An artificial neural network for five different assay systems of prostate-specific antigen in prostate cancer diagnostics

Carsten Stephan, Henning Cammann*, Hellmuth–Alexander Meyer, Christian Müller†, Serdar Deger, Michael Lein and Klaus Jung

Department of Urology, *Institute for Medical Informatics, and †Department of Clinical Chemistry and Laboratory Medicine, Charité–Universitätsmedizin Berlin, Berlin, Germany

Accepted for publication 29 February 2008
All authors are members of the Biomarkers of Prostate Cancer Group, CCM

INTRODUCTION

Measuring the PSA level has limitations as a screening test for prostate cancer (PCa). Especially in the 2–10 ng/mL PSA level ‘grey zone’, this serum test alone cannot distinguish between PCa and BPH [1]. Measurements of the two major molecular forms of PSA, free PSA (fPSA) and complexed PSA (cPSA), have been shown to improve specificity compared with total PSA (tPSA) alone [2,3]. Using the ratio of fPSA to tPSA (%fPSA) in the tPSA range of 4–10 ng/mL, 20–25% of unnecessary biopsies can be avoided [4,5]. For tPSA values of <4 ng/mL the use of the ratio of PSA level to prostate volume [6] and %fPSA have also been reported to increase specificity [7,8]. Furthermore, cPSA has been shown to be a better first-line test than tPSA but only the ratio of cPSA/tPSA (%cPSA) performs equivalently to the %fPSA [9].

As the %fPSA and tPSA are influenced by factors such as prostate volume [10–12] and age [10,11], different multivariate logistic regression models [13,14] and artificial neural networks (ANN) [15–20] have been introduced to improve cancer specificity. The use of different ANN models to detect PCa has been reviewed recently [21]. All ANN and logistic regression models show an improvement in specificity when compared with %fPSA, but they use partially different input data such as prostate volume indexes or DRE status. Most importantly, these models are not comparable due to the use of different PSA assays and PSA ranges [13,15,16,20]. It is highly recommended, that %fPSA should not be calculated when the tPSA and fPSA levels are obtained from different manufacturers because interpretive criteria for %fPSA may not apply to unproven combinations [5,22,23]. This could have unintended consequences for the number of unnecessary prostate biopsies, especially with %fPSA values near the thresholds [24,25]. Stephan et al. [26] recently published a comparison in almost 4500 men using five different PSA assay-specific ANN models on separate populations showing better performance than the first ‘ProstataClass’ ANN model, which was built with one assay only [20]. However, only parallel measurements and comparisons of the different assays using the same cohort of samples could answer the question of whether PSA assay-specific ANN models are
necessary, or if one general ANN may be established for use with any PSA assay combination [26]. Therefore, we used data from our previous study [27] that compared five different tPSA and fPSA assay systems to elaborate on how PSA assay-specific ANN models answer the following questions: (i) Are there differences of diagnostic accuracy of ANN models dependent on the PSA assay systems?; (ii) Are these differences comparable to those between tPSA and %fPSA?; and (iii) Is it possible to create one ANN for all PSA assays or is there a need for separate, assay-specific ANNs?

PATIENTS AND METHODS

In all 798 patients, 465 with PCa and 333 with no evidence of malignancy (NEM), within the tPSA range of 0.49–27 ng/mL based on the Access Hybritech PSA value were evaluated. We retrospectively investigated archival sera collected between 2001 and 2004. Data from 596 of these patients with tPSA concentrations of 0.49–10 ng/mL have been published [27].

Patients with PCa and NEM were histologically confirmed by 8–10 core prostate biopsies. All patients were urologically referred. Blood sampling and handling were performed as described previously [27]. The samples were taken before any diagnostic or therapeutic procedures, and sera were stored at −80 °C until analysed. Prostate volume was determined by TRUS using the prostate ellipse formula. A DRE finding not suspicious for cancer was defined as negative, and a finding suspicious for cancer as positive. The study was carried out in accordance with the standards of the local ethics board and the Helsinki Declaration of 1996.

The tPSA and fPSA (cPSA) were measured using the following commercially available kits: AxSYM (Abbott), ADVIA Centaur (Bayer Diagnostics; cPSA instead of fPSA), Access (Beckman Coulter), Immulite 2000 systems (Diagnostic Products Corp.) and Elecsys 2010 (Roche Diagnostics) as already published [27].

A methodological problem produced concentrations of cPSA > tPSA when measured with the ADVIA Centaur analyser in 18 patients. Consequently these patients were eliminated from further analyses. Thus, 780 patients (455 PCa and 325 NEM) were finally used for all analyses. The between-run imprecision profiles of the measurements were estimated by use of control materials supplied by the manufacturers, commercial control materials, and in-house serum pools; all interassay coefficients of variation were <8% [27].

ANN MODELS

All 10 ANN models for the five assays and respective two tPSA ranges were constructed with the MATLAB Neural Network Toolbox (The Mathworks, Natick, MA, USA). Feed-forward back-propagation networks were built, in which the input layer consisted of five neurones for the variables tPSA, %fPSA, age, prostate volume and DRE status, with three neurones as hidden layer and one output neurone, ranged from 0 (low PCa risk) to 1 (high PCa risk). To get the best generalization of the ANN we used Bayesian regularization as used by Finne et al. [16]. Each ANN model was validated by the leave-one-out method (LOO), which has been previously described in detail [28,29]. In brief, the LOO method is the extreme method of cross validation with the maximum number for the training cohort because n times the training is performed in n−1 patients and then it is tested in the remaining patient. The output values of the training run were then used to build the receiver operating characteristic (ROC) curve. Because the difference between the mean value of all outputs of the ANN from LOO training (n ANN models where each was built on n−1 patients) and the output of one ANN built with all data (n patients) was negligible, we used the latter for all calculations.

To test our ANN classification factor (number of correctly classified patients), the patients were in another step divided into a training (75%) and test group (25%) including the patients with values within the tPSA range of 0–10 ng/mL (based on measurements with the Access Hybritech PSA system, Beckman Coulter) are shown in Table 1. Between the five tPSA assays, the number of patients within these three tPSA ranges differed markedly. Based on important PCa incidence data obtained by Thompson et al. [30], which indicated a similar PCA detection rate at tPSA levels of 2–4 ng/mL compared with other data for the 4–10 ng/mL tPSA range, we used only a common tPSA range of 0–10 ng/mL for constructing the ANN models.

The data distribution of the final 780 patients for tPSA, %fPSA, volume, age, and percentage of positive DREs in patients with PCa and NEM within the clinically interesting tPSA ranges of <4, 4–10 and >10 ng/mL (based on measurements with the Access Hybritech PSA system, Beckman Coulter) are shown in Table 2. The data column named ‘true’ includes only those patients with values within the tPSA range of the respective assay. However, for a direct comparison between all assays and the subsequently established ANN models it was necessary to use exactly the same patients within both tPSA ranges for the five assays.

We used software SPSS version 14.0 for Windows (SPSS), MedCalc version 9.3.6.0 (MedCalc), and GraphPad Prism version 4.03 for Windows (GraphPad). The Mann-Whitney U-test and Kruskal-Wallis test of variance were used. The diagnostic validity of tPSA, %fPSA, and the different ANN models was evaluated by ROC curve analysis. The areas under the ROC curves (AUC) and the specificities at 90% and 95% sensitivity were compared by a nonparametric method using the software GraphROC 2.1 for Windows. Statistical significance was defined as P < 0.05.

RESULTS

The different numbers of patients with tPSA values of 0–10 and >10 ng/mL is primarily a function of the standardization of specific tPSA assays as shown in Table 2. The data column named ‘true’ includes only those patients with values within the tPSA range of the respective assay. However, for a direct comparison between all assays and the subsequently established ANN models it was necessary to use exactly the same patients within both tPSA ranges for the five assays.
Patients in the 0–10 and the five tPSA concentrations was within the 'cumulative' group was that at least one of the ANN for the tPSA ranges (106 of 780, 13.6%) where used twice to build categorization. Therefore, some patient data tPSA concentrations were used for instead of 591 or 189 when only the Access tPSA 10–27 ng/mL; (the ANN models in both tPSA ranges show small differences for tPSA, %fPSA and between the 'true' and 'cumulative' groups which were significantly lower than the AUC (cumulative group, tPSA range 0–10 ng/mL). There was a significant difference for the respective AUC (P = 0.045 at tPSA 0–10 ng/mL). There was a significant improvement for all ANN models compared with %fPSA or %cPSA.

As also seen in Table 2, two 'cumulative' ANN models (Abbott and Bayer) for the 10–27 ng/mL tPSA range had significantly lower AUCs than the other ANN models. Likewise, they show the respective largest AUC in the 'true' group, which indicates a clear dependence on the lower tPSA results in each selected patient group. The AUC for the ANN with the mean values of all five assays (tPSA 0–10 ng/mL) was 0.906 and had no differences from all other ANN models.

The specificities at the 90% and 95% sensitivity thresholds for the whole tPSA instead of using a high specificity at low tPSA values for %fPSA in the ANN models (see text below Table 3), but all ANN models showed only a small difference from each other when compared at both sensitivity thresholds. The specificities at the 90% and 95% sensitivity thresholds for the ANN with the mean values of all five assays (tPSA 0–10 ng/mL) were 60% and 71% with no significant differences from all other ANN models. However, for instance at 95% sensitivity the absolute values differed markedly between 18% (DPC) and 28% (Abbott) for %fPSA and between 0.24 (Beckman) and 0.37 (Roche) for the ANN. The same behaviour is seen for the higher tPSA range, reinforcing the questionable use of the ANN built with the mean values of the five assays.

In Table 4 the number of correctly classified patients within the 0–10 ng/mL tPSA range for tPSA, %fPSA (100% of data) and the respective ANN models (25% test population) is given (cumulative group). The ANN data were generated by using those ANN models built with 75% of all data including the ANN created for all assays by using the mean of the five different assays. The ANN created for all assays had on average the lowest rate of correctly classified patients, but still better values than %fPSA. The rate of correctly classified patients is also lower (data not shown) when using the ANN for one assay on the data of the other four assays.

The numbers of correctly classified patients within the 10–27 ng/mL range (not shown) at the 90% and 95% sensitivity thresholds for the five ANN models from the 25% of unknown patients (n = 58) were 88% and 89% and the ANN built with the mean values of the five assays identified somewhat less patients with 84.5%.

Based on these data, we developed a new version of the 'ProstataClass' ANN (Appendix) with the possibility to choose the respective assay. The difference of the clinical usable program to our former 'ProstataClass' is the extended tPSA range of 1–27 compared with 2–20 ng/mL and the change to 90% and 95% sensitivity for the whole tPSA instead of using a high specificity at low tPSA concentrations of 2–4 ng/mL [20]. Another ANN model, which was not built on the mean values of all five assays but built on all data of the five assay combination by using each patient five times in each ANN showed a comparable AUC with 0.904 to the other

<table>
<thead>
<tr>
<th>Variable</th>
<th>tPSA range, ng/mL</th>
<th>0–4</th>
<th>4–10</th>
<th>10–27</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients with Pca</td>
<td>N</td>
<td>68</td>
<td>242</td>
<td>145</td>
</tr>
<tr>
<td>Median (range)</td>
<td>tPSA, ng/mL</td>
<td>2.81* (0.86–3.96)</td>
<td>6.88 (4.04–9.39)</td>
<td>14.4 (10.0–24.0)</td>
</tr>
<tr>
<td></td>
<td>%fPSA</td>
<td>14.1* (3.36–38.4)</td>
<td>11.8* (3.10–35.8)</td>
<td>9.97* (3.58–49.9)</td>
</tr>
<tr>
<td></td>
<td>Prostate volume, mL</td>
<td>33.4 (14–60)</td>
<td>37.5* (15–110)</td>
<td>39.1* (10–100)</td>
</tr>
<tr>
<td></td>
<td>Age, years</td>
<td>63 (46–77)</td>
<td>63* (43–79)</td>
<td>63* (44–77)</td>
</tr>
<tr>
<td></td>
<td>Positive DRE, %</td>
<td>62*</td>
<td>59*</td>
<td>63*</td>
</tr>
<tr>
<td>Patients with NEM</td>
<td>N</td>
<td>143</td>
<td>138</td>
<td>44</td>
</tr>
<tr>
<td>Median (range)</td>
<td>tPSA, ng/mL</td>
<td>2.09 (0.49–3.89)</td>
<td>6.38 (4.02–9.97)</td>
<td>15.4 (10.1–27.0)</td>
</tr>
<tr>
<td></td>
<td>%fPSA</td>
<td>23.6 (6.74–69.4)</td>
<td>17.7 (4.96–54.9)</td>
<td>15.6 (2.52–56.2)</td>
</tr>
<tr>
<td></td>
<td>Prostate volume, mL</td>
<td>42.2 (15–140)</td>
<td>58.3* (13–132)</td>
<td>71.7 (25–180)</td>
</tr>
<tr>
<td></td>
<td>Age, years</td>
<td>65 (40–85)</td>
<td>66 (49–84)</td>
<td>67 (54–80)</td>
</tr>
<tr>
<td></td>
<td>Positive DRE, %</td>
<td>7</td>
<td>11</td>
<td>7</td>
</tr>
</tbody>
</table>

* P < 0.001 vs patients with NEM.
assay-specific ANN models. However, the problem of very different %fPSA values was visible again making it impossible to establish a useful and clinically acceptable ANN for all these five assays. A further problem was the inclusion of cPSA, instead of fPSA which can be solved in a separate ANN but not in one valid ANN.

**DISCUSSION**

The reduction of unnecessary prostate biopsies has gained new attention by the use
of %fPSA, which improves specificity by 10–20% compared with using tPSA levels [4,11], while other molecular PSA forms or kallikreins have shown limited clinical success [4,11]. While other molecular PSA forms or %fPSA, which improves specificity by 10–20% compared with using tPSA levels [4,11], while other molecular PSA forms or kallikreins have shown limited clinical success [4,11].

Also, at tPSA concentrations of <4 ng/mL, the value of %fPSA decreases while tPSA alone is already a strong predictor of PCa. Furthermore, the investigated population (screened or referred) with a possible unequal tPSA distribution may also lower the ability of %fPSA to discriminate between PCa and NEM [26]. Regardless of the assays used or the tPSA range, ANN models based on tPSA, %fPSA, age, prostate volume (indexes), and the DRE status, significantly enhance the performance of tPSA and %fPSA [15–18,20].

Another difference to the first ANN model are the thresholds used at low tPSA concentrations (<4 ng/mL). About 20–30% of patients with PCa show PSA concentrations in this low range. About 15% of 2950 biopsied men after 7 years follow-up with PSA concentrations of <4 ng/mL had a diagnosis of PCa [30]. Because there are comparable PCa detection rates at tPSA of <4 and 4–10 ng/mL, we decided to use the 90% and 95% sensitivity thresholds for the whole tPSA within the new ANN models. With our used data from 0.5 to 10 ng/mL and a PCa detection rate being almost equal from 2 to 10 ng/mL, we decided to set the clinical threshold for the lowest tPSA to use the ANN ‘ProstataClass’ at 1 ng/mL.

A comparison of two ANN models in two different populations (screened or referred) showed that both ANN models could not improve specificity of %fPSA in the screening population, while they did in the referred group [33]. Evaluating an ANN model based on the results of different populations is only possible with limitations, as differences in tPSA and %fPSA distributions, as well as different PSA assays or differences in the PCa detection rate from the number of cores per biopsy (6–10) may be more responsible for differences than the ANN itself [26]. The optimal number of correctly classified patients could only be achieved with the respective assay-specific ANNs. A comparison of the respective ANN models to logistic regression regarding AUC, 95% and 90% sensitivity was performed but not shown, as both methods perform equally in studies with large cohorts [34].

One point of discussion in the practical use of our program ‘ProstataClass’, is that each ANN acts independently of the others. This could lead to a patient with one data set getting different results for each assay combination.

Problems with higher cPSA than tPSA concentrations have recently been published by another group [25]. These methodological problems, as seen in 18 of the present 798 patients must be solved for cPSA to be accepted as a better first-line marker than tPSA. Also, the advantage for cPSA compared with tPSA regarding AUC disappeared at 90% and 95% sensitivity.

New possibilities to further improve ANN models are the inclusion of new serum markers, e.g. different kallikreins [21,28].

TABLE 4 The number of correctly classified patients (%) within the tPSA range of 0–10 ng/mL at 90% and 95% sensitivity if using tPSA, %fPSA (n = 651) and the respective assay-specific ANN model with 25% of tested patients (n = 162)

<table>
<thead>
<tr>
<th>Assay</th>
<th>ANN (mean of five assays)</th>
<th>Abbott</th>
<th>Bayer</th>
<th>Beckman</th>
<th>DPC</th>
<th>Roche</th>
</tr>
</thead>
<tbody>
<tr>
<td>%fPSA</td>
<td>90% sensitivity n.a.</td>
<td>68.4</td>
<td>67.9</td>
<td>68.2</td>
<td>69.3</td>
<td>67.9</td>
</tr>
<tr>
<td>%fPSA</td>
<td>95% sensitivity n.a.</td>
<td>65.0</td>
<td>65.4</td>
<td>65.1</td>
<td>66.2</td>
<td>64.7</td>
</tr>
<tr>
<td>ANN (n = 162)</td>
<td>90% sensitivity n.a</td>
<td>70.5</td>
<td>71.1</td>
<td>72.5</td>
<td>71.3</td>
<td>70.8</td>
</tr>
<tr>
<td>ANN (n = 162)</td>
<td>95% sensitivity n.a</td>
<td>67.1</td>
<td>67.6</td>
<td>68.5</td>
<td>66.2</td>
<td>66.2</td>
</tr>
<tr>
<td>90% sensitivity</td>
<td>79.6</td>
<td>83.3</td>
<td>83.3</td>
<td>84.0</td>
<td>84.0</td>
<td>82.1</td>
</tr>
<tr>
<td>95% sensitivity</td>
<td>74.7</td>
<td>78.4</td>
<td>79.0</td>
<td>72.2</td>
<td>81.5</td>
<td>75.3</td>
</tr>
<tr>
<td>n.a., not available.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
macrophage inhibitory cytokine 1 and migration inhibitor factor [29] or proPSA [35]. Nonetheless, these studies with new markers were only partially successful in limited tPSA ranges [28], limited patient groups [35] or only with the additional consideration of prostate volume [29]. Hopefully, proPSA which showed initial promising results [36] may further improve ANN models by including this serum marker.

To conclude, ANN models are helpful to assess the patient’s risk for PCa and to decide whether a biopsy is indicated. Only assay-specific ANN models can safely optimize the number of correctly classified patients while a general ANN improves the performance of %fPSA but to a smaller extent than assay-specific ANN. Therefore, we could not establish one ANN for all assays. We developed a new version of the ANN named ‘ProstataClass’ to provide clinicians with an easy way to use different tPSA and fPSA assays. However, only external use of our program can show its reliability for wider clinical use. The large %fPSA variability we currently see can be decreased with the use of this ANN.

ACKNOWLEDGEMENTS

We gratefully acknowledge Charles F. Weinzierl for helpful corrections. This work was partly supported by the Berliner Sparkassenstiftung Medizin, the ProFIT-Program, the Mildred-Scheel-Foundation (Grant 70-3295-ST1), and the Monika-Kutzner-Stiftung.

CONFLICT OF INTEREST

None declared.

REFERENCES

6. Stephan C, Stroebel G, Heinau M et al. The ratio of prostate-specific antigen (PSA) to prostate volume (PSA density) as a parameter to improve the detection of prostate carcinoma in PSA values in the range of <4 ng/mL. Cancer 2005; 104: 993–1003
7. Catalona WJ, Partin AW, Finlay JA et al. Use of percentage of free prostate-specific antigen to identify men at high risk of prostate cancer when PSA levels are 2.5–4 ng/mL and digital rectal examination is not suspicious for prostate cancer: an alternative model. Urology 1999; 54: 220–4
13. Carlson GD, Calvanese CB, Partin AW. An algorithm combining age, total prostate-specific antigen (tPSA), and percent free PSA to predict prostate cancer: results on 4298 cases. Urology 1998; 52: 455–61
22. Patel D, White PA, Milford Ward A. A comparison of six commercial assays for total and free prostate specific antigen (PSA): the predictive value of the ratio of free to total PSA. BJU Int 2000; 85: 686–9


32 Roehl KA, Antenor JA, Catalona WJ. Robustness of free prostate specific antigen measurements to reduce unnecessary biopsies in the 2.6–4.0 ng/mL range. *J Urol* 2002; 168: 922–5


34 Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer* 2001; 91 [Suppl.]: 1636–42


Correspondence: Carsten Stephan, Department of Urology, Charité–Universitätsmedizin Berlin, Charitéplatz 1, D-10117 Berlin, Germany. e-mail: carsten.stephan@charite.de

Abbreviations: ANN, artificial neural network; PCa, prostate cancer; NEM, no evidence of malignancy; ROC, receiver operating characteristic (curve); tPSA, total PSA; fPSA, free PSA; cPSA, complexed PSA; %cPSA, percentage cPSA; %fPSA, percentage free/total PSA; AUC, area under ROC curve; LOO, leave-one-out.

**APPENDIX**

The program ‘ProstataClass’ with the input parameters: age, tPSA, %fPSA (or cPSA if Bayer is selected), prostate volume and DRE status.