Assessment of Predictive Values of Tumor Markers

Joseph E. Roulston

1. Introduction

Reviewing the literature, it would appear that tumor markers have often flattered to deceive. Early promise does not often seem to be borne out in extended trials. Despite apparently high specificity, very few markers are capable of assisting in a screening process.

This brief review attempts to put the roles of tumor markers in perspective and explain how their misapplication has led to misunderstanding of their potential value in a clinical context. It also considers the theoretical basis for their use and highlights how misunderstanding of these can lead to flawed studies and application.

Cancer has been known to mankind since ancient times. There is an early Egyptian papyrus describing how one should differentiate between breast cancer and mastitis. The ancient Greeks and Romans also have left us with writings in which various treatment options are discussed (1). Disease processes and causes were not well understood however; the humoral pathology established by the ancient Greeks of the school of Galen in the 2nd century AD was to survive virtually intact until the mid-19th century. It is perhaps all the more remarkable then that the first tumor marker—Bence Jones’ protein in multiple myeloma—should come to light in what was still, by and large, the prescientific medical culture prevailing in 1845.

Multiple myeloma was fully described and named by von Rustizky (2) in 1873, but it was Kahler (3) who related the disease to Bence Jones’ proteinuria and thereby brought a specific tumor marker to medical attention, a marker that is still used to this day to assist in diagnosis.

Despite the lesson of Bence Jones’ protein, in which a marker specific for a particular cancer was discovered, many researchers still sought a general test.
for early diagnosis of all cancers. Homberger (4) reviewed more than 60 tests that had been suggested in the previous 20 yr (1930–1950). Many of these tests were based upon the physicochemical properties of serum proteins and sought to show a difference between precipitation of serum proteins from normal subjects and cancer patients.

With the benefit of hindsight, it is easy to write off such efforts as misplaced; the biochemical techniques available were crude and not always applied with logic. Bodansky (5) points out the problems with many early studies. Technically, the tests were deficient because they were based on a gross and nonspecific measurement—the change in a large fraction of the serum protein pool. Second, these investigations were usually carried out in samples from patients with advanced disease, whereas control groups of similarly aged patients with serious nonmalignant diseases were not studied. When these controls were looked at later, the false-positive rate was as high as the true-positive rate in the neoplastic group.

Apart from technical shortcomings, there is also a major assumption in the presupposition that cancers will produce some unique feature that non-neoplastic diseases will not, and for this, there is not a shred of evidence (6).

As the biochemical tools and techniques available have grown ever more sophisticated, it has enabled more precisely focused studies to be conducted. The advent of immunoassay techniques in the 1960s and their refinement during the following decades with nonisotopic labels and, especially, the development of monoclonal (“hybridoma”) technology has brought levels of analytical sensitivity and specificity that were orders of magnitude better than those available to previous generations of researchers.

2. Theoretical Considerations

In order to assess and apply tests in an appropriate and discerning manner, it is necessary to consider what the aims and objectives are and how one monitors and assesses one’s efforts. At first thought, it appears very simple. First, it is intended to apply a test to discriminate between the normal and the diseased subject, to assist in diagnosis and possibly to screen populations for occult disease. Second, one may wish to apply a test to monitor the course of the disease in a noninvasive way in order to assess the efficiency of therapy, to watch for drug resistance, and to predict outcome. Third, one may wish to monitor patients in remission to ensure that they remain disease-free and to get a valuable lead time to relapse.

In order to achieve these aims, several points must be made clear. First, one must have confidence in the analytical accuracy and precision of the test(s). However, in order to translate the analytical data into clinically meaningful information, it is essential to be aware of what the objectives are. “Is this result
normal?” is a question often asked by a requesting clinician and it is worth considering, at the outset, what the word “normal” may or may not mean.

### 2.1. What Is “Normal”?

The first problem presenting to workers in clinical medicine is the statistical definition of normal because it is widely misunderstood and even more widely misapplied. Gauss’ law of errors applies to repeated measurements on the same subject or object, not a series of measurements of the same analyte in different subjects. Gauss’ law proposes that if the same measurement were repeated over and over again in the same subject, the results’ spread would fit a bell-shaped distribution symmetrical about the mean. Abnormal results may then be defined as those outside the 95% confidence limit—in other words, the 2.5% of values at the top and bottom end of the range. There is, however, no a priori reason why this law of distribution should apply to measurements in more than one subject; it was never derived to describe the distribution of a variable (disease related or otherwise) in a population of subjects. Although it is common practice in laboratories to define a reference range for an analyte as being the limits within which 95% of the healthy population’s results fall, these limits per se give no indication of morbidity or mortality. Indeed, by definition, 5% of this population will be “abnormal” although disease-free if we assume a 95% (i.e., mean value plus or minus two standard deviations) reference range. It also follows that the more tests performed on each specimen, the greater the likelihood of at least one of the results being “abnormal”—from 5% for 1 test, to 40% for 10 tests (the chance of 10 tests on a sample all being “normal” is $0.95^{10}$, which is 0.6 or 60%, leaving a 40% chance for an “abnormal”). The percentage error figure rises to 99% $[(1–0.95^{90}) \times 100\%]$ for 90 tests.

It is for this reason that most laboratories today eschew the phrase “normal range” and prefer the alternatives “reference range” and “referent value,” in order to make clear that the range or cutoff cited is not of necessity a range or cutoff that encompasses or defines the limit of the values of the analyte in all disease-free and excludes all diseased subjects. Often, 95% reference ranges, based on the mean value plus or minus two standard deviations, are employed as the reference limits because they have been found empirically to provide cutoffs at clinically useful and discriminant values.

For tumor markers, however, there is less concern whether a reference range based on a symmetric distribution is ideal; in practice, the optimal cutoff value is sought, a point that discriminates “normal” from “elevated.” There is no lower limit to the “reference range.”

In order to establish this cutoff value empirically, it is necessary to discover the value that discriminates the best between disease and nondisease—in other words, produces the fewest misclassifications. In order to do this, large num-
bers of measurements in the disease group under study and a suitably matched control population must be made. Inevitably, there is an overlap between the range of results produced by the control group and the range produced by the diseased group. There will, therefore, be some false-positive results (elevated analyte in control subject) and some false-negative results (“normal” result in diseased subject), and just how many depends on the test and population in question. In order to proceed further, therefore, it is vital to determine how good the test is in an objective manner.

2.2. How Good Is the Test?

2.2.1. Sensitivity and Specificity

The two most usually applied criteria in order to assess a test are those of sensitivity and specificity. Sensitivity is a measure of how good a test is at picking up the disease in question by giving a positive result. It is expressed in a population with the disease as the number giving a true-positive result divided by the sum of true-positive and false-negative results—in other words, what percentage of the diseased cohort was identified correctly by the test.

It is obvious that a test that is 100% sensitive will score perfectly. A test that is 90% sensitive, however, will generate 10 false negatives in each 100 positives.

As well as correctly identifying the presence of disease, a good test must also correctly classify the disease-free subject by giving a negative result. The measure of a test to so discriminate is called the specificity. This is established by measurements in a disease-free population, and specificity is defined as the number of true negatives divided by the sum of true negatives plus false positives.

It can be seen that sensitivity and specificity are entirely “test-based” parameters; they take no account of the prevalence of the disease in the population, the sensitivity is calculated by study of a group who are all disease positive, and specificity is calculated from a group who are all disease-free. This, as will become apparent, is a serious limitation to the application of these parameters because disease prevalence has serious effects upon the clinical usefulness of tests in certain circumstances.

Furthermore, the choice of cutoff, which effectively determines the sensitivity and specificity, cannot improve both sensitivity and specificity simultaneously; moving the chosen cutoff point to a higher referent value will increase specificity, but correspondingly reduce sensitivity. The optimal choice of cutoff, therefore, depends on whether it is deemed more desirable to optimize sensitivity at the expense of specificity or vice versa, and this consideration, in turn, is influenced by the disease prevalence in the population under study.
2.2.2. Incidence and Prevalence

By definition, incidence relates to the frequency of occurrence of an event and is therefore a rate per unit time. For a disease, the incidence rate is the number of new cases per 100,000 of the population per year. The prevalence, by contrast, is the number of patients per 100,000 of the population who have the given disease at the time of the study; therefore, prevalence is a snapshot of the status quo.

The incidence of epithelial carcinoma of the ovary is of the order of 15 per 100,000 per year. If the average duration of the disease is 5 yr, it follows that the prevalence, assuming a steady-state situation in the population, must be 75 per 100,000. As a general rule, therefore,

\[
\text{Prevalence} = \text{Incidence} \times \text{Duration}
\]

The clinical usefulness of a test in a given situation will depend on the prevalence of the disease in the cohort under study; high sensitivity and specificity, although vital, are not enough to guarantee “usefulness.” For example, a test that was 100% specific and 99% sensitive seems to have impressive credentials, but it would fail dismally as a screening test for ovarian cancer. Screening 100,000 women would yield 99% out of the positives, (i.e., 74 or 75 women), which is an acceptable “pick-up rate,” but it would also generate 1% false positives (i.e., 1,000 nondiseased women). Therefore, a positive test result would correctly identify disease presence in less than 7% \([75/(1000 + 75) = 6.98\%]\) of the subjects studied. Therefore, it is necessary to use assessment procedures that take into account the prevalence of the disease in the population under study.

2.2.3. Bayes’ Theorem and the Predictive Value Model

In 1975, Galen and Gambino (7) introduced the predictive value model to clinical laboratories. The theoretical basis was hardly new; coming as it did from a posthumous publication in 1763 (8). What Bayes’ theorem allows is the calculation of the \textit{a posteriori} probability of disease being present in an individual given that the patient has a positive test result. By definition, the \textit{a priori} probability that a patient will have the disease (i.e., before the test) is equal to the prevalence of the disease. From prevalence, sensitivity, and specificity, Bayes calculated the \textit{a posteriori} probability—the so-called predictive value of a positive result or positive predictive value.

Let sensitivity be \(a\), specificity be \(b\), and prevalence be \(p\), then one can describe the positive predictive value (PPV) as follows:

\[
\text{PPV} = \frac{pa}{pa + (1 - b)(1 - p)}
\]
This simplifies to

$$PPV = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

because $pa$ is the prevalence of the disease multiplied by the sensitivity of the test for the disease (i.e., the true positive).

Similarly, $(1 - b)(1 - p)$ is the prevalence of nondisease multiplied by the probability of a positive result in such a disease-free person. $1 - b$, which is $1 - \text{specificity}$ is sometimes, albeit incorrectly, referred to as the false-positive rate.

The benefit of the predictive value model is apparent immediately; if a test has a 95% PPV in a given area of use, then the clinician may assume that in a patient with a positive result, there is a 95% chance that the patient has the disease. The same conclusion cannot be made from sensitivity and specificity values, as they take no account of disease prevalence. Table 1 shows how the predictive value of a positive test varies from virtually zero to 100% as a result of changing disease prevalence even when sensitivity and specificity are high. This gives an important insight to screening procedures; where disease prevalence is low, it is necessary to have tests with greater than 99% sensitivity and specificity to achieve an acceptable positive predictive value.

### Table 1
Positive Predictive Values as a Function of Prevalence

<table>
<thead>
<tr>
<th>Disease prevalence (%)</th>
<th>Test sensitivity and specificity = 95% PPV</th>
<th>Test sensitivity and specificity = 99% PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>0</td>
<td>2.0</td>
</tr>
<tr>
<td>0.1</td>
<td>1.9</td>
<td>9.0</td>
</tr>
<tr>
<td>1.0</td>
<td>16.1</td>
<td>50.0</td>
</tr>
<tr>
<td>2.0</td>
<td>27.9</td>
<td>66.9</td>
</tr>
<tr>
<td>5.0</td>
<td>50.0</td>
<td>83.9</td>
</tr>
<tr>
<td>50.0</td>
<td>95.0</td>
<td>99.0</td>
</tr>
</tbody>
</table>

3. Screening for Disease

Screening has been defined as “the presumptive identification of unrecognised disease or defect by the application of tests, examinations, or other procedures that can be applied rapidly” (9). By definition, therefore, a screening test is applied to asymptomatic subjects and is not diagnostic per se, confirmatory tests being required. The idea that early warning leads to a better outcome is not easily translated into a practical program. The economic difficulties of testing large numbers of apparently healthy individuals in order to pick up a small number with the disease are enormous. Second, there are diffi-
cult ethical considerations when one is investigating healthy subjects without symptoms or any substantive probability of finding disease.

### 3.1. Population Screening

The oncology literature contains many reports of apparently promising markers that fail subsequently to claim a routine clinical role. There are many reasons that contribute to this, but the commonest is overextrapolating or illogically applying the results. Consider a study in which an investigator tests a novel tumor marker for a particular cancer that has a prevalence of 100/100,000 in the general population and finds that in 100 patients with the tumor under investigation, 99 have a positive test result; that is, the test has a sensitivity of 99%. Equally, when tested on 100 disease-free subjects, only 1 is test-positive—a specificity of 99%. Owing to this excellent discrimination, it is decided to introduce the test as a screen in the general population in order to detect this tumor at an earlier stage to improve therapeutic efficacy and patient outcome. The results are disastrous; the test appears to have lost its earlier discrimination and is generating lots of false positives—Why?

In the pilot study, the disease prevalence was 50% by design; there were 100 patients and 100 controls and the positive predictive value was 99%. In the screening exercise, the prevalence would be 100/100,000, which is 0.1%. Therefore, as well as correctly identifying 99 out of the 100 true positives, the test will also under these circumstances misclassify 1000 as false positive, giving us a positive predictive value of 99/(99 + 1000) (i.e., 9.0%). In other words, a test that in the pilot investigation yielded 99% correct results, gives, in a screening situation, a 91% a posteriori probability that elevated results are not associated with the disease. The marker sensitivity and specificity remain unchanged, the fall in positive predictive value from 99% to 9% was entirely caused by the change in prevalence in the cohort under study from 50% to 0.1%.

If a test is genuinely and completely useless (i.e., it yields positive and negative results in a truly random manner), then the positive predictive value will be the same as the prevalence: the a priori probability of disease in the patient equals the a posteriori probability of disease. Furthermore, for a test to be random, it is not necessary for sensitivity and specificity each to equal 50%; a test may have 90% sensitivity and still give random results if the specificity is only 10%.

Randomness requires only that:

$$\sum_{i}(\text{Sensitivity} + \text{Specificity}) = 100\%$$

These findings can be derived simply from Eq. (1):
PPV = \frac{pa}{pa + (1 - p)(1 - b)}

In a random test, the percentage of true positives in the diseased group will equal the percentage of false positives in the well group—by definition; that is,

\[ p_{alp} = \frac{[(1 - p)(1 - b)]/(1 - p)}{p_{alp}} = \frac{[(1 - p)(1 - b)]/(1 - p)}{p_{alp}} = (1 - Specificity) \]

Also, by definition, \( p_{alp} \) is the sensitivity of the test and

Therefore, in a random test, sensitivity equals 1 – Specificity, which is to say the sum of sensitivity and specificity equals unity (or 100%).

This relationship is of value in the graphical representation of marker performance. When sensitivity is plotted as a function of 1 – Specificity, an immediate visual impression of the marker’s discrimination is obtained. This graph is termed the receiver operating characteristic (ROC) plot. A random test will give a straight-line graph at 45° to the axes, whereas a good, highly discriminatory test will give a curve of steep slope from the origin, showing a high sensitivity even at high specificity. Therefore, the greater the area under the curve, the better the test. ROC plots are particularly useful in that they remove the influence of the “cutoff” point from the marker evaluation.

### 3.2. Optimization

If screening is to be considered, it is necessary to know the disease prevalence and to have tests with high sensitivity and specificity in order to calculate whether an acceptable positive predictive value can be achieved. However, it is impossible to optimize simultaneously both sensitivity and specificity—increasing one automatically decreases the other. Considerations regarding optimization strategies will vary with the natural history of the disease under study (\textit{vide infra}).

The simplest case will be considered; the situation where there is a screening procedure to be optimized and a false-negative result carries an equivalent penalty to a false-positive result. Under these circumstances, we may define our “index of misclassification,” \( f \), as the sum of the false-negative and false-positive results.

\[
f = FN + FP
\]

False negatives, FN, can be calculated as the lack of sensitivity \((1 - a)\) multiplied by disease prevalence, \(p\). Similarly, false positives, FP, can be calculated by multiplying lack of disease specificity \((1 - b)\) by the prevalence of nondisease in the population under study. Therefore,

\[
f = p(1 - a) + (1 - b)(1 - p)
\]
For most cancers, prevalence of disease in a general population screen will be tend to zero. Therefore,

\[ f = 1 - b \] (4)

It follows, therefore, that under the conditions and assumptions outlined—very low prevalence and equality of penalty for false-negatives and false-positives—one should increase specificity at the expense of sensitivity to minimize misclassifications.

3.3. Targeted Screening

The most frequently cited example of successful screening using a tumor marker is the use of human chorionic gonadotropin (hCG) in choriocarcinoma, and it is instructive to consider briefly why hCG has worked so wonderfully well when no other tumor markers are as competent.

Choriocarcinoma is rare; it accounts for 0.02% of all cancer deaths and is almost exclusively confined to women who have had a hydatidiform mole, of whom about 8% go on to develop choriocarcinoma. The single key fact that makes the screening program workable is the application of the test to a predetermined group in which the disease is present at a high prevalence.

If we assume that hCG has a sensitivity \( (a) \) of 99% and a specificity \( (b) \) of 99% and choriocarcinoma has a prevalence \( (p) \) of 8% in our screening group, then we can calculate the positive predictive value of hCG in this context:

\[
\text{PPV} = \frac{pa}{pa + (1 - b)(1 - p)}
\]

\[
= 0.08 \times 0.99 / [(0.08 \times 0.99) + (1 - 0.99)(1 - 0.08)] = 89.6\%
\]

By contrast, if one attempted to screen for choriocarcinoma all women whose pregnancies had achieved full term (prevalence 0.01%), the positive predictive value would be vanishingly small:

\[
\text{PPV} = 0.0001 \times 0.99 / [(0.0001 \times 0.99) + (1 - 0.99)(1 - 0.0001)] = 0.98\%
\]

It is, therefore, apparent that for screening to be effective, a high-prevalence group must be identified in order to keep the number of false positives to an acceptable level.

4. Clinical Utility

Clinical effectiveness demands that the early intervention afforded by a successful screen is translated into an increased rate of cure or improved survival time. Objective quantification of improvement in survival time is not quite as simple as it first might appear, as studies are subject to various forms of methodological bias.
4.1. Lead-Time Bias

Survival is measured from the date of diagnosis to death, rather than from the date of inception to death. The date of diagnosis may therefore vary considerably, depending on the methods of detection used, without altering the true length of survival from the date of inception. Lead time generated by screening, or the period from detection while the woman is still asymptomatic until the appearance of clinical symptoms, which would permit conventional diagnosis, may increase the apparent survival without, in fact, the individual having benefited from screening. In such circumstances, the patient has to live longer with the knowledge of the disease.

4.2. Length Bias

A series of cases diagnosed at screening will be atypical of those arising clinically, because it will contain a disproportionate number of patients with slowly developing tumors, probably with a better prognosis. Patients with rapidly progressing tumors are more likely to present with symptoms before the initiation of, or in the interval between, screening tests. This bias is more likely to be manifest at the initiation of screening and is, therefore, especially important in studies of short duration.

4.3. Selection Bias

Selection bias results from entry of a cohort into a screening trial who have a different probability of developing and dying from the disease than the population at large. In self-selected populations, it is common to find a higher than normal proportion of individuals presenting for screening because of a positive family history. These individuals are more motivated to present for screening because they are more educated in this respect and are more likely to benefit from it. This has been well demonstrated in breast and cervical screening programs.

5. Optimization Strategies

It was demonstrated earlier that when prevalence was very low (tending to zero), if false negatives and false positives carried equal penalty, then to minimize misclassifications, one should maximize specificity. In addition, one should maximize specificity in situations where the disease is serious but cannot be treated or cured and for which, therefore, any false-positive result would lead to psychological trauma. Some occult cancers would clearly fall into this group, as well as diseases such as multiple sclerosis. Such incurable diseases should not be subject to population screening, as there is usually no benefit to patient or society at large in early diagnosis. In this section, the other available
options will be considered and under which circumstances it would be appropriate to use them.

5.1. Sensitivity

Sensitivity should be maximized in situations where although the disease is serious and should not be missed, it is treatable and, therefore, false positives are less psychologically damaging. Most treatable infectious diseases would fall into this category, as do phaeochromocytoma and phenylketonuria. Cervical cancer, for which the screening program is effective and confirmatory tests are available prior to an effective therapeutic intervention program, is an example of a malignancy that may fall into this category. Furthermore, the concern caused by the presence of abnormal cells upon a cervical smear can in large measure be offset by the patient being aware of the success of early treatment.

5.2. Positive Predictive Value

The positive predictive value should be maximized in any situation where treatment of a false positive could be seriously damaging. Where the treatment indicated involves major surgery and radiotherapy, such as certain occult carcinomas, instigating treatment in someone who did not have the disease would be a major catastrophe.

5.3. Accuracy (or “Efficiency”)

Accuracy of a very high order is required when a disease is both serious and treatable and false-positive and false-negative results carry equal penalty. Myocardial infarction has usually been cited as the classical example of where the tests should be optimized for accuracy \([(TP + TN)/(TP + TN + FP + FN)]\); however, a case for optimizing accuracy could be made in testing for certain leukemias and lymphomas.

6. The Use of Multiple Markers

The idea of using a group of markers in order to complement the sensitivity and specificity of each other seems logical enough and can be extremely beneficial. There are certain rules that can be defined and applied, and certain pitfalls to avoid.

There are two distinct approaches to multiple testing. The first, as described in the above example, is so-called series testing; the various tests are performed one after the other depending on the result of the previous test. In series testing, therefore, a “test-positive” patient is one who has scored positive in all of the tests. A secondary consideration here is defining the order in which the tests are to be performed to maximize efficacy, although considerations of cost and
patient compliance also need to be included in any trial design. In parallel testing, all tests are performed on all patients, a “test-positive” patient in these circumstances is one who is positive on any one (or more) of the tests.

It is usual in a screening exercise for series testing to be preferred because it maximizes specificity at the expense of sensitivity which, as discussed earlier, is a rational approach when disease prevalence is low. Calculation of the PPV for parallel and series regimes bear this out (10).

For series testing, as not all tests are performed on all samples, there is the option of the order in which the tests are to be performed. There are many considerations: the relative cost of the tests involved, the degree of invasiveness, and the relative sensitivities and specificities of the tests involved. If variables such as cost are set aside, it can be shown that the sensible option is to test in series rather than parallel, as the positive predictive value is far higher and the total number of tests performed is much less. Also, although the PPV is independent of the order of testing, the number of analyses that have to be performed varies considerably, being minimized by application first of the test with the higher (or highest) specificity of those in the panel.

**6.1. Series Testing**

In an abstract (11), a research group reported the results of screening 1010 postmenopausal women for epithelial ovarian cancer using the serum marker CA125 followed up by ultrasonography. The group found a level of greater than 30 units/mL (their cutoff level) in 31 women. These 31 were then given ultrasonography; 3 were deemed abnormal and sent for surgery. One had an early-stage ovarian cancer. The authors concluded that CA125 had a high specificity for ovarian cancer, that they could increase the sensitivity by lowering the cutoff from 30 to 23 units/mL (the widely accepted cutoff value is, in fact, 35 units/mL), and that CA125 warranted further investigation for early diagnosis.

Their data are shown in Table 2. It is apparent from these data that there is no good reason to lower the cutoff from 30 to 23, as the sensitivity is already 100%. How reliable that figure is, however, is open to question, as there is only one true positive in the study. Furthermore, false negatives—here reported as zero—invariably take longer to emerge from any study and tend to be the most difficult to follow up; for these reasons then, the reported sensitivity may be an overestimate. The one true-positive patient had a CA125 level of 32 units/mL. Therefore, if these workers had followed the axiom of optimizing specificity at the expense of sensitivity, they would, in all probability, have missed the one patient who was to benefit directly from the trial. Their reason for opting for a higher sensitivity in this case was that they had a highly efficient second test
(ultrasonography) to filter out the majority of the false positives generated by the CA125 alone and did not wish to miss any cases. It can be seen from Table 2 that despite a sensitivity of 100%, a specificity of 97%, and an overall accuracy of 97%, the PPV was only 3.1% for CA125, hopelessly inadequate as a single selector for exploratory surgery. It is also true to say that knowing the sensitivity and specificity of the test and the disease prevalence, one could have calculated this PPV without having to do the trial, saving considerable expense. (“Since Isaac Newton, we no longer have to chart the fall of each apple”—Sir Peter Medawar.)

However, when ultrasonography is added in as a second-line test, the PPV improves by an order of magnitude to 33% (1/3) which is perhaps an acceptable pickup rate considering the high mortality rate of the disease if not diagnosed early. In effect, the use of CA125 in this and other studies generates a subgroup of the population under study who are at higher risk than the population at large; it defines a high-prevalence group thereby enabling a second-line test of similar sensitivity and specificity to produce a PPV that is far higher.

### 6.2. Panel Testing

Evaluation of a panel of tests is, of course, subject to all of the same provisions as for the assessment of a single test; particularly, the prevalence of the disease in the study group must be typical of the prevalence in the population to which it is intended to apply the test(s).

In a study of ovarian cancer by Ward et al. (12) in 1987, it was reported that by using three markers, the sensitivity in samples from pretreatment patients with stage 1 and 2 disease had increased from 18% using CA125 alone to 64% using human milk-fat globulin II (HMFG2) as the second assay and placental

### Table 2
Data From 1010 Postmenopausal Women Screened for Epithelial Ovarian Cancer (EOC) Using CA125

<table>
<thead>
<tr>
<th></th>
<th>EOC positive</th>
<th>EOC negative</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CA125 positive</strong></td>
<td>1 (TP)</td>
<td>31 (FP)</td>
<td>32 (TP + FP)</td>
</tr>
<tr>
<td><strong>CA125 negative</strong></td>
<td>0 (FN)</td>
<td>978 (TN)</td>
<td>978 (TN + FN)</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>1 (TP + FN)</td>
<td>1009 (TN + FP)</td>
<td>1010 (all)</td>
</tr>
</tbody>
</table>

*Abbr:* TP, true positive; TN, true negative; FP, false positive; FN, false negative.

Sensitivity = TP/(TP + FN) = 1/1 = 100%
Specificity = TN/(TN + FP) = 978/1009 = 97%
Prevalence = (TP + FN)/(TP + TN + FP + FN) = 1/1010 = 0.1%
Accuracy = (TP + TN)/(TP + TN + FP + FN) = 979/1010 = 97%
Positive predictive value = TP/(TP + FP) = 1/32 = 3.1%
alkaline phosphatase (PLAP) as a third marker. That is to say, CA125 had picked up 2/11 of the diseased group and HMFG2 and PLAP had picked up a further 5 of the CA125 negative group, taking the total to 7/11. However, as all the subjects under study were disease-positive, it can be seen that neither CA125, HMFG2, nor PLAP performed significantly differently from random chance. They also studied the marker panel in patients with advanced disease. In the 26 patients with advanced (stage 3 and stage 4) disease, 25 had elevated CA125 (96%) and the 26th had an elevated PLAP. Therefore, all patients with advanced carcinoma of the ovary were positive for at least one of these three markers. These results are not quite as promising as one might at first believe: Using such a group of patients where prevalence is 100% (whether early-stage or advanced disease), one could achieve apparently excellent sensitivity by four consecutive coin flips at considerably less cost! (Each flip will have a 50% sensitivity; therefore, in series, the cumulative sensitivity will become 50%, 75%, 87.5%, and 93.75%.)

7. Conclusions

Disease prevalence is of fundamental importance in the rational application of tumor marker assays. By and large, cancer prevalence is too low in the population to permit effective screening even if the financial and ethical constraints could be overcome. In ovarian cancer, there is, therefore, a large amount of current research directed at the identification of possible high-risk groups—the so-called cancer families—in which prevalence is significantly higher than in the population at large because of genetic predisposition.

The use of tumor markers to monitor disease progress or remission, to track therapeutic efficacy, or to give a lead time to relapse are much more successful. Here, the markers either are being applied to a group in order to quantify a disease known to be present or to pick up a relapse in a group where relapse and, therefore, disease prevalence will be high. The routine application of tumor markers in a clinical context has been reviewed elsewhere (13,14).

Acknowledgments

The author is indebted to Dr. Cathie Sturgeon for her critique of this chapter. He is also grateful to Churchill Livingstone for permission to use excerpts from his textbook: Serological Tumour Markers: An Introduction (10).

References
