

Comparing Expert Systems for Identifying Chest X-ray Reports that Support Pneumonia

Wendy Webber Chapman, Peter J. Haug

Department of Medical Informatics, University of Utah, Salt Lake City, Utah 84132

ABSTRACT

We compare the performance of four computerized methods in identifying chest x-ray reports that support acute bacterial pneumonia. Two of the computerized techniques are constructed from expert knowledge, and two learn rules and structure from data. The two machine learning systems perform as well as the expert constructed systems. All of the computerized techniques perform better than a baseline keyword search and a lay person, and perform as well as a physician. We conclude that machine learning can be used to identify chest x-ray reports that support pneumonia.

INTRODUCTION

Current health information systems (HIS) contain patient information that can be used by decision support systems to improve patient care. Chest x-ray reports are stored in most HIS's, but because they are in free-text form instead of coded form they are seldom used by computer-aided decision tools. The information contained in a chest x-ray report can be useful for point-of-care diagnosis, quality assurance, resource management, and clinical guideline processes.

One method for obtaining coded data from chest x-ray reports is natural language processing. At LDS Hospital we have developed a natural language understanding system (NLUS) called SymText [1] that encodes most diseases and findings contained in a chest x-ray report. Our long term goal is to use coded data from SymText as the substrate upon which expert systems will work to infer a variety of diagnostic states. In this paper we test four computerized methods that interpret SymText's coded output and store the interpretation of whether or not a chest x-ray report supports acute bacterial pneumonia. We attempt to select the best expert system for identifying pneumonia given correctly encoded data from chest x-ray reports.

There are three steps in determining the quality of a disease inference made from SymText's output of chest x-ray reports: 1) Test how well the NLUS encodes specific concepts related to pneumonia; 2) Given accurate input, select the best inferencing method for detecting the presence of pneumonia; and 3) Analyze how the inferencing algorithm performs given parsed data, which invariably contains errors. For results on the first step, see [2]. This paper focuses on the second step by testing various inferencing algorithms on correctly parsed data. The third step is partially covered in [3] and will be part of a larger experiment where uncorrected data directly from the parser provides the input to the expert system.

In this paper we examine two hypotheses:

1) A machine learning algorithm can perform as well

as an expert constructed system in inferring the presence of pneumonia.

2) The best computerized technique can perform as well as a physician can.

In order to isolate the performance of the inferencing techniques from that of the parser, we test them all on manually encoded, correct data.

MOTIVATION

We selected the disease pneumonia for several reasons. First, the appearance of pneumonia and related findings is frequent enough in x-ray reports to provide a reasonable test set. Second, pneumonia is the most frequent infectious disease cause of death in patients of all ages, and the sixth overall cause of death in the U.S. [4]. Identifying pneumonia is very important because of the very distinct therapeutic and prognostic features of the disease, and a key element in the diagnosis of pneumonia is the chest radiograph [5].

Third, two computer-aided decision support systems in use at LDS Hospital require information regarding pneumonia's presence or absence in the chest x-ray. This piece of information is often the only piece of information they can not gather directly from the HIS.

The first of these decision support systems is the Antibiotic Assistant [6], a program in use at LDS Hospital that assists physicians in the use of anti-infective therapy. To find patients who might need antibiotics, the program screens all chest x-ray reports for concepts related to pneumonia. When a report contains any of these concepts, other patient data such as white-cell count, temperature, microbiology reports, etc. are collected from the HELP system and analyzed to determine a need for anti-infective therapy. The Antibiotic Assistant currently uses a sophisticated keyword search that is compared to SymText in [2].

The second decision support system is a Bayesian network being implemented in the LDS Hospital emergency room as part of a larger system to assist physicians in the management of pneumonia patients [7]. One of the 26 nodes in the network represents whether or not the chest x-ray supports pneumonia.

The common thread between these two systems and other applications is the need to recognize whether or not a chest x-ray supports pneumonia. In addition to decision support tools, information about pneumonia in a chest x-ray report can be used to perform quality assurance for radiologists' reporting of pneumonia or to perform utilization reviews of chest x-rays ordered to rule out pneumonia.

BACKGROUND

Other NLUS output used for decision support

Others have tested the performance of an NLUS in

supporting real medical processes. Some NLUS's were built to extract specific information required by a decision support system [8,9]. General purpose NLUS's like SymText have also been tested for specific decision support purposes [10,11,12].

Our work is largely based on that of Wilcox and Hripsak [13] in which the input of their general purpose parser was fed to both hand-crafted rules and a machine learning algorithm called C5.0. They tested the two different inferencing algorithms on six diseases (including pneumonia). Their results show that the hand-crafted rules outperform C5.0. We hope to have learned from their mistakes in training the machine learning system by limiting the training sets to fewer variables to avoid excess noise in the training. Our experiment differs from theirs by only testing one disease state and by isolating the evaluation of the expert systems tools by using the output of a "perfect" parser.

Why use an inferencing algorithm?

Readers may wonder why coded data from an NLUS is not sufficient for inferring pneumonia in the report. After all, several of the concepts we model are types of the disease pneumonia, including *aspiration pneumonia*, *bacterial pneumonia*, *viral pneumonia*, and *atypical pneumonia*. The simplest way to decide if a report suggested acute bacterial pneumonia would be to search for the coded concept of *bacterial pneumonia*. Easier still, one could search for the word "pneumonia" in the report. If the word appeared without any negatives preceding it, the report suggests pneumonia.

Unfortunately, chest x-ray reports are not always straight forward. Often terms such as "localized alveolar infiltrate" are used to imply pneumonia. Even infiltrates are not called infiltrates in many reports but are described with phrases such as "hazy" or "patchy opacities." Determining whether or not the patient has pneumonia from the report itself, without other knowledge about the patient's condition, requires inferencing from findings that are mentioned in the report.

METHODS

Input

Figure 1 represents the overall process of identifying pneumonia in chest x-ray reports.

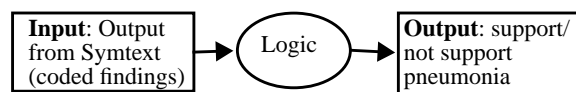


Figure 1. Process of identifying chest x-ray reports that support pneumonia

The input to the inferencing algorithm is a list of coded findings and diseases present in the report. SymText models 167 observations (findings and diseases) found in chest radiographs, along with characteristics describing the observations such as state, location, severity, change, etc. Figure 2 shows the output for the sentence "The infiltrate in the right upper lobe has increased in

size."

Observation: localized upper lobe infiltrate
State: present
Topic: infiltrate
Location: right upper lobe
Severity: null
Change with time: increased

Figure 2. Partial Symtext output for the sentence "The infiltrate in the right upper lobe has increased in size."

Although SymText's output for one observation contains information about the location, severity, etc., for this project we use only the concepts that are in bold in Figure 2, i.e., the observation concept and its state.

In this paper we test the computerized inferencing algorithms on data that is identical to SymText's output except that all of it is correct. Data was manually encoded by one of the authors with a tool used to correct output from SymText. We began with SymText's coded observations and corrected those that were both incorrect and related to pneumonia; all encodings that were not related to pneumonia or that were related to pneumonia and were already correct were simply accepted as they were output. In this way, the manually encoded data uses the same concepts SymText models. In addition, any information relating to the patient's clinical history was removed by the manual coder. Using manually encoded data allows us to isolate the inferencing algorithms from the variability of parser input in determining which of the methods compared performs the best.

Inferencing Algorithms

We will test five inferencing algorithms.

1. Expert crafted rules

A physician in our group created the following rules to determine whether or not a report supports pneumonia.

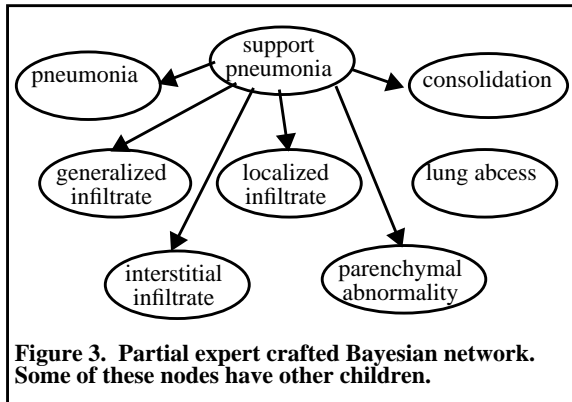
If any of the following concepts are present, the report supports pneumonia:

- *pneumonia*
- *aspiration pneumonia*
- *localized consolidation*
- *consolidation (nos)*
- *localized infiltrate (nos)*
- *localized upper lobe infiltrate*
- *localized lower lobe infiltrate*
- *perihilar infiltrate*
- *generic infiltrate*
- *local peripheral infiltrate*

2. Expert crafted Bayesian net (Bayesian Network 1)

A Bayesian network is a directed acyclic graph in which variables are nodes and conditional probabilities between those variables are links between corresponding nodes [15]. We use the software application Netica™ for creating and interpreting Bayesian networks. With input from a radiologist, the same physician who created the rules created a Bayesian network (see Figure 3) in which the top node is support pneumonia: yes or

no.



To choose a threshold for the binary decision “presence or absence of pneumonia,” we constructed an ROC curve with varying thresholds for presence of the head node. We selected .30 for our test threshold, because it maximized both sensitivity and specificity. Thus, if *support pneumonia* is present with .30 probability or above, it is considered true.

3. Machine Learned Decision Tree/Rules

We used Quinlan’s See5.0 Decision Tree software (See5.0/C5.0 RuleQuest Research Pty Ltd, Australia) to learn rules from data [14]. Given a vector of attribute-value (a-v) pairs for every training case, the algorithm calculates the information gain every pair contributes and chooses the one with the highest value to be the first branch of a tree. The winning a-v pair is not considered again and the information gain for the remaining a-v pairs are calculated, selecting the pair with the highest score and placing it at the next branch of the tree. The process repeats until all training cases can be classified successfully. The tree can then be translated into a list of sequential rules.

We trained the system on 298 chest radiograph reports. Every report was represented as an a-v vector containing 168 attributes all with binary values of *absent* or *present*. The first 167 attributes were every possible disease and finding concept represented by SymText; their values were the state of the attribute within that report, i.e., *present* or *absent*. The default value for concepts not discussed in the report was *absent*. The last attribute was the correct classification of *support/not support pneumonia* which was encoded by agreement of two physicians with a radiologist ruling on disagreements. From the 298 reports See5.0 created nine rules, as shown below.

1. If *consolidation nos = present*, support pneumonia
2. If *localized consolidation = present*, support pneumonia
3. If *pneumonia = present*, support pneumonia
4. If *localized infiltrate = present*, support pneumonia
5. If *localized upper lobe infiltrate = present*, support pneumonia
6. If *aspiration pneumonia = present*, support pneumonia
7. If *localized lower lobe infiltrate = present*, support pneumonia
8. If *perihilar infiltrate = present*, support pneumonia

9. If concepts 1-8 = *absent*, not support pneumonia

4. Bayesian network created from decision tree rules (Bayesian Network 2)

We used the attributes See5.0 deemed important to create another Bayesian network. The idea behind this system is to combine the strength of See5.0’s rules, i.e., learning structure from data, with the strength of Bayesian networks, i.e., the ability to use probabilistic reasoning. This Bayesian network has the same root node as the one in Figure 3, with children nodes representing every attribute listed by See5.0’s rules. After varying the thresholds of an ROC curve, we selected .40 as the best probability for positive support of pneumonia for this network.

5. Simple Keyword Search

In testing his Bayesian network for diagnosing pneumonia, Aronsky [7] used a simple keyword search to classify chest x-ray reports according to whether or not they support pneumonia. We adapted his algorithm as a baseline for our more complex methods. The keyword search counts all occurrences of pneumonia related words/word fragments (“bronchogram,” “consolidat,” “pneumoni, opaci,” “infiltrat,” and “multilobar”) in a report. It then adds that number to a count of negative terms in the report (“not” and “no evidence”). If their sum is greater than or equal to two, the report is classified as supporting pneumonia.

Test Set

The test set is comprised of 150 chest x-ray reports obtained from the HELP system at LDS Hospital. Half of the reports were obtained from patients with a primary discharge diagnosis of pneumonia, while the other half were randomly selected. All chest x-ray reports (not just the first report) were possible candidates for the test set. Sixty (40%) of the reports support pneumonia.

Gold Standard

The gold standard (GS) is comprised of majority opinion of three physicians (two internists and one radiologist) who read all 150 reports independently and classified every report as to whether or not it supported pneumonia. We chose not to use the discharge diagnosis as the gold standard, because the output from our system will be used by other systems that need to know only whether or not the chest x-ray supports pneumonia, not whether or not the patient actually had pneumonia.

Test Metrics

Figure 4 describes metrics we use to measure accuracy.

| | |
|--------------|---|
| Recall: | $\frac{\text{number of correct pneumonia inferences}}{\text{number of GS pneumonia inferences}}$ |
| Precision | $\frac{\text{number of correct pneumonia inferences}}{\text{number of attempted pneumonia inferences}}$ |
| Specificity: | $\frac{\text{number of correct no-pneumonia inferences}}{\text{number of GS no-pneumonia inferences}}$ |

Figure 4. Metrics for measuring accuracy in classifying pneumonia in a chest radiograph.

Although we report recall, precision, and specificity for the techniques, our statistical test, McNemar's, doesn't test any single one of these measures. Instead, McNemar's is a paired test that looks at every instance of disagreement to give an overall statistic.

RESULTS

Hypothesis 1: A machine learning algorithm can perform as well as an expert constructed system.

Table 1 compares the performance of the five algorithms.

Table 1: Comparison of Five Algorithms

| System | Recall | Precision | Specificity |
|--------------------|----------------|----------------|----------------|
| Rules | 53/60 (87%) | 52/61 (85%) | 81/90 (90%) |
| Bayesian Network 1 | 56/60 (93%) | 56/66 (85%) | 80/90 (89%) |
| Decision Tree | 52/60 (87%) | 52/60 (87%) | 82/90 (91%) |
| Bayesian Network 2 | 52/60 (87%) | 52/58 (90%) | 84/90 (93%) |
| Keyword Search | 32/60 (53%) | 32/46 (70%) | 76/90 (84%) |

According to McNemar's test with Bonferroni corrections, none of the systems differ from one another except the keyword search which is worse, $p < .001$.

Hypothesis 2: The best computerized technique can perform as well as a physician can.

To test the hypothesis that a computerized system can perform as well at classifying reports for pneumonia as a physician can, we compared the results of Bayesian Network 1 shown in Table 1 against an individual physician and a lay person. The results are shown in Table 2.

Table 2: Comparison of Computerized System, Physician, and Lay Person

| System | Recall | Precision | Specificity |
|--------------------|----------------|----------------|----------------|
| Bayesian Network 1 | 56/60 (93%) | 56/66 (85%) | 80/90 (89%) |
| Physician | 54/60 (90%) | 54/71 (76%) | 73/90 (81%) |
| Lay Person | 32/60 (53%) | 32/41 (78%) | 81/90 (90%) |

A McNemar's test with a Bonferroni correction showed the Bayesian network performed the same as the M.D. and different than the lay person ($p < .001$). Yet, the M.D.

and the lay person did not differ in their overall performance ($p = .039$).

DISCUSSION

Hypothesis 1: A machine learning algorithm can perform as well as an expert constructed system.

All computerized techniques were statistically equivalently except for the keyword search which was designed to be conservative and was used as a baseline comparison. Using only the observations and states modeled by SymText, See5.0 was able to create rules that performed as well as expert written rules. Machine learning methods offer great possibility for creating rules without expert knowledge.

Bayesian Network 1 had the highest recall. High recall is most important in a tool used for screening, like the x-ray report identifier in the Antibiotic Assistant.

The results of Bayesian Network 2, created from See5.0's rules, are hopeful. Because Bayesian Network 2 had higher precision and specificity than Bayesian Network 1, it could be more useful to a system used for alerts, generating fewer false positives than Bayesian Network 1 while still being very sensitive. However, Bayesian Network 2 acts almost like a rule-based system with only two nodes using non-binary probabilities, *left upper lobe infiltrate* and *perihilar infiltrate*. It does perform slightly better than the Decision Tree itself, though not statistically different.

Hypothesis 2: The best computerized technique can perform as well as a physician can.

None of the computerized techniques except for the keyword search performed differently from the physician. Statistically Bayesian Network 1 performed the same as the physician but differently than the lay person. However, the physician performed the same as the lay person. McNemar's looks at the overall disagreement among observers, but we can see that the M.D. performed better in recall (90%) compared to the lay person (53%). A Z-test for proportions confirms this ($p < .001$). It makes sense for the lay person to fall short of the physician in recall where expert knowledge is needed to identify pneumonia. The fact that the lay person performed better than the physician in specificity can be explained by the tendency to classify reports negatively when you are not familiar with evidence to indicate otherwise.

We have shown that classifying a chest x-ray report for pneumonia can be accomplished with a computerized system.

The Gold Standard

Other studies have shown disagreement among experts in reading chest radiographs [16] and radiograph reports [12]. We also see disagreement among our gold standard experts. Of the 150 reports, 103 were unanimously classified. Kappa tests among the experts show moderate agreement with significant kappas ranging from .52 to .61. Disagreement ranged between 17% and 25%, which reinforces our claim that chest x-ray reports are not easily classified as to pneumonia. All computerized methods and the individual physician agreed on seven cases that were incorrect according to the GS. Examining these cases reveals ambiguity in the words used within

report itself (e.g., is the term "air space disease" definitely indicative of pneumonia?) and disagreement on the significance of a given finding in indicating pneumonia (e.g., "an area of atelectasis or consolidation" was sometimes classified by the experts as supporting pneumonia and other times not). Perhaps the big difference among the experts was their threshold for classifying a report as supportive of pneumonia. The computerized systems all take the approach a screening algorithm would take: if there is any possibility this patient has pneumonia, label it as positive. The GS experts were more conservative in labeling a report positive.

Computerized Techniques and Humans

The four computerized techniques were almost identical to each other in the reports they missed but were different from the physician. In fact, fifteen of the classifications the computerized techniques missed were missed by all four techniques. How does a computerized method compare to a human? Computerized techniques are more consistent than humans - if a report says there is *consolidation*, the report suggests pneumonia - whereas a human might not come to the same conclusion in every case. Several hand-written comments from our experts indicated that "even though the report mentioned consolidation, I think it's ARDS." Such a discrepancy between computers' consistency and humans' intuition makes it hard to compare the two.

Future Work

Once we have completed both the analysis of SymText's accuracy for pneumonia related concepts and the analysis of the expert systems tools, we will perform a final test to determine if the uncorrected output of the parser combines with the expert system to provide an adequate assessment of support for pneumonia in chest x-ray reports. For this test we will increase the test set to 300 reports and will test the results on both correctly coded data and data parsed by SymText. Also, our GS will be a panel of five physicians instead of three. Eventually, the pneumonia-related concepts encoded by SymText and the inference of whether or not the report suggests pneumonia will be stored on our HIS for use by decision support systems and other computer-aided tools.

CONCLUSION

We have shown that a computerized system can perform as well as a physician in identifying chest x-ray reports that support pneumonia. In this realm machine learning can replicate the results of hand crafted systems. Like others we struggle with creating a reliable gold standard in a field where language, like the opacities it describes, is vague, hazy, and ill-defined.

Acknowledgments.

We would like to thank Dominik Aronsky, Bruce Bray, David Chapman, Marcelo Fiszman, Phillip Frederick, Greg Patten, and Ken Zollo for their collaboration on this project. This work was supported by NLM grant 1 R01 LM 06539-02.

References

[1] Koehler, SB. SymText: Understanding Natural Language Medical Text. Dissertation, University of Utah, 1998.

- [2] Fiszman M, Haug PJ, Automatic Identification of Pneumonia Related Concepts on Chest x-ray Reports. Submitted to 1999 AMIA Conference.
- [3] Chapman WW, Fiszman M, Haug PJ. Correct vs. Parsed Data for Inferring Pneumonia in Chest x-ray Reports that Support Pneumonia. Submitted to 1999 AMIA Conference.
- [4] Meehan TP, *et. al.* Quality of Care, Process, and Outcomes in Elderly Patients With Pneumonia. JAMA (1997);278:2080-2084.
- [5] Metlay JP, Kapoor WN, Fine MJ. Does This Patient Have Community-Acquired Pneumonia? Diagnosing Pneumonia by History and Physical Examination. JAMA (1997);278:1440-1444.
- [6] Evans RS, *et. al.* A Computer-Assisted Management Program for Antibiotics and Other Antiinfective Agents. NEJM (1998);338(4):232-238.
- [7] Aronsky D, Haug PJ. AMIA Proceedings (1998), pp 632-6.
- [8] Lin R, Lenert LA, Middleton B, Shiffman S. A Free-Text Processing system to Capture Physical Findings: Canonical Phrase Identification System (CAPIS). Sixteenth Annual Symposium on Computer Applications in Medical Care (1992), pp 168-172
- [9] Zingmond D, Lenert LA. Monitoring Free-Text Data Using Medical Language Processing. Computers and Biomedical Research (1993);26:467-481.
- [10] Borst F, Lyman M, Nhan NT, Tick LJ, Sager N, Scherrer JR. TEXTINFO: A Tool for Automatic Determination of Patient Clinical Profiles Using Text Analysis. Sixteenth Annual Symposium on Computer Applications in Medical Care (1992), pp 63-67.
- [11] Jain NL, Knirsch CA, Friedman C, Hripcsak G. Identification of Suspected Tuberculosis Patients based on Natural Language Processing of Chest Radiograph Reports. Twentieth Annual Symposium on Computer Applications in Medical Care (1996), pp 542-546.
- [12] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing. Annals of Internal Medicine (1995);122:681-688.
- [13] Wilcox A, Hripcsak G. Knowledge Discovery and Data Mining to Assist Natural Language Understanding. AMIA Proceedings (1998), pp. 835-9.
- [14] Quinlan JR. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.
- [15] Szolovits P. Uncertainty and Decisions in Medical Informatics. Methods of Information in Medicine (1995); 34:111-121.
- [16] Yerushalmy J. The Statistical Assessment of the Variability in Observer Perception and Description of Roentgenographic Pulmonary Shadows. Radiologic Clinics of North America (1969); 7;3:381-392.