Published in final edited form as:

Cancer Discov. 2013 February ; 3(2): 148–157. doi:10.1158/2159-8290.CD-12-0196.

From differences in means between cases and controls to risk stratification: A business plan for biomarker development

Nicolas Wentzensen and Sholom Wacholder

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20892

Abstract

Researchers developing biomarkers for early detection can determine the potential for clinical benefit at early stages of development. We provide the theoretical background showing the quantitative connection between biomarker levels in cases and controls and clinically meaningful risk measures and a spreadsheet for researchers to use in their own research. We provide researchers with tools to decide whether a test is useful, whether it needs technical improvement, whether it may only work in specific populations, or whether any further development is futile. The methods described here apply to any method that aims to estimate risk of disease based on biomarkers, clinical tests, genetics, environment or behavior.

Keywords

biomarker; risk stratification; screening; predictive value; likelihood ratio

Introduction

The central paradigm of cancer prevention and control is that recognition of a high risk of a future cancer, or early detection of an existing cancer before clinical symptoms lead to diagnosis, can reduce morbidity and mortality. Similarly, biomarkers for treatment decisions reduce cancer morbidity and mortality by optimizing allocation of treatment modalities to those who benefit most of them. Quantification of the potential clinical value in an early phase of biomarker development can help developers to accurately evaluate the promise of their work. Connecting laboratory findings with the clinical application early in development can help make wiser decisions about development strategies for the most promising biomarkers. The connection will also enable clinicians and policy makers to make better decisions about implementation of prevention programs and treatment modalities.

For simplicity, we restrict the theoretical discussion to disease prevention, but the fundamental approach also applies to treatment biomarkers. We suggest ways to guide development of prevention programs by identifying and focusing on the biomarkers, including imaging approaches and other putative determinants of risk, that show the greatest potential to prevent cancer in an early-detection program.

We use the term “*biomarker*” generically to refer to any test with potential to identify individuals at different levels of risk of a serious condition; *clinical significance* means usefulness in a program aimed to prevent disease or ameliorate the impact of disease. Specifically, clinical significance for a biomarker means that the difference between the risk

Correspondence: Sholom Wacholder PhD, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, Room 5050, Rockville, MD 20852-7234, USA, Phone: 301.496.3358, WacholdS@mail.nih.gov.

Conflict of interest: The authors declare that there is no conflict of interest.

of disease after a positive test is sufficiently greater than the risk of disease after a negative test to justify a different clinical decision about management of the patient; the precise difference required for clinical significance for a patient depends, of course, on the disease, the costs and efficacy of the management options and patient characteristics.

None of the statistics commonly used in early biomarker studies comparing patients with disease (cases) and controls – p-values, differences in means, odds ratios, area under the curve (AUC), ROC curves, sensitivity and specificity, likelihood ratios, Youden statistics – provide evidence about clinical significance, alone or together, without additional information or assumptions. The evaluation of clinical significance requires estimates of positive and negative predictive values, which are, respectively risk of disease after a positive test, and the complement of the risk after a negative test. We prefer to focus on the complement of negative predictive value, which we denote as $cNPV = (1 - NPV)$, the risk of disease after a negative test result, to allow easy comparison with the positive predictive value. Predictive values vary with the prior (before the test) risk of disease, which in turn can vary by demographics, clinical presentation, and other factors; predictive values for a screening program will need to use estimates of prior risk to calculate applicable predictive values.

We show each of the quantitative steps that connect the early-stage case-control comparisons with predictive values. We show the relationship between the receiver operator characteristic (ROC) curve and differences between biomarker levels found in cases and controls, how to convert points on the ROC curve to risks when prevalence of the disease is known, and how the likelihood ratios can be used to show relative change between disease prevalence in the population targeted for the screening program and risk after a positive or negative test at any point of the ROC curve. In addition, we discuss some of the important implications of the logic underlying the quantification. Finally, we provide a spreadsheet that takes prevalence of disease before the test, from age-specific Surveillance, Epidemiology and End Results (SEER) rates or from a risk model or after previous clinical tests and desired predictive values, and returns the sensitivity/specificity pairs and the difference in means of case and controls that would achieve those predictive values. Thus, this spreadsheet provides a realistic quantitative evaluation of the clinical benefit, in terms of improved screening and early detection programs, from the results of early case-control studies of the biomarker. We illustrate the approaches with examples from two cancer sites, cervical and ovarian cancer.

Sensitivity, complement of specificity, and Youden index

The ROC curve, a standard tool in biomarker development, graphs sensitivity $1 - \beta$ on the y-axis against $cSpecificity$, the complement of specificity ($(1 - (1 - \alpha)) = \alpha$) on the x-axis, where α and β are the fractions of diseased and non-diseased subjects whose test result does not correspond to the disease outcome (1). The beauty of the ROC curve, which shows the sensitivity for each value of $cSpecificity$, is its clear display of two countervailing aspects of test performance at each possible threshold of a continuous biomarker: a more relaxed definition of a positive test increases both sensitivity and $cSpecificity$; a more strenuous definition of a positive test, correspondingly, decreases both sensitivity and $cSpecificity$. The diagonal line on the ROC graph indicates a useless biomarker with sensitivity = $cSpecificity$, i.e., where the proportions of cases and controls with a positive test are equal.

Difference between mean biomarker levels in cases and controls and the ROC curve

The ROC curve and the Youden statistic (Sensitivity -cSpecificity, Supplementary material) portray test characteristics at each possible threshold. In this section, we show how assessing case-control differences in standard deviation units connects to the ROC curve and risk stratification.

We use Δ (delta), the ratio of the difference in biomarker means between cases and controls, in units of σ , the standard deviation for both cases and controls, as our measure of difference in distribution (Figure 1a; Supplemental material). Note that the overlap between the two curves is determined only by Δ and σ . For fixed Δ , increasing variation in biomarker levels within cases and controls, whether due to variation among individuals or measurement error, reduces the value of the biomarker, when the signal, or difference in average levels between cases and controls, is constant. The points on the curve in Figure 1b represent the possible pairs of sensitivity and specificity at each threshold of positivity T . The graph includes contours of sets of points with Δ equal to selected values, and the corresponding area under the curve (AUC) when biomarker levels are distributed normally (Supplemental material). For example, when the difference Δ between means of cases and controls is 2.5 standard deviation units, the AUC is 0.96 (Figure 1b). Figure 2 shows delta values from 0 to 5 with their corresponding AUC values. The ROC curve displays only sensitivity from diseased cases and specificity from non-diseased controls but does not reflect disease prevalence, which is necessary for risk stratification: a tool based on cases and controls without reflecting the underlying population can provide information about relative risk for any threshold, but not predictive values or risk stratification (2).

From ROC curve to risk

Positive and negative predictive values are just risk of disease (or its complement) conditional on value of the test based on the biomarker. Simple formulas (see Supplemental Material) based on Bayes' theorem connect predictive values directly to sensitivity and specificity and to the unconditional risk, i.e., assumed risk before the test result is available. An individual's unconditional risk π is a weighted average of PPV and the complement of NPV, $cNPV=1-NPV$, with weights equal to the proportions with positive or negative test result. In general, a substantial difference between PPV and cNPV is necessary for a useful biomarker; large PPV-cNPV means substantial risk stratification, and implies that PPV or cNPV, or both, are far from π .

Figures 3a to 3c graphically illustrate the increasingly stringent sensitivity and especially specificity required to achieve a high risk after a positive test with decreasing disease prevalence. For example, identifying individuals with a post-test risk or PPV of 10%, in a population with 5% disease prevalence does not require a test with outstanding performance; by contrast, specificity greater than 99%, even with sensitivity of 100%, is needed to achieve a PPV of 10% in a population with 0.1% prior disease risk. This is even more challenging for screening for ovarian cancer. With a population prevalence of 0.04% the PPV for a marker with a sensitivity of 100% and a specificity of 98.8% is only 10% (Table 1) (3;4). At any disease prevalence, the LR+, and the odds of PPV relative to the prior odds, double with relatively small increases in specificity from 90 to 95%, or 99% to 99.5%, or 99.9% to 99.95%, with constant sensitivity and prevalence, or from the greater absolute change in sensitivity from 40% to 80%, with specificity held constant; thus, the impact of small increases in specificity and large increases in sensitivity are equivalent. Therefore, to achieve meaningful risk stratification for a rare disease requires a biomarker test with very high specificity.

Figure 4 shows the obverse side of risk stratification: even for a common disease, a negative result from a test with high sensitivity provides strong reassurance that the patient's risk of disease, cNPV, is low enough to make even a standard intervention unnecessary; conversely, a negative result from a low sensitivity test has little clinical implication.

Impact of changes in sensitivity and specificity on likelihood ratios

The likelihood ratios, which are functions of sensitivity and specificity without regard to disease prevalence, and thus estimable, respectively, from cases and controls only, are useful measures of risk stratification on a relative scale (5). The likelihood ratio positive is the ratio of the probabilities of a case having a positive test (Sensitivity) and of a control having a positive test (cSpecificity); the likelihood ratio negative is the ratio of the probabilities of the control having a negative test (Specificity) and the case having a negative test (cSensitivity = $1 - \text{Sensitivity}$). From a pre-test overall risk of 1 per 10,000, an LR+ of 100 means a 100-fold increase in odds of disease after a positive test, leading to a PPV of approximately 1 per 100; an LR- of 100 means a 100-fold decrease in the odds of disease after a negative test to 1/1,000,000. As seen in figure 5, a small increase in specificity changes LR+ from 10 to 50 for fixed high sensitivity, but a far larger increase in sensitivity is needed to change LR+ substantially at a low fixed specificity (Figure 5a). Analogously, a small increase in sensitivity at fixed specificity increases the LR- from 10 to 50 (Figure 5b). Thus, the LR+ and LR- measures provide test-specific characteristics of risk stratification that yield estimates of absolute risk (PPV and NPV) when multiplied with the specific disease prevalence. Importantly, LR+ and LR- estimated in one population will be the same in another population, whenever sensitivities and specificities are the same in the two populations, even when disease prevalences are much different.

Achieving required predictive values

A useful clinical test determines subsequent clinical management of an individual; action taken on a positive test is justified when the benefits exceed the costs. Our spreadsheet shows the sensitivity/specificity pairs that take an individual with a given prior probability to the desired posterior probability or PPV. The spreadsheet reveals important aspects of clinical significance even with rough input of required parameters. The prior risk of disease can be obtained from publicly available databases, for example using an individual's age-specific cancer SEER rate. The required specificity to achieve the same posterior probability, or predictive value, therefore, will vary greatly by cancer site. Within cancer site, the prior can be personalized according to known individual characteristics.

A high positive predictive value in itself is necessary but not sufficient to justify clinical use. The spreadsheet also reveals the disadvantage of aiming for having a threshold that is too high. Restriction to the extremes may increase the predictive values of a test, but will reduce the yield of cases found per number of patients receiving the screen. Mass screening for rare mutations like those in BRCA1 or BRCA2 genes generate a high PPV for the ones who test positive, but the cNPV will be very close to the prior because mutations are so rare in unselected women.

Implications for development of biomarkers

Understanding the underlying requirements for a clinically useful biomarker can be very helpful in the early development stages and should guide further assay development.

Improving performance of biomarker

The simple summary statistic of difference between mean biomarker value in cases and controls is the foundation of the ROC curve, likelihood ratio, and risk stratification. The statistic Δ is the ratio of the absolute difference in average level of the biomarker between cases and control in units of standard deviation σ ; in turn, σ depends only on the population variation in cases and controls and laboratory and other forms of measurement error, and not on sample size. Researchers have two options to improve performance when Δ is too low to justify further clinical testing: use knowledge about disease natural history and pathology to improve the biomarker, or reduce laboratory variation in measuring the biomarker to lower the standard deviation.

Importance of high specificity to identify those at very high risk and of high sensitivity to identify those at low risk

More than sensitivity, high specificity determines the PPV when disease is rare (Figures 3a–c; Table 1) because otherwise too many of the positive tests arise from those without disease. Specificity close to 1 is a requirement for population-based screening for ovarian cancer to achieve even a PPV of 10%; a low specificity will mean that too many women among the vast majority who will not develop ovarian cancer will test positive, thereby lowering the PPV. By contrast, in a population at relatively high risk, high sensitivity is necessary for reassurance from a negative test (Figure 4), expressed as low cNPV, because with low sensitivity too many of those who test negative will in fact develop disease. For example, the high sensitivity of a single HPV test provides reassurance that a woman with a negative HPV test has low risk of pre-cancer and cancer for many years, and therefore may be safe with a longer interval between screens compared to screening with the less sensitive Pap smear (6).

Sample size

Increasing sample size to increase the power to detect a significant difference of the case-control comparison is pointless if Δ is known to be too low to provide the necessary difference between PPV and cNPV. Strong statistically significant differences between biomarker levels in cases and controls do not necessarily imply that a biomarker can provide useful risk stratification. However, studies require a very large number of controls to estimate specificity precisely and establish potential for high PPV. For example, large sample size is needed to distinguish specificities of 99% vs. 99.5% or 99.9%, which have large implications on LR+ and PPV.

Implications of prevalence or pre-test risk of disease

The simple relationship between the prior odds and posterior odds connects risk before the test is performed, i.e., prevalence of disease based on current information, to predictive values, which are themselves risks (PPV, cNPV) or complements of risk (NPV) after the test result is known.

The risk before testing, and thus the setting of the test, has major implications for predictive values, or risk after test results are known. Higher PPVs justify more aggressive interventions; lower cNPVs identifies those with least need or least potential for benefit, for which a standard intervention may not be justified. The PPV and cNPV are multiplicatively related to the prior odds, which approximates the prior risk when it is low. Risk stratification increases with disease prevalence: if the prior odds is 1 per million, posterior odds of 1 per 10,000 after a positive test, based on LR+ of 100, will still not justify any but the most benign intervention, and the cNPV can hardly provide additional reassurance. Because of the multiplicative relationships, testing in high risk individuals achieves greater risk

stratification after a positive test than testing in low-risk individuals when sensitivity and specificity are the same.

Implications for choosing biomarker thresholds

A test is a decision rule based on a biomarker and a threshold. General properties of a biomarker, like the AUC, do not imply good performance at all thresholds. Youden's index gives equal weight to sensitivity and specificity, which often leads to choosing cutoffs with non-optimal risk stratification, as outlined above. Choosing the appropriate threshold with a view to predictive values rather than Youden index or the point of inflection requires consideration of the very different impact on absolute risk measures of the same change in magnitude of sensitivity and specificity.

Although the ROC has full information on sensitivity and specificity at every possible threshold of positivity, the standard display of the ROC curve obscures the different effects of the same absolute change in sensitivity and specificity on likelihood ratios and risk stratification (5). Our figures that show areas of predictive values and LRs overlaid on the ROC curve may help to avoid the conventional emphasis on sensitivity at the expense of specificity, which causes inappropriate optimism and misdirection of effort when the goal is to find individuals with high PPV, or risk after a positive test. The key measures are the likelihood ratios, where changes in the denominator have more impact than the same change in the numerator: LR+, or the ratio of sensitivity to the complement of specificity, indicates the approximate relative change in PPV; LR-, the ratio of the complement of sensitivity to specificity indicates the change for NPV.

Interpretation of predictive values

The differences between predictive values, which are risks after a positive or negative test, and between risk before testing and risk after positive or negative results, are clinically useful measures when combined with a measure of population impact, such as number in the population needed to screen or treat to detect one new case of disease.

Temporality

Predictive values always have at least an implicit temporal unit. In some applications, biomarkers mainly indicate risk of prevalent disease, while others are markers of future disease. Importantly, the principles laid out above apply equally to all markers of risk of prevalent or future disease.

Application for therapeutic biomarkers

Our approaches also apply to early stages of biomarker development for prognosis and therapy decisions. In this scenario, a biomarker is used to predict whether a certain treatment should be used. Recently, standardized approaches for treatment biomarker development, mostly based on randomized trials, have been proposed (7;8). Similar to screening, our approach can also be helpful at early stages in treatment biomarker development to identify candidates that are worth moving forward. For example, participants in therapy trials can be categorized by their survival time (e.g. using 5-year disease-free survival as a cutoff). Biomarker candidates can be measured and stratification between categories of survival time can be evaluated the same way as we have demonstrated for screening biomarkers. This approach can help screening through a large set of markers and identify markers to carry forward into formal trials.

Discussion

We showed the logical and quantitative connections between early biomarker study in cases and controls and potential clinical significance. We presented the series of logical steps required to make the connection; these can provide quantitative insights that will help developing strategic ‘business’ plans for translating biomarker leads to clinical practice. We also provide a spreadsheet to help evaluate the potential clinical value of a test based on the biomarker.

Differences in distributions of levels of biomarker between cases and controls translate into variation in risk or risk stratification through the ROC curve when prior risk is known. The quantitative focus and graphical presentation here can help at an early stage in biomarker development by focusing on the biomarker’s potential to stratify risk sufficiently to become the cornerstone of a screening program. These allow underlying risk of disease, or individual variation in risk of disease, whether determined from demographics, risk factors, or results of previous clinical tests, to influence the course of research on the biomarker, including decisions on when to begin clinical testing or whether to abandon an early lead.

Early understanding of the simultaneous high predictive values and high sensitivity needed to determine thresholds in an effective screening program can save a lot of futile effort in developing a biomarker or risk predictor that will not be effective. Notably, Rossing et al. showed that awareness of the impact of low specificity on PPV could have avoided a recent consensus statement’s unrealistic suggestion for using very common and, necessarily very unspecific, symptoms of pelvic or abdominal pain or bloating or feeling full for ovarian cancer screening (9). Similarly, Schiffman et al. showed the tradeoffs in terms of the numbers of CIN3+ detected versus numbers of women referred to colposcopy from adding HPV types to a diagnostic HPV in different populations (10).

Good risk stratification alone is not helpful without effective interventions. Effective screening programs reserve more aggressive or expensive interventions for those at higher risk, and spare those at lower risk and with little chance of benefit from an intervention. The effectiveness, adverse effects and cost of an intervention are crucial for considering screening and add complexity to these calculations. Of course, further work on a biomarker can be justified in anticipation of future interventions or future reductions of assay costs.

Our results have wide applicability, but do have some limitations. For simplicity, we considered a single biomarker classified as positive or negative when more gradations are possible; intervention for prevention of disease, not reduced severity or mortality or alleviation of symptoms; normally distributed biomarkers with the same variance in cases and controls. We assumed that sensitivity and specificity for a test with a given threshold do not vary by population or setting. We have not considered important issues of bias in study design, fieldwork and interpretation (11–13). Finally, we have glossed over the distinction between risk of prevalent disease and risk of future disease.

Our approach applies to prevention of any disease where early intervention is particularly effective at reducing morbidity or mortality and to any risk stratification tool, whether a molecular biomarker; imaging, like mammography or ultrasound; a genetic risk score, or a routine clinical test. These ideas extend naturally to a wider scope, but with more complexity: risk stratification can incorporate demographic, genetic, and epidemiologic factors; multiple tests and triage; risk for future disease; arbitrary distribution of biomarker level in cases and controls. In all of these applications, performance measures that focus on risk stratification can help researchers decide which markers warrant further development as a potential risk stratification tool to prevention of incidence and mortality from cancer and other diseases.

Our quantitative approach to evaluate the risk stratification that can be achieved by screening tests can help redirect intellectual and financial investment from futile efforts to the most promising opportunities for translational impact, whether preventive or clinical. Estimations of risk stratification made from measurements in cases and controls, even with no prospective testing, can justify clinical testing.

In summary, we have introduced enhanced ROC curve displays that allow choosing thresholds based on predictive values and likelihood ratios. We have provided a spreadsheet that links biomarker performance with specific disease characteristics to evaluate the promise of biomarker candidates at an early stage. If biomarker performance is not adequate for the primary goals, we demonstrate three different main routes: (1) Work on reducing random variation from laboratory procedures, (2) Identify high risk populations that could benefit most from a given biomarker, or (3) Abandon a lead if none of these strategies works.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank David Check from the Biostatistics Branch for technical support in preparing the figures.

Funding: This research was supported by the Intramural Research Program of the NIH, National Cancer Institute.

Glossary Sidebar

Area under the receiver operator characteristic curve (AUC)	for a biomarker is the average sensitivity (or, equivalently, the integral of the sensitivity) in the interval of cSpecificity from 0 to 1 (specificity from 1 to 0), itself equal to the area between the ROC curve and the x-axis.
Biomarker	Any biological measurement that is useful in determining the prognosis or the appropriate management of a patient. A biomarker can be fixed over time, like germ-line DNA; or vary, like blood pressure, level of a chemical or protein in a biospecimen, or indication of a lesion on a diagnostic image such as a mammogram.
Complement	(of a probability or conditional probability of an event or possible fact being true) is the probability or conditional probability of no event or the possible fact being not true. The sum of a probability (or conditional probability) and its complement is by definition 1. The complement of specificity, <i>cSpecificity</i> is directly comparable to sensitivity: both are conditional probabilities of disease, but with different conditioning. Similarly, the complement of NPV, like PPV, is a conditional probability of disease. PPV is risk conditional on a positive biomarker test and cNPV is risk conditional on a negative biomarker test.
Conditional probability	is the technical term for a probability defined in a restricted subset, rather than “unconditionally, which refers to the entire population.” The conditional probability of event A conditional on (given) B is

denoted as the $\Pr(A|B)$. “Those who are positive for disease” is conditioning.

Disease

Disease can be prevalent or incident within a specified interval; based on clinical or molecular criteria at any stage of the pathologic process.

Likelihood ratios

are the ratios of the odds of disease given a test result and the prevalence odds. The *likelihood ratio positive* is the ratio of the conditional probability of a positive test result given disease and the conditional probability of a positive test result given no disease; the *likelihood ratio negative* is the ratio of the conditional probability of a negative test result given disease and the conditional probability of a negative test result given no disease.

Odds

is a way of characterizing a *probability*. The odds in terms of

probability p is $odds = \frac{p}{1-p}$, and the probability in terms of the odds is

$$p = \frac{odds}{odds + 1}$$

Predictive values.

- a. *Positive*: Probability of disease, given a positive test result from biomarker.
- b. *Negative*: Probability of no disease, given a negative test result from biomarker. We prefer to use $cNPV = 1 - NPV$ because $cNPV$ is also a probability of disease, which can be compared directly to the PPV.

PPV and $cNPV$ are both risks, and implicitly have a time element: either presence of disease at a given time (prevalence), or incidence within a specified incidence (e.g., within 5 years of the test).

Prevalence

Proportion of population with disease, or previously diagnosed with disease, at a given time.

Risk

Probability of disease, implicitly prevalent disease or incident disease within an interval.

Receiver operating characteristic (ROC) curve.

A presentation that plots a point for all possible thresholds of the biomarker, with the y-axis representing the sensitivity and the x-axis representing 1-specificity of the test. The ROC curve graphically displays the tradeoff of increased sensitivity but decreased specificity from lowering the threshold, and vice versa.

Sensitivity and Specificity.

Sensitivity is proportion whose biomarker test is positive (above the threshold) among those who are positive for disease. Sensitivity is usually symbolized as $1 - \beta$ where β is the complement of sensitivity or the chance that the biomarker does not correspond to disease status for someone with disease. Specificity is proportion whose biomarker test is negative (below the threshold) among those without disease. Specificity is usually symbolized as $1 - \alpha$ where α is the complement of specificity, or the chance that the biomarker does not correspond to disease status for someone without disease. Pepe (5) calls sensitivity “True Positive Fraction (TPF)” and specificity “false positive fraction” (FPF); we revert to the older standard nomenclature still widely used in laboratories. Although TPF and FPF are often called rates instead of

	fractions, the measures are actually probabilities or fractions, not rates, which imply a time denominator.
Threshold	is the value of the biomarker above which the test is considered positive and below which the test is considered negative.
Youden index	for the biomarker for a disease is the difference between the sensitivity and the complement of specificity, equivalent to the difference between probabilities of the biomarker being positive among those with and without disease. This definition is equivalent to the sum of sensitivity and specificity less 1.

Reference List

1. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982; 143:29–36. [PubMed: 7063747]
2. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004; 159:882–90. [PubMed: 15105181]
3. Hartge P, Speyer JL. Finding ovarian cancer. *J Natl Cancer Inst*. 2012; 104:82–3. [PubMed: 22247023]
4. Rauh-Hain JA, Krivak TC, Del Carmen MG, Olawaiye AB. Ovarian cancer screening and early detection in the general population. *Rev Obstet Gynecol*. 2011; 4:15–21. [PubMed: 21629494]
5. Pepe, M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; USA: 2004.
6. Katki HA, Kinney WK, Fetterman B, Lorey T, Poitras NE, Cheung L, et al. Cervical cancer risk for women undergoing concurrent testing for human papillomavirus and cervical cytology: a population-based study in routine clinical practice. *Lancet Oncol*. 2011; 12:663–72. [PubMed: 21684207]
7. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. *Ann Intern Med*. 2011; 154:253–9. [PubMed: 21320940]
8. Rector, TS.; Taylor, BC.; Wilt, TJ. *Systematic Review of Prognostic Tests*. 2012.
9. Rossing MA, Wicklund KG, Cushing-Haugen KL, Weiss NS. Predictive value of symptoms for early detection of ovarian cancer. *J Natl Cancer Inst*. 2010; 102:222–9. [PubMed: 20110551]
10. Schiffman M, Khan MJ, Solomon D, Herrero R, Wacholder S, Hildesheim A, et al. A study of the impact of adding HPV types to cervical cancer screening and triage tests. *J Natl Cancer Inst*. 2005; 97:147–50. [PubMed: 15657345]
11. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J Natl Cancer Inst*. 2008; 100:1432–8. [PubMed: 18840817]
12. Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer*. 2005; 5:142–9. [PubMed: 15685197]
13. Ransohoff DF. The process to discover and develop biomarkers for cancer: a work in progress. *J Natl Cancer Inst*. 2008; 100:1419–20. [PubMed: 18840815]

Statement of significance

Many efforts go into futile biomarker development and premature clinical testing. In many instances, predictions on translational success or failure can be made early, simply based on critical analysis of case-control data. Our paper presents well-established theory in a form that can be appreciated by biomarker researchers. Furthermore, we provide an interactive spreadsheet that links biomarker performance with specific disease characteristics to evaluate the promise of biomarker candidates at an early stage.

Figure 1.

(A) Graphical display of distributions of biomarker in controls (blue) and cases (pink), and threshold for positivity. Assume that biomarker levels are normally distributed; that in controls the mean level is 0 and variance is 1; that in cases the mean level is Δ and variances is 1. That is, the difference in means between cases and controls is 1 standard deviation. When biomarker level is above the threshold of positivity T in proportion $1-\beta$ (the pink area) of cases and proportion α (the vertically hashed blue area) of controls, sensitivity for threshold T is $1-\beta$ and specificity is $1-\alpha$ (vertical hash). Thus, $T = z_{1-\alpha} = \Delta - z_{\beta}$, where z_{κ} is the cumulative probability that a standard normal variable is above κ or $T = \Phi^{-1}(1-\alpha) = \Delta - \Phi^{-1}(\beta)$, where Φ is the cumulative distribution of the standard normal and Φ^{-1} is the inverse cumulative distribution of the standard normal. (B) Receiver Operator Characteristic (ROC) curve showing sensitivity and specificity pairs achievable for different thresholds T for values of Δ , the difference in mean biomarker level between cases and controls, and

corresponding AUC values. The contours for a given Δ represent the set of points that satisfy $1-\beta=1-\Phi[\Phi^{-1}(1-\alpha)-\Delta]$, or, equivalently, $\Delta=\Phi^{-1}(1-\alpha)-\Phi^{-1}(\beta)$.

Figure 2. Relationship of Area under the Curve (AUC) and difference in biomarker means between cases and controls

The graph is based on a biomarker with normally distributed values in cases and controls and equal variance.

Figure 3. Contour plots and range of positive predictive value (PPV), or risk after a positive test, with given sensitivity and specificity for various prior probabilities π ($\pi = 0.1$ (figure 3a); $\pi = 0.001$ (figure 3b) and $\pi = 0.00001$ (figure 3c)) on a graph with standard Receiver Operator Characteristic (ROC) curve format

Figures 3b and 3c are details of the standard ROC display; the inset shows the part of the full ROC curve included in the detail. The PPV is calculated from equation (1) of the supplemental material. The black lines represent the ROC curves for normally distributed biomarkers with means that are 1, 2, 3, 4 or 5 standard deviations apart.

Figure 4. Contour plots and range of the complement of negative positive predictive value (cNPV), or risk after a negative test, with given sensitivity and specificity, for prior probability π of 0.01 on a graph with standard Receiver Operator Characteristic (ROC) curve format
 The cNPV calculated according to equation (2) of the supplemental material. The black lines represent the ROC curves for normally distributed biomarkers with means that are 1, 2, 3, 4 or 5 standard deviations apart.

Figure 5. Contour plots and range of the likelihood ratio positive (A) or negative (B) with given sensitivity and specificity on a graph with standard Receiver Operator Characteristic (ROC) curve format.

The contours for a given LR^+ (A) can be calculated as $1-\beta=\alpha DLR^+$ and for given DLR^- (B)

$$\text{as } 1-\beta=1-\frac{DLR^-}{1-\alpha}.$$

Table 1

Test performance required to achieve a specific positive predictive value in low risk and high risk populations

		Required test specificity					
		Annual incidence of ovarian cancer					
		Women age 55–59 from the general population *			Women age 55–59 with BRCA mutations#		
Test sensitivity		PPV 5%	PPV 10%	PPV 20%	PPV 5%	PPV 10%	PPV 20%
0.5		0.9870	0.9938	0.9973	0.5928	0.8071	0.9143
0.6		0.9844	0.9926	0.9967	0.5114	0.7685	0.8971
0.7		0.9818	0.9914	0.9962	0.4299	0.7300	0.8800
0.8		0.9791	0.9901	0.9956	0.3485	0.6914	0.8628
0.9		0.9765	0.9889	0.9951	0.2671	0.6528	0.8457
0.95		0.9752	0.9883	0.9948	0.2263	0.6335	0.8371
0.975		0.9746	0.9880	0.9946	0.2060	0.6239	0.8328
0.99		0.9742	0.9878	0.9946	0.1938	0.6181	0.8303
0.995		0.9741	0.9877	0.9945	0.1897	0.6162	0.8294
1		0.9739	0.9877	0.9945	0.1856	0.6142	0.8286

* Based on rates for 55–59 year old white women in SEER17;

assuming 30-fold increased risk compared to the standard risk in that age group, following Modan et al.