Using Natural Language Processing to Improve Accuracy of Automated Notifiable Disease Reporting

Jeff Friedlin, DO, Shaun Grannis MD, J. Marc Overhage, MD, PhD
Regenstrief Institute, Inc, and Indiana University School of Medicine, Indianapolis, IN.

Abstract

We examined whether using a natural language processing (NLP) system results in improved accuracy and completeness of automated electronic laboratory reporting (ELR) of notifiable conditions. We used data from a community-wide health information exchange that has automated ELR functionality. We focused on methicillin-resistant Staphylococcus Aureus (MRSA), a reportable infection found in unstructured, free-text culture result reports. We used the Regenstrief EXtraction tool (REX) for this work. REX processed 64,554 reports that mentioned MRSA and we compared its output to a gold standard (human review). REX correctly identified 39,491(99.96%) of the 39,508 reports positive for MRSA, and committed only 74 false positive errors. It achieved high sensitivity, specificity, positive predicted value and F-measure. REX identified over two times as many MRSA positive reports as the ELR system without NLP. Using NLP can improve the completeness and accuracy of automated ELR.

Introduction

The reporting of notifiable diseases by health care providers is inadequate. Past studies of disease surveillance found that notifiable illnesses are greatly underreported by physicians[1] and underreporting hinders public health's ability to contain outbreaks in a timely manner.[2] Incomplete or delayed reporting stems from insufficient knowledge of reporting requirements, assumption that the laboratory reported the result, and lack of time or manpower.[1] Spontaneous reporting of notifiable diseases by clinical laboratories (whether paper-based or electronic) significantly improves the rate of reporting,[2] and *automated* electronic laboratory reporting (ELR) results in even more complete and timely reports.[3] However, limitations of ELR have been reported.[4] Laboratories often lack detailed patient demographic information required by public health departments, and for certain diseases, are unable to determine when a test result reflects a new case or chronic disease. Increasing amounts of patient medical data becoming electronic creates opportunities for further improvements in completeness and timeliness of ELR. An automated ELR system that leverages data from an integrated health information exchange (HIE) can overcome some of the above noted limitations by enhancing public health reporting with data such as recent laboratory results, enhanced patient demographics, medication history, etc. The Regenstrief Institute has maintained an operational automated ELR system[4] in the Indiana Network of Patient Care (INPC)[5], an operational regional HIE.

The automated ELR system first identifies potentially reportable candidate laboratory tests based on the Logical Observation Identifiers Names and Codes (LOINC®),[6] and transmits results classified as positive to public health. Laboratory reports with discrete consistent structure and limited answer ranges are easily detected and interpreted. However, some reports such as microbiology culture results are often reported as unstructured, freeform text, making the task of accurate interpretation by our system much more difficult. Although our ELR system implements text processing methodologies such as a longest common string comparator,[7] and basic negation detection (such as 'no' and 'negative'), it lacks sophisticated natural language processing (NLP) methods, which affects accuracy of ELR results classification. This is particularly true with unstructured, freeform, results.

Methicillin resistant Staphylococcus Aureus (MRSA) is a difficult-to-treat infection that has significantly increased in incidence over the past decade,[8] and is one such reportable condition found in unstructured reports. To determine MRSA prevalence, the Indiana State Department of Health (ISDH) recently mandated that *all* cases of MRSA (not just invasive MRSA) be reported. We sought to improve MRSA reporting accuracy using our ELR system. We hypothesized that we can improve the accuracy of our ELR system to report MRSA by using more sophisticated NLP to better identify MRSA positive reports. Studies show that NLP is effective in extracting relevant clinical data from text reports.[9, 10] To the best of our knowledge, there has been only one other study detailing the application of NLP to interpret microbiology reports.[11]

Methods

NLP System

We used the Regenstrief EXtraction tool (REX), described previously,[12] for this study. Briefly, REX is a rule-based NLP system written in Java. It has successfully extracted patient data and concepts from

radiology reports,[12] admission notes,[13] and pathology reports. REX has a modular design that can easily be adapted for specific NLP tasks. It uses regular expressions to detect where in the text keywords or phrases related to a concept are found. It then examines a window of words before and after the keyword(s) to determine the context (i.e. negated, ambiguous, historical, etc) in which they are found. A system based on regular expressions has successfully detected negations in text reports.[14] We chose this design because it closely mimics how humans determine the meaning of freeform text.

REX classifies Health Level 7 (HL7) result messages with MRSA keywords present into four categories:

1. Positive MRSA results
2. MRSA keyword present in a non-positive context (such as negated)
3. No MRSA keyword in result
4. Messages excluded from interpretation

Category 3 messages are tests in which no MRSA keyword is present in the result sections of interest (e.g. OBX-3, OBX-5 or NTE-3 fields). For example, the patient may live on '*Aureus Blvd'* with no MRSA keyword in the result. Category 4 messages are excluded using business rules. Examples of these rules include: a) not originating from a known source of laboratory results messages b) containing no result (HL7 OBX or NTE) segment, or c) the result is pending according to the OBX-11 component.

We focused on laboratory messages and excluded other kinds of reports that may contain MRSA, such as discharge summaries. While the presence of MRSA in a discharge summary may be valuable information for public health purposes, our specific aim in this project was to improve the accuracy of our notifiable disease reporting, and in this context public health agencies are primarily interested in *acute* MRSA infections that are chiefly detected in laboratory results. Additionally, we excluded pending results and processed only final results. For public health reporting purposes, only category 1 messages are transmitted. We further categorized the negative messages to characterize the accuracy of the process.

Data

Training Set

For our training set we selected all INPC laboratory HL7 messages that contained the keywords 'methicillin', 'MRSA' or 'Aureus' during the month of December 2007. The training set totaled 3,957 messages from 25 unique message sources. Approximately 97% of the tests were culture results.

Accurately localizing *where* in a message the target concept is found is crucial in our system because the concepts' position serves as the focal point around which subsequent contextual detection processes (such as negation) rely on. To improve concept localization, we used the training set as well as a literature search to identify frequently occurring words and phrases that express the concept of MRSA infection. We also used the training set to modify REX's contextual detection processes. We observed that common negation words and phrases found in most medical documents are not negations in the context of a culture result. For example, the word 'negative' usually conveys negation in reports such as admission notes. However, in culture results, 'negative' is commonly seen in a non-negation context such as '*gram negative rods were isolated'* or '*coagulase negative Staph.*' We therefore modified REX's context detection processes to account for these instances. REX can recognize false negations, and we added patterns to account for those seen in culture results. For example, a culture report may have the phrase '*positive for MRSA, not called to floor'*. In this instance, we must avoid interpreting 'not' as a negation term applied to MRSA, even though in most cases it would be. By adding patterns to REX's false negation database, we minimize false negative errors.

The training set revealed significant variation in laboratory systems' use of the HL7 standard to report their results. Some sources transmit one large report with each OBX-5 field representing one line in the report, while for other sources both the OBX-3 and the OBX-5 fields are part of a line of a report. Still other sources report their results in the NTE segment of the message and bypass the OBX-5 field completely. Lastly, some sources use varying combinations of the above three formats. REX required modification to account for these variations.

After initially modifying the software and knowledge base, we processed these reports to find false positive and false negative errors and made subsequent adjustments to REX. We performed several iterations of this testing-analysis-modification cycle to minimize errors.

Test Set

After modifying REX using the training set, we applied it to our test set, which consisted of all INPC HL7 messages containing the keywords 'MRSA', 'methicillin' or 'Aureus' during a one year period from January 25, 2007 to January 25, 2008 inclusive. The test set totaled 232,776 messages from 131 data sources and included 53,338 unique patients.

We performed a manual review (JF) of REX's output (64,554 messages) after processing the test set.

We reviewed all reports identified by REX as category 1 (positive for MRSA) and category 2 (MRSA keyword present but negated), and a random sample of category 3 and 4 reports. The final positive or negative determination was made by the physician reviewer. This became our gold standard. For both category 1 and category 2 messages we calculated REX's performance by comparing its output to the gold standard. Further, to evaluate the new algorithm's effect on MRSA reporting rates we randomly selected a subset of true positive messages from the test set and processed them with the existing operational notifiable condition processor that has limited NLP capability. We compared the two processes to evaluate any difference in reporting rates.

Results

A total of 232,776 HL7 messages containing MRSA keywords were sent to the INPC during the 12 month period. These reports contained over 12 million lines, and represented nearly 53,000 unique patients. All reports were input into REX for processing. A total of 116,551 (50.07%) messages were excluded (category 4), approximately 82% because they were not laboratory messages. A total of 51,671 (22.20%) messages did not contain MRSA keyword(s) in the results sections of interest (category 3). This left 64,554 (27.77% of the total) messages for interpretation. The remaining reports originated from 41 sources, represented 19,034 unique patients, and 97% were culture results. Of these, REX interpreted 39,565 (61.29%) as positive for MRSA and 24,989 (38.71%) as MRSA found but negated (category 2). Of the initial total of 232,776 messages, REX interpreted 17.00% of the reports as positive.

We manually reviewed all 64,554 category 1 and 2 reports, comparing them to the gold standard. We calculated specificity, sensitivity, positive predictive value (precision), and F-measure. Table 1 displays REX's performance vs. gold standard. REX achieved a sensitivity of 99.96%, a specificity of 99.71%, and a positive predictive value of 99.81%.

Of the 39,565 reports interpreted as positive by REX, we found only 74 false positives. Most of these (73) were due to a report format which caused improper sentence detection. For example, in the phrase '*MRSA DNA……NEGATIVE'* REX interpreted the '.' as a sentence delimiter thereby overlooking the negation term *'NEGATIVE'*. REX failed to identify only 17 positive MRSA reports. A spelling error caused one of these errors ('note' was misspelled 'not' in the report causing an erroneous negation). The other 16 false negatives were all caused by a report format which placed the phrase

	REX	
Manual Review	**Positives**	**Negatives[a]**
Positives	39,491	74
Negatives	17	24,972

Table 1. Results of MRSA interpretation of 64,554 HL7 messages by REX compared to gold standard
[a]includes only MRSA that was negated

'*rule out MRSA*' in the same sentence and close to a MRSA phrase in a positive context ('*rule out MRSA-culture wound positive for MRSA.'*). REX interpreted these reports as MRSA negative because the phrase 'rule out' is listed as a negation phrase. In these sentences, the first MRSA instance is negated, but the second is not. REX interpreted both MRSA instances as negated. REX correctly identified the negative qualifier in greater than 99% of the reports. REX achieved an F-measure of 0.9989 for the test data.

Manual review of a random sample of 5,000 category 3 messages and 10,000 category 4 messages revealed no laboratory messages positive for MRSA.

To compare the accuracy of REX to our existing ELR system, we collected a random sample of 454 blood culture result messages interpreted by REX and the gold standard as positive for MRSA. We compared the results of processing these messages with our existing ELR processor. REX accurately identified almost three times as many MRSA positive blood cultures as did our current ELR system. Our current ELR processor interpreted only 166 (36.56%) of these messages as positive.

Discussion

REX correctly interpreted the majority of HL7 laboratory messages containing MRSA. It misclassified few positive messages and its false positive rate reflected a small percentage of the total positive messages.

We originally developed REX to extract patient data from chest x-ray reports and admission notes, whose structure and content vary substantially from culture reports. We devoted approximately 30 man-hours to adapt the code and knowledge base to make REX operational for this project.

Because of the vagaries of real-world laboratory results, accurate NLP methods are required to automate the interpretation of these mostly free-text culture reports. This is evidenced by the large number of messages that contained MRSA keywords but were not true positives (24,972). By improving notifiable condition detection accuracy,

we minimize public health's burden to cull spurious results.

A chief barrier to accurate NLP classification was that many free text results disregarded grammatical rules. These reports frequently contained incomplete sentences, spelling errors, and/or lacked proper punctuation. REX, like the majority of medical NLP systems, processes text in the context of a single sentence. If text lacks the punctuation or structure to correctly delimit sentences, errors are more likely. Similar difficulties have been reported with NLP systems processing physician-typed clinic notes, which notoriously contain poor punctuation, non-standard abbreviations and spelling errors.[15]

Other challenges added to the complexity of this project. First, there are substantial variations in the way laboratories use the HL7 standard. This required the creation of multiple algorithms to account for these variations. We used HL7 messages for this study because the INPC and our current notifiable disease processor receive HL7 as input and we needed our NLP system to fully integrate with this message flow. Second, we found that many reports contained phrasing very similar to our target concept of MRSA. For example, several reports contained '*Methicillin resistant Staphylococcus Epidermidis*,' and *'Methicillin Resistant Staphylococcus Hominis'.* Simply searching for the keywords 'methicillin resistant' would cause false positive errors in these cases.

A limitation of this study is that the developer of the software also acted as the gold standard and evaluator of the data extraction process. In future studies, several trained experts not part of the development team will perform the evaluation of the data extraction process.

We found negation detection more difficult in these messages than in previously encountered medical documents. Negation terms in a sentence with MRSA often referred not to MRSA but to other items. For example, several reports contained phrasing such as '*few GPC MRSA no anaerobes seen*' and '*rare MRSA negative for inducible Clindamycin resistance*' and '*no collection times given, organism MRSA'* Also, several reports contained both negated and positive phrasing for MRSA in the same report as in: '*MRSA is positive in this specimen. The previous report of negative for MRSA was in error*.' It would be unusual to see similar phrasing such as '*I see evidence of CHF. When I said earlier that I saw no CHF, I was wrong.*' in a radiology report.

Finally, several reports were worded ambiguously, rendering it difficult and/or impossible to interpret for the human reviewer. Three examples are below:

'CULTURE STAPH only. Negative for Methicillin Resistant Staphylococcus aureus (MRSA). Staphylococcus aureus (MRSA) to date.'

'MRSA by PCR POSITIVE. No viable methicillin resistant S. aureus (MRSA) for isolation and/or susceptibility studies.'

'NO METHICILLIN-RESISTANT STAPH AUREUS (MRSA) ISOLATED. METHICILLIN-RESISTANT STAPH AUREUS (MRSA) ISOLATED.' For this last example, we called the laboratory to assist us in deciphering the message. They acknowledged that the message appeared to be reported in error.

Measures can be taken to augment automated interpretation of these types of messages, and laboratories can play a key role in increasing rates for reporting of notifiable diseases. By producing structured and coded messages easily interpreted by automated processes, the need for complex NLP methods would likely diminish. Standards are readily available to produce fully structured and coded HL7 laboratory reports. We will discuss the details of such a system.

First and foremost, an HL7 laboratory message should conform to the Health Information Technology Standards Panel (HITSP)[16] standard. HITSP provides guidelines for reporting laboratory results to ambulatory electronic medical records. For example, results of a culture report should be placed in the OBX-5 component, not in the NTE segment. Each segment should include a *single* result (results such as '*No shigella, salmonella or E.coli isolated'* in a single OBX-5 should be avoided). Also, an abnormal result should be flagged as such in the OBX-8 component.

In addition to including local codes for test *names*, laboratories should include LOINC codes in the HL7 message. This simplifies the task of identifying the specific test name, particularly in the context of a heterogeneous HIE. Similarly, the test *result* (OBX-5) should also be standardized. We recognize the need to include free text in the OBX-5 result section of a message for human readability. But the OBX-5 component should also contain the SNOMED-CT codes that map to the result (i.e. the organism or condition name). By structuring HL7 messages in this manner, accurate, automated interpretation is simply a matter of database calls and table look-ups rather than complicated processing of free text.

Using REX significantly improved our rate of reporting MRSA infections. In addition to public health uses, the identification of patients with notifiable conditions is also valuable for other

applications such as clinical decision support systems and research. We are currently using REX to augment our notifiable disease reporting to the ISDH and plan to add other conditions (i.e. Shigella, Salmonella) to REX's database.

Conclusion

Automatically classifying free-text culture reports for the purpose of public health reporting is complex and challenging. Part of the complexity relates to the use of the messaging standard itself. Although the messages we receive in our HIE are ostensibly in standardized HL7 format, there are substantial variations. The complexity of automatically classifying these messages can be mitigated by proper, consistent use of the HL7 standard. Additional complexity arises from the fact that information contained *within the message* is largely non-standardized and highly variable. Using standard result codes such as SNOMED encoding of microbiology results would reduce variation and improve accuracy of detecting notifiable conditions. We recognize the inherent complexities associated with improving message and content standards, but remain hopeful that ongoing national HIT standardization efforts will move healthcare toward these ends.

In spite of these complexities, improving accuracy of identifying notifiable conditions for public health by using incrementally advanced NLP methods is feasible within the context of an operational HIE. We significantly improved the reporting accuracy and completeness of our ELR in reporting MRSA infections using NLP.

References

1. Ward LD, Spain CV, Perilla MJ, Morales KH, Linkin DR. Improving disease reporting by clinicians: the effect of an internet-based intervention. J Public Health Manag Pract. 2008 Jan-Feb;14(1):56-61.
2. Schramm MM, Vogt RL, Mamolen M. The surveillance of communicable disease in Vermont: who reports? Public Health Rep. 1991 Jan-Feb;106(1):95-7.
3. Overhage JM, Grannis S, McDonald CJ. A Comparison of the Completeness and Timeliness of Automated Electronic Laboratory Reporting and Spontaneous Reporting of Notifiable Conditions. Am J Public Health. 2008 Jan 2.
4. Overhage JM, Suico J, McDonald CJ. Electronic laboratory reporting: barriers, solutions and findings. J Public Health Manag Pract. 2001 Nov;7(6):60-6.
5. McDonald CJ, Overhage JM, Barnes M, et al. The Indiana network for patient care: a working local health information infrastructure. An example of a working infrastructure collaboration that links data from five health systems and hundreds of millions of entries. Health Aff (Millwood). 2005 Sep-Oct;24(5):1214-20.
6. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem. 2003 Apr;49(4):624-33.
7. Friedman C, Sideli R. Tolerating spelling errors during patient validation. Comput Biomed Res. 1992 Oct;25(5):486-509.
8. Maree CL, Daum RS, Boyle-Vavra S, Matayoshi K, Miller LG. Community-associated methicillin-resistant Staphylococcus aureus isolates causing healthcare-associated infections. Emerg Infect Dis. 2007 Feb;13(2):236-42.
9. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc. 1994 Mar-Apr;1(2):161-74.
10. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004 Sep-Oct;11(5):392-402.
11. Hripcsak G, Knirsch CA, Jain NL, Pablos-Mendez A. Automated tuberculosis detection. J Am Med Inform Assoc. 1997 Sep-Oct;4(5):376-81.
12. Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annu Symp Proc. 2006:269-73.
13. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc. 2006:925.
14. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301-10.
15. Barrows Jr RC, Busuioc M, Friedman C. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. Proc AMIA Symp. 2000:51-5.
16. Health Information Technology Standards Panel Web site. Available at: www.ansi.org/hitsp/ Accessed Feb 25, 2008.