Comparing Diagnostic Decision Support Systems for Pneumonia

Charles Lagor, MD, Dominik Aronsky, MD, Marcelo Fiszman, MD, Peter J. Haug, MD
Dept. of Medical Informatics, LDS Hospital/ University of Utah, Salt Lake City, Utah

OBJECTIVE
To compare a Bayesian network (BN) and an artificial neural network (ANN) in diagnosing community-acquired pneumonia.

BACKGROUND
In the past decade, BNs and ANNs have been increasingly used as decision support methodologies in medicine. Their clinical applications include diagnosis, imaging, signal processing, analysis of laboratory data, and pharmacology. Despite this broad range of applications, it is not always clear whether a BN or an ANN is more appropriate for a particular clinical situation. In evaluating different decision support methods, comparisons between BNs and ANNs or between BNs, ANNs, and other decision support systems are helpful. Such comparisons, however, are still scarce. The goal of this study, therefore, was to compare a BN and an ANN. The networks were designed to diagnose community-acquired pneumonia (CAP). We investigated the networks under the following aspects: the behavior of the networks at a sensitivity of 95%, and the overall diagnostic accuracy of the networks.

METHODS
We used a database that contained data on 32,662 adult patients (18 years or older) collected over a two-year period (May 26, 1995, to June 14, 1997). The database captured 24 variables that are routinely available in the hospital information system, known as the HELP (Health Evaluation through Logical Processing) system. The target disease of the decision support systems was CAP (ICD-9 code: 480-486). Of the 32,662 patients, 498 patients (1.5%) had a primary discharge diagnosis of CAP. We used a set of 21,775 randomly selected patients (two-thirds of the data) to train both networks. For the ANN we additionally subdivided this training set into a learning set (14,517 patients or two-thirds of the training set) and a verification set (7,258 patients or one-third of the training set). The internal verification set prevented over-fitting of the ANN. We evaluated the ability of the trained networks to identify CAP patients in the remaining 10,887 patients (one-third of the data). For each testing case, the BN generated a probability and the ANN generated an activation value. Both output types require thresholds to decide whether CAP is present or absent. In each network we set the threshold to achieve a sensitivity of 95% and then calculated the specificity, the positive predictive value (PPV), and the negative predictive value (NPV). Over the range of all thresholds, we also plotted sensitivity against 1 – specificity to get receiver operating characteristic (ROC) curves for the BN and ANN. We determined the area under the receiver operating characteristic curve (AUC) as a measure of overall accuracy for each network. We tested for statistical difference between the AUCs using the correlated area z statistic.

RESULTS
At a sensitivity of 95%, the values for specificity, PPV, and NPV were:

<table>
<thead>
<tr>
<th></th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN</td>
<td>92.3%</td>
<td>15.1%</td>
<td>99.3%</td>
</tr>
<tr>
<td>ANN</td>
<td>94.0%</td>
<td>18.6%</td>
<td>99.9%</td>
</tr>
</tbody>
</table>

The BN had an AUC of 0.9795 (95% CI: 0.9736, 0.9843). The ANN had an AUC of 0.9855 (95% CI: 0.9805, 0.9894). The difference between the AUCs was statistically significant (p=0.0044).

DISCUSSION
A decision support system for CAP must detect as many cases as possible, because CAP has a low prevalence. Therefore, we investigated the behavior of the networks at a sensitivity of 95%. At this sensitivity, both networks had low PPVs. The PPVs of 15.1% and 18.6% indicate that out of 100 pneumonia classifications, approximately 12 are appropriate, whereas 88 are false. This high number of misclassifications is justifiable, because the consequences of missing patients with CAP are worse than those of misdiagnosing patients without CAP. Overall, both networks achieved excellent accuracies; the ANN achieved a slightly better overall accuracy. Nevertheless, we do not conclude that the ANN is superior to the BN, because the difference in the AUCs of 0.9855 – 0.9795 = 0.006 may be statistically significant, but compared to the AUC of either network it plays a minor role.

ACKNOWLEDGMENTS
The first author is supported with a grant from the Austrian Federal Ministry Science and Transport (GZ 558.011/103-I/19a/98).