Journal of Biomedical Informatics **34,** 4–14 (2001)
doi:10.1006/jbin.2001.1000, available online at http://www.idealibrary.com on IDEAL®



# Identifying Chest X-Ray Reports That Support Pneumonia

Wendy Webber Chapman,* Marcelo Fizman,† Brian E. Chapman,‡ and Peter J. Haug†

**Center for Biomedical Informatics and ‡Department of Radiology, University of Pittsburgh, Pittsburgh, Pennsylvania 15213; and †Department of Medical Informatics, University of Utah/LDS Hospital, Salt Lake City, Utah*

*Received December 20, 2000; published online March 13, 2001*

We compared the performance of expert-crafted rules, a Bayesian network, and a decision tree at automatically identifying chest X-ray reports that support acute bacterial pneumonia. We randomly selected 292 chest X-ray reports, 75 (25%) of which were from patients with a hospital discharge diagnosis of bacterial pneumonia. The reports were encoded by our natural language processor and then manually corrected for mistakes. The encoded observations were analyzed by three expert systems to determine whether the reports supported pneumonia. The reference standard for radiologic support of pneumonia was the majority vote of three physicians. We compared (a) the performance of the expert systems against each other and (b) the performance of the expert systems against that of four physicians who were not part of the gold standard. Output from the expert systems and the physicians was transformed so that comparisons could be made with both binary and probabilistic output. Metrics of comparison for binary output were sensitivity (sens), precision (prec), and specificity (spec). The metric of comparison for probabilistic output was the area under the receiver operator characteristic (ROC) curve. We used McNemar's test to determine statistical significance for binary output and univariate *z*-tests for probabilistic output. Measures of performance of the expert systems for binary (probabilistic) output were as follows: Rules—sens, 0.92; prec, 0.80; spec, 0.86 ($A_z$, 0.960); Bayesian network—sens, 0.90; prec, 0.72; spec, 0.78 ($A_z$, 0.945); decision tree—sens, 0.86; prec, 0.85; spec, 0.91 ($A_z$, 0.940). Comparisons of the expert systems against each other using binary output showed a significant difference between the rules and the Bayesian network and between the decision tree and the Bayesian network. Comparisons of expert systems using probabilistic output showed no significant differences. Comparisons of binary output against physicians showed differences between the Bayesian network and two physicians. Comparisons of probabilistic output against physicians showed a difference between the decision tree and one physician. The expert systems performed similarly for the probabilistic output but differed in measures of sensitivity, precision, and specificity produced by the binary output. All three expert systems performed similarly to physicians. 

## INTRODUCTION

Computerized clinical guidelines and decision support systems have been developed to help physicians diagnose and manage disease. Success of these automated systems depends largely on the availability of computable clinical data in medical information systems. Computerized patient records currently contain a variety of patient-specific data, including lab values, nursing notes, admit and discharge diagnoses, and radiology reports. Some of the computerized data are represented in a computable format that can be automatically accessed and manipulated by decision support systems. However, a large portion of the patient data, including history and physical exams, discharge notes, and radiology reports, is stored as narrative reports. Narrative reports are available for review or for printing but are not accessible to computerized guidelines or decision support systems. Thus computerized systems do not have access to important



1532-0464/01 $35.00

clinical information such as physical symptoms, imaging observations, and clinical assessments. Access to the missing information would enhance the performance and increase the usefulness of automated systems.

Computerized clinical guidelines and decision support systems have been implemented to help physicians diagnose and manage pneumonia and other infectious diseases [1–3]. Because information from the chest X-ray is a key component in diagnosing pulmonary diseases, these systems need automatic access to observations from chest X-ray reports. Often the information from chest X-ray reports is the only important piece of information the systems cannot automatically access [4].

Some groups have developed natural language processing (NLP) methods for extracting information from chest X-ray reports and representing that information in a computable format [5–13]. Coded output from an NLP system can be stored in a hospital database, providing encoded data not previously available to computerized methods.

NLP systems differ in the nature of their output and in the degree to which the systems generate structured, semantically interpretable output. One goal of natural language processing is to model the underlying idea represented by various combinations of words. Therefore, output from an NLP system contains inferences made from the phrases in the text. An extension to the local inferencing made from the text is global inferencing in which information from the coded output is combined to classify the document.

In this paper we compare three computerized methods that interpret the coded output of our NLP system to determine whether a chest X-ray report contained enough information to support a diagnosis of acute bacterial pneumonia. We examine a rule based system, a probabilistic system called a Bayesian network, and a machine learning system called a decision tree. To isolate the performance of the inferencing techniques from that of an imperfect NLP system, we corrected mistakes in the NLP system's output before testing the expert systems.

## BACKGROUND

Our study addresses the disease pneumonia for two reasons. First, pneumonia-related information is frequent enough in chest X-ray reports to provide a reasonable test set. Second, two computerized decision support systems currently in use at LDS Hospital in Salt Lake City, Utah, require information regarding pneumonia's presence or absence in a chest X-ray [2–4].

Other researchers have used the output of an NLP system to support real medical processes. Some NLP systems were built to extract specific information required by a decision support system [12]. Other NLP systems have been created as more general purpose systems that attempt to extract all diseases and findings that are commonly discussed in a chest X-ray report. Hripcsak and co-workers have tested the ability of their NLP system (MedLee) combined with medical logic modules to identify suspected tuberculosis patients [1, 14]. MedLee has also been tested for accuracy in recognizing six other clinical concepts [7].

### *Natural Language Processing with SymText*

We have created an NLP system called SymText that is in use at LDS Hospital [13]. SymText is an NLP system created for use on chest X-ray reports [15]. SymText has also been applied to admit diagnoses [16] and ventilation/perfusion scan reports [17]. SymText is comprised of a syntactic and a semantic component. The syntactic component contains an augmented transition network [18] combined with a system for grammatical transformations. The semantic component is comprised of Bayesian networks that model the domain of interest. For chest X-rays we model all findings, diseases, and devices found on a chest radiograph. The input to SymText is a sentence. SymText's output is an attribute-value template that has been instantiated based on relevant words in the sentence. Figure 1 represents SymText's output for the sentence "The hazy opacity in the right upper lobe has increased in size."

```
Observation: *localized upper lobe infiltrate (.689)
State: *present (.856)
Topic concept: *poorly-marginated opacity (infiltrate) (.999477)
   Topic term: opacity (1.0)
   Topic modifier: hazy (1.0)
Anatomic concept: *right upper lobe (0.999991)
Anatomic link concept:*involving (0.999477)
   anatomic link term: in (1.0)
   anatomic location term: lobe (1.0)
   anatomic location modifier: null (0.936417)
   anatomic modifier side: right (1.0)
   anatomic mod. superior/inferior: upper (1.0)
Severity: *null (0.9999)
Change with time: *increased (0.9778)
   change term:increased (1.0)
```

**FIG. 1.** Partial Symtext output for the sentence "The hazy opacity in the right upper lobe has increased in size." Numbers in parentheses are probabilities. Values with a probability of 1.0 are words from the sentence. Values with an asterisk (*) are concepts SymText has inferred from the words.

We have previously tested SymText's ability to extract specific pneumonia-related concepts from chest X-ray reports [19]. Once the pneumonia-related concepts are correctly encoded, the concepts can be used as input to an expert system that will determine whether the information in the report supports pneumonia.

### *Classification Algorithms*

Determining whether a report supports pneumonia is a classification problem. Several classification algorithms exist that analyze specific attributes describing an object to classify that object into a predefined category. We compare three classification algorithms: a rule based system, a Bayesian network [20, 21], and a decision tree [22, 24].

Two distinctions among these algorithms are particularly relevant to this project. The first distinction relates to how the algorithm knows which values predict specific classifications. An expert system can learn classification patterns from data or from experts. In our case, the Bayesian network and the decision tree learned which combinations of attributes predict a given classification from patterns in training cases. Conversely, the rule-based system learned which attributes predict which classification from experts.

The second distinction relates to creation of the expert system's structure. The structure of an expert system can also be either learned from data or created by an expert. In this study, the structure of the decision tree was derived directly from training cases, whereas the structures of both the rules and the Bayesian network were created by experts.

For this project we compared one algorithm that was completely created by expert input (rules), one algorithm whose structure was created by experts but whose classification patterns were learned from data (Bayesian network), and one supervised machine learning technique that derived both classification patterns and structure from data (decision tree). Wilcox and Hripcsak have examined practical aspects of applying classification algorithms to the output of an NLP system to infer a variety of clinical conditions [25, 26]. In this paper we compare three classification algorithms that classify reports as supporting or not supporting pneumonia.

## METHODS

Figure 2 represents the overall process of identifying pneumonia in chest X-ray reports.

Below we describe input to the expert system, the gold standard, the training and test sets, the three inference algorithms we compared, and our test metrics.

### *Input to the Inferencing Algorithms*

Input to the inferencing algorithms was a list of finding and disease templates encoded by SymText for a chest X-ray report. SymText models a combination of 167 findings and diseases described in chest X-ray reports, along with characteristics describing the observations such as state, location, severity, and change (see Fig. 1). For this project we used only the observation concept and its state, e.g., *localized upper lobe infiltrate − present*.

We tested the computerized inferencing algorithms on a corrected version of SymText's output. Two of the authors reviewed separate portions of SymText's coded observations, correcting observations 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.

Because our purpose was to determine if an automated system could detect reports that indicated pneumonia, we removed any information in the report that did not relate to the X-ray itself. Therefore, the patient's clinical history was removed from the beginning of the report by the manual coder.

### *Gold Standard*

A study involving interpretation of radiology reports requires a gold standard of physicians who read and interpret the reports. On one hand, radiologists may understand the intent of a radiology report better than clinicians. The radiologists who dictated the reports would be the best source

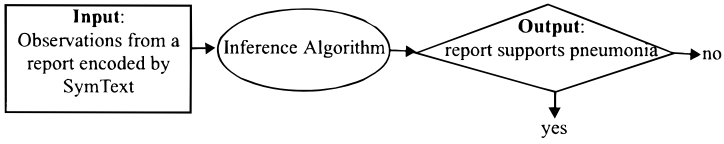


**FIG. 2.** Process of identifying chest X-ray reports that support pneumonia.

for identifying the perceived state of the patients; however, convincing the original radiologists to read their past reports and remember the patients is impractical. On the other hand, internists are the end users of chest X-ray reports whose interpretations we seek to imitate with automated methods. We have more access to internists willing to read and interpret reports. Hripcsak *et al.* [7] compared radiologist and clinician interpretations of chest X-ray reports and found no significant difference in their performance. Therefore, in this study the gold standard was comprised of the majority opinion of three internists who read the full text of all 292 reports independently to determine whether the reports contained radiologic evidence supporting pneumonia.

We assessed the reliability of the gold standard with a methodology based on generalizability theory proposed by Shavelson and Webb [27] and adapted to the NLP domain by Hripcsak *et al.* [28]. Following this methodology, a generalizability coefficient is computed. The coefficient, which ranges from zero to one, quantifies the agreement among the experts who generated the standard. A generalizability coefficient of 0.7 or higher is considered adequate if the gold standard is going to be used to estimate the overall performance of a system [28].

### *Training Set*

A set of 298 chest X-ray reports was used to train the Bayesian network and the decision tree. Reports were collected from all patients having a chest X-ray report dictated before September 1998. We randomly selected 150 from the subset of reports for which the patient had a hospital discharge diagnosis of bacterial pneumonia. We randomly selected 148 from the mutually exclusive subset of reports for which the patient did not have a diagnosis of pneumonia. Two physicians read the reports to determine if the reports supported pneumonia. Cases on which the physicians disagreed were decided by a radiologist. One hundred seventy-three of the training cases (58%) supported pneumonia.

### *Test Set*

Our test set was comprised of 292 chest X-ray reports dictated at LDS Hospital. Three-quarters of the reports (217) were randomly selected from all chest X-ray reports produced between October and December 1998. To increase the prevalence of pneumonia-related reports in our sample we randomly selected 75 additional reports from patients with a known primary hospital discharge diagnosis of bacterial pneumonia. The reports from pneumonia patients were dictated between January and March 1999. We did not restrict the selection to patients' first chest X-ray reports, so more than one report for the same patient was sometimes included in our test set.

### *Inferencing Algorithms*

Below we describe the three inferencing algorithms we compared.

*Expert-crafted rules.* An internist who completed his internal medicine residency created the following rule to determine whether a report supports pneumonia.

If any of the following encoded 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
- Localized peripheral infiltrate
- Localized parenchymal abnormality

The encoded concepts in the rule already contain phrasal inferences made by SymText. For example, the concept *localized lower lobe infiltrate* is inferred by SymText from phrases in the text such as "hazy opacity in the right lower lobe." Because the diagnostic task was to determine radiologic support for acute bacterial pneumonia, only infiltrates highly predictive of bacterial pneumonia, i.e., localized infiltrates, were included in the rule.

*Expert-crafted Bayesian network.* A Bayesian network is a directed acyclic graph representing joint probabilities among concepts [20,21]. With input from an experienced radiologist, the same physician who created the rules created a Bayesian network (see Fig. 3) in which the top node was "support pneumonia." Nodes in the network are concepts from the parser, and links connecting nodes represent conditional dependence among the concepts. We used the Bayesian network software application Netica (Norsys Software Corp., Vancouver, British Columbia, Canada). Output of the network is a probability between 0 and 1 that the report supports pneumonia.

In creating the Bayesian network we attempted to model findings that commonly rule in and rule out acute bacterial pneumonia. In this way, the posterior probability of *support*

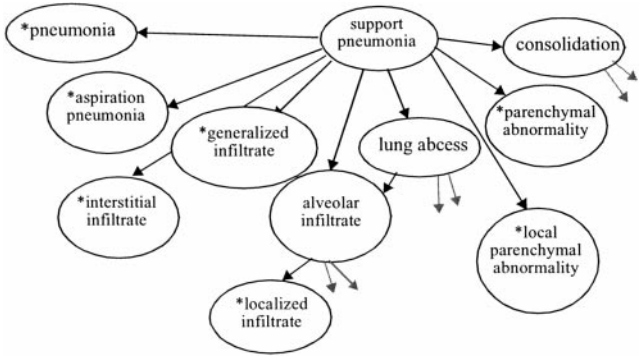


**FIG. 3.** Part of the expert-crafted Bayesian network. Nodes with an asterisk (*) are concepts encoded by SymText. Nodes without an asterisk are super concepts whose children are concepts modeled by SymText. Arrows represent conditional dependence between the connected nodes. The root node makes an inference from SymText's output.

*pneumonia* decreases from the prior probability when a finding rarely associated with acute bacterial pneumonia is instantiated. Hence, findings such as *interstitial infiltrate* (which is more commonly associated with viral pneumonia) or *generalized infiltrate* (which is more commonly associated with adult respiratory distress syndrome or congestive heart failure) decrease the probability of *support pneumonia*.

*Decision tree.* We used Quinlan's See5.0 Decision Tree software (RuleQuest Research Pty Ltd, Australia) to learn a decision tree from data [22]. 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 most of the training cases are classified successfully.

From the training set 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 represented by SymText with the possible values of *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*. The final decision tree created from training cases contained 20 nodes (Fig. 4).

## *Comparison Metrics*

We examined two questions: (1) which of the three expert systems performs better at identifying reports that support pneumonia; and (2) do the three expert systems perform as well as physicians perform? To address the second question, four physicians not involved in creation of the gold standard also determined whether the reports in the test set supported pneumonia.

Classification of a report as supporting or not supporting pneumonia can be accomplished with a binary classification, as the decision tree produces, or a probabilistic classification, as the Bayesian network provides. Comparison of binary and probabilistic output is not straightforward. One possible method of comparison is to transform the binary output into probabilistic output or the probabilistic output into binary output. Such a transformation involves some loss of information. Since each transformation may favor one or another methodology we considered both transformations (see [23] for another example of transforming output for the sake of comparison).

Below we describe transformation of the output data required to make valid comparisons among the subjects. We then describe how we evaluated the two questions listed above.

*Transformation of the output data.* We transformed output from the Bayesian network into a binary response by

```
consolidation (nos) = present: yes (79/81)
consolidation (nos) = absent:
:...localized consolidation = present: yes (42/43)
    localized consolidation = absent:
    :...pneumonia = present: yes (17/18)
        pneumonia = absent:
        :...localized infiltrate = present: yes (11/12)
            localized infiltrate = absent:
            :...aspiration pneumonia = present: yes (5/6)
                aspiration pneumonia = absent:
                :...localized upper lobe infiltrate = present: yes (3/3)
                    localized upper lobe infiltrate = absent:
                    :...perihilar infiltrate = present: yes (3/3)
                        perihilar infiltrate = absent:
                        :...localized lower lobe infiltrate = present:[S1]
                            localized lower lobe infiltrate = absent:
                            :...generalized consolidation = present: yes (1/1)
                                generalized consolidation = absent:
                                :...
SubTree [S1]
enlargement of the pulmonary vessels = present: no (1/1)
enlargement of the pulmonary vessels = absent: yes (6/6)
```

**FIG. 4.** Part of the decision tree for classifying reports by attribute *support pneumonia*. Parentheses contain the proportion of training cases the branch correctly classified.

assigning a specific probability to be the threshold for a yes/no classification. Reports with a computed probability less than the threshold were classified as not supporting pneumonia; reports with a probability greater than the threshold were classified as supporting pneumonia. A threshold of 0.50 might seem appropriate, but sensitivity is more important than specificity in identifying reports that might support pneumonia. Therefore, we selected 0.20 as the threshold value because the value maintained high sensitivity with a relatively small loss of specificity.

The four physicians classified the test reports into four categories: (1) definitely not pneumonia, (2) possibly pneumonia, (3) probably pneumonia, and (4) definitely pneumonia. Their ordinal responses were transformed into binary responses by classifying any answer with possible support for pneumonia (2)–(4) as supportive of pneumonia.

Output from the expert crafted rules was transformed into ordinal output identical to that of the physicians, as shown in Table 1.

We altered the decision tree to provide a probability instead of a binary classification by using the actual proportion of training reports successfully classified in each leaf node of the tree as the probability of the classification. As an example, consider the decision tree shown in Fig. 4. The first node of the tree indicates that if consolidation (nos) is present, the classification should be "yes." In parentheses we see that 79/81 of the training cases classified by this node were correctly classified. Therefore, the transformation of the binary classification of "yes" for this node is the probability that the classification is "yes": 79/81 = 0.98.

*Comparison of three expert systems.* Once the output data were transformed into both binary and probabilistic output, we were able to analyze the performance of the three expert systems. We tested the performance of the rules, the Bayesian network, and the decision tree using both binary output and probabilistic output.

To compare the expert systems based on binary output we calculated sensitivity, precision (positive predictive value), and specificity as follows:

Sensitivity: Number of reports correctly classified by expert system as supporting pneumonia/Number of reports classified by gold standard as supporting pneumonia

Precision: Number of reports correctly classified by expert system as supporting pneumonia/Total number of reports classified by expert system as supporting pneumonia

Specificity: Number of reports correctly classified by expert system as not supporting pneumonia/Number of reports classified by gold standard as not supporting pneumonia

We then used McNemar's test [29] to determine if any expert system significantly differed from the others. Because we made three comparisons we applied Bonferroni corrections by dividing our $\alpha$ by 3 ($\alpha = 0.05/3 = 0.017$).

Probabilistic output can be analyzed by considering the sensitivity and specificity of the output at various probability levels with a receiver operating characteristic (ROC) curve [30, 31]. The area under the ROC curve ($A_z$) is an overall measure of performance accounting for variation in true-positive and false-positive rates that depend on the threshold used to classify an item. A measure of $A_z$ ranges between 0.50 (chance classification) and 1.0 (perfect classification).

We used ROCkit (University of Chicago, Chicago, IL, 1998) to calculate the $A_z$ for all three expert systems with a maximum-likelihood estimate equation. We also used ROCkit to test the statistical significance of differences between the correlated ROC curves with a univariate *z*-score test. Again, we made Bonferroni corrections for three comparisons.

*Similarity of expert systems to physicians.* Correctness of an expert system is not the only consideration in evaluating a system's usefulness. An important comparison in evaluating expert systems is how well the expert systems imitate human experts. In our case, physicians are the human experts to imitate. Therefore, we compared the performance of the three expert systems against the performance of four physicians at the same task. The performance of physicians indicates the highest level of performance we might reasonably expect from an expert system. We also measured the performance of three lay persons with no medical experience and of a simple keyword search. Performance of the lay people and key-word search provide a baseline against which we can evaluate the expertise demonstrated by the expert systems.

Again, we used different comparative methods for binary

TABLE 1
Transformed Expert-Crafted Rules That Provide Ordinal Output

| Concepts detected in report | Rule output |
|---|---|
| *Pneumonia*<br>*Aspiration pneumonia* | 4<br>Definitely pneumonia |
| *Consolidation* (localized or nos)<br>Infiltrate (localized or nos) | 3<br>Probably pneumonia |
| *Localized parenchymal abnormality* | 2<br>Possibly pneumonia |
| No pneumonia-related concepts | 1<br>Definitely not pneumonia |

and probabilistic output. Using binary output we plotted sensitivity by one minus specificity points on ROC axes for the expert systems, three lay persons, a simple keyword search, and four physicians. We used McNemar's test to determine if any statistically significant differences existed between the physicians and the other subjects. Because we divided the $\alpha$ of 0.05 by the 28 comparisons (four physicians vs three expert systems, three lay persons, and one keyword search), the statistically significant $\alpha$ value was 0.0018.

Probabilistic output was not obtained from the lay persons or the key-word search. We compared the $A_z$ of the three expert systems with that of the four physicians using univariate *z*-score tests with an $\alpha$ of 0.004 (0.05/12 comparisons).

## RESULTS

The generalizability coefficient for our gold standard was 0.89. According to our gold standard, 112 (38%) of the reports supported pneumonia. Below we present results for binary and probabilistic output that address (1) how well the expert systems identify chest X-ray reports that support pneumonia and (2) how well the expert systems imitate physicians.

### *Comparison of Three Expert Systems*

We tested the performance of the rules, the Bayesian network, and the decision tree using both binary and probabilistic output. Figure 5 compares sensitivity, precision, and specificity of the three expert systems. Sensitivity was highest for the rules (0.920) and lowest for the decision tree (0.857). The decision tree performed better than the rules and the Bayesian network in precision and specificity.

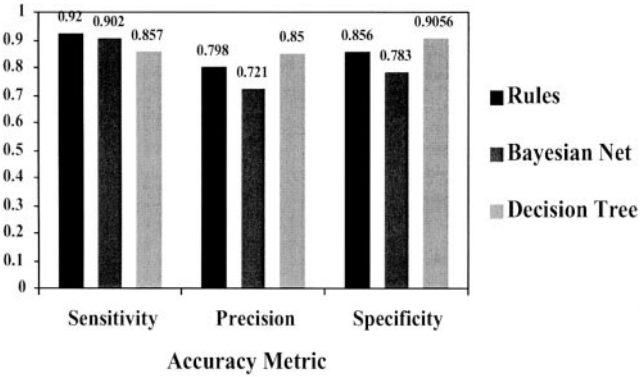


**FIG. 5.** Performance accuracy of the three expert systems.

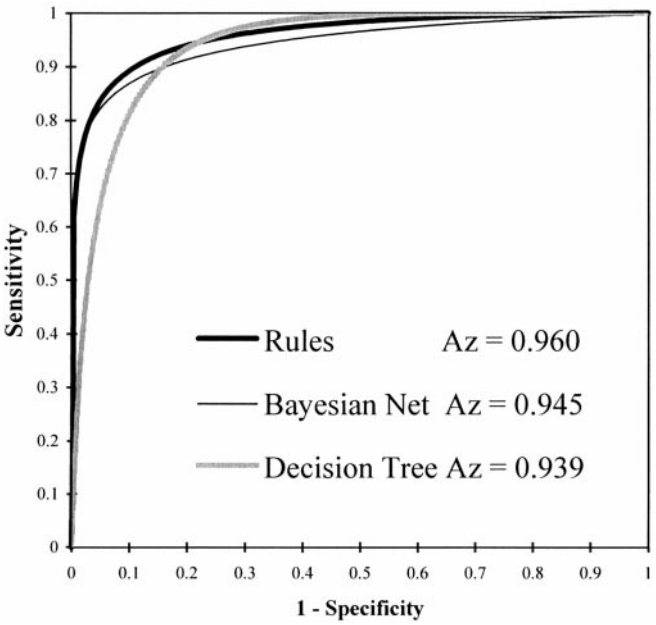


**FIG. 6.** Maximum-likehood estimated ROC curves for three expert systems. The rules demonstrated an area under the curve ($A_z$) of 0.960 (95% CI: 0.927–0.980), the Bayesian network an $A_z$ of 0.945 (95% CI: 0.906–0.970), and the decision tree 0.939 (95% CI: 0.909–0.962).

We found a statistical difference between the rules and the Bayesian network ($P = 0.0071$) and between the Bayesian network and the decision tree ($P = 0.0079$). However, there was no difference between the rules and the decision tree ($P = 0.7237$).

The ROC curves generated for the expert systems are presented in Fig. 6. The rules demonstrated the highest $A_z$ (0.960), followed by the Bayesian network (0.945) and the decision tree (0.939). The decision tree curve crossed the curves of the rules and the Bayesian net, making a comparison of areas difficult to interpret. However, no statistical differences existed among the $A_z$'s for the three expert systems.

### *Similarity of Expert Systems to Physicians*

We also compared the performance of physicians to that of the expert systems for binary and probabilistic output. Using binary output, we plotted the sensitivity and corresponding specificity points on an ROC plot (Fig. 7). A subject with perfect performance would appear at the upper-left corner of the plot. The physicians are grouped near that

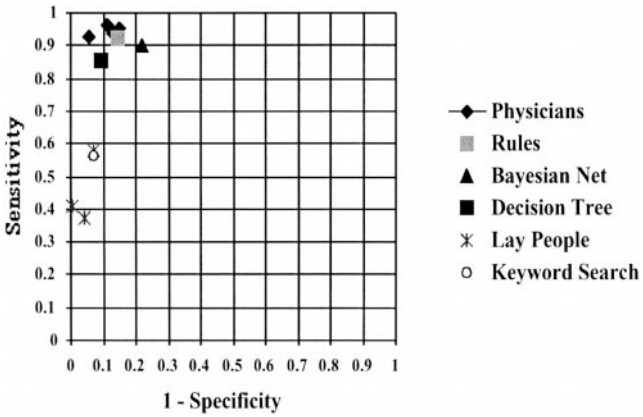


**FIG. 7.** ROC plot of sensitivity and corresponding specificity points for physicians, expert systems, lay people, and the key-word search.

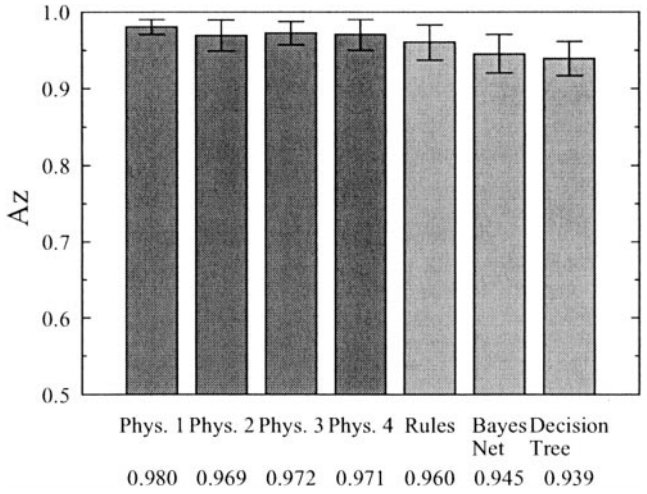


**FIG. 8.** Areas under the ROC curve ($A_z$) for the four physicians and the three expert systems. Bars represent 95% confidence intervals. The only significant difference occurred between the decision tree and physician 1 ($P = 0.0029$).

corner, along with the three expert systems. The lay people and the key-word search demonstrated high specificity but low sensitivity. Statistical comparisons against the physicians reveal a significant difference between physicians and lay people with $P$ values ranging from 0.0001 to 0.0004. The physicians and the key-word search also differed significantly with $P$ values ranging from 0.0001 to 0.0003. The Bayesian network significantly differed from two of the physicians, whereas the rules and decision tree did not differ from any of the physicians.

The $A_z$'s from the expert systems were compared against those of all four physicians (Fig. 8). Physicians' $A_z$'s ranged from 0.969 to 0.980. At the $P < 0.004$ level, no statistical differences were found except between the decision tree and physician 1. Table 2 lists the $P$ values for comparisons between the binary and probabilistic outputs of the expert systems and the physicians.

## DISCUSSION

We compared the performance of a rule set, a Bayesian network, and a decision tree at identifying chest X-ray reports that support pneumonia. We also measured how well the expert systems performed compared to physicians. Below we discuss differences in the performance of the three expert systems. To select the best expert system we also discuss the performance of the systems in the context of (a) how the system was derived and (b) how useful the system's output would be to automated decision support systems.

### *Performance of the Expert Systems*

The three expert systems performed well at identifying chest X-ray reports that support pneumonia. The three systems did not differ in overall performance. However, specific metrics such as sensitivity and specificity were different for the systems. For our purposes, sensitivity is the most important measure. In general, identification of a report that supports pneumonia will comprise only one of several data points analyzed by a decision support system. Therefore, falsely identifying reports is less of a problem than missing reports that support pneumonia.

Results from the binary output of the expert systems (Fig. 5) suggest that a screening application aiming to capture as

TABLE 2
*P* Values from Statistical Tests on Binary and Probabilistic Output

| | Physician 1 | Physician 2 | Physician 3 | Physician 4 |
|---|---|---|---|---|
| Binary | | | | |
| Rules | 0.0031 | 0.5485 | 0.0278 | 0.398 |
| Bayesian net | **0.0010** | 0.0044 | **0.0001** | 0.0039 |
| Decision tree | 0.0112 | 0.8788 | 0.1390 | 0.6394 |
| Probabilistic | | | | |
| Rules | 0.1577 | 0.6190 | 0.5021 | 0.3555 |
| Bayesian net | 0.0316 | 0.1701 | 0.1030 | 0.1206 |
| Decision tree | **0.0029** | 0.0426 | 0.0183 | 0.0329 |

*Note.* Values in bold were statistically significant after Bonferroni corrections

many reports that support pneumonia as possible might be best served by the expert crafted rules (sensitivity: 0.920). The Bayesian network also demonstrated quite high sensitivity (0.902). The decision tree demonstrated lower sensitivity (0.857) but higher precision and specificity than either of the other systems. Because of higher precision and specificity, the decision tree did not differ significantly from the rules. The rules and decision tree did differ significantly from the Bayesian net, however.

No significant differences were found among the $A_z$'s for the expert systems. At a high sensitivity level the rules appear to maintain a higher specificity than the other methods (Fig. 6). However, curves that cross are difficult to interpret. We were surprised that no significant difference existed between the decision tree and the other systems in a test for which the decision tree's binary output was transformed into probabilities.

We also analyzed the amount of expertise demonstrated by the expert systems. Figure 7 shows the sensitivity and specificity from the binary output of various subjects. All three expert systems performed similarly to the physicians. At the 0.0018 significance level, the Bayesian network differed significantly from two of the physicians, but the rules and decision tree did not differ from any physicians. Comparisons on probabilistic output show a difference at the 0.004 level between the decision tree and one physician but not between the other systems and the physicians.

The only expert systems that differed from any physicians were those that had been altered to produce different output. The Bayesian network's binary classifications and the decision tree's probabilistic classifications differed from some of the physicians. In their natural state, however, none of the expert systems performed differently from physicians. The rules did not differ from physicians whether they produced binary or ordinal output.

### *Selecting the Best Expert System*

Correctness is not the only factor contributing to selection of the best expert system. Two other considerations are how easily the system is expanded and the type of output the system provides.

We compared one algorithm that was completely created by expert input (rules), one algorithm whose structure was created by experts but whose classification patterns were learned from data (Bayesian network), and one algorithm that derived both classification patterns and structure from data (decision tree). Although the expert crafted rules consistently performed more like physicians, a system that could be trained or derived from actual reports would be advantageous for several reasons. First, the ability to train a system on real data allows the system to evolve as the mix of radiologists evolves or as the common way of describing images evolves. Second, our study only addressed the disease pneumonia. Expanding the inferencing method to other chest diseases is necessary to maximize the usefulness of coded information from chest X-ray reports. An algorithm that derives classification patterns and structure from data could be more easily adapted to other diseases. Third, our parser will be applied to other radiology reports such as computed tomography or magnetic resonance imaging. Algorithms that are easily ported to other domains will be required. The decision tree performed slightly worse than the other expert systems. However, the small loss in accuracy might be offset by the ability to automatically expand to other diseases and domains.

We transformed binary and probabilistic output so that we could compare different systems. The format of a system's output is also important in selecting an inferencing system whose output will be used by decision support tools. For instance, the Antibiotic Assistant requires data from X-ray observations to be labeled as either present or absent. The Bayesian network that derives the probability a patient has pneumonia [4] triggers a guideline that assists emergency department physicians in managing pneumonia patients. Probabilistic evidence of pneumonia from a chest X-ray report would contribute more meaningfully to the pneumonia Bayesian network than binary output would.

### *Limitations and Future Work*

Our purpose was to test how well expert systems determine whether radiologic support for pneumonia exists in a coded chest X-ray report. Because the focus was on the content of the current report, we removed any information about the clinical history from the beginning of the reports. One could argue, however, that because the clinical history influenced the radiologist in interpreting the examination, the clinical history should have been included.

Transforming the output of the Bayesian network and decision tree decreased the systems' performance. We know of no better way to transform probabilistic data into binary responses. However, a better method for transforming binary decision tree classifications into probabilities exists. A method proposed by Quinlan [32] accounts for small numbers of instances in leaf nodes and for statistical relationships between leaf nodes and variables higher in the tree. Using this method might have improved the decision tree's performance with probabilistic output. Still, a decision tree is

designed to give binary output, and transforming the values may inevitably result in lower performance.

Accuracy of information encoded from chest X-ray reports is dependent not only on the expert system that makes inferences, but also on the accuracy of the parser. Any data we actually store on the hospital information system will originate with SymText's parsed output. For this experiment we used manually corrected output so that we could isolate the performance of the expert systems from SymText's mistakes. We describe elsewhere how well one of the expert systems (the rules) performed on SymText's uncorrected parsed data [33]. Using parsed data as input to the rules produced accuracy similar to that of physicians. We are currently storing all pneumonia-related information from chest X-ray reports on LDS Hospital's HELP System [34].

A major limitation in this and similar studies is the methodology for analyzing how physician-like an expert system performs. We compared our expert systems to four physicians. Some of the expert systems differed from one or two physicians but not from the other physicians. Because physicians also vary in their performance, determining whether an expert system behaves like a physician is not easily accomplished. Is an expert system that differs from a few of the physicians different from physicians? If we had compared against different physicians, our results might have been different.

Moreover, the fact that an expert system does not differ from physicians does not indicate that the system performs the same as the physician. "Absence of evidence is not evidence of absence" [35]. Detection of statistical differences between an expert system and physicians is dependent on which physicians are used for comparison and on the study's power. Larger sample size and more physicians are desirable in these types of studies, but both are expensive and difficult to secure. We would like to experiment with different testing methodologies where an expert system can be compared against a distribution of physician performance rather than against individual physicians. One such method is equivalency testing [36] which is commonly used in the pharmaceutical field.

## CONCLUSION

We compared the performance of three inferencing algorithms that automatically identify chest X-ray reports that support acute bacterial pneumonia. The expert systems demonstrated similar accuracy, with the expert-crafted rule set performing slightly better than the Bayesian network and the decision tree. All of the expert systems performed similarly to physicians.

## ACKNOWLEDGMENTS

We acknowledge the physicians who read the reports: Bruce Bray, M.D., Alexandra P. Edelwein, M.D., Philip Frederick, M.D., Chuck Mullett, M.D., Gustavo Oderich, M.D., Greg Patton, M.D., and Ken Zollo, M.D. We also thank Dominik Aronsky, M.D. for helpful suggestions on this project. This work was supported by NLM Grant 1 R01 LM 06539-02.

## REFERENCES

1. Hripcsak G, Knirsch CA, Jain NL, Pablos-Mendez A. Automated tuberculosis detection. J Am Med Inform Assoc 1997; 4:376–81.
2. Evans RS, Pestotnik SL, Classen DC, *et al.* A computer-assisted management program for antibiotics and other antiinfective agents. N Engl J Med 1998; 338:232–8.
3. Aronsky D, Haug PJ. Automatic identification of patients eligible for a pneumonia guideline. Proc AMIA Symp 2000; 12–6.
4. Aronsky D, Haug PJ. Diagnosing community-acquired pneumonia with a Bayesian network. Proc AMIA Symp 1998; 632–6; Fiszman M, Haug PJ. Using medical language processing to support real-time evaluation of pneumonia guidelines. Proc AMIA Symp 2000; 235–9.
5. Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999; 74:890–5.
6. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1994; 1:142–74.
7. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 1995; 122:681–8.
8. Hahn U, Schnattinger K, Romacker M. Automatic knowledge acquisition from medical texts. Proc AMIA Symp 1996; 383–7.
9. Taira RK, Soderland SG. A statistical natural language processor for medical reports. Proc AMIA Symp 1999; 970–4.
10. Rassinoux AM, Lovis C, Baud RH, Scherrer JR. Versatility of a multilingual and bi-directional approach for medical language processing. Proc AMIA Symp 1998; 668–72.
11. Ceusters W, Spyns P, DeMoor G. From natural language to formal language: when Multi-TALE meets GALEN. Stud Health Technol Inform 1997; 43:396–400.
12. Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Comput Biomed Res 1993; 26:467–81.
13. Koehler SB. SymText: A natural language understanding system

for encoding free text medical data. Ph.D. dissertation, University of Utah, 1998.

14. Jain NL, Knirsch CA, Friedman C, Hripcsak G. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. Proc AMIA Annu Fall Symp 1996; 542–6.
15. Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff SM. Experience with a mixed semantic/syntactic parser. Proc Annu Symp Comput Appl Med Care 1995; 284–8.
16. Gundersen ML, Haug PJ, Pryor TA, *et al.* Development and evaluation of a computerized admission diagnoses encoding system. Comput Biomed Res 1996; 29:351–72.
17. Fiszman M, Haug PJ, Frederick PR. Automatic extraction of PIOPED interpretations from ventilation/perfusion lung scan reports. Proc AMIA Symp 1998; 860–4.
18. Woods WA. Transition network grammars for natural language analysis. Commun ACM 1970; 13:591–606.
19. Fiszman M, Chapman WW, Evans SR, Haug PJ. Automatic identification of pneumonia related concepts on chest x-ray reports. Proc AMIA Symp 1999: 67–71.
20. Szolovits P. Uncertainty and decisions in medical informatics. Methods Inf Med 1995; 34:111–21.
21. Jensen FV. Introduction to Bayesian networks. New York: Springer Verlag, 1996.
22. Quinlan JR. C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann, 1993.
23. Long WJ, Griffith JL, Selker HP, D'Agostino RB. A comparison of logistic regression to decision-tree induction in a medical domain. Comput Biomed Res 1993; 26:74–97.
24. Mitchell TM. Machine learning. Boston, MA: McGraw-Hill, 1997.
25. Wilcox A, Hripcsak G. Classification algorithms applied to narrative reports. Proc AMIA Symp 1999; 455–9.
26. Wilcox A, Hripcsak G. Medical text representations for inductive learning. Proc AMIA Symp 2000; 923–7.
27. Shavelson RJ, Webb NM. Generalizability theory: a primer. Newybury Park, CA: Sage, 1991.
28. Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology reports. J Am Med Inform Assoc 1999; 6:143–50.
29. McNemar Q. Psychological statistics. New York: Wiley, 1969.
30. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.
31. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8:283–98.
32. Quinlan JR. Improved estimates for the accuracy of small disjuncts. Mach Learn 1991; 6:93–98.
33. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000; 7:593–604.
34. Gardner RM, Pryor TA, Warner HR. The HELP hospital information system: update 1998. Int J Med Inf 1999 Jun; 54:169–82.
35. Altman DG, Bland JM. Absence of evidence is not evidence of absence. Br Med J 1995; 311:485.
36. Jones B, Jarutz P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the impact of rigorous methods. Br Med J 1996; 313:36–9.