Variance Function as Basis for Assessment of Test Performance: Methodological Studies with Two Assays of Prostate-Specific Antigen

Edgar Hänseler,1,3 Bernhard Keller,2 and Herbert Keller1

We evaluated the usefulness of a recently described procedure to assess the analytical performance of an assay. To demonstrate the advantages of this approach, we compared the performance of two analytical systems for determining prostate-specific antigen (PSA). Triplicate measurements of PSA with the IMx (Abbott) and the ACS 180 (Ciba Corning) were used to calculate the variance function. This function was the basis for the definition of the critical limit (LC), the limit of detection (LD), the power of definition (PD), and the lower limit of the quantification interval. The standard deviation of the blank was extrapolated by means of the variance function. LC was calculated as the concentration at which the normal distribution of the blank intersects an adjacent normal distribution (with a defined overlap, e.g., 5%). The mean of the adjacent normal distribution represents the LD. The PD is a new mathematical approach to describe the analytical sensitivity of an assay in different ranges of the quantification interval. The procedure is statistically well defined and allows one to obtain the data on the test performance directly from patients' samples, without artificial zero controls. Therefore, use of the variance function could be a general model for the assessment of the analytical performance of an assay.

Indexing Terms: critical limit/limit of detection/power of definition/lower limit of quantification/precision profile/statistics

We previously described a method for assessing the limit of detection (LD), lower limit of quantification (LLQ), and the so-called power of definition (PD) (1).4 The procedure is based on the variance function according to Ekins (2) and Sadler and Smith (3). With the program developed by Sadler and Smith, not only can the variance function be determined, but the imprecision profile with its confidence limits can be plotted. We have proposed [according to Bayer (4)] to define the lower (and upper) limit of the measuring interval by the point at which a line (representing an acceptable CV) intersects the confidence limit. For determination of the LD we set the analyte concentration to zero in the variance function, plot the corresponding normal distribution, and search for the adjacent normal distribution overlapping the zero distribution by a defined amount, using an iterative procedure (1).

The PD, representing the number of normal distributions overlapping each other by a defined amount (e.g., 5%) in a distinct concentration interval (e.g., the reference interval), can be constructed in a similar manner. The PD determines in each optional interval of the analytical range the smallest difference of subsequent values that can be discriminated with a defined statistical confidence. This characteristic is especially relevant for follow-ups to determine if an increased (or decreased) value indeed represents an increase (or decrease) of the analyte concentration, or if it could still be the result of the analytical variation.

A further crucial advantage of this procedure is that data for the assessment of the variance function originate from patients' sample material only, and one need not use data from artificial zero-standards of unknown composition, or from atypical sources.

The reliable knowledge of the LD, the limits of quantification, and the PD are especially useful for the interpretation of laboratory results for prostate-specific antigen (PSA). PSA measurements are used in combination with other diagnostic procedures such as imaging methods and digital rectal examination for the detection of existing prostate disorders. However, for postoperative monitoring of minimal residual disease or recurrence of the tumor, procedures such as these are not sensitive enough and are therefore of little use. After total prostatectomy, PSA concentrations should become undetectable within several weeks by the current immunoassays (5, 6). Since very low values are used for clinical decisions, the term "undetectable" should be precisely defined. This means that one must know the probability for a positive result when the analyte (PSA) is absent from the sample (error of the 1st kind, or α-error, false-positive result) and the probability for a negative result when the analyte is present in the sample (error of the 2nd kind, β-error, false-negative results). As a consequence, lower limits of detection (LLD) and LLQ must be defined for each procedure measuring PSA. The LLD of PSA assays has been determined either by adding 2 SD (7) or 3 SD (8) to the zero-mean obtained by serial measurements of a zero calibrator or a female patient's sample. LLD values in the literature for PSA vary widely from assay to assay, ranging from 0.002 μg/L (9) to 0.3 μg/L [reviewed in (10)].

For prostatectomized patients the use of so-called ultrasensitive assays has been suggested by several authors (7, 9, 11–14). The biological detection limit (BLD)
was first used by Vessella et al. (7, 15), who observed that false-positive diagnostic results were obtained from prostatectomized patients when the LLD was used for the Tandem- and IMx-PSA assays. The BLD was obtained by adding 2 SD (gained from determinations of diluted patients' samples) to the LLD (7). With the BLD, fewer false-positive results resulted. Nevertheless, this term does not conform with International Federation of Clinical Chemistry recommendations (16), since the detection limit depends solely on the analytical performance of the method used and should not be mixed up with biological criteria as described by Solberg (17), Fräkebeck (18), and others. Thus the procedures used to obtain these limits were inconsistent and not standardized.

The multiple terms used for the LD (lower limit of detection, lowest limit of detection, detection sensitivity, sensitivity, biological limit of detection) emphasize the need for a standardized terminology and a uniform statistical approach by which this important characteristic of the analytical performance of an assay can be obtained. We, like Currie (19) and others (20, 21), use LD to refer to “limit of detection.”

Here, we address the limits of the analytical performance of two widely used PSA methods with regard to the clinical application of the measurement. Like Currie (19), we describe the performance of the assays as follows:

1) Determine the critical limit (LC), or decision limit, at which the probability of the presence or absence of the analyte is 0.5.

2) Calculate a LD at which the analytical procedure leads to a detection with a known probability of, e.g., ≥95%.

3) Calculate the PD on the same statistical basis. The PD determines in each optional interval of the analytical range the smallest difference of subsequent values that can be discriminated with a defined statistical confidence.

4) Finally, define a LLQ at which the results of a method will not exceed a predefined limit of imprecision e.g., 5%.

Materials and Methods

Patients' Sera

Sera from 120 randomly selected patients were stored frozen (−20°C) until analyzed. After thawing, the samples were split in half and analyzed simultaneously with both analytical systems in triplicate. Patients' samples used in the study were collected according to standards issued by the ethical committee in our institution.

Analytical Systems

The IMx (Abbott Laboratories, North Chicago, IL) is a batch analyzer for use with a microparticle immunoassay technique for analyses of PSA. The instrument and the PSA assay have been described in detail elsewhere (22). The ACS 180 (Ciba Corning, Medfield, MA) is a random-access analyzer for use with a two-site chemiluminescent sandwich assay. A detailed description of the ACS 180 is given in ref. 23.

Mathematical Procedure

The regression line in Fig. 1 was calculated according to the procedure described by Passing and Bablok (24). The method is based on considerations initially reported by Theil (25).

The procedure described by Sadler and Smith (3, 26) was used for construction of the precision profile, which was derived from a three-parameter variance function \[ \sigma^2(U) = (b_1 + b_2U)^2 \], where \( U \) denotes the analyte concentration and \( b_1, b_2, \) and \( j \) are parameters. We used the original program, which is available from the authors on request.

The lower and upper limits of the measuring interval were defined as the points where the arbitrarily chosen CV line (5%) intersects the confidence limit of the precision profile (not the intersection with the curve itself). The LC, LD, and PD were calculated as described previously (1). Programs for the method comparison according to Passing and Bablok, and the calculation of LC, LD, and PD, have been adapted for the Microsoft (Redmond, WA) Excel version 4.0 and are available from the authors on request.

Validation of the Procedure

To validate the procedure described here, especially at low analyte concentrations, we chose two experimental approaches. First, we used a procedure proposed by Abbott that is based on Clinical Laboratory Improvement Amendments (CLIA) recommendations. According to this procedure, one uses "raw units" instead of processor-calculated analyte concentrations for the determination of the minimum detection limit (27). In case of the IMx, "rates of fluorescence" had to be used; they are printed out with every result without any modifications of the instrument. From 10 repetitive measurements with the zero calibrator the standard deviation was cal-

![Fig. 1. Regression analysis of PSA in sera (n = 120) determined by the ACS 180 and the IMx assay systems.](image)
culated and the minimum detection limit extrapolated via regression analysis.

In a second approach, a human serum sample with a PSA value of 2.2 μg/L was diluted in six steps with a serum sample from a woman to a theoretical concentration of 0.06 μg/L (Table 1). The standard deviations obtained by 10 repetitive measurements of each dilution were compared with the standard deviations calculated by the variance function for the respective concentrations for each system.

**Results**

**Method Comparison**

The regression line shown in Fig. 1 has a slope of 1.706 (1.68–1.75). The y-intercept is -0.07 (-0.24–0.10). The results of the two methods showed an excellent linear correlation (r = 0.998) and are thus comparable, but because of the high systematic differences they are not compatible. Since the procedure used to calculate the regression line is very robust, an undue influence of single data points or the uneven distribution of the data over the measuring interval can be excluded. Values >120 μg/L are not shown in Fig. 1, but were included for the calculation of the regression line.

**Precision Profile/Limit of Quantification**

The variance function for the precision profile curves in Fig. 2 was calculated as:

- curve A (IMx): $s^2(U) = (0.023 + 0.027U)^{2.178}$
- curve B (ACS): $s^2(U) = (0.056 + 0.021U)^{2.172}$

where $U$ denotes the PSA concentration. In addition the confidence interval was determined for each curve. From these curves an interval can be determined, limited by the LLQ and the upper limit of quantification (ULQ), wherein a defined CV is not surpassed (e.g., 5%). The lower the limit for an acceptable CV is set, the higher the LLQ and the narrower the interval between LLQ and ULQ will be. The LLQ (assuming a CV of ≤5%) was 0.8 μg/L (A in Fig. 2) for the IMx and 1.9 μg/L for the ACS (B in Fig. 2), resulting from where the 5% CV line intersects the upper limits of the confidence intervals.

The point at which the upper limit of the confidence interval of the precision profile intersects, theoretically, the 5% CV line again (not shown in Fig. 2) is >1000 μg/L for both the ACS and the IMx. Therefore, for these two assays, it is not the imprecision that determines the upper limit of the quantification interval, but rather the systematic deviation from the calibration line, which the manufacturers report to be 100 μg/L (IMx) and 200 μg/L (ACS).

An upper biological reference value (cutoff) of 4 μg/L has been recommended by both manufacturers of the reagents used in this study. At that cutoff point a CV = 3.0% is expected for the ACS method, whereas a CV of =2.7% was found for the IMx assay. Both methods exhibited an excellent performance within the whole range of linearity (CV ≤ 4%).

**Table 1. Comparison of CVs determined experimentally and predicted by the variance function.**

<table>
<thead>
<tr>
<th></th>
<th>Mean, μg/L</th>
<th>Found</th>
<th>Experimental</th>
<th>Predicted</th>
<th>s</th>
<th>CV, %</th>
<th>s</th>
<th>CV, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMx</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.20</td>
<td>2.19</td>
<td>0.062</td>
<td>2.82</td>
<td>0.065</td>
<td>2.98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.10</td>
<td>1.00</td>
<td>0.026</td>
<td>2.63</td>
<td>0.038</td>
<td>3.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.55</td>
<td>0.51</td>
<td>0.035</td>
<td>6.92</td>
<td>0.027</td>
<td>5.38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.18</td>
<td>0.009</td>
<td>5.36</td>
<td>0.020</td>
<td>11.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.12</td>
<td>0.04</td>
<td>0.015</td>
<td>39.07</td>
<td>0.017</td>
<td>44.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.06</td>
<td>Nonlinear</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.50</td>
<td>3.45</td>
<td>0.118</td>
<td>3.42</td>
<td>0.108</td>
<td>3.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.75</td>
<td>1.42</td>
<td>0.042</td>
<td>2.97</td>
<td>0.069</td>
<td>4.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.83</td>
<td>0.81</td>
<td>0.032</td>
<td>3.90</td>
<td>0.058</td>
<td>7.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.41</td>
<td>0.52</td>
<td>0.042</td>
<td>8.11</td>
<td>0.053</td>
<td>10.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.21</td>
<td>0.41</td>
<td>0.032</td>
<td>7.71</td>
<td>0.050</td>
<td>12.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>0.32</td>
<td>0.042</td>
<td>13.18</td>
<td>0.040</td>
<td>15.40</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

n = 10.

**Fig. 2. Precision profiles, as determined by the method of Sadler and Smith (3).**

Sera of 120 patients were measured in triplicate. Hatched areas are the confidence limits of each curve. Inset: A and B represent the lower limits of the quantification interval, determined as those values at which the line representing a CV of 5% intersects the upper limit of the confidence interval.
were detected at 0.5 > P < 0.95, but their determination was an unavoidable uncertainty (horizontally hatched area in Fig. 3).

**Discussion**

As a logical continuation of the procedure shown in Fig. 3, a normal distribution can be constructed adjacent to each other normal distribution by using the standard deviation derived from the variance formula and overlapping the preceding normal distribution by 5% (Fig. 1). Thus a sensitivity profile can be obtained over the interval chosen (I).

The PD was determined in five arbitrarily chosen intervals to give a PD profile over the relevant part of the quantification interval. The results are summarized in Table 2. The most striking differences between the two systems were found for PSA concentrations < 3 ng/mL. Whereas the IMx assay is able to discriminate 18 values (with P ≥ 0.95) in the range 0.5–3.0 ng/mL, the PD or the ACS in the same range is 10. To better explain the meaning of the PD, additional examples are given in Table 2: For an assumed value A, a value B is given, which is the first value that can be discriminated from A with P ≥ 0.95. This means, e.g., that if a patient's result determined with the IMx was 5.0 ng/mL (A), a second value (B) measured with the same instrument would be significantly different (P ≥ 0.95) from A only if B was ≥ 5.5 ng/mL, whereas with the ACS 180 a value of 5.6 ng/mL could be discriminated from 5.0 ng/mL with a probability of 95%. For values > 45.0 ng/mL the performance of the ACS remains slightly superior to the IMx with respect to the PD.

**Validation Studies**

The minimum detection limit determined according to a method described by Abbott (based on CLIA recommendations) was 0.03 ng/mL. This value is nearly identical to the LC of the IMx, determined with the variance function.

The results of the second validation study are summarized in Table 1. Like the results of the method comparison, the mean values for each dilution step with the ACS were significantly higher than those with the IMx. The standard deviations obtained by an experimental procedure (10-fold measurements of a dilution series in six steps) were compared with the standard deviations calculated by the variance function for the respective concentrations for each system. Four (IMx) or five (ACS) of the six concentrations of the dilution series were below the respective LLQs and therefore had a CV ≤ 40%. If one takes these high CVs into account, the correlation of the experimental with the predicted standard deviations was sufficient.

**Discussion**

The comparison of patients' results determined in our study with both the ACS (γ) and the IMx (x) assay revealed a significant difference of the values (γ = 1.706x – 0.07). The deviation of the results found with the two systems may be caused by a difference in specificity of the antibodies used in the assays for the different forms of PSA (free, α1-antichymotrypsin-, α2-macroglobulin-bound) (28) and (or) the different calibrators provided for the determination of the calibration curve (29). An impact on the cutoff values should be the consequence of this finding, but, astonishingly, both manufacturers list the same value of 4.0 ng/mL for > 95% of healthy men monitored. Regarding the significant difference we observed in the method comparison, their recommendations seem inappropriate.

The LLQ reported in the literature for different PSA assays range from values as low as 0.002 ng/mL (immu-
The nofluorometric method (9), 0.03 μg/L (IMx) (7), and up to 0.3 μg/L (Tosoh) (30). These detection limits were not determined according to a standardized procedure. Most authors used repetitive measurements of either zero-calibrator or zero-PSA serum derived from female patients. The LD is then obtained by the formula \( LD = y_b + K_s \), where \( y_b \) denotes the reading of the blank sample, \( K \) represents a factor, and \( s_b \) the standard deviation at an analyte concentration of zero. For \( K \) the values described in the literature vary significantly: 2 (7), 3 (8), 6 (31), and 10 (21). Haeczel et al. (20) proposed a formula by which \( K \) can be calculated for any individual application. We propose here a procedure for the assessment of LLD and, additionally, LC and LLQ, that can be referred to as standardized inasmuch as the procedure is not restricted to immunoassays. Indeed, data from very different analytical measurements can be processed by the Sadler and Smith program without any restrictions, provided that an adequate number of different measuring points is given.

By using the data of Vessella et al. [Table 1 in (7)], one can assess the following variance function: \( s^2(U) = (0.0390 + 0.0785U)^2 \). On the basis of this function, one might extrapolate the LD according to the method described: \( LD = 0.32 \) μg/L (and LC = 0.015 μg/L), a value very competitive with Vessella's limit, which was experimentally determined. Nevertheless, the imprecision at this point is high (CV = 28%), and might not be acceptable for diagnostic decisions. The LLQ, also assessed according to the method described, is 0.4 μg/L, obviously lower than the one we found (0.8 μg/L). The difference might be explained by the fact that we take the confidence interval of the variance function into account.

A comparison of the approach described here with a frequently used procedure (27) has shown that the experimentally determined minimal detection limit (0.03 μg/L) corresponds well to the LC determined with the variance function (0.03 μg/L). But as discussed in more detail below, LC should be used neither as LD nor as LLQ. Not unexpectedly, the LD values found in this study for the ACS (0.18 μg/L) and the IMx (0.07 μg/L) were slightly higher than those published by others [ACS 0.13 μg/L (32); IMx 0.03 μg/L (7)].

The second approach to verify the use of the variance function (summarized in Table 1) showed that, despite the fact that most of the concentrations of the dilution series were below the limit of quantification, one could obtain a reasonable correlation between the experimentally determined SDs and the ones calculated by the variance function.

From a theoretical point of view, only an inclusion criterion (and not an exclusion criterion) can be verified when one is defining an LD. In analytical chemistry, and especially in clinical chemistry, the hypothesis that an analyte is not present in the system under investigation cannot be verified. The presence of an analyte, on the contrary, can be proved (e.g., in instrumental analysis) when the measuring signal is distinguishable from the noise of the system. If the signal is not observed, one can state that the presence of the analyte cannot be proved by the method used, but one cannot state that the analyte is absent. Therefore, the procedure of extrapolation of the analytical variability at low concentrations to a theoretical zero-concentration is advantageous compared with the experimental measurement of so-called zero-standards.

Quantitative results should be reported to the clinician only if the laboratory can guarantee an adequate analytical precision for the range in which the determinations have been performed. In principle, LC < LLQ < DCO (diagnostic cutoff) is valid provided that an appropriate analytical method is available. If there is no appropriate method available for a given analytical diagnostic problem, and if the quantification is performed below a clearly defined LLQ, the value of the result obtained is reduced by the largely undefined and high degree of imprecision. Despite the fact that CV can be calculated for very low analyte concentrations, the use of CVs in the very low range seems to be questionable. Measurements between LD and LLQ are problematic because they are performed in an analytical gray zone and exhibit in most cases high variability.

The question of which CV might be diagnostically acceptable is answered controversially in the literature [see, e.g., Skendzel and Barnett (33) and Fraser (34)]. Several reports on this issue, including studies of preanalytical, analytical, and biological (intraindividual variations, have proposed valuable approaches for the determination of an adequate CV for each analyte (35-39). In any event, different medical situations (screening, monitoring, etc.) require different analytical efficiencies.

The PD, introduced by Gautschi et al. (1), describe the analytical sensitivity of a method for a certain interval of the measuring range. For a given overlap (e.g., 5%) an exact number of normal distributions can be calculated for such an interval. Table 2 shows that the IMx system is superior in the lower range (<3.0 μg/L) whereas the ACS exhibits a higher PD for values >10 μg/L. But in general the differences of both systems will respect to the PD are small.

Despite the fact that differences in the PSA values >90 μg/L are diagnostically less critical than those in the range of 4 or 10 μg/L, it is important to know for each assay whether differences of subsequent measurements reflect a significant increase or decrease. For an assumed value of, e.g., 100 μg/L determined with the IMx, a decrease or increase of a subsequent result is significant only if the second result is ≥112.3 or ≤89.8 μg/L, respectively. The corresponding values for the ACS are ≥109.5 or ≤91.4 μg/L, respectively.

In our study, at low analyte concentrations (≤2 μg/L) the performance of the IMx assay was superior with regard to LC, LD, and PD compared with the ACS. At analyte concentrations near the cutoff, the performance of the assays was very similar. At analyte concentrations >15 μg/L the ACS exhibited a slightly better PD.

Our argument for strict distinctions between LC, LD LLQ, and DCOs is represented in Fig. 5. On the one
and, the LC (taken for LD or LLD by many authors) is a fictitious exclusion criterion with respect to a supposed zero concentration, whereas LC and LD are, on the other hand, objective and reproducible findings if genuine patients' material is used as a basis for the calculation (Fig. 3).

Modern analytical systems, including the IMx, very often do not print out negative results of blank samples. A correct calculation of SD at the zero concentration is therefore not possible in most cases. The assessment by extrapolation of the variance function seems to be the only reliable and for every laboratory practicable way.

In the PSA example described here, it is difficult to judge whether the state of the art (for the determination of low PSA concentrations) is optimal with respect to diagnosis and intervention. With more sensitive methods, earlier detection of increased PSA might be possible and, therefore, better therapeutic results might be obtained. In view of the present results we recommend the following regime for monitoring prostatectomized patients: The first value > LC indicates the need for more frequent medical tests, but not more. The first > LD value to be found indicates that the “PSA concentration is increasing,” but gives no numerical declaration. If later a value > LLQ is registered, then the exact PSA concentration can be reported to the physician. Changes (increases or decreases) in the concentration are considered significant only if they amount to the differences defined by the PD. With knowledge of the potential biological variation in PSA concentrations, the physician will then be able to decide what a change in a patient's concentration of PSA (or any other analyte) implies: a true increase or decrease, or a random error.

We thank G. H. Hungerecker and Y. Neumaier for their excellent technical assistance and K. Rentzch for critically reading the manuscript.

References