Clinical Chemistry 57:7
995–1004 (2011)

Between-Method Differences in Prostate-Specific Antigen Assays Affect Prostate Cancer Risk Prediction by Nomograms

Carsten Stephan,[1,2*†] Kerstin Siemßen,[1†] Henning Cammann,[3] Frank Friedersdorff,[1] Serdar Deger,[1] Mark Schrader,[1] Kurt Miller,[1] Michael Lein,[1,2,4] Klaus Jung,[1,2] and Hellmuth-Alexander Meyer[1]

BACKGROUND: To date, no published nomogram for prostate cancer (PCa) risk prediction has considered the between-method differences associated with estimating concentrations of prostate-specific antigen (PSA).

METHODS: Total PSA (tPSA) and free PSA were measured in 780 biopsy-referred men with 5 different assays. These data, together with other clinical parameters, were applied to 5 published nomograms that are used for PCa detection. Discrimination and calibration criteria were used to characterize the accuracy of the nomogram models under these conditions.

RESULTS: PCa was found in 455 men (58.3%), and 325 men had no evidence of malignancy. Median tPSA concentrations ranged from 5.5 μg/L to 7.04 μg/L, whereas the median percentage of free PSA ranged from 10.6% to 16.4%. Both the calibration and discrimination of the nomograms varied significantly across different types of PSA assays. Median PCa probabilities, which indicate PCa risk, ranged from 0.59 to 0.76 when different PSA assays were used within the same nomogram. On the other hand, various nomograms produced different PCa probabilities when the same PSA assay was used. Although the ROC curves had comparable areas under the ROC curve, considerable differences were observed among the 5 assays when the sensitivities and specificities at various PCa probability cutoffs were analyzed.

CONCLUSIONS: The accuracy of the PCa probabilities predicted according to different nomograms is limited by the lack of agreement between the different PSA assays. This difference between methods may lead to unacceptable variation in PCa risk prediction. A more cautious application of nomograms is recommended.

Prostate cancer (PCa)[5] detection relies on the measurement of prostate-specific antigen (PSA) concentrations *(1, 2)*. An increased PSA value is directly associated with a higher probability of having PCa *(3–5)*, but benign prostate hyperplasia or prostatitis can also cause increases in serum PSA *(6)*. Of all the molecular variables involving the total PSA (tPSA) concentration, only the ratio of free PSA (fPSA) to tPSA [i.e., the percentage of free PSA (%fPSA)] is clinically relevant and capable of avoiding unnecessary biopsies *(7)*. Yet, the low specificity of PSA and %fPSA remains problematic.

Multivariate models, such as artificial neural networks or logistic regression–based nomograms, improve PCa risk prediction by combining tPSA, %fPSA, age, digital rectal examination (DRE) results, and/or prostate volume *(8, 9)*. The frequent use of nomograms as PCa-classification models and for recurrence prediction has recently been reviewed *(8, 10)*. The inclusion of %fPSA in nomograms has improved the accuracy of PCa diagnosis *(11, 12)*. The nomograms show an improvement in specificity *(13)* compared with the use of %fPSA alone, but these prediction models were developed with data from different populations, used various tPSA intervals (e.g., 0–20, 4–10, or 0–50 μg/L) and applied different PSA assays. To the best of our knowledge, no one has analyzed whether the use of different PSA and fPSA assays has an effect on nomogram-based PCa prediction. Clinicians use some nomograms that are available online in patient coun-

[1] Department of Urology, Charité – Universitätsmedizin Berlin, Berlin, Germany; [2] Berlin Institute for Urologic Research, Berlin, Germany; [3] Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Berlin, Germany; [4] Department of Urology, University Teaching Hospital, Offenbach, Germany.
* Address correspondence to this author at: Department of Urology, Charité – Universitätsmedizin Berlin, CCM, Charitéplatz 1, D-10117 Berlin, Germany. Fax +49-30-450-515904; e-mail carsten.stephan@charite.de.
† These authors contributed equally to the study.

Received June 3, 2010; accepted March 21, 2011.
Previously published online at DOI: 10.1373/clinchem.2010.151472

[5] Nonstandard abbreviations: PCa, prostate cancer; PSA, prostate-specific antigen; tPSA, total PSA; fPSA, free PSA; %fPSA, percentage of free PSA; DRE, digital rectal examination; NEM, no evidence of malignancy; cPSA, complexed PSA; Siemens-c, Siemens test using cPSA; Siemens-f, Siemens test using fPSA; AUC, area under the ROC curve; ICC, intraclass correlation coefficient.

seling, but without considering the inadequate comparability of PSA results obtained with the various assays and the effect on the probability calculated from the nomograms. Despite the introduction of tPSA and fPSA assays that are calibrated against WHO PSA reference materials, PSA values still cannot be used interchangeably *(14, 15)*. Besides tPSA, %fPSA values also differ between assays *(15)*. External nomogram validation studies have not considered *(11, 16)* or have considered only partially *(17, 18)* the influence of PSA and %fPSA assays.

The aim of this study was to evaluate the effect of assay-dependent variation in PSA and %fPSA values on nomogram-based PCa prediction. To this end, we measured the PSA values of 780 patients simultaneously with 5 different PSA assays and used the results from each of the assays to calculate the probability of PCa from 5 different nomograms. The 2 nomogram validation criteria—discrimination and calibration—were used to assess the effect of PSA and %fPSA variation on the predictive results obtained with the nomograms. This study also provided an external validation of the nomogram models.

Materials and Methods

STUDY GROUPS AND SAMPLES

The study population consisted of 455 PCa patients and 325 men with no evidence of malignancy (NEM) who had a tPSA value within the interval of 0.54–23.4 μg/L as measured with the AxSYM assay (Abbott Diagnostics). The study population was previously described, in detail, in another report *(19)*. All men had been referred to the Department of Urology or the affiliated outpatient department at the Charité – Universitätsmedizin Berlin, and samples of archival sera collected between 2001 and 2004 were investigated retrospectively. Disease status was confirmed histologically for all patients by prostate biopsy (8–10 cores). A total of 364 patients treated by radical prostatectomy had pathologic stages pT1/pT2 (n = 263) or pT3/pT4 (n = 101) and Gleason scores ≤6 (n = 148) or ≥7 (n = 215; results not available for 1 patient). The clinical stages for the remaining 91 patients were as follows: T1/T2, n = 58; T3, n = 33. The biopsy Gleason scores were ≤6 for 49 patients, ≥7 for 34 patients, and unavailable for 8 patients.

Blood samples were taken before any procedures involving the prostate and at least 3–4 weeks after prostate manipulation. All samples were aliquoted and stored at −80 °C. Two aliquots of unthawed samples were analyzed in 2004 for the parallel measurement of tPSA and fPSA by 5 different assays, as previously described *(15, 19)*. Prostate volume was measured by transrectal ultrasound examination and using the prolate ellipse formula. All patients underwent DRE.

PSA ASSAYS AND MEASUREMENTS

tPSA and fPSA [or complexed PSA (cPSA)] concentrations were measured with the following analyzers: AxSYM (Abbott Diagnostics), ADVIA Centaur [Siemens Healthcare Diagnostics; this assay measures cPSA instead of fPSA (Siemens-c)], Access (Beckman Coulter), Immulite 2000 [Siemens Healthcare Diagnostics; measures fPSA (Siemens-f)], and Elecsys 2010 (Roche Diagnostics). The assays that used the AxSYM, ADVIA, and Elecsys analyzers were calibrated against the WHO PSA standards (96/668 and 96/670), whereas the Access and Immulite assays used their own calibrators. The differences between these assays have previously been described in detail *(15, 19)*. %fPSA values were determined from measurements obtained with these 5 assays. The PSA and %fPSA values used in a previous study *(19)* were applied to nomogram predictions.

APPLICATION OF NOMOGRAMS

The data from all patients with complete sets of values for the 5 PSA assays, age, prostate volume, and DRE status were then used with the 5 different nomograms. Table 1 lists the characteristics of the 5 nomograms. Nomogram I is based on age, DRE, and tPSA. Nomogram II uses these variables but also includes %fPSA *(11)*. Nomogram III, which is available at http://www.nomogram.org/, was constructed by combining age, DRE, tPSA, %fPSA, and sampling density *(16)*. The other 2 nomograms include age, DRE, tPSA, and fPSA (nomogram IV) and the additional factors of transrectal ultrasound and prostate volume (nomogram V) *(17)*.

STATISTICAL ANALYSIS

Statistical calculations were performed with SPSS 17.0 for Windows (IBM SPSS). Specifically, the Friedman test was used to detect significant differences in multiple samples. The Wilcoxon test was used for pairwise comparison of groups, assays, or nomogram outputs. *P* values <0.05 were considered statistically significant. We used the Bonferroni correction to correct the results for pairwise comparisons involving multiple tests.

We used ROC curve analyses to evaluate the discrimination capability of the nomograms, and we used the test of Hanley and McNeil to compare areas under the ROC curve (AUCs). MedCalc software (version 10.4.8.0; MedCalc Software) was used to compare sensitivity and specificity data for the 5 nomograms at various cutoffs.

The nomograms were calibrated according to Harrell et al. *(20)* by using calibration plots in a diagram. The results were used as a performance measure

Table 1. Characteristics of the nomograms used and predicted probabilities of PCa.

Characteristics	Nomogram I	II	III	IV	V
Input variables	Age, DRE, PSA	Age, DRE, PSA, %fPSA	Age, DRE, PSA, %fPSA, sampling density[a]	Age, DRE, PSA, %fPSA	Age, DRE, PSA, %fPSA, prostate volume, TRUS[b]
Patients, n[c]	4193	1762	1162	1509	1509
PSA range, μg/L	0–50	0–50	Not given	0–20	0–20
PSA assay	Not given	Not given	Not given	Tandem R[d] and AxSYM	Tandem R[d] and AxSYM
Reference	Karakiewicz et al. *(11)*	Karakiewicz et al. *(11)*	Chun et al. *(16)*	Kawakami et al. *(17)*	Kawakami et al. *(17)*
Predicted PCa probabilities of the study cohort according to PSA assay[e]					
Abbott	0.25 (0.11–0.72)[f]	0.42 (0–0.92)[g]	0.59 (0.02–0.95)[g]	0.59 (0.02–0.97)[g]	0.48 (0.01–1.00)[g]
Siemens-c	0.24 (0.11–0.70)[f]	0.45 (0–0.93)[g]	0.59 (0.02–0.96)[g]	0.61 (0.02–1.00)[g]	0.50 (0.01–0.99)[g]
Beckman Coulter	0.26 (0.11–0.73)	0.49 (0–0.93)	0.65 (0.02–0.97)	0.68 (0.01–0.98)	0.56 (0.01–1.00)
Siemens-f	0.26 (0.11–0.74)	0.54 (0.01–0.94)	0.70 (0.06–0.97)	0.76 (0.03–1.00)	0.62 (0.03–1.00)
Roche	0.26 (0.11–0.77)	0.50 (0.02–0.94)	0.66 (0.07–0.98)	0.71 (0.03–0.98)	0.57 (0.02–1.00)

[a] Sampling density is the ratio between prostate volume and biopsy cores.
[b] TRUS, transrectal ultrasound.
[c] Number of patients used for constructing the nomogram.
[d] Tandem R is the radioimmunoassay of Hybritech, Inc. (now Beckman Coulter).
[e] Data are for the respective individual PCa probabilities of all PCa patients and are expressed as the median (range). The pairwise Wilcoxon test yielded *P* values <0.001 (statistically significant) for all assay comparisons, except as noted for the Abbott and Siemens-c PSA assays.
[f] $P = 0.76$.
[g] $P < 0.05$.

of the agreement between the predicted probabilities and observed outcomes. The points of these calibration plots are constructed with the predicted probability of a positive biopsy result on the *x* axis and the observed frequency of PCa-positive biopsies on the *y* axis. For this purpose, the 780 patients were subdivided into 20 groups (i.e., each group being 5% of the entire study group) of 39 men, according to the order of their respective predicted PCa probabilities. The mean observed outcomes and predicted probabilities were calculated for each group. A cubic smoothing spline was computed to suppress random fluctuations in the graphical representation and to expose the relationship between the predicted probabilities and observed outcomes. Figures were developed with MATLAB (MathWorks). To determine the consistency between these pairs, we computed the intraclass correlation coefficient (ICC), where a value of 1 is ideal. The ICC is a measure of consistency, which is obtained by multiplying the Pearson correlation coefficient by a correction factor that is based on the means and SDs of the observed outcomes and predicted probabilities [see Lin et al. *(21)*].

Results

NOMOGRAM-BASED PREDICTIONS OF PCa ACCORDING TO PSA ASSAYS

Table 2 summarizes the clinical data and PSA values measured with the 5 assays for each of the study groups (58.3% PCa, 41.7% NEM). The median tPSA values obtained for the 5 assays were always significantly different except for the 2 lowest tPSA values for the Abbott and Siemens-c assays (Table 2). The highest values were observed with the Siemens-f assay. The largest median differences in %fPSA results were detected between the Siemens-c and Siemens-f assays.

Table 1 summarizes the predicted PCa probabilities, which are the median values of the respective individual PCa probabilities of all patients. For every nomogram, the pairwise comparison of the predicted probabilities was significantly ($P < 0.0001$) dependent on the PSA assay, except for the Abbott and Siemens-c assays. Nomogram IV had the most diverse results, with median PCa probabilities of 0.59 and 0.76 for the Abbott assay and the Siemens-f assay, respectively. In addition, we observed remarkable differences between

Table 2. Characteristics of the study groups.

Characteristic	All patients	PCa patients	NEM patients
Patients, n	780	455	325
Age, years[a]	64 (40–85)	63 (43–79)	66 (40–85)
Prostate volume,[a] cm^3	38 (10–180)	34 (10–110)	46 (13–180)
Positive DRE, n	304 (39%)	276 (61%)	28 (8.6%)
tPSA, $\mu g/L$[a,b,c]			
Abbott	5.6 (0.54–23.4)	6.63 (0.67–23.4)	3.98 (0.54–22.9)
Siemens-c	5.5 (0.69–24.5)	6.54 (0.70–20.7)	4.01 (0.69–24.5)
Beckman Coulter	6.54 (0.49–27.0)	7.68 (0.86–24.0)	4.61 (0.49–27.0)
Siemens-f	7.04 (0.76–24.9)	8.48 (0.96–24.3)	5.15 (0.76–24.9)
Roche	6.53 (0.56–29.5)	7.69 (1.18–25.7)	4.71 (0.56–29.5)
%fPSA, %[a,b,c]			
Abbott	16.2 (3.7–73.8)	12.7 (3.9–71.7)	22.5 (3.7–73.8)
Siemens-c[d]	16.4 (0.5–72.4)	12.9 (0.5–63.1)	22.2 (2.2–72.4)
Beckman Coulter	13.1 (2.5–69.4)	10.4 (3.1–49.9)	18.4 (2.5–69.4)
Siemens-f	10.6 (0.7–54.3)	8.2 (0.7–45.3)	14.6 (2.3–54.3)
Roche	12.5 (2.1–51.8)	10.3 (3.1–36.9)	16.9 (2.1–51.8)

[a] Data are presented as the median (range).
[b] Within each study group, the Friedman test showed significant ($P < 0.05$) differences between the assays for tPSA and %fPSA. Multiple comparisons of the particular assays showed significant differences (P values at least <0.05) except for the difference in tPSA results between the Abbott and Siemens-c assays. These results of multiple comparisons were confirmed by pairwise comparisons of the assays (Wilcoxon test; P values <0.0001 for all comparisons except for differences in tPSA between the Abbott and Siemens-c assays (all, $P = 0.86$; PCa, $P = 0.16$; NEM, $P = 0.14$).
[c] PCa and NEM patients had significantly different median tPSA and %fPSA values ($P < 0.0001$, Wilcoxon test) for each assay.
[d] %fPSA for the Siemens-c assay was calculated as: 100 − (cPSA/tPSA × 100).

the predicted probabilities obtained with the various nomograms, which ranged from 0.24 for nomogram I to 0.76 for nomogram IV, despite the fact that nomograms I and II and nomograms IV and V were established by the same group. Three patients, with tPSA values of approximately 2, 7, and 16 $\mu g/L$, exemplify the fact that different PCa probabilities are obtained when the results of different PSA assays are used, irrespective of the other variables (Table 3). Whereas the difference between algorithms in the probability of PCa between the lowest and highest tPSA increases with higher tPSA values, the difference in probabilities between the lowest and highest %fPSA values appears to be large for all 3 patients.

DISCRIMINATIVE ABILITY OF THE NOMOGRAMS ACCORDING TO PSA ASSAYS

The AUC as an overall discriminative criterion. The ability of a nomogram to distinguish between PCa and NEM patients is termed discrimination, which is generally assessed by AUC analysis. A comparison of AUCs for the nomograms with data from the same PSA assay (Table 4) revealed significantly ($P < 0.001$) lower AUCs for nomogram I (0.79–0.80) compared with the other nomograms (0.82–0.87), with the exception of the comparison of nomograms I and IV for the Siemens-c assay ($P = 0.033$). No differences between the other nomograms were observed. The reason for the lower AUCs for nomogram I may be because nomogram I is the only one that does not include %fPSA.

Comparing the AUCs for the 5 PSA assays within the same nomogram showed clinically irrelevant differences (AUCs ≤0.03).

Assessment of prediction ability according to various cutoffs. Although AUC values represent the overall measure of the discriminative ability of a given model, it is most important to analyze the sensitivity and specificity values from ROC curve analyses from the nomograms over a certain range of outputs. Therefore, we applied data for different PSA assays to the nomograms and compared the sensitivity and specificity curves (Fig. 1) as a function of sensitivity or specificity on the *y* axis and the respective cutoff probability used for the nomogram on the *x* axis. The curves are obviously different. For example, the specificities obtained with the Siemens-f and Abbott assays for nomogram IV (Fig. 1D) at a chosen nomogram probability were 43% and 73%, respectively, whereas the sensitivity varied from

Table 3. Predicted PCa probabilities according to the nomogram and the PSA assay for 3 exemplary patients with different PSA concentrations.

	PCa probability for nomogram				
Assay (PSA, μg/L; %fPSA)	I	II	III	IV	V
Patient A: PSA ≅ 2 μg/L; NEM patient; age, 66 years; prostate volume, 45 cm^3; nonsuspicious DRE result[a]					
Abbott (2.36; 36.4%)	0.18	0.08	0.20	0.16	0.11
Siemens-c[b] (2.71; 33.6%)	0.18	0.10	0.23	0.18	0.13
Beckman Coulter (3.39; 23.3%)	0.19	0.21	0.38	0.32	0.28
Siemens-f (3.4; 19.7%)	0.19	0.265	0.43	0.38	0.23
Roche (3.53; 22.95%)	0.19	0.21	0.38	0.32	0.22
Patient B: PSA ≅ 7 μg/L; PCa patient; age, 68 years; prostate volume, 72 cm^3; nonsuspicious DRE result[c]					
Abbott (5.98; 18.7%)	0.22	0.31	0.4	0.49	0.19
Siemens-c[b] (6.35; 17.2%)	0.22	0.34	0.43	0.54	0.22
Beckman Coulter (7.25; 14.2%)	0.23	0.41	0.49	0.63	0.27
Siemens-f (7.88; 10%)	0.24	0.5	0.57	0.76	0.35
Roche (7.11; 13.5%)	0.23	0.42	0.5	0.64	0.28
Patient C: PSA ≅ 16 μg/L; PCa patient; age, 54 years; prostate volume, 41 cm^3; suspicious DRE result[d]					
Abbott (14.2; 19.8%)	0.54	0.69	0.74	0.63	0.61
Siemens-c[b] (13.7; 24.4%)	0.53	0.6	0.67	0.54	0.55
Beckman Coulter (16.0; 13%)	0.56	0.81	0.83	0.78	0.72
Siemens-f (17.5; 7.3%)	0.58	0.88	0.89	0.91	0.82
Roche (16; 14.2%)	0.56	0.79	0.82	0.76	0.7

[a] NEM patient with assay-specific PSA measurements and calculated %fPSA values.
[b] %fPSA for the Siemens-c assay was calculated as: 100 − (cPSA/tPSA × 100).
[c] PCa patient with assay-specific PSA measurements and calculated %fPSA values.
[d] PCa patient with assay-specific PSA measurements and calculated %fPSA values.

81% (Siemens-c assay) to 95% (Siemens-f assay). On the other hand, at the clinically important sensitivity cutoff of 95% (see Table 1 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol57/issue7), the specificity shows large variation within 1 nomogram when the 5 assays are compared. These data (Fig. 1; see Table 1 in the online Data Supplement) demonstrate that

Table 4. ROC curve analysis: comparison of resulting AUCs according to the nomogram and the PSA assay.[a]

	Resulting AUC for nomogram[b]				
Assay	I	II	III	IV	V
Abbott	0.79 (0.016)	0.87 (0.012)	0.86 (0.013)	0.84 (0.014)	0.85 (0.013)
Siemens-c	0.79 (0.016)	0.86 (0.013)	0.85 (0.013)	0.82 (0.015)[c]	0.85 (0.014)
Beckman Coulter	0.79 (0.016)	0.87 (0.012)	0.86 (0.013)	0.85 (0.014)	0.85 (0.014)
Siemens-f	0.80 (0.016)	0.87 (0.013)	0.86 (0.013)	0.85 (0.013)	0.85 (0.013)
Roche	0.79 (0.016)	0.86 (0.013)	0.86 (0.013)	0.84 (0.014)	0.84 (0.014)

[a] The between-nomogram AUC comparisons (i.e., within rows) indicated always significantly lower ($P < 0.001$) AUCs for nomogram I, compared with all of the other nomograms, except for the comparison between nomograms I and IV for the Siemens-c assay ($P = 0.033$).
[b] Data are presented as the AUC (SE).
[c] The between-assay AUC comparisons (i.e., within columns) shows significantly lower values ($P < 0.05$) only for nomogram IV for the Siemens-c PSA assay, compared with the AUCs obtained with the Abbott, Beckman Coulter, and Siemens-f assays (all significance levels are after the Bonferroni correction).

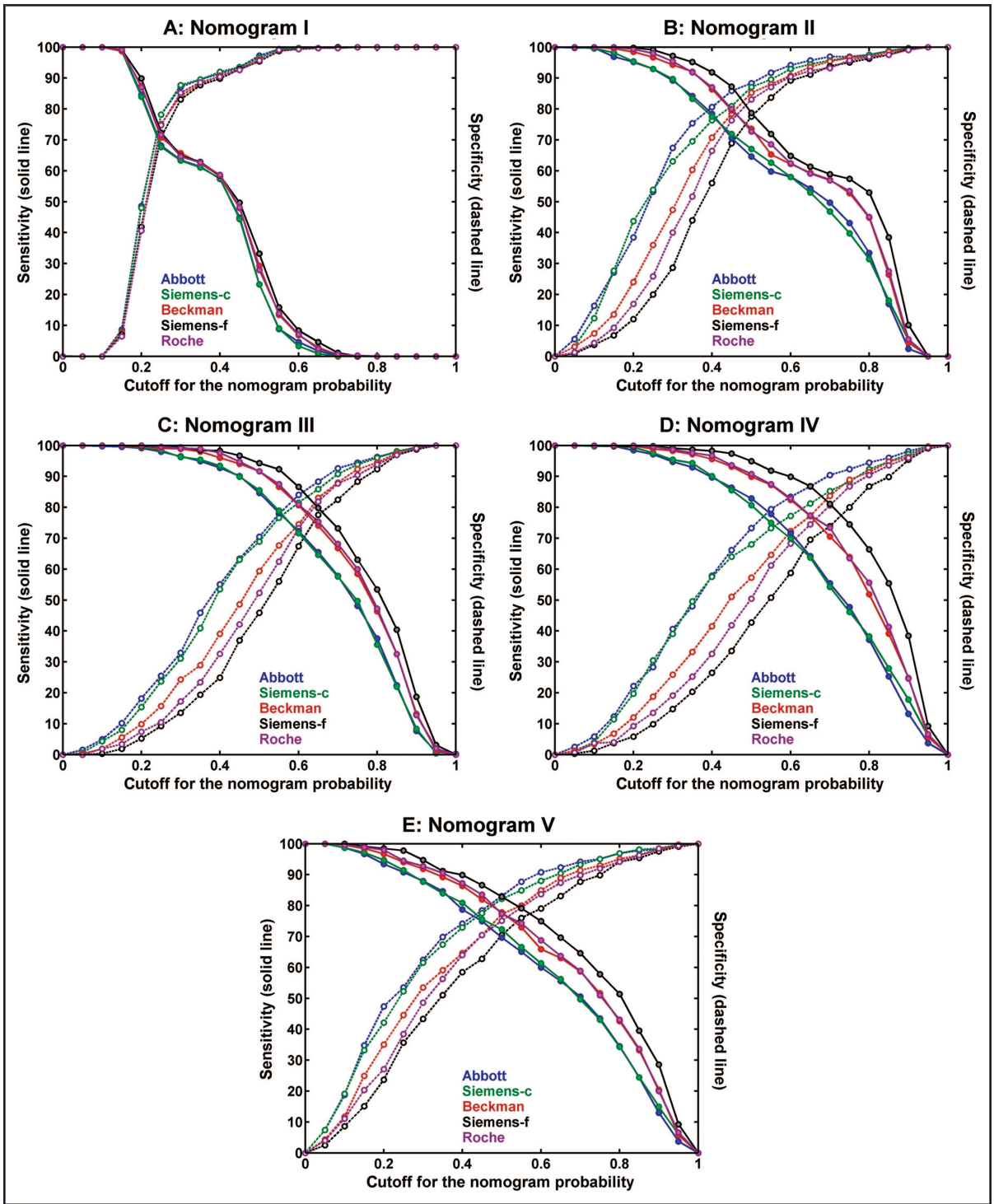

Fig. 1. (A–E), Sensitivity and specificity broken down by the chosen PCa probability cutoffs for the 5 different assays in each of 5 nomograms showing different discrimination between PCa and NEM.

sensitivity and specificity data must be considered as important criteria, rather than only the global AUC measurements, in characterizing the effect of PSA interassay variation (Table 4).

CALIBRATION OF THE NOMOGRAMS ACCORDING TO THE PSA ASSAYS

The concordance between the PCa probability predicted by the nomograms and the real (observed) rate of PCa can be visually represented in a calibration plot, which is considered a measure of a model's quality. With total concordance, there is no difference between predicted probabilities and observed rates, and all points lie on the 45° line. Fig. 2 shows the calibration differences due to assay-dependent PSA values for all of the nomograms. In general, differences between observed PCa rates and predicted PCa probabilities depend on both the PSA assay used and the corresponding nomogram. Only the Siemens-f assay shows excellent performance in nomogram V, with an ICC of almost 1 (Fig. 2E). In contrast, Fig. 2A shows large differences between the observed PCa rates and predicted PCa probabilities, with up to 2-fold underestimation of the PCa rate, regardless of the PSA assay that is applied to the nomogram. These data indicate a weak performance of nomogram I, regardless of the PSA assay that is used. The main reason for this inferior validity for nomogram I is the absence of the %fPSA value.

Discussion

Numerous nomograms have been developed to predict PCa risk and to facilitate the process of prostate biopsy decision-making for the clinician *(22)*. All of these tools use the PSA value as a decisive variable for risk stratification, in addition to such other variables as a suspicious DRE result and prostate volume. External validations of various nomograms have most often been performed by applying the nomograms to different populations *(16, 18)*. Table 1 shows that the different results obtained with the different PSA assays greatly affect the reliability of PCa risk prediction with nomograms.

It is well established that the use of different PSA assays generally yields different tPSA and %fPSA values *(14, 23)*, despite the introduction of WHO PSA standards to improve the interchangeability of PSA results among the various assays *(14)*. Thus, assay calibration is only partially responsible for the differences between the assays in PSA estimation *(15)*. An analysis of PCa probabilities that used nomograms IV and V in a separate cohort of approximately 640 men *(24)* revealed lower medians with the WHO-calibrated data than with the Hybritech-calibrated data (see Table 2 in the online Data Supplement). Aside from the variation in tPSA, differences in %fPSA are also responsible for the variation in PCa probabilities. For example, the PCa probability ranged from 0.49 for the Abbott assay to 0.76 for the Siemens-f assay when predictions were made with nomogram IV, as seen in 1 patient (patient B in Table 3). The predicted PCa probabilities obtained with nomogram IV for the other 2 patients also demonstrated large variation (between 0.16 and 0.38, patient A in Table 3; between 0.54 and 0.91, patient C in Table 3).

When we applied fixed PCa probability cutoffs, the differences between different assays in sensitivity and specificity increased when we used nomograms that take %fPSA into account (Fig. 1, B–E). Therefore, tPSA assay variation seemed to have a more moderate impact on PCa probability values than the variation in %fPSA, a result that has already been shown *(25)*. This conclusion is documented in the examples of 3 patients with different tPSA concentrations (Table 3). The specificities also demonstrated large variation at a given PCa probability, such as 0.5. Specificities for the Siemens-f and Abbott data were >30% different at a 0.5 PCa probability when nomogram IV was used (Fig. 1D). These data suggest that the nomograms showed large differences in discrimination power, whereas overall AUC values did not (Table 4). Predictions based on ROC curve analyses are based on rank-order statistics *(26)*. This approach is insensitive to systematic errors in calibration, an issue that has recently been reviewed *(27)*; therefore, AUC comparisons alone are not appropriate for the validation of risk calculators. ROC curve analysis has been criticized when it is used as the only tool to differentiate between 2 cohorts *(28, 29)*, and results can be misinterpreted *(18)*. Our results, however, confirm reviewed data *(8, 22)* that show a general advantage of %fPSA-based multivariate models for PCa detection.

Calibration differences in PCa probabilities are also important, as Fig. 2 demonstrates. Yet nomogram I (Fig. 2A), which does not account for %fPSA results, showed only marginal variation among the PSA assays for predicted PCa probabilities and observed PCa rates but had the lowest overall ICC value. In addition, nomogram I showed the largest difference between observed rates and predicted PCa probabilities, with an approximately 2-fold higher observed PCa rate. This detection rate clearly improved in nomograms that included %fPSA in the calculations. On the other hand, the variance between the assays was much larger for nomograms that included %fPSA, a finding that is especially evident with nomogram IV. Thus, external validation of multivariate models requires a thorough assessment of the potential contributions of calibration analysis when attempting to estimate the concordance

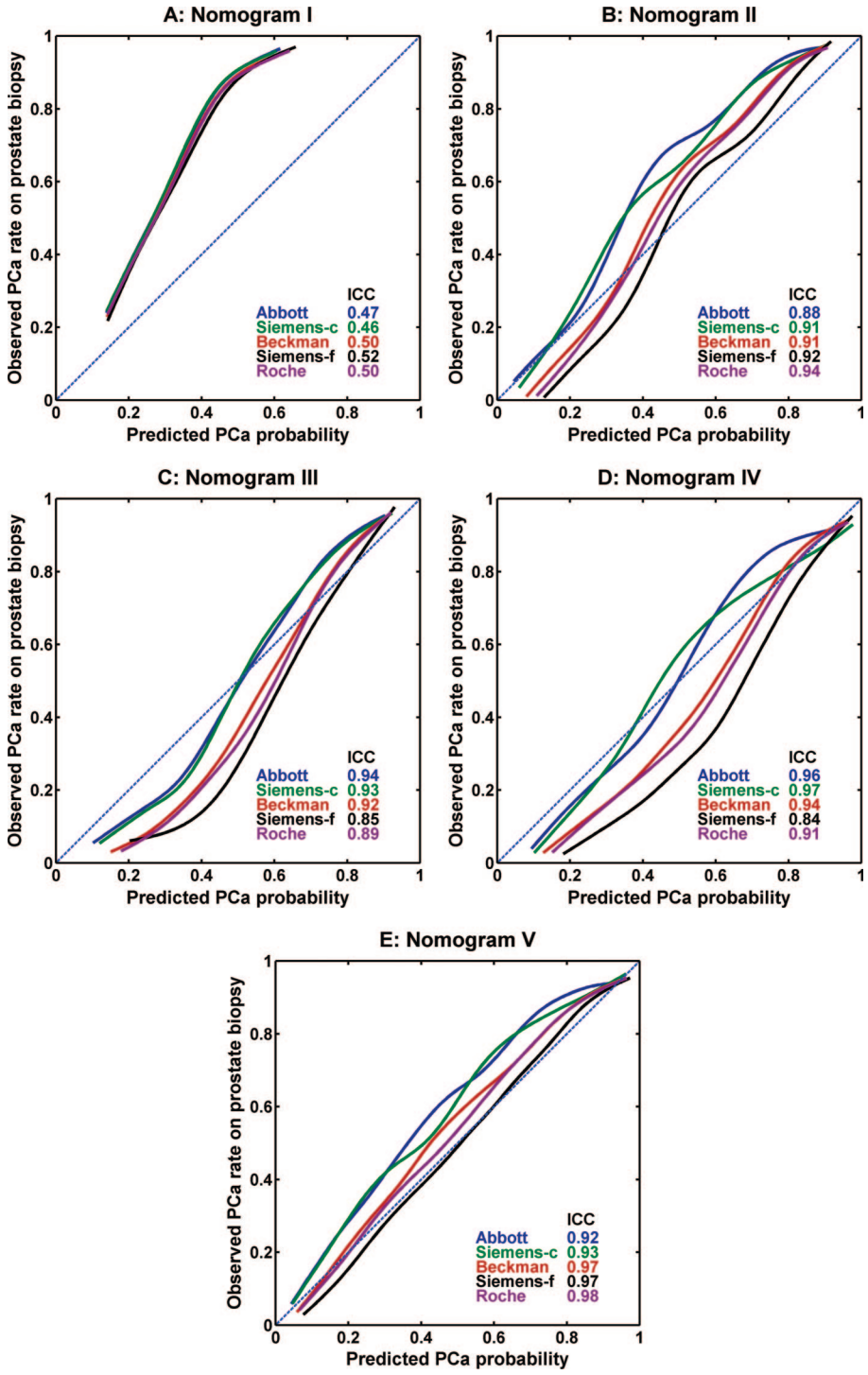

Fig. 2. (A–E), Calibration plots with cubic smoothing splines for the respective observed rates and predicted PCa probabilities for 5 different assays.

Plots show the assay effect on accuracy and performance characteristics. The ICC, a measure of the consistency of the observed and predicted values, is given for the data for each assay. To calculate the ICC, we subdivided the 780 patients into 20 groups in the order of predicted PCa probability. The mean observed rates and predicted probabilities were then computed for each group.

between predicted PCa probabilities and observed PCa rates.

Interestingly, nomograms IV and V, which were developed with data from the Beckman Coulter and Abbott PSA assays, had relatively low ICC values (Fig. 2, D and E) for these 2 assays. This result indicates that the effect of the assays on PCa prediction seems to be superimposed on other effects. These contradictory results are most likely caused by differences in the characteristics of the cohorts used to build the nomograms, compared with our cohort. In addition to the effect of the variation contributed by the different PSA assays, 39% of the patients in our cohort had suspicious DRE findings (Table 2). This rate is different from the rates for the cohorts upon which the nomograms were based *(11, 16, 17)*. Although Kawakami et al. *(17)* and Chun et al. *(16)* found only 17.5% and 20.3% of all patients, respectively, with a suspicious DRE result, the 3 cohorts used to develop nomograms I and II had higher rates of suspicious DRE findings (31.1% overall) *(11)*.

Additionally, the prevalence of PCa in our population (58.3%) was higher than the prevalences in the cohorts used to develop the 5 nomograms, which varied from 35.2% to 41.9% *(11, 16, 17)*. These differences may also have an impact on nomogram performance in external validations.

The study is limited by both the inclusion of only 5 available nomogram-based models and the retrospective study design. Unfortunately, none of the other nomograms available for PCa risk prediction were suitable for our data *(30–33)*. In some cases, nomograms were available, but they had been developed with small cohorts *(34)*.

In summary, the present study has provided 2 main conclusions. First, our results demonstrate that nomogram-based PCa prediction is influenced by the type of PSA assay that is used. Second, AUC comparison alone is insufficient, and calibration analysis is recommended for validation of models. The dependence on PSA assay calls into question the general applicability of these models without considering the suitability of a specific PSA assay for a given model.

Author Contributions: *All authors confirmed 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: *No authors declared any potential conflicts of interest.*

Role 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

1. Lilja H, Ulmert D, Vickers AJ. Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. Nat Rev Cancer 2008;8:268–78.
2. Catalona WJ, Smith DS, Ratliff TL, Dodds KM, Coplen DE, Yuan JJ, et al. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. N Engl J Med 1991;324:1156–61.
3. Postma R, Schröder FH, van Leenders GJ, Hoedemaeker RF, Vis AN, Roobol MJ, van der Kwast TH. Cancer detection and cancer characteristics in the European Randomized Study of Screening for Prostate Cancer (ERSPC)—Section Rotterdam. A comparison of two rounds of screening. Eur Urol 2007;52:89–97.
4. Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, Parnes HL, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level ≤4.0 ng per milliliter. N Engl J Med 2004;350:2239–46.
5. Loeb S, Catalona WJ. Prostate-specific antigen in clinical practice. Cancer Lett 2007;249:30–9.
6. Stephan C, Jung K, Lein M, Diamandis EP. PSA and other tissue kallikreins for prostate cancer detection. Eur J Cancer 2007;43:1918–26.
7. Catalona WJ, Partin AW, Slawin KM, Brawer MK, Flanigan RC, Patel A, et al. Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease: a prospective multicenter clinical trial. JAMA 1998;279:1542–7.
8. Stephan C, Cammann H, Meyer HA, Lein M, Jung K. PSA and new biomarkers within multivariate models to improve early detection of prostate cancer. Cancer Lett 2007;249:18–29.
9. Stephan C, Rittenhouse H, Cammann H, Lein M, Schrader M, Deger S, et al. New markers and multivariate models for prostate cancer detection. Anticancer Res 2009;29:2589–600.
10. Chun FK, Karakiewicz PI, Briganti A, Walz J, Kattan MW, Huland H, Graefen M. A critical appraisal of logistic regression-based nomograms, artificial neural networks, classification and regression-tree models, look-up tables and risk-group stratification models for prostate cancer. BJU Int 2007;99:794–800.
11. Karakiewicz PI, Benayoun S, Kattan MW, Perrotte P, Valiquette L, Scardino PT, et al. Development and validation of a nomogram predicting the outcome of prostate biopsy based on patient age, digital rectal examination and serum prostate specific antigen. J Urol 2005;173:1930–4.
12. Virtanen A, Gomari M, Kranse R, Stenman UH. Estimation of prostate cancer probability by logistic regression: free and total prostate-specific antigen, digital rectal examination, and heredity are significant variables. Clin Chem 1999;45:987–94.
13. Filella X, Alcover J, Quintó L, Molina R, Bosch-Capblanch X, Carretero P, Ballesta AM. Evaluation of a multivariate prostate-specific antigen and percentage of free prostate-specific antigen logistic regression model in the diagnosis of prostate cancer. Tumour Biol 1999;20:312–8.
14. Kort SA, Martens F, Vanpoucke H, van Duijnhoven HL, Blankenstein MA. Comparison of 6 automated assays for total and free prostate-specific antigen with special reference to their reactivity toward the WHO 96/670 reference preparation. Clin Chem 2006;52:1568–74.
15. Stephan C, Klaas M, Müller C, Schnorr D, Loening SA, Jung K. Interchangeability of measurements of total and free prostate-specific antigen in serum with 5 frequently used assay combinations: an update. Clin Chem 2006;52:59–64.
16. Chun FK, Briganti A, Graefen M, Montorsi F, Porter C, Scattoni V, et al. Development and external validation of an extended 10-core biopsy nomogram. Eur Urol 2007;52:436–44.
17. Kawakami S, Numao N, Okubo Y, Koga F, Yamamoto S, Saito K, et al. Development, validation, and head-to-head comparison of logistic regression-based nomograms and artificial neural network models predicting prostate cancer on initial extended biopsy. Eur Urol 2008;54:601–11.
18. Utsumi T, Kawamura K, Suzuki H, Kamiya N, Imamoto T, Miura J, et al. External validation and head-to-head comparison of Japanese and Western prostate biopsy nomograms using Japanese data sets. Int J Urol 2009;16:416–9.
19. Stephan C, Cammann H, Meyer HA, Müller C, Deger S, Lein M, Jung K. An artificial neural network for five different assay systems of

prostate-specific antigen in prostate cancer diagnostics. BJU Int 2008;102:799–805.
20. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.
21. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–68.
22. Schroder F, Kattan MW. The comparability of models for predicting the risk of a positive prostate biopsy with prostate-specific antigen alone: a systematic review. Eur Urol 2008;54:274–90.
23. Stephan C, Kramer J, Meyer HA, Kristiansen G, Ziemer S, Deger S, et al. Different prostate-specific antigen assays give different results on the same blood sample: an obstacle to recommending uniform limits for prostate biopsies. BJU Int 2007;99:1427–31.
24. Stephan C, Kahrs AM, Klotzek S, Reiche J, Müller C, Lein M, et al. Toward metrological traceability in the determination of prostate-specific antigen (PSA): calibrating Beckman Coulter Hybritech Access PSA assays to WHO standards compared with the traditional Hybritech standards. Clin Chem Lab Med 2008;46:623–9.
25. Stenman UH, Leinonen J, Zhang WM, Finne P, Wu P. The clinical importance of free prostate-specific antigen (PSA). Curr Opin Urol 1998;8:393–9.
26. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74.
27. Obuchowski NA, Lieber ML, Wians FH Jr. ROC curves in clinical chemistry: uses, misuses, and possible solutions. Clin Chem 2004;50:1118–25.
28. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928–35.
29. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27:157–72.
30. Benecchi L, Pieri AM, Melissari M, Potenzoni M, Pastizzaro CD. A novel nomogram to predict the probability of prostate cancer on repeat biopsy. J Urol 2008;180:146–9.
31. Kawamura K, Suzuki H, Kamiya N, Imamoto T, Yano M, Miura J, et al. Development of a new nomogram for predicting the probability of a positive initial prostate biopsy in Japanese patients with serum PSA levels less than 10 ng/mL. Int J Urol 2008;15:598–603.
32. Nam RK, Toi A, Klotz LH, Trachtenberg J, Jewett MA, Appu S, et al. Assessing individual risk for prostate cancer. J Clin Oncol 2007;25:3582–8.
33. Finne P, Finne R, Bangma C, Hugosson J, Hakama M, Auvinen A, Stenman UH. Algorithms based on prostate-specific antigen (PSA), free PSA, digital rectal examination and prostate volume reduce false-positive PSA results in prostate cancer screening. Int J Cancer 2004;111:310–5.
34. Rochester MA, Pashayan N, Matthews F, Doble A, McLoughlin J. Development and validation of risk score for predicting positive repeat prostate biopsy in patients with a previous negative biopsy in a UK population. BMC Urol 2009;9:7.