produce their own false positives, thereby lowering specificity, increasing false alarms, and ultimately wasting limited public health resources. Most importantly, monitoring symptoms adequately in the presence of such variability requires ongoing, costly, labor-intensive maintenance.

The problem of linguistic variation in surveillance systems designed for early detection of covert attacks deserves attention. Increasingly sensitive statistical methods are available with the potential to detect an outbreak affecting a local geographic area (e.g., a single hospital). These methods are most effective when clean data are provided. In addition, syndromic surveillance systems have been used not only for outbreak detection but for case finding and outbreak monitoring; these functions can also be compromised when substantial numbers of cases are missed. Even if a surveillance system contains minimal errors when used in the site where it was developed, word usage can vary substantially among sites, making algorithms developed for one site inadequate for others. Efforts to combine systems for extensive regional surveillance need to be able to detect and address performance differences caused by word variation from one site to another.

This paper examines the extent of word variation in the text of ED chief complaints. It then reviews different approaches for managing word variation, discusses their limitations, and outlines a new approach to text normalization on which three approaches to handling linguistic variation are based. The performance of these approaches when combined is then compared to a common approach in free-text surveillance systems based on word-stem matching.

Extent of Word Variation in Chief-Complaint Databases

Chief-complaint databases from the New York City (NYC) Department of Mental Health and Hygiene (DOHMH), Emergency Medical Associates of New Jersey (EMA), and Boston Beth Israel Deaconess Medical Center (AEGIS) were used in these studies. Collectively, the data consist of the chief complaints from approximately 6 million patient encounters at 54 hospitals over a period of 1–7 years, depending on the hospital.

Types of Word Variation

The word variation in these approximately 6 million chief complaints can be grouped into two types. The first, orthographic variation, includes variations in spelling attributable either to different grammatical forms of the same word (e.g., coughs, coughed, or coughing) or to spelling errors, transcription errors, or typographic errors. In principle, orthographic variation might be addressed, at least in part, through the use of string-matching algorithms that group similarly spelled words.

The second type of word variation, nonorthographic (or semantic) variation, unfortunately cannot be managed merely

by looking at the arrangement of letters in a word. The same chief complaint can usually be described in multiple ways by using acronyms, word truncations, idiosyncratic abbreviations, or legitimate synonyms, all of which can differ from one hospital to another. For example, spelling-correction or string-matching algorithms cannot be expected to discover that the 869 chief complaints of *NV* in the DOHMH database should be regarded as instances of *nausea and vomiting*. Such cases in which only a limited number of letters are retained from the original word are better treated as synonyms rather than orthographic variations and are referred to here as examples of nonorthographic or semantic variation.

Orthographic Variation

Substantial orthographic variation was found among words commonly included in chief complaints (e.g., *diarrhea*, *nausea*, or *abscess*) (Table 1). These numbers were derived from the DOHMH database, but results for the EMA and AEGIS databases were similar. A word as simple as *vomiting* was misspelled at least 379 ways (Table 2).

TABLE 1. Variability in strings used to denote selected words in free-text emergency department chief-complaint data — New York City, November 2001–November 2002

Word	No. of variations	No. of instances	Incorrect (%)
Abscess	92	3,419	45.4
Diarrhea	349	4,006	11.1
Vomiting	379	16,288	16.7
Nausea	137	4,143	18.8
Headache	196	1,771	3.4

Source: New York City Department of Health and Mental Hygiene chief-complaint database.

TABLE 2. Examples of different strings* used to denote vomiting in free-text emergency department chief-complaint data

1. Andvomiting	100. Vomitedx5today	300. Vommioting
2. Bomiting	101. Vomiteing	301. Vommitted
3. Cvomiting	102. Vomites	302. Vommitting
—	103. Vomiteded	303. Vommming
15. V0mitting	104. Vomitfever	304. Vommitintig
16. Vamiting	105. Vomitg	305. Vommitit
17. Vbomiting	—	—
18. Vfomiting	200. Vomitint	325. Vomti
19. Vimit	201. Vomitintg	326. Vomtied
20. Vimited	202. Vomitiny	327. Vomtig
—	—	—
50. Vomiging	250. Vomitting3xdays	377. Vvomitting
51. Vomihing	251. Vomittinga	378. Womiting
52. Vomiig	252. Vomittingab	379. Womitting

Source: New York City Department of Health and Mental Hygiene chief-complaint database.

* N = 379

Spelling-correction programs are often based on the observation that 80% of spelling errors are usually caused by a single insertion, substitution, or deletion of a letter in the word (2). That study, based on the performance of computer transcriptionists in 1964, did not reflect the conditions of the typical modern-day ED. By contrast, in the present study, the modal number of errors per misspelled word was two, and in 31% of instances the misspelled words contained ≥ 3 errors.

Nonorthographic Variation

Nonorthographic variation in the study data was common. In each of the three databases, >20% of all nonstop words (i.e., words other than common articles, conjunctions, and prepositions [e.g., *the*, *and*, *a*, or *or*]) in the chief complaints were nonstandard acronyms, abbreviations, or truncations. This number was obtained *after* first excluding such standard medical abbreviations as CHF, ECG, HCT, HBV, HIV, SOB, WBC, and 43 others. This observation necessitated this study's efforts to address the nonorthographic or semantic variation found in medical free text.

Substantial differences in usage among sites were present. Approximately 55% of the word strings in the EMA database were not contained in the DOHMH data, and 35% of the strings in the AEGIS database were not present in the DOHMH data even though the AEGIS database is only 8% the size of the DOHMH database. The words *rigors* and *myalgias* were used in the AEGIS database 211 and 76 times more frequently than in the DOHMH and EMA databases, respectively. Those words occurred so rarely in the NYC chief complaints that they were not included in, and would not have been detected by, the DOHMH algorithms. Similarly, 3,392 instances of skin rashes described in the EMA hospitals using the string *erupt* would not have been retrieved because the truncation *erupt* was used only rarely in NYC and therefore not included in their algorithms. The acronym *DIB* for *difficulty in breathing* appeared in 2,679 chief complaints from New York City but only twice in the >3.5 million chief complaints recorded elsewhere. Such differences highlight the need for more systematic, preferably automated, methods for managing site customization.

Methods for Managing Linguistic Variation

The need to clean textual data has been recognized in every discipline in which textual data is processed, and corresponding methods to deal with the problem have been developed (3). The majority of these methods have addressed only orthographic word variation. This paper describes the limita-

tions of the three most commonly used methods (phonetic spelling correction, word-stem algorithms, and edit distances) and proposes the need for a fourth, more powerful approach for managing medical text.

Phonetic Spelling Correction Methods

Phonetic spelling-correction methods include algorithms such as Soundex, Editex, or Phonix (4). Soundex has been used for more than a century and is often used in medical applications. These methods recognize that words can be misspelled when certain letters that sound alike, such as *d* and *t* (as in *jauntice* or *pregnand*) or *g* and *j* (as in *conjested*) are substituted for one another.

Unfortunately, multiple exceptions to these pairings exist (e.g., *g* does not sound like *j* in *cough* and thus misspellings of *cough* would not be detected). More importantly, among the chief complaints examined in this study, typing and transcription errors were more common than phonetic errors. The letters *r* and *y* were substituted for *t* 5 times more frequently than the letter *d* because they are located on either side of *t* on the keyboard.

Keyword or Word-Stem Methods

The idea behind this current method in free-text syndromic surveillance is that most words contain a unique string, usually the first few letters, that is specific enough to identify the word and that is unlikely to be misspelled. For example, this method assumes that although *breathing* might be spelled 147 ways in chief-complaint data, searching for all words beginning with *breat* would capture the majority of them. Unfortunately, this strategy did not find the 56 (38%) spellings of *breathing* in the DOHMH database that did not begin with *breat*.

Relying on a word-stem approach not only misses cases but also requires an untenable level of labor-intensive maintenance. For example, for a system to recognize cases not beginning with *breat*, other word stems (e.g., *brath*, *bereath*, and *DIB*) need to be added. However, this strategy results in multiple false positives (e.g., *mandibular* fractures, *dibetes*, or the use of a *dibfulator*). Further logic is required to avoid retrieving mentions of any therapeutic *breathrough*. Eliminating such new false positives requires making further ad-hoc modifications and a continuing spiral of time-consuming maintenance and increasingly unreadable, error-prone code.

Even if a temporary state is reached in which false positives and negatives are minimal, new strings will keep arriving, making the previous logic inadequate. In the present study, even after 2 million chief complaints had been processed in the DOHMH system, approximately 750 new strings

appeared each week. Furthermore, when separate systems are joined, the complexity of the algorithmic logic must be increased again (e.g., although *diarrhea* was spelled 349 different ways in the DOHMH database, the EMA database contained an additional 154 spellings).

A system is far more maintainable if the medical logic regarding which concepts best represent a syndrome can be kept at a conceptual level, separate from the underlying technical intricacies of text processing.

Edit-Distance Methods

A third common method for matching strings is the edit-distance approach, which measures similarity as the minimum number of operations (e.g., insertions, deletions, substitutions, or transpositions) required to transform one string into another. Multiple modifications of this approach have focused primarily on computational efficiency in matching long strings (5). Edit distances, however, often give results inconsistent with human intuition. For example, the method would score both *azma* and *stomac* as equally close to *asthma*. Health professionals would not find this useful.

Generalized Edit-Distance Method

The text-normalization method developed for and used in this project is a generalization of the edit-distance approach — it models the similarity between two words as the minimum number of typographic errors, phonetic spelling errors, transcription errors, medical affixes (suffixes and prefixes), and concatenations that could transform one word into another. Because the method attempts to create the most plausible model of how a misspelled string could be generated, it is designed to represent the psychological distance between two strings rather than the computational distance.

As an example of the capabilities of this approach, the string *coughvomitingdiarre*, which actually appeared in a chief complaint, would be recognized by the text normalization software as an instance of the string *vomiting* (as well as of *cough* and *diarrhea*). Programs based on phonetic matching, edit distance, keywords, or the majority of other algorithms would not recognize the first string as a plausible instance of the second string.

Because the distances between words produced by the algorithm make intuitive sense (i.e., they correspond closely to the judgments about word similarity that would be made by humans), users can more easily work interactively with the computer or rely on the algorithm to make good decisions when run fully automatically. In one configuration, text-normalization software can be used as a pre-processor that passes normalized chief complaints or medical records as

input into a separate program dedicated to the higher-level task of recognizing syndromes and analyzing their frequency.

Applications of Text Normalization

To improve system performance, the text-normalization method was applied to the chief-complaint databases in three ways. The first use was a straightforward application of text normalization to automatically remove typographical errors, misspellings, word concatenations, and other forms of orthographic variation in chief complaints. The other two methods used text normalization as an essential tool for vocabulary expansion, in particular to search for overlooked abbreviations, acronyms, and other relevant vocabulary. Each application is described briefly.

Normalization of Chief Complaints

Chief complaints were presented to a text-normalization program, which compared each word in each chief complaint to a list of 68 key concepts that had been identified as useful for syndrome identification in the DOHMH syndromic surveillance algorithms. For example, the list included the words *pulmonary*, *pleuritic*, *cough*, *gasp*, and *dyspnea* for respiratory syndrome identification. Words sufficiently close to a key concept were matched with that concept.

To compare the performance of the text-normalization approach to orthographic variation with the often-used word-stem approach, the DOHMH word-stem algorithm for diarrheal syndrome was applied to the EMA chief-complaint data, both with and without prior text normalization. Each instance retrieved by one algorithm but not by the other was reviewed to determine which approach was correct. Of the 38,956 cases of diarrhea in the EMA database identified by either approach, 5,217 (13%) were recorded in a nonstandard way. When previously trained on the DOHMH chief complaints, the text-normalization program was able to identify all but five of these cases, an improvement of 896 when compared with cases recognized without normalization, while incurring only 17 false positives. Orthographic normalization alone improved performance by 2.3% when compared with the word-stem approach.

Using ICD-9-CM Codes To Uncover Overlooked Vocabulary

Although orthographic normalization generated substantial improvement, the possibility remained that additional words were being used to indicate symptoms and were being missed. Two additional approaches based on text normaliza-

tion were used to uncover overlooked vocabulary (e.g., unanticipated abbreviations, acronyms, and truncations).

The first approach was to use *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM) codes to select chief complaints most likely to contain vocabulary relevant to a particular syndrome. For example, the chief complaints of patient encounters assigned one of the ICD-9-CM codes in CDC's ESSENCE grouping for gastrointestinal (GI) syndrome (6) could be examined as a likely source for overlooked words that indicate GI syndrome. Specifically, these chief complaints were analyzed to see which words occurred more frequently in this ICD-9-CM GI syndrome group than in the cases not in that group. The procedure, in effect, searched for words with the highest relative risk of occurrence in the selected group as a means to detect words useful for designating GI syndrome.

In practice, this strategy is compromised because the seemingly innumerable misspellings and corruptions of words in the chief complaints result in an unmanageable list of word strings (e.g., 349 variations of *diarrhea*) whose relevance cannot be distinguished from the numerous irrelevant words that also occur with low frequency in the group. Used in this case, text normalization removes much of the noise and allows the relevant concepts to emerge.

The EMA database was used for this experiment because it contained both ED chief complaints and discharge ICD-9-CM codes for each case. Using the ICD-9-CM codes for GI syndrome with text normalization uncovered a number of relevant words not previously included in the DOHMH word-stem algorithms, including *cramps* (4,415 instances), *runs*, *NVD*, *LBM*, *Shigella*, *noninf* (1,689 instances, as in *noninf gastroenteritis*) and others.

Choosing a different subset of ICD-9-CM codes (e.g., only those codes that reflect intestinal rather than upper GI disease) might have uncovered yet additional words. The best strategies and criteria for choosing productive codes for synonym generation remain to be investigated. The potential benefits of using more precise and comprehensive coding schemes (e.g., SNOMED CT[®]) might also be explored (7).

Using Context To Uncover Overlooked Vocabulary

A second approach to retrieving overlooked vocabulary, as well as site-specific idiosyncratic vocabulary requiring customization, is adapted from the dictum in computational linguistics that "a word is known by the company it keeps" (8). This approach seeks to retrieve words with similar meanings by finding words that occur in similar contexts. Words

that co-occur with the same other words tend either to have similar meanings or at least to be closely related.

In this approach, for each word in the chief-complaint database, the words that most specifically occurred with that word were tabulated, resulting in a co-occurrence profile of closely associated words for each word. Each word was then compared with every other word to identify those with the most similar co-occurrence profiles. Similarity was assessed by using rank-order correlation between profiles. Examples of word strings of ≤ 5 letters uncovered by this method that would have been overlooked when using current word-stem algorithms are provided (e.g., 4,970 hive-like rashes would have been overlooked because *hives* was not previously a search term) (Table 3).

Detection Performance With and Without Text Normalization

Fortified with normalized text and additional vocabulary, a syndrome classifier operating on text that has been normalized can demonstrate greater sensitivity and specificity than a word-stem algorithm operating without normalization. Even though the two approaches will agree in the majority of cases, the cases where they differ are revealing.

Word-stem algorithms with and without text normalization were applied to detect instances of lower respiratory illness syndrome. On this particular task, in 3.3 million chief complaints, 201,327 instances were retrieved, and the sensitivity of the keyword and text-normalization approaches differed by 5.6% (11,252 instances). When the word-stem algorithm without normalization indicated presence of a lower respiratory illness but the algorithm using text normalization did not, the text-normalization approach was correct in 96.4% of cases. In the instances in which the text-normalization

TABLE 3. Expanding keyword vocabulary by locating words that appear in similar contexts in free-text chief-complaint data

Key concept*	Word strings of ≤ 5 letters with similar contexts [†]
Black	Dark, [§] brown, drk, [§]
Cough	Plegm, [§] cgh, [§]
Enteritis	Age [§]
Fever	Fevr, feve, fev, cough
Nausea	N, [§] NVD, [§] NV [§]
Pneumonia	RLL, pneu, [§] exac [§]
Rash	Rashes, hives [§]
SOB	DIB
Stool	Urine, dark, [§] brown, black, drk, [§] tarry, [§] BRBPR [§]

Source: New York City Department of Health and Mental Hygiene chief-complaint database.

* Key concepts used in free-text syndromic surveillance.

[†] Word strings with similar contexts, shown in order of computed similarity.

[§] Word strings that would have been overlooked when using current word-stem algorithms.

approach declared a syndrome to be present and the word-stem algorithm alone did not, human review of the chief complaints determined that text normalization was correct in 99.8% of cases. Use of text normalization thus substantially reduced the number of both false positives and false negatives (Table 4).

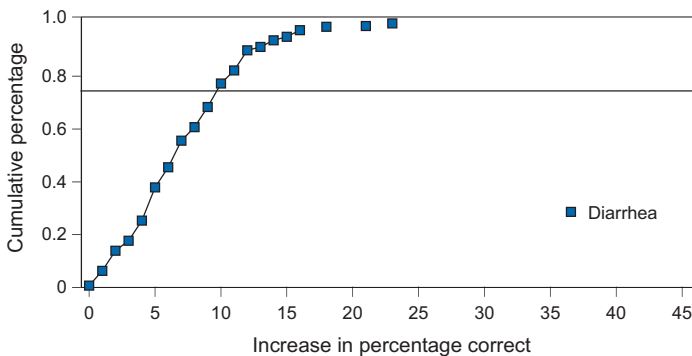
If text normalization were applied daily or in hospitals or surveillance systems with far fewer visits than 3.3 million, the differences between the two approaches might be greater still. The two approaches generated substantial differences when tracking diarrhea or bloody diarrhea syndrome in New York City hospitals with ≥ 100 visits for diarrhea per week (Figure). In approximately 25% of instances, the sensitivity of the word-stem approach was improved by 10%–20% when used with text normalization. In no case was the specificity decreased.

TABLE 4. Comparison of accuracy of word-stem algorithm with and without text normalization as applied to chief-complaint data in 12,270 instances in which the two approaches disagreed

Algorithm decision*	Reviewer determination	
	Present	Absent
Text normalization: present	11,238	14
Word stem (without text normalization): absent		
Text normalization: absent	37	981
Word stem (without text normalization): present		

*Decision of the algorithm regarding presence or absence of a given syndrome.

FIGURE. Effect of text normalization on free-text chief-complaint data in emergency departments with ≥ 100 diarrhea cases/week



A similar analysis was performed for fever/influenza syndrome (excluding upper respiratory illness), which comprises approximately 16.5% of New York City ED encounters. In 12% of instances, text normalization resulted in a 10%–20% improvement in sensitivity over the word-stem approach in tracking the number of fever/influenza chief complaints.

Conclusions

Incomplete vocabulary and word errors can have a substantial impact on the retrieval performance of free-text syndromic surveillance systems. Certain methods based on text normalization can greatly reduce the impact of these problems. New, increasingly sensitive methods of analysis will be most effective with careful attention to the quality of the data on which they rely.

Acknowledgments

Rick Heffernan and Farzad Mostashari of the New York City Department of Health and Mental Hygiene, Dennis Cochrane and John Allegra of Emergency Medical Associates of New Jersey, and Karen Olsen and Kenneth Mandl of Boston Children's Hospital provided portions of their chief-complaint data for this study.

References

- Mandl KD, Overhage JM, Wagner MM, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. *JAMA* 2004;11:141–50.
- Damerou F. A technique for computer detection and correction of spelling errors. *Communications of the ACM* 1964;7:171–6.
- Kukich K. Techniques for automatically correcting words in text. *ACM Computing Surveys* 1992;24:377–439.
- Zobel J, Dart P. Phonetic string matching: lessons from information retrieval. In: *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, 1996.
- Navarro G. A guided tour to approximate string matching. *ACM Computing Surveys* 2001;33:31–88.
- CDC. Syndrome definitions for diseases associated with critical bioterrorism-associated agents. Atlanta, GA: US Department of Health and Human Services, CDC, 2003. Available at <http://www.bt.cdc.gov/surveillance/syndromedef/>.
- McClay JC, Campbell J. Improved coding of the primary reason for visit to the emergency department using SNOMED. *Proc AMIA Symp* 2002;499–503.
- Firth JR. *Modes of meaning*. In: *Papers in linguistics*. London, UK: Oxford University Press, 1957.