Multicenter Evaluation of \([-2]\text{proprostate-specific antigen} \text{ and the Prostate Health Index for Detecting Prostate Cancer}\

Carsten Stephan, Sébastien Vincendeau, Alain Houilgatte, Henning Cammann, Klaus Jung, and Axel Semjonow

BACKGROUND: Total prostate-specific antigen (tPSA) is flawed for prostate cancer (PCa) detection. \([-2]\text{proprostate-specific antigen} \text{ (p2PSA), a molecular isoform of free PSA (fPSA), shows higher specificity compared with tPSA or percentage of free PSA (\%fPSA). The prostate health index (Phi), a measure based on p2PSA and calculated as p2PSA/fPSA (\%fPSA). The prostate health index (Phi), a measure based on p2PSA and calculated as p2PSA/fPSA (\%fPSA).}

CONCLUSIONS: The results of this multicenter study show that Phi, compared with tPSA or \%fPSA, demonstrated superior clinical performance in detecting PCa at tPSA 1.6–8.0 \(\mu\)g/L (i.e., approximately 2–10 \(\mu\)g/L in traditional calibration) and is better able to detect aggressive PCa.

Received September 3, 2012; accepted November 2, 2012.
Previously published online at DOI: 10.1373/clinchem.2012.195784

Nonstandard abbreviations: PSA, prostate specific antigen; PCa, prostate cancer; tPSA, total PSA; fPSA, free PSA; proPSA, proprostate-specific antigen; p2PSA, \([-2]\text{proprostate-specific antigen}; \text{Phi, prostate health index; AUC, area under the ROC curve; ANN, artificial neural network; TRUS, transrectal ultrasound; DRE, digital rectal examination; DCA, decision curve analysis; } \tau_c \text{, Spearman rank correlation coefficient.}
specificity within a multivariate model (10). Intact PSA, a further subtraction of fPSA, when combined with tPSA and fPSA, has demonstrated a discriminative ability between patients with and without PCa (11) but further studies are lacking. Further PSA complexes and their limitations have been reviewed elsewhere (12).

Recently, several studies demonstrated significant improvement in PCa detection using a PSA isoform [−2]proprostate-specific antigen (proPSA) (p2PSA) and its percentage derivative %p2PSA [p2PSA/ \( tPSA \times 1000 \) \( \times 100 \)], especially within multivariate models (13, 14). A further improvement in PCa diagnosis was seen with the additional inclusion of tPSA to %p2PSA by developing the prostate health index (Phi), calculated as p2PSA/PSA \( \times \sqrt{tPSA} \) (8, 15–17). In screening populations (8, 17) and in referred populations (15, 16), %p2PSA (based on 2 markers) and even more Phi (based on 3 markers) demonstrated better diagnostic performance than tPSA and %fPSA for PCa detection. This was indicated by larger areas under the ROC curve (AUC) and better specificities at high sensitivities, results that indicate a potential reduction of unnecessary biopsies. A relationship between %p2PSA or Phi with biopsy or pathological Gleason score also suggested that these biomarkers may more accurately detect aggressive PCa (8, 15, 16).

In the multicenter study presented in this report, the clinical performance of p2PSA and its derivatives, such as %p2PSA and Phi, was assessed for the detection of PCa at initial and repeat biopsies. The data were obtained from a large cohort of patients with tPSA of 1.6–8 \( \mu \)g/L (calibration against a WHO PSA reference material) corresponding to the former 2–10 \( \mu \)g/L Hybritech calibration. The potential reduction in unnecessary biopsy with %p2PSA and Phi and the preferential detection of aggressive PCa of these biomarkers were evaluated.

Materials and Methods

STUDY POPULATION

At each of the 4 participating institutions, patients referred to the urology department were recruited prospectively and retrospectively between 2003 and 2012 on the basis of having tPSA results between 1.6 and 8.0 \( \mu \)g/L (calibration against a WHO PSA reference material). The numbers of patients in each center were as follows: Münster, n = 544 with 266 PCa; Berlin, n = 393 with 167 PCa; Rennes, n = 227 with 132 PCa; and Paris, n = 198 with 103 PCa. Patients scheduled for initial or repeat prostate biopsy were included in the study. All patients enrolled underwent a transrectal ultrasound (TRUS)-guided needle biopsy (10 or more cores). Several exclusion criteria were defined, including clinical acute/chronic prostatitis or infection of the urinary tract, use of medications that interfere with tPSA serum concentration (e.g., 5α reductase inhibitors, androgen therapy), previous history of PCa, and previous transurethral resection of the prostate for benign prostatic hyperplasia. The final study population included 1362 men; 681 patients (50%) were included for initial biopsy and 280 patients (21%) were scheduled for a repeated biopsy, and for the remaining 401 patients (29%) this information was missing.

STUDY DESIGN

The study was a multicenter, nonrandomized case control trial to evaluate the diagnostic performance of p2PSA, %p2PSA, and Phi in comparison with the current gold standard markers (tPSA, %fPSA) and other parameters. Participants and investigators were blinded to p2PSA results and the personnel involved in testing were blinded to patients’ clinical information. The study was approved by the local hospital ethics committees and reported in accordance with the Standards for the Reporting of Diagnostic Accuracy (18). All patients gave informed consent to participate in the study.

METHODS

Blood was drawn at each participating site and serum samples were prepared and frozen at −20 to −80 °C within 3 h of blood collection according to recommendations for preanalytic tPSA and fPSA (19) and p2PSA as previously published (20). All blood samples were obtained before any manipulations involving the prostate and at least 3 weeks after a digital rectal examination (DRE). Serum samples were analyzed on Access2 (Münster and Berlin) or DxI800 instruments (Beckman Coulter) using WHO standard calibrated Access tPSA and fPSA immunoassays or the automated p2PSA assay. Measurements were performed within 1–2 weeks after blood draw or in series. The analytical performance of the measurements assessed with control materials (Beckman Coulter) showed values within the recommended limits.

Prostate volume was determined using TRUS and calculated with the prostate ellipse formula. TRUS-guided biopsies of 10–22 cores were performed on all included patients according to the standard clinical practice routinely used at each participating site. Pathological analysis and Gleason grading were performed at each institution separately according to the updated Gleason grading from the consensus conference in 2005 (21). Samples from PCa patients without the updated Gleason grading or only WHO grading information were not included in analyses regarding Gleason grade (n = 157).
STATISTICAL ANALYSIS

Statistical analyses were performed with MedCalc version 12.2.1 (MedCalc Software). Differences between 2 groups were assessed with the nonparametric Mann–Whitney U-test. Multiple groups (4 centers) were compared by use of the nonparametric Kruskal–Wallis test. In cases of nominal data (differences in specificity at 90% and 95% sensitivity in ROC analyses) we used the nonparametric McNemar test. AUCs were estimated according to DeLong et al. (22). ROC curves were used to compare specificities at given sensitivities. Two-sided P values <0.05 were considered statistically significant.

Phi, %p2PSA, and p2PSA were added to an artificial neural network (ANN) and to binary logistic regression models with the classic variables age, tPSA, %PSA, prostate volume, and DRE status (14, 23) to evaluate their ability to improve specificity. The models were constructed with the MATLAB Neural Network Toolbox (Mathworks). Each of the ANN models had 3 layers: 1 input layer with 2–8 neurons and 1 hidden layer with 3 or 2 neurons and 1 output neuron, ranging from 0 (low) to 1 (high) PCa risk. To avoid overfitting during training we used the Bayesian regularization algorithm (24). With this method there is no necessity for a special validation data set, and the whole training data set can be used for training. For the respective best models and the single parameters, the calibration plots and decision curve analysis (DCA) (25) were used.

Results

For the whole population of 1362 men (any biopsy), Phi, %p2PSA, p2PSA, age, and tPSA were significantly higher in patients with PCa, whereas prostate volume, fPSA, and %pPSA were significantly higher in patients without PCa (Table 1). Only the p2PSA-based parameters %p2PSA and Phi maintained the highest significance level (P < 0.0001) when the patient data were subdivided into those with negative DRE or with initial or repeat biopsy.

ROC ANALYSIS

The 4 centers showed similar AUCs for each parameter, with AUCs ranging from 0.55 to 0.56 for tPSA, 0.59 to 0.64 for %PSA, 0.70 to 0.74 for %p2PSA, and 0.72 to 0.74 for Phi. As evident in Table 2, in the whole population of 1362 men the overall performance of Phi was superior for all parameters and significantly better than that of %p2PSA, the nearest competing measure (AUC 0.74 vs 0.725; P = 0.022). Both Phi and %p2PSA significantly outperformed (P always <0.0001) all other laboratory (p2PSA, %PSA, tPSA, fPSA) and clinical parameters (age, prostate volume, DRE status) as judged by AUC comparison (See Fig. 1 in the Data.
Table 2. Univariate and multivariate analyses for parameters and models.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Any biopsy</th>
<th></th>
<th>Initial biopsy</th>
<th></th>
<th>Repeat biopsy</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Univariate AUC (95% CI); P</td>
<td>Multivariate AUC (95% CI)</td>
<td>Univariate AUC (95% CI); P</td>
<td>Multivariate AUC (95% CI)</td>
<td>Univariate AUC (95% CI); P</td>
<td>Multivariate AUC (95% CI)</td>
</tr>
<tr>
<td>Age</td>
<td>0.56 (0.53–0.59); &lt;0.0001</td>
<td>0.60 (0.56–0.64); &lt;0.0001</td>
<td>0.62 (0.55–0.68); 0.00015</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prostate volume, cm³</td>
<td>0.62 (0.59–0.64); &lt;0.0001</td>
<td>0.62 (0.58–0.67); &lt;0.0001</td>
<td>0.60 (0.54–0.67); 0.0019</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRE</td>
<td>0.58 (0.56–0.60); &lt;0.0001</td>
<td>0.58 (0.55–0.62); &lt;0.0001</td>
<td>0.56 (0.52–0.61); 0.0008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tPSA, μg/L</td>
<td>0.56 (0.53–0.59); &lt;0.0001</td>
<td>0.56 (0.51–0.61); &lt;0.0001</td>
<td>0.54 (0.47–0.61); 0.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%fPSA</td>
<td>0.61 (0.59–0.64); &lt;0.0001</td>
<td>0.60 (0.56–0.64); &lt;0.0001</td>
<td>0.59 (0.52–0.65); 0.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%p2PSA</td>
<td>0.72 (0.70–0.75); &lt;0.0001</td>
<td>0.72 (0.69–0.76); &lt;0.0001</td>
<td>0.74 (0.68–0.80); &lt;0.0001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phi</td>
<td>0.74 (0.71–0.76); &lt;0.0001</td>
<td>0.73 (0.69–0.77); &lt;0.0001</td>
<td>0.74 (0.68–0.80); &lt;0.0001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base model</td>
<td>0.69 (0.66–0.72)</td>
<td>0.69 (0.65–0.73)</td>
<td>0.74 (0.67–0.80)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base model + %p2PSA</td>
<td>0.75 (0.72–0.77)</td>
<td>0.73 (0.69–0.77)</td>
<td>0.79 (0.74–0.84)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base model + Phi</td>
<td>0.75 (0.72–0.78)</td>
<td>0.73 (0.69–0.77)</td>
<td>0.80 (0.74–0.85)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain in predictive accuracy for %p2PSA</td>
<td>0.06</td>
<td>0.04</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain in predictive accuracy for Phi</td>
<td>0.06</td>
<td>0.04</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Base model includes: age, prostate volume, DRE, tPSA, %fPSA.
Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol59/issue1. As also shown in online Supplemental Fig. 1, an ANN including %p2PSA or Phi did not significantly enhance the AUC over Phi alone (0.75 vs 0.74; \( P = 0.41 \)). However, the gain in predictive accuracy of Phi compared with a base model (based on tPSA, %fPSA, age, prostate volume, and DRE status) was visible in all subcohorts with observed increases in AUC of 0.04 – 0.06 (Table 2). The specificities at 90% and 95% sensitivity also demonstrated an advantage for Phi (35.4% and 15%) and %p2PSA (33.6% and 13.3%) compared with p2PSA (20.8% and 10.8%), %fPSA (15% and 7.5%), and tPSA (16.1% and 8.5%).

The calibration curves for both models with (model 1) and without Phi (model 2) were quite similar and excellent (see online Supplemental Fig. 2).

**ASSOCIATION OF PHI WITH POSITIVE BIOPSY RATE AND AGGRESSIVE PCA**

Increasing rates of a PCa-positive biopsy sample with increasing Phi values were seen in all (Fig. 1A), initial (Fig. 1B), and repeat biopsy (Fig. 1C) patients. Differences between tumors with Gleason scores \( <7 \) vs scores \( \geq 7 \) were seen for all laboratory values (Table 3). Phi provided the best discrimination between aggressive and nonaggressive PCAs. The number of tumors with Gleason scores \( \geq 7 \) increased with increasing Phi values, reaching approximately 50% of all detected cancers when Phi was >50 (Fig. 2). There was a positive relationship between Phi and biopsy Gleason score with a Spearman rank correlation coefficient \( (r_s) \) of 0.154 (CI, 0.06 – 0.24; \( P = 0.0005 \)). Classification of all 1362 patients into quartiles based on Phi (5–31.7; 31.8 – 43.1; 43.2– 60.6; 60.7–260) revealed an increasing likelihood of harboring PCAs, with PCa rates of 21.5% in the lowest Phi quartile, 40.8% in the 2nd quartile, 57.1% in the 3rd quartile, and 76.8% of patients in the highest Phi quartile. When we analyzed the 511 PCa patients with an available Gleason score (Phi median, 56), the tumors with Gleason scores \( \geq 7 \) accounted for 39.5% of all cancers when Phi was below the median, whereas above the median the percentage of aggressive PCa cases (Gleason score \( \geq 7 \)) increased to 49.8%. It was difficult to obtain a practicable Phi cutoff for detection of aggressive PCa. With the use of data only from tumors with Gleason scores \( <7 \) (n = 283) and \( \geq 7 \) (n = 51) the Phi value at 90% sensitivity was 40 and at the highest efficacy Phi was 49.5. However, these cutoffs did not allow significant improvement in the differentiation between the Gleason 4 + 3 = 7 and 3 + 4 = 7 tumors. With 40 as a cutoff, 84% of the more aggressive Gleason 4 + 3 tumors were above this value but only 18% of the Gleason 3 + 4 tumors were below this value. With the use of the 49.5 cutoff for Phi, 76% of Gleason 4 + 3 tumors were above this value and 47% of the Gleason 3 + 4 tumors were below this value.

**DECISION CURVE ANALYSIS**

The DCA was performed for the models with (model 1) and without p2PSA-based parameters (model 2, Fig. 3A) as well as for the parameters tPSA, %PSA, %p2PSA, and Phi separately (Fig. 3B). There appeared to be a higher net benefit when we used model 1 with Phi compared to the base model (model 2) based on tPSA, %PSA, age, prostate volume, and DRE status (Fig. 3A). Phi considered separately also provided the highest net benefit compared with %p2PSA and especially with %fPSA and tPSA (Fig. 3B).
Discussion

The introduction of Phi as a p2PSA-based formula clearly increased the specificity of tPSA and %fPSA for the detection of PCa in cohorts from large multicenter trials (8, 15, 26) and prospective trials (16, 26) and cohorts with only initial biopsies (16) and only repeat biopsies (27). Earlier studies using only %p2PSA but not Phi had indicated that proPSA enhances PCa detection in comparison with tPSA and %fPSA (13, 14, 17). Such studies confirmed the first predictions that p2PSA seemed to be the most promising proPSA subform for PCa detection and especially for detection of aggressive PCa (28–31).

The current study assessed the largest population cohort so far, consisting of 1362 patients in the 1.6–8 µg/L tPSA range. The study findings confirmed %p2PSA and Phi as the strongest discriminating parameters between patients with and without PCa (Table 1). The AUC analysis across all 1362 patients showed similar results for Phi (0.74) as were observed in 2 earlier studies (15, 16) that used the same tPSA range of 2–10 µg/L (Hybritech calibration), which translates into the WHO standard calibrated tPSA range of 1.6–8 µg/L. Although Catalona et al. (15) found an AUC of 0.65 for %fPSA vs 0.7 for Phi in 892 patients, this difference was much larger in the other prospective study involving 268 patients with an AUC of 0.58 for %fPSA and 0.76 for Phi (16). However, both studies included only DRE-negative patients. When we compared these data to those for our 1059 patients with negative DRE, the AUCs for tPSA (0.56), %fPSA (0.62), p2PSA (0.62), %p2PSA (0.73), and Phi (0.74) were almost identical with those obtained for the whole population. This finding might suggest that the future use of Phi could be recommended without considering the DRE status. ROC comparisons at 95% sensitivity revealed specificities of 15% for Phi and 7.5% for %fPSA in this study, nearly identical to the results of Catalona et al. (15), who found specificities of 16% for Phi and 8.4% for %fPSA. By using absolute Phi thresh-

<table>
<thead>
<tr>
<th>Group/parameter</th>
<th>Gleason &lt;7 PCa (n = 283)</th>
<th>Gleason ≥7 PCa (n = 228)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years</td>
<td>63 (62–65)</td>
<td>66 (64–67)</td>
<td>0.02</td>
</tr>
<tr>
<td>Volume, cm³</td>
<td>38 (35–40)</td>
<td>33 (31–37)</td>
<td>0.031</td>
</tr>
<tr>
<td>tPSA, µg/L</td>
<td>4.83 (4.57–5.15)</td>
<td>5.33 (5.03–5.77)</td>
<td>0.003</td>
</tr>
<tr>
<td>%fPSA, %</td>
<td>13.3 (12.5–14.3)</td>
<td>11.9 (11–12.8)</td>
<td>0.014</td>
</tr>
<tr>
<td>p2PSA, ng/L</td>
<td>14.6 (13.5–15.4)</td>
<td>15 (13.7–17.3)</td>
<td>0.28</td>
</tr>
<tr>
<td>%p2PSA</td>
<td>2.34 (2.21–2.49)</td>
<td>2.68 (2.51–2.85)</td>
<td>0.011</td>
</tr>
<tr>
<td>Phi</td>
<td>53.1 (47.5–56.6)</td>
<td>59.7 (55.3–62.1)</td>
<td>0.002</td>
</tr>
</tbody>
</table>

*a All values are given as median (95% CI).

Fig. 2. Proportion of all prostate cancer detected with Gleason score ≥7 in relation to Phi intervals.

Fig. 3. Decision curve analysis (DCA).
(A), DCA for the models with Phi (model 1) and without Phi (model 2); (B), DCA for tPSA, %fPSA, %p2PSA, and Phi separately.
The risk of having PCa with a Gleason score increased PCa risk in patients with a Phi value above 5.5 in a large multicenter study there was an almost 5-fold increase in PCa risk in those patients in the highest Phi quartile. More importantly, the likelihood of Gleason ≥7 tumors increases to more than three-fourths of all patients in the lowest Phi quartile harbor PCa, the percentage of tumors with Gleason ≥7 tumors increased to almost 50%. In another study, even stronger correlations were observed with the Gleason score for Phi (r = 0.39) and %p2PSA (r = 0.3) (16), but both measures failed to predict Gleason ≥7 tumors. However, the same group of investigators, in a study of 350 men who underwent radical prostatectomy, found that %p2PSA and Phi both could predict pathological Gleason ≥7 PCa, pT3 tumor stage, and a tumor volume of >0.5 mL, and that p2PSA-based parameters were the most accurate predictors for final pathology results (33). These important data suggest a role for p2PSA-based markers not only for PCA diagnosis but also for prediction of aggressiveness and possibly also for prognosis. A further study in a total of 756 men found AUCs for Phi of 0.75 and 0.71 in 2 populations and increased specificities at 90% and 95% sensitivity compared with tPSA and %fPSA (8). In that study the p2PSA-based model also missed the fewest of the Gleason ≥7 tumors at 90% and 95% sensitivity (8).

In addition to investigations of the potential role of p2PSA and Phi for PCa diagnosis (8, 15, 16, 26, 34) and staging (33), other potential uses have also been explored. Longitudinal changes of p2PSA with increasing age suggest a role for p2PSA and its derivatives as useful predictors of PCa development (35). Furthermore, a baseline p2PSA value in the upper quartile results in an almost 8-fold risk of developing a PCa (36). For active surveillance, p2PSA-based parameters including Phi appear to provide improved prediction of biopsy reclassification during follow-up (37). Lazzeri et al. (27) demonstrated in 222 patients with 1–2 repeat biopsies that with the use of Phi more than 50% of biopsies could have been avoided and that %p2PSA and Phi were the most accurate PCA predictors. These findings are in accord with our results in those 280 patients with repeat biopsies, for whom %p2PSA and Phi reached the highest significance levels (Table 1) and showed the largest AUCs with 0.74 (Table 2).

A recent comparison of Phi with another promising PCA biomarker, PCA3, revealed a larger AUC for Phi compared with PCA3, but the difference was not statistically significant (38). In an analytical multicenter study by Sokoll et al. (39), the p2PSA assay showed a clinically acceptable analytical performance with excellent precision and reproducibility and had negligible interference from other PSA isoforms.

It is also important to note that our present study is one of the first to use the WHO standard calibrated tPSA and fPSA assays in conjunction with p2PSA. Other studies on smaller populations still used the traditional Hybritech calibration (15, 16) or did not provide this information (27). We believe that researchers should pay attention to this aspect because median values for Phi tend to be higher with the use of the WHO calibration. Whereas Guazzoni et al. (16) found median values of 44.3 and 33.1 in PCa and non-PCa patients, Jansen et al. reported medians of 36.7 and 26.1. The present study has higher values, with 53.8 and 35.5 for PCa and non-PCa patients.

In summary, in the largest cohort of patients reported so far who had tPSA in the WHO standard calibrated range of 1.6–8 μg/L, we found %p2PSA and Phi were the best parameters to predict PCa. These two parameters provided the largest AUCs of 0.72–0.74 in all investigated cohorts, with a small overall advantage for Phi. With increasing Phi values, the percentage of PCA-positive biopsy results increased. Phi was significantly higher in patients with Gleason score ≥7 tumors compared with Gleason score <7 tumors (P = 0.0018).
The proportion of aggressive PCa (Gleason score ≥7) increased with the Phi score.

Author Contributions: All authors confirmed that they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors’ Disclosures or Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

Roles of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

References


