ARTIFICIAL NEURAL NETWORK MODEL FOR THE ASSESSMENT OF LYMPH NODE SPREAD IN PATIENTS WITH CLINICALLY LOCALIZED PROSTATE CANCER

JOSEPH T. BATUELLO, EDUARD J. GAMITO, E. DAVID CRAWFORD, MISOP HAN, ALAN W. PARTIN, DAVID G. MCLEOD, AND COLIN O’DONNELL

ABSTRACT

Objectives. To develop an artificial neural network (ANN) model to predict lymph node (LN) spread in men with clinically localized prostate cancer and to describe a clinically useful method for interpreting the ANN’s output scores.

Methods. A simple, feed-forward ANN was trained and validated using clinical and pathologic data from two institutions (n = 6135 and n = 319). The clinical stage, biopsy Gleason sum, and prostate-specific antigen level were the input parameters and the presence or absence of LN spread was the output parameter. Patients with similar ANN outputs were grouped and assumed to be part of a cohort. The prevalence of LN spread for each of these patient cohorts was plotted against the range of ANN outputs to create a risk curve.

Results. The area under the receiver operating characteristic curve for the first and second validation data sets was 0.81 and 0.77, respectively. At an ANN output cutoff of 0.3, the sensitivity achieved for each validation set was 63.8% and 44.4%; the specificity was 81.5% and 81.3%; the positive predictive value was 13.6% and 6.5%; and the negative predictive value was 98.0% and 98.1%, respectively. The risk curve showed a nearly linear increase (best fit $R^2 = 0.972$) in the prevalence of LN spread with increases in raw ANN output.

Conclusions. The ANN’s performance on the two validation data sets suggests a role for ANNs in the accurate clinical staging of patients with prostate cancer. The risk curve provides a clinically useful tool that can be used to give patients a realistic assessment of their risk of LN spread.

© 2001, Elsevier Science Inc.

The accurate assessment of clinical stage is crucial to the successful treatment of prostate cancer. Inaccuracies in staging may result in inappropriate treatment and, consequently, poor patient outcomes. Published data reveal that as many as 59% of patients who undergo radical prostatectomy for treatment of presumed clinically localized prostate cancer are found to have cancer beyond the prostate gland on pathologic examination of the surgical specimens.1–7 Furthermore, between 2% and 12% of patients with prostate cancer with presumed localized disease are found to have lymph node (LN) involvement.1,8–11 An accurate method of assessing the true clinical stage is needed so that physicians and their patients can choose stage-appropriate treatments.

A number of investigators have used the clinical and pathologic data recorded in large institutional databases to develop statistical models to predict the pathologic stage (true clinical stage at the time of surgery).11–15 The most widely used and accepted of these models are the nomograms developed by Partin et al.11 These nomograms use the clinical TNM stage, biopsy Gleason sum, and preoperative prostate-specific antigen (PSA) level to predict the pathologic stage in men with clinically localized prostate cancer. Most models currently in
use, including the Partin nomograms, were developed using traditional statistical methods such as multinomial log-linear regression analysis. Previously, Crawford et al. used decision-tree analysis to predict LN spread in men with clinically localized prostate cancer. In the present study, we use an artificial neural network (ANN) to develop a model using clinical stage, biopsy Gleason sum, and PSA level to predict LN spread in men with clinically localized prostate cancer. In addition, we describe a method by which the ANN model’s predictions can be given a clinically useful meaning.

In contrast to statistical modeling, ANNs use a software construct inspired by the neural structure of the brain consisting of simple processing units called nodes to simulate neurons and weighted interconnections between the nodes to simulate dendrites and axons. The interconnection weights serve as coefficients (multipliers) that simulate the connection strengths in the biologic model. An ANN is not programmed like a conventional linear computer model, but learns by experience. Typically, an ANN is developed in two phases, a training phase and a validation phase. During training, cases that include inputs and known outputs are presented to the ANN sequentially and repeatedly. A training algorithm automatically adjusts the connection weights, consequently changing the output values, to reduce the errors between the actual ANN outputs and the expected outputs. Over time, a matrix of connection weights emerges that allows for the largest number of correct predictions or classifications for the given training data set. Once trained, the ANN is validated on cases not used during the training process. The trained ANN will function in a manner similar to a mathematical function with inputs analogous to independent variables and outputs analogous to dependent variables. The ANN’s performance on the validation set can then be quantified using measures such as sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve.

MATERIAL AND METHODS

Training, testing, and validation data sets were derived from a database containing clinical and pathologic information for 6135 men who had undergone bilateral, pelvic lymphadenectomy and radical retropubic prostatectomy between 1985 and 1998 for clinically localized prostate cancer at the Johns Hopkins Medical Institutions (JHMI). None of the patients had received prior radiation or hormonal treatment. The patients were staged by the operative surgeon according to the 1992 American Joint Committee on Cancer classification and staging guidelines with digital rectal examination findings and serum PSA measurements (Tandem-R and Tandem-E, Beckman-Coulter, San Diego, Calif and Tosoh Medics PSA assays, San Francisco, Calif). Biopsy and surgical specimens, including prostate gland, seminal vesicles, and pelvic LNs, were examined for the presence of prostate cancer at the JHMI.

A validation data set of 1840 patients (30%) was drawn at random (Microsoft Excel random number generation) from the JHMI database. A test set was then created by randomly selecting 20 LN-positive and 20 LN-negative cases from the remaining 4295 patients. This set of 40 patients was used to test the ANN periodically during training. Finally, a training set was derived from the remaining 4235 patients by selecting all of the LN-positive cases (n = 184) and randomly selecting 552 LN-negative cases. The resulting training set consisted of 736 cases with an enriched LN-positive prevalence of 23% to allow the characteristics of the LN-positive cases to have a significant impact on the ANN’s weight matrix.

A second validation set of 319 cases was derived from a database from the Walter Reed Army Medical Center (WRAMC) containing clinical and pathologic information for men who had undergone radical retropubic prostatectomy and pelvic lymphadenectomy between 1988 and 1998 for clinically localized prostate cancer. These cases represent those patients at WRAMC who had not received radiation or hormonal treatment before surgery. Each patient was staged by the operative surgeon. Biopsy and surgical specimens were examined centrally by the Armed Forces Institute of Pathology. A comparison of the descriptive statistics for the patients in the training and validation sets can be found in Table I.

A three-layer, feed-forward ANN was designed using a commercially available software package (BrainMaker Professional, California Scientific Software, Nevada City, Calif). A back-propagation algorithm was used to train the ANN. Three clinical variables, clinical TNM stage, Gleason sum, and PSA, were used as the input parameters and the LN status for each patient was used as the output parameter. The ANN was validated using the two validation sets.

In addition to the standard analysis of the ANN’s performance on the validation sets, an exploratory method for interpreting the raw ANN outputs was used. The entire JHMI database (n = 6135) was run through the ANN. The resulting raw ANN outputs were continuous numbers between 0 and 1. The prevalence of LN-positive cases was calculated for a data window of 614 patients (10%) with raw outputs closest in value to each tenth of the raw ANN output (0.1, 0.2, 0.3, and so on) up to 0.7. As only 103 patients had ANN outputs of 0.8 or higher, it was necessary to use smaller windows to calculate the prevalences for patient groups with ANN outputs of 0.8 or higher. The number of patients in each data window is shown in Figure 1. The sizes of the windows were selected to include the largest number of cases that gave a stable estimate of prevalence. The prevalence of LN-positive cases for each data window was plotted against the raw output to generate a risk curve.

RESULTS

The area under the ROC curve for the JHMI validation set was 0.81. At an ANN output cutoff of 0.3, 80% of the patients (1464 of 1840) in the JHMI validation set were identified as being at low risk of LN spread, with a 2.0% false-negative rate (29 of 1464). The sensitivity and specificity at this cutoff was 63.8% and 81.5%, respectively. The positive predictive value was 13.6% and the negative predictive value was 98.0%. The ANN’s performance on the WRAMC database was similar, with 81% of the patients (257 of 319) considered at low risk of LN spread and a 1.9% false-negative rate (5 of 257) at the 0.3 cutoff. The sensitivity was 44.4%, specificity 81.3%, positive predictive value 6.5%, and
negative predictive value 98.1%. The area under the ROC curve was 0.77. The risk curve generated by the data window analysis (Fig. 1) indicates a nearly linear relationship between the ANN outputs and the prevalence of LN spread.

**COMMENT**

The ANN was able to identify patients at low risk of LN involvement at the time of surgery with an accuracy of 98% (less than 2% false-negative results) using three clinical parameters: biopsy Gleason sum, clinical stage, and preoperative PSA level. The area under the ROC curve for the JHMI and WRAMC validation sets was 0.81 and 0.77, respectively. This model was optimized for high negative predictive values. This is reflected in the resulting negative predictive values of 98.0% and 98.1% on the validation sets. Conversely, the corresponding positive predictive values were low (13.6% and 6.5%). Confidence in the ANN’s ability to generalize is supported by its consistent performance on two validation data sets from separate institutions.

We determined that all of the false-negative cases (ie, the patients who were positive for LN spread,

\[
\text{TABLE I. Comparison of descriptive statistics among original, training, and validation data sets (testing data set not shown)}
\]

<table>
<thead>
<tr>
<th></th>
<th>Full JHMI Database</th>
<th>Training Set (JHMI)</th>
<th>Validation (JHMI)</th>
<th>Validation (WRAMC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients (n)</td>
<td>6135</td>
<td>736</td>
<td>1840</td>
<td>319</td>
</tr>
<tr>
<td>Mean age (yr)</td>
<td>58.7 ± 6.5</td>
<td>58.9 ± 6.8</td>
<td>58.7 ± 6.4</td>
<td>62.2 ± 6.4</td>
</tr>
<tr>
<td>Gleason sum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>2–10</td>
<td>3–10</td>
<td>2–9</td>
<td>2–10</td>
</tr>
<tr>
<td>Mean</td>
<td>6.1 ± 0.8</td>
<td>6.3 ± 0.8</td>
<td>6.3 ± 0.8</td>
<td>5.4 ± 1.4</td>
</tr>
<tr>
<td>Median</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>PSA (ng/mL)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>0.2–151.0</td>
<td>0.3–84.1</td>
<td>0.2–66.4</td>
<td>0.0–73.6</td>
</tr>
<tr>
<td>Mean</td>
<td>8.5 ± 7.6</td>
<td>10.1 ± 9.8</td>
<td>8.5 ± 7.0</td>
<td>9.4 ± 10.0</td>
</tr>
<tr>
<td>Median</td>
<td>6.6</td>
<td>7.3</td>
<td>6.7</td>
<td>7.0</td>
</tr>
<tr>
<td>Stage (n)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cT1a</td>
<td>55 (0.9)</td>
<td>4 (0.5)</td>
<td>17 (0.9)</td>
<td>4 (1.3)</td>
</tr>
<tr>
<td>cT1b</td>
<td>126 (2.1)</td>
<td>14 (1.9)</td>
<td>37 (2.0)</td>
<td>3 (0.9)</td>
</tr>
<tr>
<td>cT1c</td>
<td>2904 (47.3)</td>
<td>311 (42.3)</td>
<td>853 (46.4)</td>
<td>148 (46.4)</td>
</tr>
<tr>
<td>cT2a</td>
<td>1652 (26.6)</td>
<td>166 (22.6)</td>
<td>512 (27.8)</td>
<td>55 (17.2)</td>
</tr>
<tr>
<td>cT2b</td>
<td>1015 (16.5)</td>
<td>159 (21.6)</td>
<td>310 (16.8)</td>
<td>85 (26.6)</td>
</tr>
<tr>
<td>cT2c</td>
<td>307 (5.0)</td>
<td>55 (7.5)</td>
<td>89 (4.8)</td>
<td>21 (6.6)</td>
</tr>
<tr>
<td>cT3a</td>
<td>96 (1.6)</td>
<td>27 (3.7)</td>
<td>22 (1.2)</td>
<td>3 (0.9)</td>
</tr>
<tr>
<td>LN positive count (%)</td>
<td>284 (4.6)</td>
<td>184 (25.0)</td>
<td>80 (4.3)</td>
<td>9 (2.8)</td>
</tr>
</tbody>
</table>

Key: JHMI = Johns Hopkins Medical Institutions; WRAMC = Walter Reed Army Medical Center; PSA = prostate-specific antigen; LN = lymph node.

Numbers in parentheses are percentages.

![Risk Curve](image.png)

**FIGURE 1.** Risk curve generated by calculating the prevalence of lymph node spread for each data window over the full JHMI database (solid line). Best-fit linear regression line (dotted line) added for reference (y-intercept fixed at zero, \( R^2 = 0.972 \)). Bin sizes for each data window are shown in columns on the right.

UROLOGY 57 (3), 2001 483
but had been classified as negative by the ANN) had clinical parameters identical or very similar to several patients who were LN negative. This indicates that additional input parameters would be needed to distinguish the false-negative cases from the true negative cases.

Two exploratory methods to the application of ANNs in medicine were implemented in this study. The first was to increase the prevalence of LN-positive cases in the data set used to train the ANN. The rationale for this approach was to allow the ANN to recognize the characteristics of LN-positive individuals. This recognition would have been difficult to achieve with the data set's natural prevalence of LN-positive cases (4.6%). The ANN training algorithm tends to minimize the overall error in the ANN's predictions during training. Because most cases were LN negative (greater than 95%), the training algorithm was prone to treat the scarce positive cases as “noise” and achieve an apparent minimal error by classifying all cases as negative. Increasing the number of LN-positive cases to 25% of the training set allowed the characteristics of the LN-positive cases to have a significant impact on the ANN's weight matrix, resulting in higher LN-positive recognition. The prevalence of 25% was chosen empirically to be a good balance between achieving a sufficiently enriched training set and allowing for a sufficiently large validation set. Regardless of the techniques used to train the ANN, the true test of its validity is its performance on the validation sets. Since the first validation data set was drawn at random from the original database before creating the enriched training set, the results obtained for the first validation set are those that would be expected for the entire original database. Both validation sets served as controls against any bias that may have been introduced in the enrichment process.

The second method introduced was the analysis of the relationship between the raw ANN output and the prevalence of LN-positive cases to provide a clinically useful interpretation of the ANN output scores. The prevalence of LN-positive cases was calculated for a series of data windows framing patients with similar ANN outputs. The prevalence for each data window was then plotted against the raw output to generate a risk curve. The single assumption made in developing this method was that patients having similar ANN outputs would have similar risks of LN disease. Thus, all patients with similar outputs were assumed to be part of a cohort. The risk of LN disease for a given patient in that cohort is simply the prevalence in those patients having similar scores. The clinical utility of this method is that a patient may reasonably be told: “Of all patients with similar risks as determined by computer modeling, x% will have cancer in the lymph nodes.”

The ANN analysis used in this study refined the decision rules previously presented by us. The earlier study produced boundaries for preoperative PSA level, biopsy Gleason sum, and clinical stage by which patients could be classified as being at low or high risk of LN metastasis. The decision rules thus derived did not recognize the spectrum of risks present in either group. A single probability of LN disease was assumed for all patients classified as “low risk,” despite the range of clinical parameters represented in that group.

The greater precision afforded by ANN analysis is illustrated by considering patients whose clinical parameters matched the cutoffs derived for the decision-tree analysis (ie, PSA level 10.6 ng/mL, Gleason sum of 6, and clinical Stage T2a. These patients would be at the boundaries of the patient population classified as at low risk of LN spread using decision-tree analysis, but would reasonably be expected to have a higher risk than patients with lower PSA levels or Gleason sums. The ANN analysis confirms this, giving an estimated risk of LN spread of 5% for these patients. As a reference, the tables presented by Partin et al. give an estimated risk of 4% for this patient population.

We have demonstrated a method by which each individual patient can be assigned a more precise estimate of risk than the binary classification presented earlier. Using the ANN model presented here, 80% of the patients in the JHMI database were categorized as low risk, with a false-negative rate of 2.0%. The decision rules published previously classified 62% of the patients as being at low risk, with a 1.7% false-negative rate. The patients used to derive the decision-tree rules were a subset of those used for the ANN model.

The advantage of the approach presented here over decision-tree analysis will become more obvious as the number of clinical variables included in the decision model increases. The three variables used in this study represent a manageable and easily assimilated set of decision variables. As molecular and genetic markers in prostate cancer proliferate, and more sophisticated imaging technologies are developed, integration of these into the process of clinical decision making will likely become far more complex. The ANN analysis presented here should, given sufficient training data, be robust enough to accommodate such complexity.

We should emphasize that the ANN has modeled the data contained in the training set, and thus reflects the epidemiologic conditions that existed when the training set was collected. Since the advent of widespread PSA screening around 1990, a significant shift has occurred in the clinical presen-
tation of prostate cancer, with a trend toward an earlier stage at diagnosis.\textsuperscript{2,19} The training data contained no time-referenced information. The ANN thus had no way of incorporating time-dependent trends into its predictions. Updating the ANN to address any limitations imposed by these facts requires retraining it with data more representative of the current population of patients with prostate cancer.

CONCLUSIONS

The results of this study suggest a role for ANNs in the clinical staging of patients with prostate cancer. The use of ANNs to analyze large sets of patient data recorded in institutional databases represents a powerful method for supplementing more conventional methods of analysis. ANNs can be used to create detailed models derived from the clinical experience of large numbers of patients, against which the clinical data of a particular patient may be compared. These models do not rely on a priori assumptions about the relationships that may exist between different clinical variables. This study has demonstrated how ANNs may be used to derive clinically relevant estimates of risk. How this information affects a given clinical decision depends on the unique aspects of each clinician and patient.

REFERENCES