Analytical Performance Goals for Measuring Prostate-Specific Antigen

Herbert A. Fritsche1,3 and Richard J. Babaian2

We have assessed the feasibility of using fixed-limit criteria based on medical relevance and biological variation for evaluating the analytical performance of the prostate-specific antigen (PSA) test. The estimated within-subject variation of serum PSA is on the order of 10–20% at clinical decision points. The calculated performance goals of 5–10% CV are attainable with current immunoassay technology and agree with precision goals based on clinical experience and the current clinical use of the test. However, new clinical applications of PSA may require a degree of analytical performance that current methods may not be able to provide. The PSA model demonstrates the need for biologically based fixed-limit criteria for all tumor-marker tests.

Indexing Terms: biological variation • within-subject variation • proficiency testing • fixed-limit criteria • tumor markers • cancer

Tests for tumor markers such as prostate-specific antigen (PSA) in serum are becoming increasingly important for the diagnosis and clinical management of the cancer patient. These tests are useful additions to the clinical procedures commonly used to establish the diagnosis of cancer. When these tests are used in a serial manner during follow-up, they can provide an early assessment of treatment response. For those patients who achieve a complete remission, patient monitoring with tumor markers can give an early signal of disease recurrence. The “lead time” provided by the early detection of disease recurrence is especially valuable when effective therapeutic options are available for extending survival time. More recently, it has been shown that some tumor markers may have a significant role for cancer screening in those individuals who are at high risk for developing neoplastic disease.

The increasing medical relevance of tumor-marker tests requires that clinical laboratories utilize stringent quality-assurance measures to ensure high-quality, long-term analytical performance. This is accomplished with quality-control procedures and interlaboratory comparisons via proficiency-testing programs. Quality-control procedures ensure the “day-to-day” precision of laboratory measurements by requiring that the test results for control sera fall within the mean ± 2 SD limits that have been established by that laboratory. Proficiency testing ensures overall analytical performance by requiring that the laboratory measurements of a reference preparation agree with a target value. Significant deviation from the target value indicates method bias, poor test precision, or a combination of these constant and random errors. Historically, agreement with the target value has been gauged by peer performance based on the mean ± 3 SD limits, and such is the case currently for the tumor-marker tests. More recently, fixed-limit criteria have been developed for some analytes, which are based on the medical significance of the test results.

Here, we explore the possibility of assigning fixed-limit criteria for evaluating proficiency-testing performance of tumor-marker tests. We have selected the PSA test as a model for tumor markers because of its multiple clinical roles as a diagnostic aid, indicator of therapy response, and monitor for recurrence of prostatic cancer.

Prostate-Specific Antigen

PSA is a glycoprotein produced specifically by epithelial cells of the prostate gland (1). It is composed of a single 240-residue polypeptide chain and is 6.9% carbohydrate; the resulting molecular mass is 34 000 Da (2, 3). A serine protease, PSA shares sequence homology with the kallikrein family of proteases (4, 5). In the circulation, the predominant form of PSA is an 80–90 kDa complex with α1-antichymotrypsin (6). PSA is present at low concentrations in the sera of healthy men, and at increased values in ~25% of men with benign prostatic hypertrophy (BPH) and in most patients with adenocarcinoma of the prostate (7). Although the concentration of serum PSA increases with tumor burden, PSA evaluation cannot be used to establish the stage of the disease for the individual patient (8, 9). Serial PSA values reflect the clinical course of prostate cancer. After prostatectomy, if serum PSA values fall to “undetectable” concentrations (values <0.2 μg/L), this is suggestive of complete resection of the tumor (10). Patients with residual tumor after surgery, as indicated by measurable PSA values, require additional therapy (11). Serial PSA values that show an increase after curative therapy are suggestive of local recurrence or metastatic progression of the disease (11, 12).

Medically Relevant PSA Values

The analytical performance for any assay should be optimized at the analyte's decision points (13). For the PSA test, three decision points are required—one each for distinguishing between BPH and cancer, for assessing curative treatment, and for detecting disease recur-
rence. Each of these decision points can be defined by statistical methods, such as receiver-operating characteristic curve analysis, which allows the selection of a decision value with optimal clinical sensitivity and specificity (14). However, clinical objectives help to define the desired clinical sensitivity. For example, an optimum decision limit that maximizes clinical sensitivity and specificity could be lowered to a value that correctly classifies the desired number of cancer patients. If the objective of this classification is to direct patients to the next diagnostic modality, the increased number of false-positives observed with the increased number of true-positives does not result in a significant medical problem. Such is the case with the PSA test when it is used as part of the diagnostic profile including digital rectal exam (DRE) and transrectal ultrasonography (TRUS). In our experience, the cancer detection rate in men with an abnormal DRE and (or) TRUS increases with the degree of increase of serum PSA (17.5% for PSA <4.0 μg/L, 68.7% for PSA >4.0 μg/L, 83.3% for PSA >20.0 μg/L). In addition, a serum PSA value >4.0 μg/L improves the positive predictive value of the DRE or TRUS when results for either or both of these tests are abnormal (15). Thus, PSA >4.0 μg/L suggests a significant risk for prostate cancer and identifies those patients who require further evaluation and possible biopsy.

A PSA decision limit for assessing curative treatment is defined by the detection limit of the assay. Because PSA is produced only by epithelial cells of the prostate gland, a total prostatectomy should result in a “zero” concentration of serum PSA, if the disease is confined to the prostate. In one study (16), 93% of 36 patients who had postprostatectomy PSA values <0.2 μg/L remained disease-free during 6 to 70 months of follow-up. In contrast, all of the patients whose PSA values exceeded 0.4 μg/L developed recurrence. Other studies have confirmed the postprostatectomy decrease of PSA below the assay detection limit in 80–90% of organ-confined prostate cancer patients and the appearance of measurable PSA values with disease recurrence (17, 18). These observations have led several groups to develop assays that have lower detection limits than the conventional immunoassays (19, 20); however, clinical application of these assays will require consideration of the possibility of remnants of prostatic tissue remaining after surgery and the PSA production capability of the periurethral glands (21). Until these issues are resolved, it would seem appropriate to use a decision value just slightly above the assay’s detection limit (0.3 μg/L) for assessing curative therapy. In our clinical experience, 0.5 μg/L is a reasonable estimate of the decision value for detecting disease recurrence after radical prostatectomy.

Fixed-Limit Criteria and Biological Variation of PSA

A widely accepted approach for defining analytical imprecision is based on the biological variation of the analyte (22). One estimate suggests that the analytical precision, defined as the coefficient of variation (CV), should be ≤1/2 of the within-subject CV (23). Few published reports address the biological variation of PSA and even fewer data are available for assessing within-subject variation at the medical decision values of 4.0, 0.5, and 0.3 μg/L. Schifman et al. (24) determined the biological variation of PSA measured twice daily for three days for 10 prostate cancer patients. The PSA values ranged from 4.1 to 428 μg/L and the respective within-subject CVs ranged from 9.4% to 3.1%. The 10% estimate of biological variation for PSA at values near 4.0 μg/L is substantiated in a report from Maatman (25). Only one study (26) has addressed the biological variation of PSA at values <1.0 μg/L. In that report, for PSA values determined four times over a 24-h period, the within-subject variation for BPH patients and normal subjects was ~20% for the mean values of 0.6 and 0.2 μg/L, respectively. If these limited data are accepted as reliable estimates of biological variation, the analytical CVs should be ≤5% at a decision value of 4.0 μg/L and 10% for decision values of 0.3 and 0.5 μg/L.

The long-term precision (CV) for the PSA test in our laboratory, determined with commercial quality-control sera at mean values of 2.7, 0.97, and 0.81 μg/L, was 4.7% (n = 396), 8.3% (n = 259), and 8.2% (n = 487), respectively. These data suggest that the long-term precision of the PSA test in our laboratory would meet the fixed-limit criteria based on the biological variation data. A review of the data from a recent College of American Pathologists' proficiency-testing survey indicates that most laboratories would achieve acceptable performance if the biological-variation-based fixed limits were used. In that survey, with 322 laboratories reporting, a CV of 9% at a mean value of 2.28 μg/L was achieved. A calculated fixed-limit range based on a CV of 5% is 2.04–2.48 μg/L (2 SD) and 1.92–2.60 μg/L (3 SD). This compares well with the peer 3 SD range of 1.68–2.88 μg/L. Finally, a generally accepted empirical criterion for assessing clinically significant changes of serial tumor-marker values is a change of 25%. The CV calculated from the 25% change at a mean value of 4.0 μg/L is 12.5%, which agrees with the 10.0% CV biological variation at that decision point.

Considerations for the Future

The use of fixed-limit criteria for assessing analytical performance has been validated for a number of analytes and appears to have promise for the tumor-marker tests. The clinical significance of tumor-marker tests continues to increase, and their role in clinical decision-making is becoming more important. As a result, the clinical demands of these tests may soon exceed current analytical capabilities and laboratories will need guidelines for analytical performance that extend beyond peer-performance criteria.

This preliminary assessment of biological-variation-based fixed-limit criteria for evaluating the analytical performance of the PSA test suggests the feasibility of this approach. However, it is clear that performance criteria will need to be defined for all decision points for each tumor marker and that the biological variation of the analyte in cancer patients should be determined at
these medical decision points. Standard reference materials would help minimize method bias related to calibration differences and perhaps that attributable to different molecular forms of the analyte. It would also seem appropriate to consider a standard approach for reporting within-subject variation, i.e., average CV, maximum CV, or pooled CV.

As previously stated by Batsakis (27), the analytical goal for any test should be a composite of total analytical error, biological variation of the analyte, and medical judgment. These performance goals must meet medical needs, should be flexible to accommodate new knowledge, be attainable to most laboratories to ensure a national standard of patient care, and enhance the accuracy of medical decision-making. The cooperation of the clinician, laboratory scientist, and industry is of paramount importance in accomplishing this endeavor.

References


Discussion

Neal Dawson: The PSA problem is an excellent example of both a tough clinical issue and some of the difficulties with media coverage. The media claims sometimes overstretch the limits of what tests can actually do. I was glad to see you suggested that PSA is not a good screening test. Unfortunately, as a clinician, I get mail from many organizations or people who want to sell me things, who want to treat PSA like a screening test.

This problem highlights several issues with respect to methodological issues for diagnostic test development and evaluation. The first concerns selection of study patients. Most people studying this particular problem are urologists. Yet, most men who are at risk for this particular disease do not see urologists every year. We need to think about the reasons why men go to urologists. How are the men who urologists see different from the vast majority of asymptomatic men who are at risk for this particular disease?

Second, when a test like PSA is used in usual clinical practice, it is at significant risk for verification bias or work-up bias, in that not everybody gets the same evaluation. In fact, the rest of the evaluation depends on the test results. That is how usual clinical practice works. When we set up studies to evaluate these sorts of tests, it is important to make sure (to the degree we can) that everybody gets the same evaluation. Again, usual clinical practice would generally limit the number of people who get biopsies—that’s one problem with respect to detection of disease. And, some people who have cancer are going to have a benign course, which also may affect detection.
Finally, I want to respond to an earlier statement that sensitivity and specificity are independent of prevalence. I understand entirely that mathematically they should be or perhaps are. Empirically, however, sensitivity and specificity can be shown not to be independent of prevalence. The difficulty, I believe, is that there is covariation between comorbidity and prevalence. Groups of people who have a low prevalence of disease have a different disease burden or a mix of diseases than those who are at a higher risk of disease. This can be seen with prostate cancer. Men who are at higher risk of prostate cancer are older; older men tend to have larger prostates (and thus may have false-positive PSAs). Older men also have lots of other things going wrong—more lung disease, more heart disease, more diabetes, more hypertension. They have many conditions that may influence how well this test performs with respect to its sensitivity and specificity. The values for sensitivity and specificity may be different in a younger cohort with smaller prostates, lower comorbidity, and lower prevalence of disease. Could either speaker comment about that?

Richard Babaian: Remember, PSA is not a standalone screening test. The upcoming American Cancer Society recommendation concerning guidelines for early detection of prostate cancer is going to include PSA and a digital rectal examination.

It is very important to educate the primary-care physicians concerning PSA. I don't think the primary-care physician should be making the judgment about whether a cancer is clinically significant or insignificant—urologists have enough trouble trying to do that. In any event, that sort of judgment needs to come from the next level of specialization, the urologist. The potentially benign course of the disease should not impact our detection of the disease. The disease is the second-most common cause of cancer deaths in American men, and its clinical expression is as the most common cancer among American men. This is not to be confused with the prevalence rate of prostate cancer, which is that 1 of every 10 men will have prostate cancer sometime during his lifetime; the statistics about its being the number one cancer and the second cause of cancer death are related to its clinical incidence, not its prevalence rate. That needs to be carefully understood.

Neal Dawson: For the reasons I have stated, we need to very seriously consider whether we have the answer in hand. I don't believe we have sufficient data to take the steps being suggested by the national bodies. Now, I'm not saying that I know for a fact that the recommendations are wrong; I am suggesting that there is still considerable uncertainty about the appropriateness of screening for prostate cancer. The important piece that is missing is the relationship between these practices—be they clinical recognition or screening practices—and outcomes. Frankly, we don't have a very clear idea about this relationship. We have a pretty good idea about the clinical course once the affected people are recognized, but the relationship of outcome to recognition by screening for the entire group of people who are at risk is less well known.

Richard Babaian: The population that was studied in the National Prostate Cancer Detection Project was a screening population. That is, they were asymptomatic individuals who were no different from people who would go to their family-care physician or their primary-care physician. The data are striking, being able to detect clinically significant cancer in an organ-confined state. That is one of the reasons that the recommendations are being put forward.

Outcome is a key issue for screening, there is no doubt, but for prostate cancer I wonder if we are asking the right question about the outcome. Let us put it in terms of the person who comes to the physician's office, is treated for prostate cancer, and has an improved survivorship. There is no doubt that early detection improves survivorship. The question is, is it going to change mortality due to the disease? If someone has 10 more years of good-quality life, I think that is important, particularly in prostate cancer, where the tumor can be slow growing. I'm sure it makes a difference to that man if he dies of his prostate cancer at age 70 or at age 60, especially if he had a good quality of life for those 10 years. So, I'm not sure for prostate cancer if mortality is the right question.

John Ross: The basic issue has to do with a desire for an individual patient, with the help of his or her physician, to be able to weigh the risks and benefits of a given approach at a given time. To do that, we have to have a much better idea as to what, in fact, those risks and benefits are for individual people who are likely to have different courses of disease.

I would like to mention two things, purely from the analytical goal-setting point of view. Dr. Fritsche appears to suggest two different models. One model appears to be based on a clinically significant change in variability based on an individual's own baseline. This is an interesting perspective; however, it is not a perspective that is widely available in clinical chemistry. It is very useful, and people in this audience have published extensively on the usefulness of individual-specific normal ranges. However, a goal-setting process that develops from that perspective would be limited and is quite different from a goal-setting perspective that involves a decision limit. Incidentally, I would suggest that decision limits are another model in which bias, as well as imprecision, becomes critical in developing performance specifications for the analytical quality of the measurement process.

Second, a previous slide showed a confusion between intralaboratory precision and interlaboratory precision. A CAP survey CV includes the bias component of the analytical error, whereas the intralaboratory estimates will not include that same component. To compare them will lead to the wrong conclusions.

Herbert Fritsche: I wasn't suggesting that we compare the results of the precision studies. But I think that you can use a fixed-limit criteria for your day-to-day quality control, in much the same way you use it to evaluate the
performance for 322 laboratories on the CAP survey, can you not?

John Ross: I don't think so. The estimate dispersion among those 322 laboratories includes the fixed- and short-term bias and is peculiar to those laboratories at that particular analytical run. Now, exactly how that relates to some estimate of intralaboratory imprecision is a complex issue; it looks simple, but it is not. It depends on whether you are referring to the within-run component or some longer-term component of intralaboratory variation. But to fold those back in will always lead to a smaller estimate of the intralaboratory component than the interlaboratory component, no matter what your other assumptions are.

Herbert Fritsche: I believe fixed-limit criteria are useful and necessary for evaluating your own laboratory's performance, independently of your performance in comparison with other laboratories.

Callum Fraser: The studies quoted take multiple samples over a very short time. Therefore, the values over time will quite likely be serially correlated and the estimate will not be very applicable to the main use of the test, which is monitoring, when samples are taken over longer periods. Browning and McFarlane have published quite extensively on the biological variation of tumor markers [J Nucl Med Allied Sci 1990;34(Suppl):89–91] including carcinoembryonic antigen (CEA), CA 15.3, PLAP, and PSA. PSA in healthy individuals in samples taken at longer-term intervals has about 18% variation. If the critical difference between two results is used to generate analytical goals, it is very important to think in specifics rather than in global terms. PSA has a variability of 18%; therefore, a change of 25% will actually have a very low probability of significance. The same change (25%) will have different probabilities of being significant for CEA, which has a biological variation of 9%, or CA 15.3, for which the variation is 5%. Education of the clinician is warranted about the significance of changes for different markers; they should not be put into the same basket, because they are very different.

It is important that tumor markers are very individual. It is not surprising, because individuals are so individual, that markers that change within the reference interval can still indicate a significant change. Moreover, individuals can change over time from above to below reference intervals (or cutoff points), or vice versa. Whether the change is significant or not depends on analytical, preanalytical, and biological variation, and this should be taken into account when thinking about the change.

Richard Babaian: We have not observed an 18% variability in PSA within individuals in our practice. That seems a bit high and, from the data reported in several articles in this country, I don't think that is within acceptable standards in the US. There are reasons for variability in PSA values, but some of them have to do with manipulation and infections and so on—the things we try to exclude when we are looking at variability. If we are talking about the real relative and absolute changes in PSA, and I don't think it is as high as 18%.

Mario Werner: This and the two preceding presentations were not only scholarly and convincing but also aptly chosen to contrast three model situations. The models differed in the character of the target for testing, and represented three broad clinical categories—cancer, endocrinological disease, and risk factors. Can one address the three specific situations chosen as models by deriving the necessary analytical quality from some physiological interval, such as the normal range, or must one work with a target of clinical relevance? To me, to ask the question seems to answer it. Cancer presents obvious targets: detection, monitoring, recurrence, and so forth. Endocrinology also presents clear-cut targets, and the precision required to deal with them is not defined by anything happening in healthy people but by issues linked to diseases. When it comes to cholesterol, in contrast, the required precision indeed has to be derived from some physiological benchmark, but the question arises as to whether hypercholesterolemia is a true target, in that it has only a possible statistical link to cardiovascular disease, which may or may not be applicable for a particular patient. Thinking about an analogous situation, Joslin 40 years ago proposed similarly tight and uninterrupted control of blood glucose, but this is no longer a universal standard today. Several changes have also occurred in the use of blood lipid data. Everything being said today about the prediction of pathology has first been said about total blood cholesterol but apparently is no longer wholly true. Subsequently, similar claims were made about high-density lipoprotein (HDL) cholesterol, but that is apparently not totally true any longer either. However, we are now assured that blood low-density lipoproteins are going to provide the reliable indicator. To the question, "Where is the evidence?" the answer presented was a synthetic case of a hypothetical man who lost 30 pounds [13.5 kg] but whose HDL wouldn't change, and who, therefore, poses a diagnostic quandary. Is the mere fact of such an important weight loss not going to deeply influence his survival chances by itself? With a real patient, would we not also want to know the blood pressure as well as many other facts before committing to a prognosis? So, we come back to this: If you really want to define clinically relevant precision, you need a specific clinical target, where the cost of missing the target is calculable. Allowable precision is based on intraindividual variability when no such target exists and the measured variable has only some statistical relation to pathology.