

Automatic Identification of Patients Eligible for a Pneumonia Guideline: Comparing the Diagnostic Accuracy of Two Decision Support Models

Charles Lagor, Dominik Aronsky, Marcelo Fiszman, Peter J. Haug

Department of Medical Informatics, LDS Hospital/ University of Utah, Salt Lake City, Utah, USA

Abstract

Background: In busy clinical settings, physicians often do not have enough time to identify patients for specific therapeutic guidelines. As a solution, decision support systems could automatically identify eligible patients and trigger computerized guidelines for specific diseases. Applying this idea to community-acquired pneumonia (CAP), we developed a Bayesian network (BN) and an artificial neural network (ANN) for identifying patients who have CAP and are eligible for a pneumonia guideline. *Objective:* The aim of this study was to determine whether the diagnostic accuracy of these two decision support models differs in terms of identifying CAP patients. *Methods:* We trained and tested the networks with a data set of 32,662 adult patients. For each network, we (1) calculated the specificity, the positive predictive value (PPV), and the negative predictive value (NPV) at a sensitivity of 95%, and (2) determined the area under the receiver operating characteristic curve (AUC) as a measure of overall accuracy. We tested for statistical difference between the AUCs using the correlated area z statistic. *Results:* At a sensitivity of 95%, the respective values for specificity, PPV, and NPV were: 92.3%, 15.1%, and 99.9% for the BN, and 94.0%, 18.6%, and 99.9% for the ANN. The BN had an AUC of 0.9795 (95% CI: 0.9736, 0.9843), and the ANN had an AUC of 0.9855 (95% CI: 0.9805, 0.9894). The difference between the AUCs was statistically significant ($p=0.0044$). *Conclusions:* The networks achieved high overall accuracies on the testing data set. Because the difference in accuracies is statistically significant but not clinically significant, both networks are equally suited to drive a guideline.

Keywords:

Diagnosis, Computer-Assisted; Decision Support Techniques; Expert Systems; Artificial Intelligence; Bayes Theorem; Models, Statistical; Neural Networks (Computer)

Introduction

Medical decision support systems have existed since 1945

[1]. The applied methodologies of these systems are diverse: rule-based systems, fuzzy logic, decision trees, Bayesian networks, and artificial neural networks, to name a few. Of these, Bayesian networks (BN) and artificial neural networks (ANN) have been increasingly used in the past decade. Both networks can model the uncertainty inherent in medical reasoning [2] and make decisions based on incomplete data. Their clinical applications include diagnosis, imaging, signal processing, analysis of laboratory data, and pharmacology [3-6]. In view of this broad range of applications, comparisons between different models may be helpful in understanding the characteristics of different methodologies in specific clinical domains. Comparisons between BNs and ANNs [7] or between BNs, ANNs, and other decision support systems [8, 9] have been made, but such comparisons are still scarce.

In this paper, we compare a BN and an ANN, which were both designed to identify patients with community-acquired pneumonia (CAP). Cooper has also compared BNs and ANN in the domain of pneumonia [10]. The models in his study, however, predict the mortality of pneumonia, whereas the models in our study predict the likelihood of CAP. Implemented in a routine clinical setting, our models could use routinely available electronic data that are present during a patient encounter to automatically identify patients eligible for a computerized pneumonia guideline, and thereby save physicians the effort of identifying the patients themselves. In view of this utility, we examined the diagnostic behavior of the networks, the diagnoses of patients who were misclassified as having CAP, and the overall diagnostic accuracy of the networks.

Background

Community-Acquired Pneumonia

Pneumonia is a common and potentially life-threatening lung infection that is either acquired within the hospital or outside in the community. It has been reported as the sixth leading cause of death in the United States, with CAP having an incidence of 2.66 cases per 1000 adults per year [11]. Its annual costs have been estimated to be \$23 billion

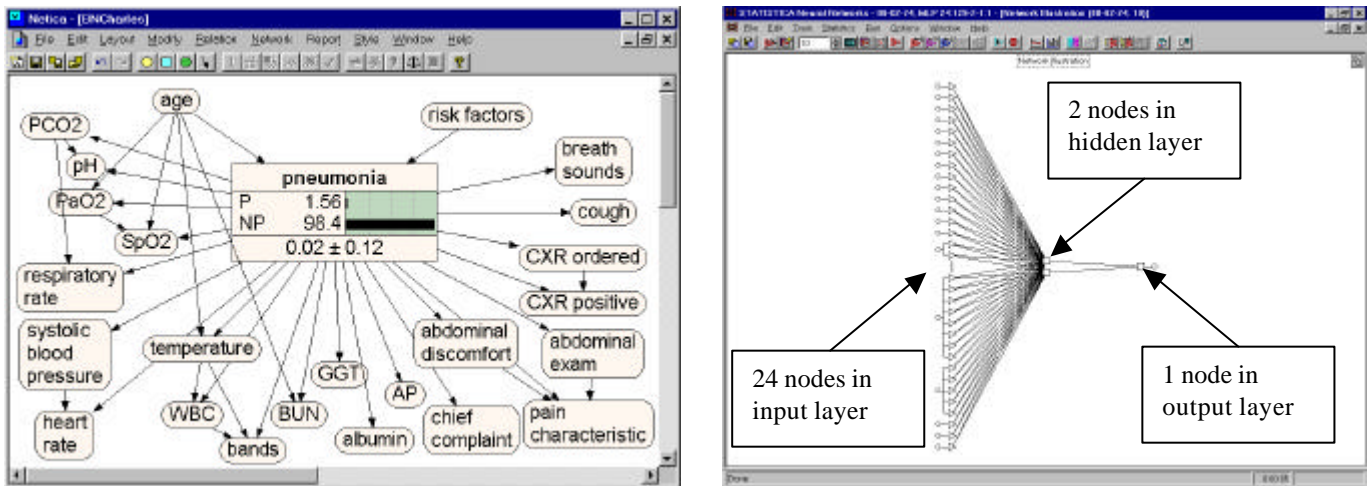


Figure 1 - The structures of the Bayesian network (left) and the artificial neural network (right). The 24 variables in the input layer of the artificial neural network correspond to the 24 input variables of the Bayesian network. The node in the output layer of the artificial neural network corresponds to the outcome variable (CAP present).

USD [12]. Although most patients with CAP present only with respiratory symptoms, a proportion of patients may present with additional non-respiratory symptoms such as abdominal pain. A few patients may even lack respiratory symptoms [12]. Consequently, some cases of CAP may be challenging to diagnose.

Bayesian Networks

Bayesian networks are probabilistic systems that model a clinical problem with nodes, directed links, and conditional probability tables [13, 14]. The nodes represent variables; of these variables, one or more represent the output variable. The directed links connect the nodes and capture the dependencies among the variables. A conditional probability table is associated with each node. When the BN processes new information, the probability at each node is revised using the information stored in the conditional probability tables.

Although the structures of simple BNs can now be created automatically, experts have traditionally created complex BNs “manually”. Once the structure is established, a network’s conditional probability distributions can be trained using a clinical data set. As one of their main advantages, properly designed BNs can offer clinicians explicit explanations for their decisions.

Artificial Neural Networks

Artificial neural networks are connectionist systems consisting of interconnected nodes [15]. The nodes are organized into an input layer that captures the variables of interest, an output layer that gives the result, and one or more hidden layers that connect the two. ANNs learn automatically from clinical data by encoding the knowledge as numeric weights between the nodes.

The main strength of ANNs is their ability to learn autonomously. Their weakness is their inability to offer explanations for a particular decision, because they are

black boxes. Nevertheless, ANNs have been applied to clinical problems that do not require specific explanations [16].

Materials and Methods

Data

We used an existing database from the emergency department at LDS Hospital, a tertiary care hospital with 520 beds. The database contained data on 32,662 adult patients (18 years or older) collected over a two-year period (May 26, 1995 to June 14, 1997). It captured 81 variables, which are routinely available in the hospital information system known as the HELP (Health Evaluation through Logical Processing) system [17]. The proportions of missing values for each variable varied considerably. For example, 94.9% of values for paO_2 were absent, because only few patients require an arterial blood gas. The data set reflects the large variation of available data typical of an adult emergency department population. The target disease of the database was CAP (ICD-9 code: 480-486). Of the 32,662 patients, 498 patients (1.5 %) had a discharge diagnosis of CAP. We randomly assigned two-thirds of the data (21,775 patients) to a training set and one-third to an independent testing set (10,887 patients).

Networks

We took a manually derived structure of a diagnostic BN for CAP [18] (Figure 1) and trained it with the software tool Netica™. The 25 variables of the BN are listed in Table 1.

Using the same 25 variables, we developed an ANN (Figure 1) with the software tool STATISTICA™ Neural Networks (Release 4.0 C). For the ANN, we randomly assigned two-thirds of the training set (14,517 patients) to a learning set and one-third to an internal verification set (7,258 patients) to prevent over-fitting.



Table 1 - Variables used to train both networks

Variable	Variable Type
age	continuous
respiratory rate	continuous
systolic blood pressure	continuous
heart rate	continuous
temperature	continuous
SpO ₂ ^a	continuous
paO ₂ ^b	continuous
pCO ₂ ^c	continuous
blood pH ^d	continuous
white blood count	continuous
bands	continuous
blood urea nitrogen	continuous
gamma glutamyl transferase	continuous
albumin	continuous
alkaline phosphatase	continuous
risk factors	continuous
cough	categorical
breath sounds	categorical
chief complaint	categorical
abdominal discomfort	categorical
abdominal exam	categorical
pain characteristic	categorical
chest x-ray ordered	dichotomous
chest x-ray indicating pneumonia	dichotomous
CAP ^e present (output variable)	dichotomous

^aSpO₂ = pulse oximeter estimate of arterial oxygen saturation

^bpaO₂ = partial pressure of oxygen

^cpCO₂ = partial pressure of carbon dioxide

^dblood pH = blood acidity

^eCAP = community-acquired pneumonia

Statistical Analysis

We evaluated the ability of the BN and the ANN to identify CAP patients in the independent testing set. 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. As the thresholds change, the classification of a case having CAP might change. The test characteristics of the networks reflect this behavior. In this study, we used sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) as test characteristics (Table 2).

We set the network thresholds to detect 95% of the CAP patients (95% sensitivity) and calculated the test characteristics. Under this high sensitivity, the networks identify most patients and thus act as screening tools. In addition, we determined the most frequent diagnoses of the misclassified patients, to see whether the misclassifications made by the networks were clinically justifiable.

Over the range of all thresholds, we plotted sensitivity against 1 - specificity to get receiver operating

Table 2 - Test characteristics

	CAP present	CAP absent	
network +	a	b	PPV = a/(a+b)
network -	c	d	NPV = d/(c+d)
	sens = a/(a+c) spec = d/(b+d)		

The table shows how to calculate the following test characteristics: sensitivity (sens), specificity (spec), positive predictive value (PPV), and negative predictive value (NPV).

CAP present/ absent, whether the patients had/ had not community-acquired pneumonia (CAP); network +/-, whether the network diagnosed/ did not diagnose CAP.

characteristic (ROC curves) [19] for the BN and ANN. The area under such curves (AUC) is a measure of the overall accuracy of a decision support tool. An AUC ranges from 0 to 1, where 1 implies that the network perfectly discriminates between two classifications, and 0.5 implies that it cannot discriminate between two classifications. To determine the accuracies of the networks, we calculated the AUC with the 95% confidence intervals for each ROC curve using the software tool ROCKIT (version 0.9.1 BETA). To investigate if the AUCs were significantly different, we used the correlated area z statistic [20].



Results

At a sensitivity of 95%, the specificities and the NPVs of the

Table 3 - Test characteristics at a sensitivity of 95%

	Specificity	PPV	NPV
BN	92.3%	15.1 %	99.9%
ANN	94.0%	18.6%	99.9%

PPV, Positive predictive value; NPV, Negative predictive value; BN, Bayesian network; ANN, artificial neural network.

networks were high, and the PPVs were low (Table 3).

At the same sensitivity, the patients, whom the networks misclassified as having CAP, had diagnoses for pulmonary, cardiac, and infectious diseases (Table 4). These diagnoses captured 50% of the misclassified cases. For both networks, the three diagnoses most frequently misclassified as CAP were asthma, bronchitis, and congestive heart failure.

The ROC curves of both networks are shown in Figure 2. The AUC of the BN was 0.9795 with a 95% confidence interval of (0.9736, 0.9843). The AUC of the ANN was 0.9855 with a 95% confidence interval of (0.9805, 0.9894). The correlated area z statistic was statistically significant (p=0.0044), indicating a difference between the AUCs.

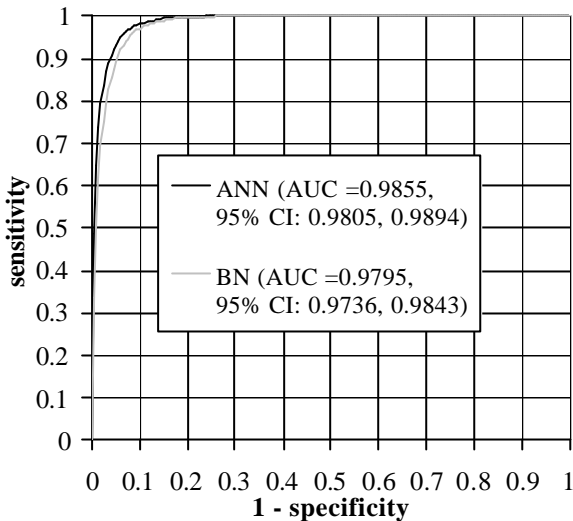


Figure 2 - Receiver operating characteristic curves for the Bayesian network (BN) and the artificial neural network (ANN). AUC, Area under the curve.

Discussion

Both the BN, with an AUC of 0.9795, and the ANN, with an AUC of 0.9855, achieved a high overall accuracy in discriminating pneumonia patients in our data set. These good results might seem astonishing considering the incompleteness of the data. Still, we were able to construct sensitive networks, because we had a large number of cases. The following example illustrates the effect of large numbers. Only 5.1% of the cases had a value for the partial pressure of oxygen (paO_2). With a smaller data set, this percentage is probably insufficient for paO_2 to help predict CAP. With our data set of 32,662 cases, however, 5.1% corresponds to 1,669 cases. This large number was sufficient for paO_2 to have an effect.

Despite the statistically significant difference between the AUCs, we do not conclude that the ANN is more accurate than the BN. We interpret the results as follows. First, the large sample size contributed to the statistically significant correlated area z statistic. Second, a difference in the overall accuracy of $0.9855 - 0.9795 = 0.006$ may be statistically significant, but compared to the overall accuracy of either network it probably plays a minor role.

For the purpose of driving a clinical guideline, both models must detect as many cases as possible. Therefore, we investigated the behavior of the networks at a high 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 made by the network, approximately 17 are appropriate, whereas 83 are false. This high number of misclassifications is justifiable for two reasons. First, in a system designed to screen patients for CAP, missing patients with CAP is worse than misclassifying patients without CAP. Second, 50% of false positively diagnosed patients had a pulmonary, cardiac, or infectious disease (Table 4). The false positive classifications are, thus, clinically plausible suggestions, consistent with the differential diagnosis of CAP.

Our study had limitations. First, we used a testing set of retrospective cases. We are about to complete a validation of both networks on a prospectively collected data set. A preliminary prospective study on the BN [21] gives reason to believe that the performance will not drop substantially. Second, we used the ICD-9 discharge codes as a gold standard for CAP. In view of the variability in physicians' agreement on physical examinations [22] and chest radiographs [23], however, an objective gold standard for CAP does not exist, and the ICD-9 codes were a feasible and economic approach for this pilot study. Third, with a different graphical structure or with a restructuring of select nodal values, the BN may show significantly better

Table 4 - Most frequent diagnoses of patients, who were misclassified as having community-acquired pneumonia

Artificial neural network (N=936 false positives)			Bayesian network (N=861 false positives)		
Diagnosis	Count	Cumulative Percentage	Diagnosis	Count	Cumulative Percentage
asthma	88	9%	asthma	75	9%
acute and chronic bronchitis	81	18%	acute and chronic bronchitis	57	15%
congestive heart failure	51	24%	congestive heart failure	51	21%
upper respiratory infection	39	28%	fever of unknown origin	46	27%
fever of unknown origin	37	32%	chest pain	40	31%
chest pain	36	35%	dyspnea	27	34%
painful respiration	30	39%	urinary tract infection	23	37%
dyspnea	27	42%	painful respiration	22	40%
coronary artery disease	22	44%	upper respiratory infection	22	42%
aspiration pneumonia	20	46%	aspiration pneumonia	21	45%
acute myocardial infarction	20	48%	sepsis	21	47%
cough	18	50%	acute myocardial infarction	17	49%
			pulmonary embolism	17	51%

accuracy. Therefore, we plan to evaluate the two decision support methods in greater depth with the complete data set.

Conclusion

We detected a statistically significant difference of 0.006 between the areas under the receiver operating characteristic curves (AUC) of the two decision support models. Nevertheless, we consider this difference to be clinically insignificant and conclude that both models are equally accurate. In view of the high AUCs of the ANN and BN (0.9855 and 0.9795 respectively) and the clinically plausible false positive classifications, we believe that both networks are equally suited to drive a pneumonia guideline.

Acknowledgements

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

References

- [1] Miller RA. Medical diagnostic decision support systems --past, present, and future: a threaded bibliography and brief commentary. *J Am Med Inform Assoc* 1994; 1; pp. 8-27.
- [2] Szolovits P. Uncertainty and decisions in medical informatics. *Methods Inf Med* 1995; 34; pp. 111-21.
- [3] Baxt WG. Application of artificial neural networks to clinical medicine. *Lancet* 1995; 346; pp. 1135-8.
- [4] Forsström JJ and Dalton KJ. Artificial neural networks for decision support in clinical medicine. *Ann Med* 1995; 27; pp. 509-17.
- [5] Kahn CE, Jr. Artificial intelligence in radiology: decision support systems. *Radiographics* 1994; 14; pp. 849-61.
- [6] Kück K, Haryadi DG, and Westenskow DR. Application of Artificial Neural Networks in Anesthesiology. In *Advances in Anesthesia*, CL Lake, LJ Rice, and RJ Sperry, Editors. 1998, Mosby: St Louis, Missouri; pp. 151-166.
- [7] Eisenstein EL and Alemi F. A comparison of three techniques for rapid model development: an application in patient risk-stratification. *Proc AMIA Annu Fall Symp* 1996; pp. 443-7.
- [8] Duh MS, Walker AM, Pagano M, and Kronlund K. Prediction and cross-validation of neural networks versus logistic regression: using hepatic disorders as an example. *Am J Epidemiol* 1998; 147; pp. 407-13.
- [9] Ohmann C, Moustakis V, Yang Q, and Lang K. Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain. Acute Abdominal Pain Study Group. *Artif Intell Med* 1996; 8; pp. 23-36.
- [10] Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, Caruana R, Fine MJ, Glymour C, Gordon G, Hanusa BH, Janosky JE, Meek C, Mitchell T, Richardson T, and Spirtes P. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med* 1997; 9; pp. 107-38.
- [11] Marrie TJ. Community-acquired pneumonia: epidemiology, etiology, treatment. *Infect Dis Clin North Am* 1998; 12; pp. 723-40.
- [12] Marrie TJ. Community-acquired pneumonia. *Clin Infect Dis* 1994; 18; pp. 501-13.
- [13] Jensen FV. *An Introduction to Bayesian Networks*. New York: Springer Verlag New York Inc., 1996.
- [14] Spiegelhalter DJ, Dawid AP, Lauritzen SL, and Cowell RG. Bayesian Analysis in Expert Systems. *Statistical Science* 1993; 8; pp. 219-283.
- [15] Gurney K. *An Introduction to Neural Networks*. 1st ed. London: UCL Press Limited, 1997.
- [16] Hart A and Wyatt J. Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. *Med Inform (Lond)* 1990; 15; pp. 229-36.
- [17] Gardner RM, Pryor TA, and Warner HR. The HELP hospital information system: update 1998. *Int J Med Inf* 1999; 54; pp. 169-82.
- [18] Aronsky D and Haug PJ. Diagnosing community-acquired pneumonia with a Bayesian network. *Proc AMIA Symp* 1998; pp. 632-6.
- [19] Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978; 8; pp. 283-98.
- [20] Meistrell ML. Evaluation of neural network performance by receiver operating characteristic (ROC) analysis: examples from the biotechnology domain. *Comput Methods Programs Biomed* 1990; 32; pp. 73-80.
- [21] Aronsky D and Haug PJ. Automatic identification of patients eligible for a pneumonia guideline. *Proc AMIA Symp* 2000; (20); pp. 12-6.
- [22] Wipf JE, Lipsky BA, Hirschmann JV, Boyko EJ, Takasugi J, Peugeot RL, and Davis CL. Diagnosing pneumonia by physical examination: relevant or relic? *Arch Intern Med* 1999; 159; pp. 1082-7.
- [23] Albaum MN, Hill LC, Murphy M, Li YH, Fuhrman CR, Britton CA, Kapoor WN, and Fine MJ. Interobserver reliability of the chest radiograph in community-acquired pneumonia. PORT Investigators. *Chest* 1996; 110; pp. 343-50.

Address for correspondence

Charles Lagor, M.D.
 Department of Medical Informatics, LDS Hospital
 8th Avenue and C Street
 Salt Lake City, Utah 84143, USA
 ldclagor@ihc.com