Correct vs.Parsed Data for Inferring Pneumonia in Chest X-ray Reports

Wendy Webber Chapman, Marcelo Fiszman M.D., Peter J. Haug M.D.
Department of Medical Informatics, University of Utah and LDS Hospital, Salt Lake City, Utah

INTRODUCTION
One possible solution to coding free-text chest x-ray reports is natural language processing. At LDS Hospital we are using a natural language understanding system called SymText that codes most diseases and findings in chest x-ray reports. In this paper we test SymText's performance on acute bacterial pneumonia.

A key piece of information in diagnosing pneumonia is the chest radiograph [1], thus automatically classifying reports that support pneumonia can be important information for computer-aided decision support tools. We have tested SymText on detecting pneumonia-related concepts, and we have tested four computerized techniques for inferring the presence of pneumonia given correctly coded input. Here we examine the robustness of the four techniques by comparing their performance with correct data to their performance with parsed data.

METHODS
Inferencing Algorithm
The four inferencing algorithms we test are:
2. Expert crafted Bayesian net (BN1).
3. Machine Learned Decision Tree. We used Quinlan’s See5.0 Decision Tree software to learn rules from data contained in 298 chest radiograph reports [2].
4. Bayesian network created from decision tree rules (BN2). We used the attributes See5.0 deemed important to create another Bayesian network.

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; the other half were randomly selected.

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.

Test Metrics
We measure accuracy with recall (sensitivity), precision (positive predictive value), and specificity.

RESULTS
Figure 1 compares the performance of the five systems with correct and parsed data.

McNemar's tests showed that BN1 performed significantly worse with parsed data (p=.046).

DISCUSSION
The only computerized technique to perform significantly worse with parsed data was BN1. BN1 examines more findings and diseases than the other methods, so more incorrect data causes more problems for this network. The expert rules performed better with parsed data, because some mistakes made by the parser aligned with interpretations of the GS panel that differed from interpretations by the manual coder.

Precision and specificity decreased for all techniques when given parsed data, whereas recall stayed the same. Redundancy facilitates recall, and there is generally enough redundancy in the reports for the parser to generate a pneumonia-related observation if it exists. Yet, a false positive pneumonia-related observation causes specificity and precision scores to drop.

CONCLUSION
Three computerized techniques were shown to perform just as well with parsed data, demonstrating that natural language processing can be used for classifying chest x-ray reports for pneumonia.

References