avoiding pitfalls in applying prediction models, as illustrated by the example of prostate cancer diagnosis

Henning Cammann,1 Klaus Jung,2,3 Hellmuth-A. Meyer,2,4 and Carsten Stephan2,3*

BACKGROUND: The use of different mathematical models to support medical decisions is accompanied by increasing uncertainties when they are applied in practice. Using prostate cancer (PCa) risk models as an example, we recommend requirements for model development and draw attention to possible pitfalls so as to avoid the uncritical use of these models.

CONTENT: We conducted MEDLINE searches for applications of multivariate models supporting the prediction of PCa risk. We critically reviewed the methodological aspects of model development and the biological and analytical variability of the parameters used for model development. In addition, we reviewed the role of prostate biopsy as the gold standard for confirming diagnoses. In addition, we analyzed different methods of model evaluation with respect to their application to different populations. When using models in clinical practice, one must validate the results with a population from the application field. Typical model characteristics (such as discrimination performance and calibration) and methods for assessing the risk of a decision should be used when evaluating a model’s output. The choice of a model should be based on these results and on the practicality of its use.

SUMMARY: To avoid possible errors in applying prediction models (the risk of PCa, for example) requires examining the possible pitfalls of the underlying mathematical models in the context of the individual case. The main tools for this purpose are discrimination, calibration, and decision curve analysis.

© 2011 American Association for Clinical Chemistry

One possible method for gaining more objectivity in medical decisions is to use mathematical models to predict diagnosis or prognosis. These models can help reduce the influence of highly subjective factors on the decision; however, prediction models can lead to mistakes in decision-making if the reliability of the input data for the model has not been sufficiently considered for the construction or use of the models. Usually, clinical chemists, clinicians, and statisticians separately provide the data for building a model. Because the most dangerous mistakes can occur when factors influencing the construction and use of models are not considered in model development, this merging of data into models requires that all individuals involved have knowledge of the reliability of the data so that wrong decisions regarding the detection, treatment, or active surveillance of a disease can be avoided. Table 1 summarizes the general points that should be considered when a model is constructed and used in daily practice.

In this report, we use the development of a model for predicting the occurrence of prostate cancer (PCa)5 as an example. The aim is to demonstrate how to obtain reliable results in daily practice when using mathematical models. PCa is diagnosed when cancer cells are detected in a prostate biopsy. The physician’s decision to use an invasive biopsy procedure is driven by clinical findings, such as suspicious digital rectal examination (DRE) results or higher serum concentrations of prostate-specific antigen (PSA). The increased use since 1985 of the PSA concentration in men has led to increasing numbers of prostate biopsies. Aside from the common drawbacks of using PSA tests for the early detection of possible PCa (1), performing a biopsy when the PSA value is in the range of 2–10 ng/mL leads to negative findings in about 60%–80% of biopsies. This lack of predictive accuracy has led to substantial costs and confusion among patients. Furthermore, a biopsy also carries a risk of severe complications. Therefore, numerous researchers have examined mathematical combinations of demographic factors (age, family history), clinical data (DRE, ultrasound findings, prostate volume), and laboratory findings

1 Institute of Medical Informatics and 2 Department of Urology, Charité – Universitätsmedizin Berlin, Berlin, Germany; 3 Berlin Institute for Urologic Research, Berlin, Germany; 4 Institute of Physiology, Charité–Universitätsmedizin Berlin, Berlin, Germany.

* Address correspondence to this author at: Department of Urology, Charité – Universitätsmedizin Berlin, CCM, Charitéplatz 1, D-10117 Berlin, Germany. Fax +49-30-450-51904; e-mail carsten.stephan@charite.de.

Received April 11, 2011; accepted August 24, 2011. Previously published online at DOI: 10.1373/clinchem.2011.166959

5 Nonstandard abbreviations: PCa, prostate cancer; DRE, digital rectal examination; PSA, prostate-specific antigen; fPSA, free PSA; tPSA, total PSA; %fPSA, percentage of fPSA; LR, logistic regression; ANN, artificial neural network; AUC, area under the ROC curve; ICC, intraclass correlation coefficient; DCA, decision curve analysis.
[e.g., PSA, free PSA (fPSA) subforms, and additional analytes, including other kallikreins or the PCA3 gene (prostate cancer antigen 3, non-protein coding)], as well as combined indices (the ratio of PSA to prostate volume, PSA density, and the change in PSA values over time, i.e., the PSA velocity) to improve the predictive power of PCa risk prediction. Thus, clinicians can select from a wide variety of potential models. Because the external validation of the models has not always been successful, critics have questioned the use of these mathematical models (2–6).

We thought it particularly important to examine the factors that are responsible for this unsatisfactory situation. The aim of this mini-review is to identify the pitfalls of prostate biopsy decision models (and, by extension, the pitfalls of prediction models in general) by an examination of the mathematical models that are frequently used in clinical practice. We recommend that this review be taken as a basis for discussions among clinicians, clinical chemists, and statisticians regarding these pitfalls, and we encourage its use in facilitating mutual understanding of the roles each discipline plays in model construction. Such discussions might prove instructive for how to avoid mistakes when building a nomogram.

**Literature Review**

We conducted MEDLINE searches for applications of multivariate models supporting the prediction of PCa risk. We critically reviewed the methodological aspects of model development and the biological and analytical variation in the parameters used for model development. In addition, we reviewed the role of prostate biopsy as the gold standard for confirming diagnoses. In addition, we analyzed different methods of model evaluation with respect to their application to different populations.

**Determinants Influencing the Quality of a Model**

To be effective in clinical practice, a model should include only the factors necessary to provide optimal information. Therefore, finding an optimal balance between potential and necessary factors is critical when creating a new model. Table 1 provides an overview of the common factors that influence the validity of predictive/prognostic models in general and their particular importance for prostate biopsy decision models.

<table>
<thead>
<tr>
<th>Important factors</th>
<th>Example for PCa detection models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference system for final diagnosis</td>
<td>Soft-criteria prostate biopsy</td>
</tr>
<tr>
<td>Predetermined (fixed) values</td>
<td>Age, weight, family history</td>
</tr>
<tr>
<td>Preanalytical factors</td>
<td>Sampling, handling, storage</td>
</tr>
<tr>
<td>Factors with analytical variation</td>
<td>PSA, %fPSA, proPSA, PCA3, other kallikreins</td>
</tr>
<tr>
<td>Factors with biological variation</td>
<td>PSA, %fPSA, proPSA, PCA3, other kallikreins</td>
</tr>
<tr>
<td>Technical factors</td>
<td>Controllable variables (sample collection and transport), uncontrollable variables (age, sex)</td>
</tr>
<tr>
<td>Population</td>
<td>Screening vs referred</td>
</tr>
<tr>
<td>Choice of model</td>
<td>Nomogram, ANN, decision tree, support vector machine, nearest neighbor</td>
</tr>
<tr>
<td>Model evaluation</td>
<td>Internal/external validation and criteria discrimination, calibration, and DCA</td>
</tr>
<tr>
<td>Implementation in clinical practice</td>
<td>Application to complete populations, use for individual patients, comparison with single parameters, uncritical use of cutoffs</td>
</tr>
</tbody>
</table>

**REFERENCE STANDARDS, END POINTS AND THE GOLD STANDARD**

Ultrasound-guided needle prostate biopsy is the current standard for PCa detection. A suspicious DRE re-
sult or increased values of PSA or PSA derivatives (such as PSA velocity and PSA density) trigger a prostate biopsy. Biopsy samples are taken according to a standardized scheme, but not all areas of the prostate can be included. The amount of sampling in biopsies has increased over time (10–21 cores or saturation biopsy), and that has improved the positive-detection rate. This fact implies that a negative biopsy result does not completely rule out the presence of PCa (7). PCa-detection rates of 18%, 17%, and 14% obtained in the second, third, and fourth extended 21-sample biopsy, respectively (8), demonstrate the limitations of a single biopsy for detecting the presence of PCa. These data indicate that a prostate biopsy is a “soft” gold standard. Thus, even if clinicians consider all of the important parameters for a biopsy decision, such as PSA velocity, %fPSA, PSA density, and other factors, doing so does not completely eliminate the uncertainty in a biopsy result. Particularly after one (or more) negative biopsy results, the use of the aforementioned parameters, including the PCA3 gene and repeat PSA measurements, should help clarify whether additional biopsies are needed. If the pathology report for the biopsy is unclear or highly suspicious for PCa (as indicated by histologic findings of prostatic intraepithelial neoplasia or atypical small acinar proliferation), a repeat biopsy should be performed within 4–12 weeks. If the PCa risk in the model still does not decrease after additional PSA measurements, the next biopsy should be performed within 3–6 months.

### tPSA and fPSA
PSA concentrations in serum are subject to biological, preanalytical, and analytical influences. The imprecision in PSA measurements is about 5%. The biological variation, however, is much larger. PSA concentrations can vary by up to 15% in blood samples collected 2 weeks apart (9). A review of 12 studies shows a mean biological variation in PSA concentration of up to 20% (10). Thus, to avoid mistakes, one should confirm any decision in response to a critical high PSA value with at least 1 or 2 repeat measurements. The well-known approach of critical difference should be applied in determining whether a true increase between sequential laboratory results has occurred (11).

Having several different commercial assays for measuring tPSA and fPSA is another source of variation. Comparative studies of 5 separate PSA assays have revealed differences in tPSA values of up to 28% (12) and differences in fPSA values of up to 16% (12), for %fPSA differences as high as 46%. Although most manufacturers of PSA assays use WHO PSA standards (13) as calibrators, distinct differences in the PSA values obtained with the various assays persist. Thus, laboratories need to inform clinicians of the assays they use to ensure that PSA values are interpreted correctly in a model (13, 14). It is obvious that the lack of uniformity in measured PSA values obtained with different assays and their uncritical use in models can lead to incorrect prediction results. The large tPSA and %fPSA differences in the described study (12) led to clinical consequences regarding the use of 5 different nomograms with patient data from different PSA assays (6). The predicted median probability of PCa varied between 0.59 and 0.76 in one nomogram. The most hazardous consequence in PCa diagnosis would be to use data from an incorrect or unknown PSA assay to decide not to perform a biopsy, leading to a subsequent delay in PCa detection.

### Prostate Volume (Ultrasound-Guided Measurement)
Prostate volume is another common parameter used in model-based prediction tools. The mean intraobserver imprecision for volume measurement is about 4.6%, and the interobserver imprecision is twice as high (15). Volume measurements differ more at higher volumes than at lower ones. In addition, larger differences exist between experienced and inexperienced examiners than between experienced investigators (16, 17).

### Digital Rectal Examination
The DRE has been the classic screening tool for PCa, even after PSA measurement became available. DRE data collected by non-urologists, however, show low correlation with urologist-collected data (18). In addition, results also vary between urologists (19). The DRE is obviously the most subjective factor, but a suspicious DRE result nevertheless requires that a biopsy be performed. A wrong interpretation of a suspicious DRE result may delay a subsequent prostate biopsy necessary for a PCa diagnosis.

---

**Table 2. Frequencies with which certain parameters have been used to construct models for predicting PCa risk [based on Shariat et al. (30)].**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Frequency, n*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>26 (96%)</td>
</tr>
<tr>
<td>PSA</td>
<td>26 (96%)</td>
</tr>
<tr>
<td>DRE</td>
<td>24 (89%)</td>
</tr>
<tr>
<td>%fPSA</td>
<td>14 (52%)</td>
</tr>
<tr>
<td>TRUS&lt;sup&gt;a&lt;/sup&gt;</td>
<td>12 (44%)</td>
</tr>
<tr>
<td>Family history</td>
<td>11 (41%)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>10 (37%)</td>
</tr>
<tr>
<td>Previous negative biopsy result</td>
<td>5 (19%)</td>
</tr>
<tr>
<td>TRUS&lt;sup&gt;+&lt;/sup&gt;</td>
<td>2 (7%)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Percentages are based on n = 27 studies.  
<sup>b</sup> TRUS, transrectal ultrasound; TRUS<sup>+</sup>, suspicious TRUS finding.
Mathematical models are developed with data sets generated from selected populations to produce answers to a specific question. The selection of the population depends on the intended use of the model. Thus, the established model in question is initially valid only for the population for which it was developed (20–22). An internal validation procedure (through bootstrapping or resubstitution) should be used as a first measure for checking the validity of a model developed from the source population; however, external validation of the model is required before it can be used in general clinical practice. For that purpose, data from another population are examined with the newly developed model. If the results obtained from the model with the 2 separate populations are consistent, the model can be expected to perform reliably in clinical practice. The external validation results will generally be weaker, however, than those obtained by internal validation.

Positive PSA and/or DRE findings are obviously more frequent in clinical diagnosis studies of patients with definitively diagnosed PCa. Such populations differ substantially from population-based screening cohorts, however. Therefore, data and models are not interchangeable or applicable between populations that are being screened and populations that have been referred to urologists.

### Choice of Model

In addition to models constructed with the long-established techniques of logistic regression (LR) and discriminant analysis (which is rare for PCa), there are also models based on more recent methods, such as artificial neural networks (ANNs), decision trees, or support vector machines (23).

The basic advantages and disadvantages of various models are often compared (24–26), although their performances can be compared only by the use of identical data sets (27). An important advantage of ANNs is their ability to model nonlinear relationships between parameters (25). ANN models are more flexible in this respect than are LR models; however, the importance of each variable can be assessed directly in an LR model but not in an ANN model, which is considered a “black box.” The choice of a model should be determined by the results of the validation, however. Standards for ANNs have been requested from the beginning (28), but systematic comparisons of different models constructed with data from the same source have rarely been conducted (23). One study found that all of the models under investigation (using linear discriminant analysis, LR, ANN, support vector machines, nearest-neighbor classifier, and decision trees) that analyzed the same data set performed well, with only marginal differences between them. This study did not include model comparisons, however (23).

We restrict ourselves to the most commonly used LR and ANN models (25, 29).

### Model Assessment (Validation)

The general mathematical performance of a model can be assessed in internal and external validation studies by means of 2 essential criteria: discrimination and calibration (21, 30). Discrimination describes a model’s ability to separate healthy and ill individuals. In model assessment, the term “calibration” refers to the differences between predicted and observed probabilities. Various tools, which generate several measures, are available for characterizing model-performance criteria (Fig. 1).

#### Discrimination

Box plots (Fig. 1A) give an impression of the distribution of the predicted probabilities for the associated groups. The median difference provides insight into the discrimination potential of the parameter in question. The discrimination slope is equal to the difference between the mean predicted probabilities of the group with no evidence of malignancy and the group with PCa (21).

Fig. 1B shows a visualization of the changes between specificity and sensitivity, including the Youden index (sensitivity + specificity − 1). The model output is a continuous range of predicted probabilities for PCa. An individual interpretation as a positive or negative finding for a patient can be obtained only by choosing a threshold (cutoff). For every chosen cutoff value, a sensitivity and specificity value can be calculated. In practical terms, the cutoff is simplified by fixed sensitivity targets that lead to fixed cutoffs and the respective specificity. Most ANNs allow for cutoffs (31, 32), whereas that is rare in nomograms (33).

A compressed representation of the discriminatory power of a model is given by the ROC curve, which characterizes the relationship between the sensitivity and the respective specificity (Fig. 1C). The measure of the discriminatory power of a model is the area under the ROC curve (AUC), or the c statistic (34, 35), AUC values can be used to compare models. The higher the AUC, the better is the separative ability of the model. ROC curves are not universally applicable, however. For example, the ROC curves of 2 compared tests with the same AUC can cross graphically and therefore differ at a given sensitivity. Thus, such ROC curves may differ in clinically relevant ways (34). Alternatively, partial areas of ROC curves may be used to assess the diagnostic value (36).

#### Calibration

In biometrics, calibration consists of a quantile–quantile plot that graphically determines whether 2 data sets have the same distribution. If the distributions are the same, the points lie on a diagonal line that bi-
sects the 90° angle. Generally, this approach of calibration is used to plot the true observed disease status (e.g., healthy or ill) against the probability predicted by the model (Fig. 1D). In our case, the known prostate biopsy results (PCa or no cancer) were plotted against the probability predicted by the model. The number of observed probability quantiles used depends on the sample size. Because the calculated pairs of observed and predicted probabilities scatter strongly, visualizing the model’s behavior is usually achieved by fitting the points to a smooth curve (e.g., a cubic spline). The resulting curve shows whether the model generally or in certain areas underestimates or overestimates the actual PCa risk. To increase the information content of the calibration plots, we advise showing the distributions of predicted probabilities in the figure separately for the healthy and ill groups. The example in Fig. 1D is based on 637 patients (30 groups) for whom the Spearman rank correlation coefficient ($r_s = 0.835$), a general value for the correlation between observed and predicted PCa probabilities, is higher than the intraclass correlation coefficient ($ICC = 0.734$). In addition to considering the correlation, the ICC also incorporates differences in absolute values. Another criterion used to describe the predictive value of a model is the sum of residuals, which is a measure of

**Fig. 1.** The validation of an ANN model [Stephan et al. (31)] with a new test population.

The measured variables are age, tPSA, %fPSA, transrectal ultrasound, and DRE; the Beckman Coulter Access test kit was used. The population to train the ANN comprised 780 patients: 455 PCa patients and 325 patients with no evidence of malignancy (NEM). The test population consisted of 637 patients: 336 PCa patients and 301 patients with NEM. The panels present the results for the test group. (A), Box plots of the predicted probabilities for the 2 groups The median and interquartile range are indicated. (B), The Youden index (sensitivity + specificity − 1) shows the best discriminatory interval for the model (from 0.2 to 0.4). (C), An ROC curve. The discriminatory ability is apparent from the AUC (0.77). For clinical applications, a preset sensitivity of 95% is often used. (D), A calibration plot. The cubic spline shows that the model underestimates the PCa risk in the interval from 0.2 to 0.6. The horizontal line at the bottom of the plot indicates the distribution of the calculated probabilities (PCa above, NEM below). (E), A Bland–Altman plot. The predicted probabilities for 95% of the cases are $\approx 0.44$ and $\approx 0.16$ of the observed probabilities. This result agrees with the calibration plot.
the deviation of the model outputs from an ideal match with the observed PCa probabilities (37). The best measure of the suitability of a model, however, remains the graphical presentation.

In addition, presentation of the data as a Bland–Altman plot (Fig. 1E) helps to clarify the relationship between the observed and predicted probabilities. We prefer this plot for comparing methods because it allows easy recognition of both the systematic error between the observed and predicted probabilities (a high degree of scatter throughout the entire scale) and error that depends on the probability values (a proportionally increased/decreased error or a similarly increased/decreased error). The correlation coefficient \( r_{ba} \) for the difference between the observed and predicted probabilities and the mean of the observed and predicted probabilities helps to determine whether a systematic bias is constant for the entire range of values.

**Clinical Practice**

Predictive or prognostic models have often been recommended in the literature. Some of these models have eventually been used in daily clinical practice despite a lack of information about the discrimination and calibration characteristics. This situation is unsatisfactory. Therefore, it is desirable that at least an internal validation process be performed and that the results of the discrimination and calibration analysis be reported. Use of an established model, however, that produces an overall improvement in the results, as characterized by better AUC values in external validations, does not prevent over- or underestimation of the actual PCa risk (38). The steps required to successfully use models that have been tested in one’s own clinical practice have rarely been described (21).

Before a model is to be used in clinical practice, it must be validated with an appropriate test population to determine whether the validation results meet the practical requirements. The application of such a model in clinical practice requires exactly the same parameters that were used for constructing the model. In addition, the use of similar intervals (e.g., for PSA concentration or age) and similar PSA and fPSA assays would be preferable in order to reduce the frequencies of possible errors.

If a validated model is already in use, a head-to-head comparison of its results with those of a newly established model is necessary with the same data sets (37, 39). In practice, calibration or Bland–Altman plots are used to compare the 2 models (Fig. 1E). A comparative assessment of models with validity criteria obtained with different data sets is misleading (27).

In clinical practice, it is important to know the PCa risk probability (i.e., whether a patient who tests positive is actually sick and whether a patient with a negative result is actually healthy). For this purpose, the positive predictive value and the negative predictive value are calculated from the sensitivity and specificity of the model and the PCa prevalence by means of the Bayes theorem (40). Fig. 2A shows the positive predictive value and the negative predictive value by prevalence. One can clearly see that in populations with high prevalences, the positive predictive value increases rapidly. Conversely, the negative predictive value falls in...
populations with high prevalences. This fact is valid for all tests, regardless of the discriminatory power of the test.

The experience of the examiner and the patient’s willingness to take risks are important for individual evaluations of decisions with clinical consequences (such as whether to perform a biopsy). With the aid of decision curve analysis (DCA) (41, 42), it is clear when a model is superior for the assumptions “all should undergo biopsy” or “no one should be biopsied” (Fig. 2B). By using DCA, one generally assumes that the threshold probability for a disease or event at which a patient opts for treatment provides information on how the patient weighs the relative harms of false-positive and false-negative predictions (42). This theoretical relationship is then used to derive the net benefit of the model across different threshold probabilities (42). Plotting the net benefit against the threshold probability yields the DCA.

Sometimes the superiority of nomograms to computer-based procedures (particularly ANNs) is due to the inavailability of computers at all locations. In such cases, the complexity of the mathematical models must be weighed against the technological conditions. The use of a model implies that the model has been validated with the clinical population for which the model is intended to be used. This process will be performed with a computer. A nomogram is useful in clinical practice when no computer is available. In all other cases, regardless of the model, the use of computers is superior to methods based on paper and pencil because of the ability to avoid input errors with computers by using plausibility criteria for the input data. For a nomogram example of a PCa risk calculator see http://deb.uthscsa.edu/URORiskCalc/Pages/calcs.jsp; for an ANN example, see http://www.charite.de/pcaberlin/ann5/ann5.html (see Fig. 3). Furthermore, a computer can easily display and save the results (31). A disadvantage of computer programs is their apparently "mystical" nature: One inputs the numbers; then a result appears (43). This disadvantage can be reduced by using an appropriate graphical presentation of the results.
that contains references to the influences of individual variables; however, one should note that aside from the pros and cons for ANNs and nomograms, every validated model performs better than PSA alone.

Conclusions

Precise information is required regarding the quality of the measured variables (exploratory statistics), the population used for model development, and the quality of the gold standard. Neglect of the factors that influence the modeling (Table 1) leads to incorrect probability results. There is a relationship between the quality of the parameters used and the final outcome of a model.

The selection of a model for application in clinical practice should be based primarily on the validation results from a hospital-specific test sample. The PCA probability calculated with a model should be assessed by ROC curve analysis and the model’s calibration and DCA curves when applied in an appropriate population (21). These 3 tools should supply the necessary information for all users of a model. All other tools are more important for the developers of models. Used correctly, models provide substantial support for complex decision-making.

It is essential for clinicians not to simply enter data into a computer or nomogram and blindly accept the result (26). Models do not replace personal knowledge and personal decisions, but they can improve certainty and support the decision-making step. Because of the biological and analytical variability of the parameters, models provide only probability results and are tailored to specific populations. There is no universal model, but rather a set of models, each of which is suitable for use in specific populations.

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: Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:

Employment or Leadership: None declared.

Consultant or Advisory Role: None declared.

Stock Ownership: None declared.

Honoraria: None declared.

Research Funding: H. Cammann, Berliner Sparkassenstiftung Medizin (no. 89838600); K. Jung, Foundation of Urologic Research (Berlin) and Wilhelm Sander-Stiftung (grant no. 2010.111.1) (Munich); H.-A. Meyer, Foundation of Urologic Research (Berlin); C. Stephan, Berliner Sparkassenstiftung Medizin (no. 89838600), Foundation of Urologic Research (Berlin), and Wilhelm Sander-Stiftung (grant no. 2010.111.1) (Munich).

Expert Testimony: None declared.

References