Anesthesiology 2010; 112:1023–40

David S. Warner, M.D., Editor

Statistical Evaluation of a Biomarker

Patrick Ray, M.D., Ph.D.,* Yannick Le Manach, M.D.,† Bruno Riou, M.D., Ph.D.,‡
Tim T. Houle, Ph.D.§

ABSTRACT

A biomarker may provide a diagnosis, assess disease severity or risk, or guide other clinical interventions such as the use of drugs. Although considerable progress has been made in standardizing the methodology and reporting of randomized trials, less has been accomplished concerning the assessment of biomarkers. Biomarker studies are often presented with poor biostatistics and methodologic flaws that precludes them from providing a reliable and reproducible scientific message. A host of issues are discussed that can improve the statistical evaluation and reporting of biomarker studies. Investigators should be aware of these issues when designing their studies, editors and reviewers when analyzing a manuscript, and readers when interpreting results.

HISTORICALLY, the term biomarker referred to analytes in biologic samples that predict a patient's disease state. However, the term biomarker has evolved over time to any biologic measurement, recently including genomic or proteomic analyses, that could also predict a response to a drug (efficacy, toxicity, or pharmacokinetics) or indicate an underlying physiologic mechanism.[1] New biomarkers exploring the cardiovascular system, kidney, central nervous system, inflammation, and sepsis are under the scrutiny of bioengineering companies, and we are witnessing a biomarkers revolution similar to the imaging technique revolution.[2] Remarkably, this revolution has already occurred for cancer drugs.[1]

Assessment of these biomarkers is complex but valuable in perioperative and critical care medicine as markers of diagnosis, disease severity, and risk. Although considerable progress has been made in standardizing the methodology and reporting of randomized trials, less has been accomplished concerning the assessment of diagnostic and prognostic biomarkers. Analysis of the literature, even in prestigious journals, has revealed that the methodologic quality of diagnostic studies is on average poor.[3] Recommendations concerning the reporting of diagnostic studies, the Standards for Reporting of Diagnostic Accuracy (STARD) initiative, have been published recently,[4] several years after the first recommendations concerning reporting of randomized trials.[5] However, these recommendations do not encompass all issues of this rapidly evolving domain.

The purpose of this article was to provide the anesthesiologist with a comprehensive introduction of the problems, potential solutions, and limitations raised by the assessment of the diagnostic properties of modern biomarkers. It is important to appreciate the available statistical methodologic tools to face the biomarker revolution, either as a clinical investigator or as a consumer of scientific literature. This is no easy task, for we must now look beyond the classic diagnostic indices (sensitivity, specificity, and predictive values) and even beyond the more widely used receiver operating characteristic (ROC) curves by integrating the principles of Bayesian theory. To appreciate these issues, the different roles of a biomarker must first be explored.

Role of a Biomarker

A biomarker may serve different roles (table 1) and, thus, need to accomplish several reporting goals. A biomarker may

* Staff Emergency Physician, Department of Emergency Medicine and Surgery, Groupe Hospitalier Pitié-Salpêtrière, Paris, France, Université Pierre et Marie Curie-Paris 6, Paris, France, and Institut National de la Santé et de la Recherche Médicale (INSERM) U956, Paris, France. † Assistant Professor, Department of Anesthesiology and Critical Care, Groupe Hospitalier Pitié-Salpêtrière. ‡ Professor of Anesthesiology and Critical Care, Chairman, Department of Emergency Medicine and Surgery, Université Pierre et Marie Curie-Paris 6, and INSERM U956. § Research Assistant Professor, Department of Anesthesiology, Wake Forest University School of Medicine, Winston-Salem, North Carolina.

Received from Department of Emergency Medicine and Surgery, Université Pierre et Marie Curie-Paris 6, Institut National de la Santé et de la Recherche Médicale U956, Department of Anesthesiology and Critical Care, Groupe Hospitalier Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, Paris France, and Department of Anesthesiology, Wake Forest University School of Medicine, Winston-Salem, North Carolina. Submitted for publication December 2, 2009. Accepted for publication December 14, 2009. Support was provided solely from institutional and/or departmental sources. Dr. Patrick Ray received honoraria from Biomérieux SA (Marcy l'Etoile, France), BRAHMS (Clichy, France), and Roche Diagnostics France (Meylan, France). Dr. Riou received honoraria from BRAHMS. James C. Eisenach, M.D., served as Handling Editor for this article.

Address correspondence to Dr. Houle: Department of Anesthesiology, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, North Carolina 27157-1009. thoule@wfubmc.edu. Information on purchasing reprints may be found at www.anesthesiology.org or on the masthead page at the beginning of this issue. ANESTHESIOLOGY's articles are made freely accessible to all readers, for personal use only, 6 months from the cover date of the issue.

Table 1. The Main Roles of a Biomarker

Role	Description	Examples
Diagnosis of a disease	To make a diagnosis more reliably, more rapidly, or more inexpensively than available methods	Troponin Ic diagnoses myocardial infarction[6] Procalcitonin diagnoses bacterial infection[7]
Severity assessment	To identify subgroup of patients with a severe form of a disease associated with an increased probability of death or severe outcome	Procalcitonin identifies severe outcome in septic patients[8] Troponin Ic identifies severe outcome in patients with pulmonary embolism[9]
Risk assessment	To identify subgroup of patients who may experience better (or worse) outcome when expose to an intervention	Brain natriuretic peptide and postoperative outcome in noncardiac surgery[10] Troponin and long term outcome in cardiac surgery[11]
Prediction of drug effects	To identify the pharmacological response of a patient exposed to a drug (efficacy, toxicity, and pharmacokinetics)	Efficacy of clopidogrel[15]
Monitoring	To assess the response to a therapeutic intervention	Procalcitonin may guide antibiotic duration[13]

provide a diagnosis or assess severity (or assess a risk). For example, cardiac troponin I is a very sensitive and specific biomarker of myocardial infarction in the postoperative period in noncardiac surgery.[6] In contrast, it is considered only as a severity biomarker in pulmonary embolism,[9] whereas procalcitonin is considered both as a diagnostic and severity biomarker of infection.[8] Biomarkers are often used for risk stratification. For example, blood lactate levels have been proposed for risk stratification of sepsis.[14] However, the purpose of diagnostic and prognostic settings markedly differ. For example, in the diagnostic setting, although unknown, the outcome (the disease) has occurred, whereas in the prognostic setting, the outcome remains to be determined and can only be estimated as a probability or a risk, and the uncertain nature of this outcome should be considered.

There are several important hierarchical steps in demonstrating the clinical interest of a biomarker:

1. Demonstrate that the biomarker is significantly modified in diseased patients as compared to control.
2. Assess the diagnostic properties of the biomarkers.
3. Compare the diagnostic properties of the biomarker to existing tests.
4. Demonstrate that the diagnostic properties of the biomarker increase the ability of the physician to make a decision; this might be difficult to analyze because timing of diagnosis may be crucial and not easy to identify. For example, although the accuracy of procalcitonin to diagnose postoperative infection after cardiac surgery was lower than that of physicians, procalcitonin enabled to make the diagnosis earlier.[7]
5. Assess the usefulness of the biomarker, which should be clearly distinguished to the quality of diagnostic information provided.[15] Assessment of the usefulness mainly involves both characteristics of the test itself such as cost, invasiveness, technical difficulties, rapidity, and characteristics of the clinical context (prevalence of the disease, consequences of outcome, cost, and consequences of therapeutic options).
6. Demonstrate that the measurement of the biomarkers modifies outcome (intervention studies). For example, several studies nicely demonstrated that a diagnostic strategy based on procalcitonin level reduces antibiotic use for acute respiratory tract infections, exacerbation of chronic obstructive pulmonary disease, and ventilator-associated pneumonia.[16] However, intervention studies are lacking for many novel biomarkers or give conflicting results for others.[17]

For all stages of this process, it is important to understand the pathophysiologic mechanisms involved in the biomarker's synthesis, production, its kinetic properties, and its physiologic effects. For example, brain natriuretic peptide (BNP) is known to be released predominantly from the left cardiac ventricles in response to increased ventricular wall stretch, volume expansion, and overload. Its physiologic role includes systemic and pulmonary arterial vasodilation, promotion of natriuresis and diuresis, inhibition of the renin-angiotensin-aldosterone system, and endothelins. In contrast, the pathophysiologic background of procalcitonin remains poorly understood.

A biomarker may also guide other clinical decisions, particularly concerning the use of drugs. This area is now widely developed in oncology in which biomarkers are used to predict an efficacy and/or a toxicity response of a drug.[1] For example, procalcitonin has been advocated to guide the clinician to decide the duration of antibiotherapy,[13] and genetic determinants of metabolic activation of clopidogrel have been shown to modulate the clinical outcome of patients treated by clopidogrel after an acute myocardial infarc-

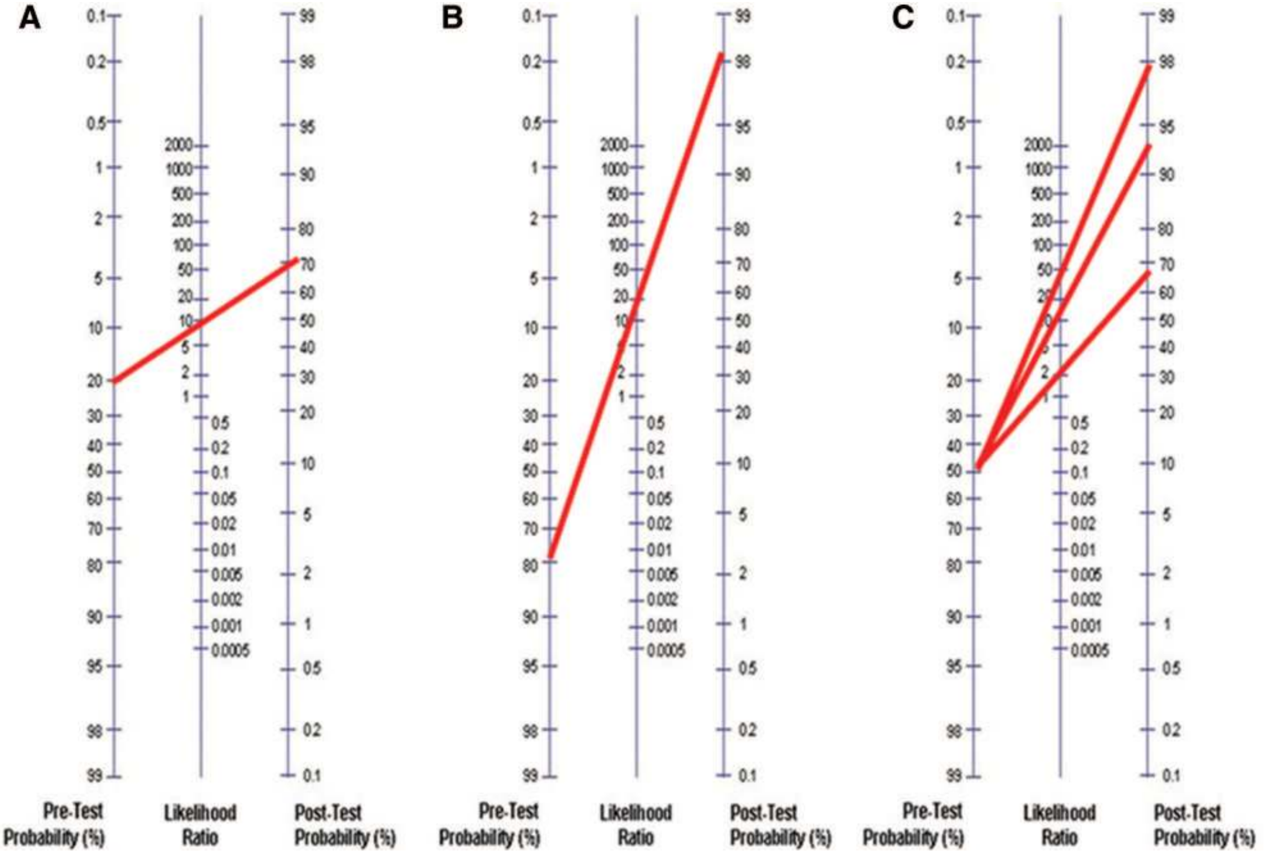

Fig. 1. Fagan nomogram using the Bayesian theory showing the pretest and postprobabilities and the likelihood ratio. (*A*) A straight line is applied for a low pretest probability (0.20) for a good biomarker with a positive likelihood ratio of 10, providing a posttest probability of (0.80); the important change in probability suggests a change for the physician (diagnostic or therapeutic). (*B*) In contrast, when the same good biomarker is applied to a patient with a high pretest probability (0.80), the posttest probability is more than 0.95, but this may not represent an important change for the physician. (*C*) The effects of several biomarkers with different likelihood ratios (2, 10, and 50) in a patient with a pretest probability of 0.50. The nomogram is reprinted with permission from Fagan.[21]

tion.[12] Finally, a biomarker can be used as a surrogate endpoint in a clinical trial,[1,18] but this issue is beyond the scope of this review.

The Bayesian Approach

One way to conceptualize the utility of a biomarker is its value for enhancing our existing knowledge in predicting the probability of some outcome (*e.g.*, disease state, prognosis). In this regard, Bayesian statistical methods provide a powerful system from which to update existing information about the likelihood of the occurrence of some disease or prognosis. A comprehensive introduction to these methods is far beyond the scope of this review, but an interested reader is referred to one of the several textbooks that have been written on applying Bayesian statistical methods to medical problems.[19]

Bayes' theorem uses two types of information to compute a predicted probability of the outcome. First, the prior probability (or pretest probability) of the outcome must be considered. For biomarker studies involving the diagnosis of a disease, this can be akin to the general prevalence of the disease in the population under study, or what we know about the base rate of the disease for this individual without any additional information. This information is combined with the predictive power of the biomarker (*i.e.*, the ability of the test to discriminate between disease states) to adjust our prediction of the likelihood of the outcome. Stated simply, the predicted probability of a patient having the disease (posttest probability) can be calculated as[20]: posttest probability = (pretest probability) × (predictive power of the evidence).

Numerical examples of this calculation are given in Likelihood Ratios, where likelihood ratios (LHRs) are discussed. However, the use of Bayes' theorem to "update" our expectation of the presence of a disease is illustrated using Fagan's nomograms (fig. 1).[21] In these examples, it can be seen how disease prevalence (pretest probability) is used in conjunction with the LHR (strength of evidence) to calculate an updated (posttest) probability of the disease.

Although a powerful method for updating assumptions given the available information, there are many instances where an appropriate estimate of the pretest probability is not known (or agreed on). In such cases, different physicians might have different estimates of the probability of the disease for a given

Table 2. The Diagnostic Matrix and the Derivation of Main Diagnostic Parameters

	Disease		
	Present	Absent	Total
Biomarker			
Positive	a (true positive)	b (false positive)	a + b
Negative	c (false negative)	d (true negative)	c + d
Total	a + c	b + d	a + b + c + d

Prevalence = (a + c)/(a + b + c + d); sensitivity = a/(a + c); specificity = d/(b + d); positive predictive value = a/(a + b); negative predictive value = d/(c + d); accuracy = (a + d)/(a + b + c + d); Youden index = sensitivity + specificity − 1; positive likelihood ratio (LHR+) = sensitivity/(1 − specificity); negative likelihood ratio (LHR−) = (1 − sensitivity)/specificity; diagnostic odds ratio = (ad)/(bc) = (LHR+)/(LHR−).

LHR = likelihood ratio.

patient. Further, information may actually be available for one or more risk factors, but the unique combination of these factors may obscure the subjective probability for a given patient. For these applications, the sensitivity of the expectation can be checked against a range of assumptions (see Reclassification Table for more details).

Statistical Tools

Decision Matrix

The diagnostic performance of a biomarker is often evaluated by its sensitivity and specificity. Sensitivity is the ability to detect a disease in patients in whom the disease is truly present (*i.e.*, a true positive), and specificity is the ability to rule out the disease in patients in whom the disease is truly absent (*i.e.*, a true negative). Calculation of these indices requires knowledge of a patient's "true" disease state and a dichotomous prediction based on the biomarker (*i.e.*, disease is predicted to be present or absent) to construct a 2×2 contingency table. Table 2 displays how the frequency of predictions from a sample of patients could be used in conjunction with their known disease state to calculate sensitivity and specificity.

Although sensitivity and specificity are the most commonly provided variables in diagnostic studies, they do not directly apply to many clinical situations because the physician would rather know the probability that the disease is truly present or absent if the diagnostic test is positive or negative rather than probability of a positive test given the presence of the disease (sensitivity). These former, more clinically interesting probabilities are provided by the positive predictive value and negative predictive value. Table 2 presents the calculation of these predictive indices.

The diagnostic accuracy of a test is the proportion of correctly classified patients (*i.e.*, the sum of true positive and true negative tests). Perhaps because it is the most intuitive index of diagnostic performance, diagnostic accuracy is sometimes reported as a global assessment of the test. However, the use of this index for this purpose is inherently flawed and produces unsatisfactory estimates under a range of situations, such as when the prevalence of the disease substantially deviates from 50%.[22] It is recommended that authors report more than just a single estimate of diagnostic accuracy.

The Youden index Y = sensitivity + (specificity − 1) represents the difference between the diagnostic performance of the test and the best possible performance[23] (sometimes called the "regret" defined as the utility loss because of uncertainty about the true state).[24] Interestingly, accuracy is actually a weighted average of sensitivity and specificity, using as weight the prevalence of the disease. It should be clear that these five indices (sensitivity, specificity, negative and positive predictive values, and accuracy) are partially redundant, because knowing three of them enables the calculation of the rest.

Influence of Prevalence

Although sensitivity and specificity are not markedly influenced by the prevalence of the disease, negative predictive value, positive predictive value, and accuracy are affected by prevalence. Figure 2 shows the influence of prevalence on the various diagnostic indices. This issue is of paramount importance because disease prevalence can markedly differ from one population to another. The prevalence of sepsis in an intensive care unit is high compared with that in the emergency department, and the positive predictive values and accuracy of procalcitonin for sepsis might markedly differ between these two settings. For example, Falcoz *et al.*[25] reported that a 1 ng/ml procalcitonin had a positive predictive value of 0.63 for predicting postoperative infection after thoracic surgery. However, in that study, the prevalence of infection was 16%. If a *post hoc* analysis was conducted restricting the scope of inclusion only to patients with systemic inflammatory response syndrome criteria, the prevalence would have been 63% and the positive predictive value 0.90.

Although the mathematical calculation of sensitivity and specificity are not necessarily altered by prevalence, certain clinical situations may foster higher estimates.[26,27] These indices can be influenced by case mix, disease severity, or risk factors for disease.[27] For example, a biomarker is likely to be more sensitive among more severe than among milder cases of the diseases. The sensitivity of procalcitonin to diagnose bacterial infection is greater in patients with meningitis than in patients with pyelonephritis.[28]

Likelihood Ratios

LHRs are another way of describing the prognostic or diagnostic value of a biomarker. Although we call them "diagnostic" LHR, these ratios are LHRs in the true statistical sense and correspond to the ratios of the likelihood of the observed test result in the diseased *versus* nondiseased populations. Two dimensions of accuracy have to be considered, the LHR for a positive test (positive LHR) and the LHR for a negative test (negative LHR). One of the most interesting features of LHRs is that they quantify the increase in knowledge about

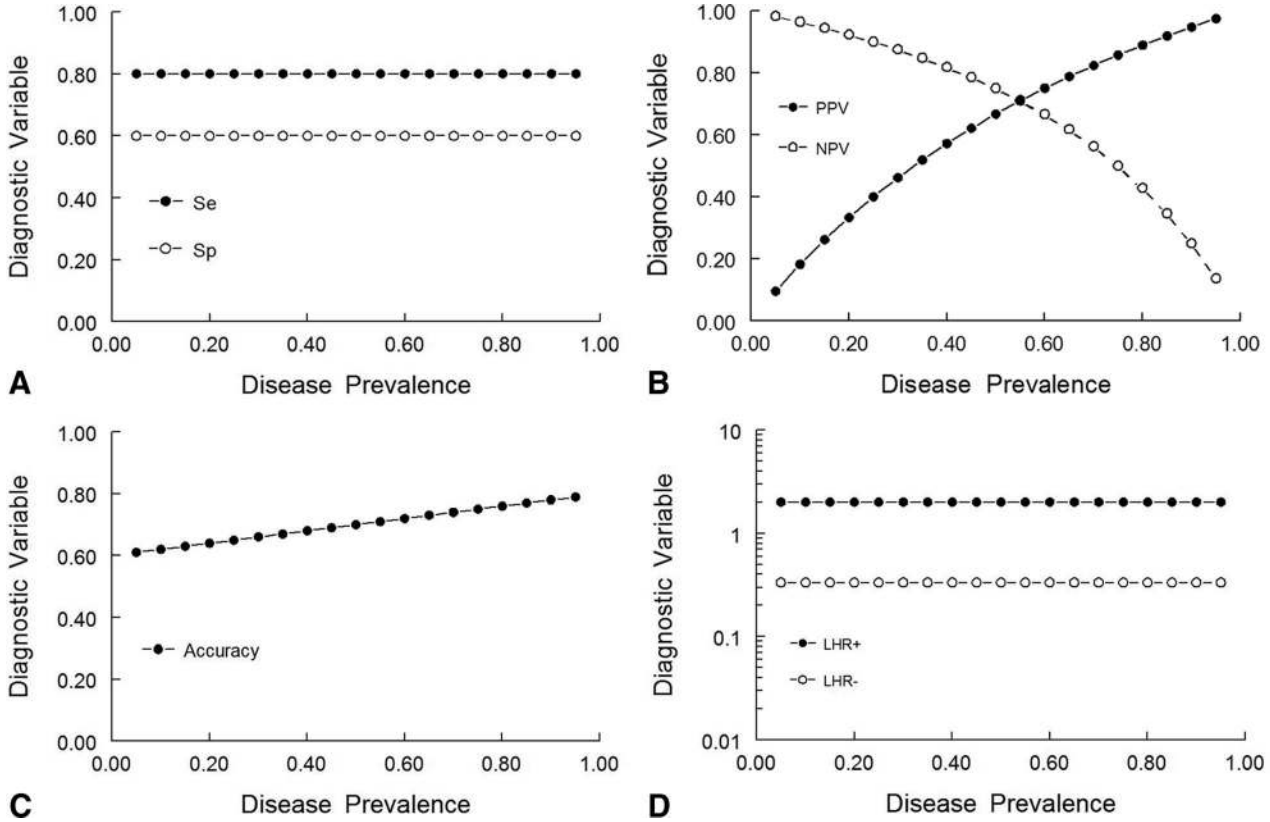

Fig. 2. Effects of prevalence of a disease on main diagnostic variables in a simulated population (n = 1,000) in an ideal world. A biomarker with a sensitivity of 0.80 and a specificity of 0.60 was considered where each point on the horizontal axis corresponds to different prevalence (from 0.05 to 0.95, step of 0.05). The effects of prevalence on sensitivity (Se) and specificity (Sp) (*A*), positive predictive value (PPV) and negative predictive value (NPV) (*B*), accuracy (*C*), and likelihood ratios (*D*) can be seen in each of the panels. LHR− = negative likelihood ratio; LHR+ = positive likelihood ratio.

the presence of disease that is gained through the diagnostic test. Thus, LHR could also be referred as Bayes factors; we could demonstrate that using the following formula: posttest probability of disease = (positive LHR) × (pretest probability of disease) or posttest probability of nondisease = (negative LHR) × (pretest probability of nondisease).

Furthermore, LHRs are not dependant on disease prevalence and, thus, are considered as a robust global measure of the diagnostic properties of a test, and they can be used with tests that have more than two possible results (see interval LHR).

More pragmatically, positive LHR ranges from 1.0 to infinity and negative LHR from 0 to 1.0. An uninformative test having no relation with the disease has LHR of 1.0, whereas a perfect test would have a positive LHR equal to infinity and a negative LHR of 0 (table 3). For example, a common sense translation of a positive LHR of 8.6 for a plasma soluble triggering receptor expressed on myeloid cells 1 value exceeding 60 ng/ml is that this value is obtained approximately nine times more often from a patient with sepsis than from a patient without sepsis.[29] Experts usually consider that tests with a positive LHR greater than 10 (or a negative LHR less than 0.1) have the potential to alter clinical decisions. Although it could be tempting to follow definitive rules-of-thumb for interpretation (such as those provided in table 3), we must primarily consider the clinical setting to determine what level of increased likelihood is clinically relevant to improve the management of patients.

Revisiting the nomograms in figure 1, several important issues become clear concerning diagnostic LHRs. First, no change in prediction (expectation) is possible without a strong LHR. As might be expected, when an LHR provides no added information (*e.g.*, LHR = 1.0), the pretest probability equals the posttest probability. Second, pretest probability greatly influences what can be learned from using even

Table 3. Rule of Thumb: Correspondence between Accuracy, Positive (LHR+) and Negative (LHR−) Likelihood Ratio, and Area under the Receiver Operating Characteristics Curve (AUC_{ROC}) and the Diagnostic Value of a Biomarker

	Accuracy	LHR+	LHR−	AUC_{ROC}
Excellent diagnostic value	>0.90	>10	<0.1	>0.90
Good diagnostic value	0.75–0.90	5–10	0.1–0.2	0.75–0.90
Poor diagnostic value	0.50–0.75	1–5	0.2–1	0.50–0.75
No diagnostic value	0.50	1	1	0.50

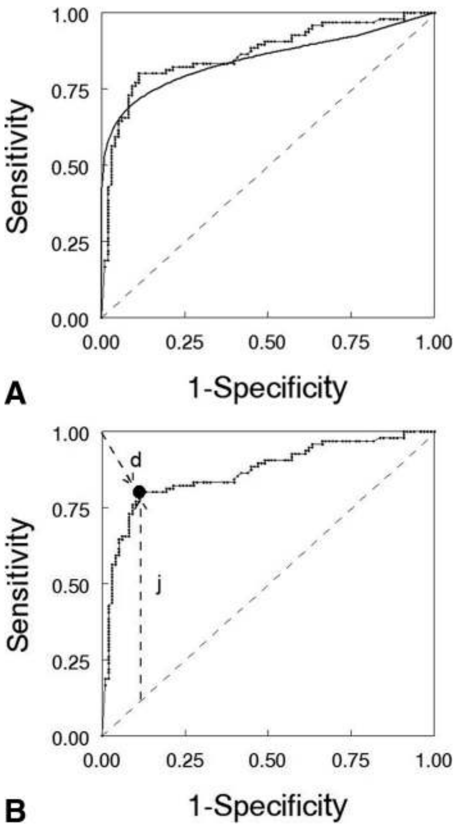

Fig. 3. Receiver operating characteristics (ROC) curve showing the relationship between sensitivity (true positive) and 1 − specificity (true negative) in determining the predictive value of Brain Natriuretic Peptide (BNP) for cardiogenic pulmonary edema in elderly patients (>65 yr) admitted to the emergency department for acute dyspnea. (*A*) The empirical ROC curve is shown by the *line containing points* that corresponds to different cutoff; the area under the empirical ROC curve was 0.87 (95% confidence interval 0.80–0.91); the ROC curve is also shown by the *continuous line*, which was fitted to the binomial distribution. The *dotted line* is the identity line. (*B*) The best cutoff was chosen as that one which minimizes the mathematical distance (d) to the ideal point (sensitivity = specificity = 1), corresponded to a concentration of BNP of 250 pg/ml with a sensitivity of 0.78 and a specificity of 0.90. But the best cutoff should be preferably chosen as that one which maximizes the distance (j) between the ROC curve and the identity line, for example, that which maximizes the Youden index (sensitivity + specificity − 1), which, in the present case provided the same cutoff value. These two options do not take into account the prevalence and the cost-benefit analysis (see text). Adapted from data from Ray *et al.*[31]

a very predictive biomarker (*e.g.*, LHR = 10.0). Very high (or low) pretest probabilities result in smaller adjusted expectations than those less extreme.

ROC Curve

Basic of ROC Curve. The ROC curve is a form of a series of pairs (proportion of true positive results; proportion of false-positive results) or (sensitivity; 1 − specificity): the sequence of points obtained for different cutoff points can be joined to draw the empirical ROC curve, or a smooth curve can be obtained by appropriate fitting, usually using the binomial distribution (fig. 3).[30,31] In other words, the positive LHRs calculated at various values of the diagnostic test can be plotted to produce a ROC curve, and thus, a ROC curve is a graphical way of presenting information presented in the table of LHRs. Graphically, the positive LHR is the slope of the line through the origin (sensitivity = 0; 1 − specificity = 0) and a given point on the ROC curve, whereas the negative LHR is the slope of the line through the point opposite to the origin (sensitivity = 1; 1 − specificity = 1) and that given point on the ROC curve.

The area under the receiver-operating characteristic curve (AUC_{ROC}) (also called the c statistics or the c index) is equivalent to the probability that the biomarker is higher for a diseased patient than a control and, thus, is a measure of discrimination. By convention, ROC curves should be presented above the identity curve (fig. 2) that represents a test without any value and which performs like chance. It is important to note that the following points belong to the identity curve: of course sensitivity = 0.50 and specificity = 0.50 but also sensitivity = 0.90 and specificity = 0.10, and sensitivity = 0.10 and specificity = 0.90. This enables us to understand that sensitivity cannot be interpreted without specificity. The AUC_{ROC} should be reported with confidence intervals (CIs) to allow statistical evaluation *versus* the identity line or statistical comparison *versus* other diagnostic tests (see Comparison of ROC Curves). Usually, biomarkers are considered as having good discriminative properties tests when AUC are higher than 0.75 and as excellent more than 0.90 (table 3). The ROC curve is a global assessment of the test accuracy but without any *a priori* hypothesis concerning the cutoff chosen, is relatively independent on prevalence, and is a simple plot that is easily appreciated visually. However, the cutoff point and the number of patients are not typically presented (although a small sample size is easily detected by a jagged and bumpy ROC curve). The generation of a ROC curve is no longer cumbersome because most statistical software provides the calculation and display of the relative parameters.

There are three common summary measures for the accuracy described by a ROC curve. The first is simply to report the pair of values (sensitivity and specificity) associated with a chosen cutoff point. The second is the AUC_{ROC}, and the third is the area under a portion of the curve (partial area) for a prespecified range of values. Interpreting the AUC_{ROC} is somewhat problematic because of the substantial portion of variance in this index that comes from values of the biomarker of no clinical relevance. One ROC curve may have a higher proportion of false positive than another in the region of clinical interest, but the two ROC curves may cross, leading to different conclusion when curves are compared on the basis of the entire area. Therefore, it is recommended that examination of the ROC curve be conducted in the context of partial area or average sensitivity over a range of clinically relevant proportion of false positives in addition to the AUC_{ROC}.[32]

Comparison of ROC Curves. The first step for any ROC curve comparison should be a visual inspection of their graphical representation. This inspection allows the evaluation of large differences between the AUC_{ROC} and to detect the situations where ROC curves cross. However, formal statistical testing is required to assess differences between the curves. Several different approaches are possible, and all must take into consideration the nature of the collected data. When the predictive value of a new biomarker is compared with an existing standard(s), two or more empirical curves are constructed based on tests performed on the same individuals. Statistical analysis on differences between these curves must take into account the fact that one individual is contributing two scores to the analysis. Most biomarker studies collect data that are paired (*i.e.*, measurements are correlated) in nature. Parametric approaches to these comparisons assume that there is a continuous spectrum of possible values of the biomarker for both diseased and nondiseased patients (generally true with biomarkers) and that the underlying distribution is Gaussian (normal). However, this assumption is often not tenable in biomarker studies. Despite this, paired parametric methods of ROC comparison are often used to evaluate biomarkers, using an approach described by Hanley and Mc Neil.[33] An alternative nonparametric paired method, described by DeLong *et al.*,[34] is based on the Mann–Whitney U statistic. The two approaches yield similar estimates even in nonbinormal models.[35]

Two main limitations must be considered for global comparisons of the ROC curves. First, this way of comparing two ROC curves is not precise, specifically when two ROC curves cross each other. Second, many cutoffs on the ROC curves are not considered in practice because their associated specificity and sensibility are not clinically relevant. To reduce the impacts of these limitations, comparisons of partial AUC_{ROC} within a specific range of specificity for two correlated ROC curves have been developed and might be interesting to consider for some biomarkers.[36]

Finally to maximize the generalization capacity of the observed data, resampling methods have been proposed to compare ROC curves. This modern approach is actually easier to conduct with the increase of computing power and seems to provide more accurate results for small sample sizes.

Determination of Cutoff. The ROC curve is used to determine a clinical cutoff point to make a clinical discrimination. The method used to choose this cutoff is crucial but unfortunately not always reported in published studies.[37] In some situations, we do not wish to (or could not) privilege either sensitivity (identifying diseased patients) or specificity (excluding control patients), and thus, the cutoff point is chosen as that one which could minimize misclassification. Two techniques are often used to choose an "optimal" cutoff. The first one (I) minimizes the mathematical distance between the ROC curve and the ideal point (sensitivity = specificity = 1) and thus intuitively minimizes misclassification. The second (J) maximizes the Youden index (sensitivity + [specificity − 1]) and thus intuitively maximizes appropriate classification (fig. 3).[38] Interestingly, Perkins *et al.*[39] present a sophisticated argument that the J point should be preferred, because I does not solely rely on the rate of misclassification but also on an unperceived quadratic term that is responsible for observed differences between I and J.[39]

However, the use of I or even J may not be satisfactory for two main reasons. First, this equipoise decision, which does not privilege either sensitivity or specificity, is valid only in the case of a prevalence of 0.50; in other situations, the prevalence should be taken into account. Second, in many clinical situations, the researcher could privilege either sensitivity or specificity because the consequence of false-positive or false-negative results is not equivalent in terms of a cost–benefit relationship. For example, it is clear that it is more crucial to rule out bacterial meningitis than treat a patient with antibiotics who has viral meningitis, or at least those due to enterovirus.[40] The researcher should assign a relative cost (financial or health cost, from the patient, care provider, or society points of view) of a false positive to a false-negative result and consider the prevalence, and these different elements can be combined to calculate a slope m: m = (false-positive cost/false-negative cost) × ([1 − P]/P), the operating point on the ROC curve being that which maximizes the function (sensitivity − m[1 − specificity]).[15,41] Other methods includes the net cost of treating controls to net benefit of treating individuals and the prevalence.[42,43]

In any case, the following recommendations should be provided: (1) the choice of the researcher must be clearly explained and justified; (2) the choice (at least its methodology) must be *a priori* decided; (3) the ROC curve should be provided to allow the reader to make its opinion; and (4) the cutoff that maximizes the Youden index should also be indicated. It remains clear that data-driven choice of cutoff tends to exaggerate the diagnostic performance of the biomarker.[44] This bias should be recognized and probably concerns many published studies.

Surprisingly, although the cutoff point has a crucial role in the decision process, it is provided in most (if not all) studies without any CI. This may constitute a major methodologic flaw, particularly in small sample studies, because this cutoff point might be markedly influenced by the value of very few patients, although the CIs of sensitivity and specificity associated with that cutoff are reported.[45,46] The reason of the absence of CI is probably related to the fact that more sophisticated statistical methods should be used. The principle of all these methods is to perform multiple resampling of the studied population to provide a large sample of different populations providing a large sample of cutoff points and thus a mean (or a median) associated with its 95% CI. Several techniques of resampling can be used (bootstrap, Jackknife, Leave-One-Out, n-fold sampling).[47,48] In a recent study, Fellahi *et al.*[49] used a bootstrap technique to provide median and 95% CIs for cutoff points of troponin Ic in patients undergoing various types of cardiac surgery, enabling the comparison of these different cutoff points. Here again, CI rule enables the researcher and the reader to honestly communicate or understand the values presented, taking into account the sample size.

The Gray Zone. Another option for clinical discrimination is to avoid providing a single cutoff that dichotomizes the popula-

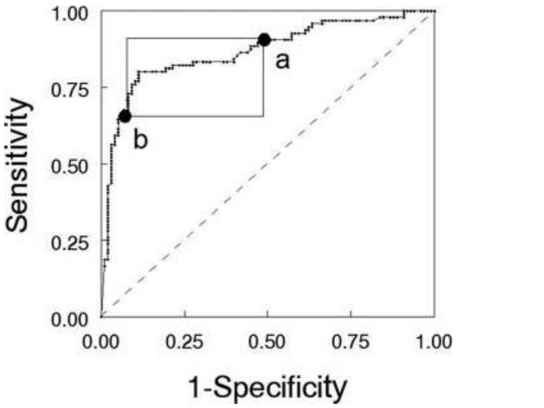

Fig. 4. Receiver operating characteristics (ROC) curve and the "gray zone." The ROC curve is the same as in figure 2 and shows the predictive value of Brain Natriuretic Peptide (BNP) for cardiogenic pulmonary edema in elderly patients (>65 yr) patients admitted to the emergency department for acute dyspnea. Two cutoffs were chosen as, one corresponding to a high value of BNP associated with certainty for diagnostic inclusion (BNP = 360 pg/ml; sensitivity = 0.66, specificity = 0.93; positive predictive value = 0.90) and the other with certainty for diagnosis exclusion (BNP = 100 pg/ml; sensitivity = 0.91, specificity = 0.51, negative predictive value = 0.85). The *square* indicates the gray zone. Adapted from data from Ray *et al.*[31]

tion, but rather to propose two cutoffs separated by the "gray zone" (fig. 4). The first cutoff is chosen to exclude the diagnosis with near-certainty (*i.e.*, privilege specificity). The second cutoff is chosen to include the diagnosis with near-certainty (*i.e.*, privilege sensitivity). When values of the biomarker falls into the gray zone between the two cutoffs, uncertainty exists, and the physician should pursue a diagnosis using additional tools. This approach is probably more useful from a clinical point of view and is now more widely used in clinical research. Moreover, the two cutoffs and gray zone comprise three intervals of the biomarker that can be associated with a respective LHR. In that case, the positive LHR of the highest value of the biomarker in the gray zone is considered to include the diagnosis and the negative LHR of the lowest value to exclude the diagnosis. This interesting option is often called the interval LHR[50] and results in less loss of information and less distortion than choosing a single cutoff, providing an advantage in interpretation over a binary outcome. This allows the clinician to more thoroughly interpret the results improving clinical decision-making.

Here again, the 95% CI of the cutoff points may be calculated using a resampling method,[47,48] and the rules for choosing the cutoffs be determined *a priori* and clearly explained and justified.

Reclassification Table

Biomarkers' abilities to predict a disease are commonly evaluated using ROC curves. The improvement in AUC_{ROC} for a model containing the new biomarker is defined simply as the difference in AUC_{ROC} calculated using models with and without the biomarker of interest. This increase, however, is often very small in magnitude. Ware *et al.*[51] and Pepe *et al.*[52] describe examples in which large odds ratios are required to meaningfully increase the AUC_{ROC}. As a consequence, many risk factors that we know to be clinically important may not affect the c-statistic very much. Thus, the ROC curves approach might be considered as insensitive to evaluate the gain of biomarkers.[52] Furthermore, ROC curves are frequently not helpful for evaluating biomarkers because they do not provide information about the actual risks or the proportion of participants who have high- or low-risk values. Moreover, when comparing ROC curves for two biomarkers, the models are aligned according to their false-positive rates (that is, different risk thresholds are applied to the two models to achieve the same false-positive rate), and this might be considered as inappropriate.[53] In addition, the AUC_{ROC} or c-statistic has poor clinical relevance. Clinicians are never asked to compare risks for a pair of patients among whom one who will eventually have the event and one who will not. To complete the results obtained by ROC curves, some new approaches to evaluate risk prediction have been proposed. One of the most interesting is the risk stratification tables. This methodology better focuses on the key purpose of a risk prediction, which remains to classify individuals into clinically relevant risk categories. Pencina *et al.*[54] have recently purposed two ways of assessing improvement in model performance using reclassification tables: Net Reclassification Index (NRI) and Integrated Discrimination Improvement.

The NRI approach enables us to assess the role of a biomarker to modify risk strata and alter clinical decisions. It requires a predefined risk stratification, which is usually expressed in several strata (>2), and the use of 3 strata (high, intermediate, and low risks) is probably the most easily handled for routine clinical management.[55] NRI is the combination of four components: the proportion of individuals with events who move up or down a category and the proportion of individuals with nonevents who move up or down a category. Because the NRI and its four components might be affected by the choice of stratification of the risks, lack of clear agreement on the categories that are clinically important could be problematic when using the NRI to assess new biomarkers. This concern is common with the Hosmer-Lemeshow test. Again, prevalence, predictive values, cost, and benefit should probably be considered to make clinically relevant decisions.[56] On the contrary, the Integrated Discrimination Improvement table does not require predefined strata, and it can be seen as continuous version of NRI with the probability of disease differences used instead of predefined strata. Alternatively, it could be defined as the difference of mean predicted probabilities of events and no events.

NRI and Integrated Discrimination Improvement tables provide an important increase in the power to detect an improvement in risk stratification associated with the use of a new biomarker. Indeed, numerous clinical situations exist where a considered small increase of AUC_{ROC} lead to substantial improvement in reclassification by the NRI and/or Integrated Discrimination Improvement table. This might suggest that very small increase of AUC_{ROC} might still be suggestive of a meaningful improvement in the risk prediction and that the exclusive use of ROC curve is not sufficient to demonstrate that a biomarker is not useful. This is clearly an evolving domain of biostatics,[53] which should be highly considered for perioperative medicine and risk stratification.

Common Pitfalls of the Evaluation of a Biomarker

Intrinsic Properties

A biomarker supposes a biologic assay that is associated with measurement errors. Thus, the precision of the measurement of the biomarker and the limit of detection should be provided (reproducibility). Moreover, the measurement of the biomarker should be sensitive, for example, detects very low concentrations, and specific, for example, provides a measure of the biomarkers itself without interferences with other molecules, particularly those related to its metabolism. All these intrinsic properties are important to report and disclose to the reader.

Moreover, because most of biomarkers are molecules produced by our cells and measured in blood, urine, or other organic fluid, the possibility that abnormal cells can produce the biomarker in a completely abnormal pathophysiologic mechanism when compared with that (or those) known should always be considered. For example, there is some evidence that KIM-1, a biomarker of kidney injury, can be secreted by the kidney cancer cells in the absence of renal injury.[‖] This point might be difficult because the physiology of a given biomarker is often incompletely known, and the abnormal cells may produce anything. Normality may differ between populations (example of troponin in cardiac surgery *vs.* other type of surgery).[49]

The analytical characteristics of any assay should be distinguished from its diagnostic characteristics.[57] The terms "limit of detection," "limit of quantitation," or "minimal detectable concentration" are synonyms used for analytical sensitivity. Polymerase chain reaction is considered as a very sensitive test because it could detect a very low number of copies of gene or gene fragment. However, despite this exquisite analytical sensitivity, its diagnostic sensitivity may not be so perfect when the target DNA is absent in the biologic material analyzed: this could be the case of a patient with endocarditis but whose withdrawn blood samples do not contain any bacteria. In the same way, polymerase chain reaction can be considered as an assay with exquisite analytical specificity, but its diagnostic specificity may not be so perfect just because of contamination.[57]

Numerical Expression of Diagnostic Variables

Most of these variables should be considered as percentages and, thus, can be expressed either using the unit "percent" or using two digits: thus sensitivity might be presented either as 89% or 0.89. The important point is to ensure coherence along a given manuscript for a given variable and among all diagnostic variables. More importantly, because these variables are percentages, a CI (95% CI) should always be associated.[58] The lower and upper limits of the 95% CIs inform the reader about the interval in which 95% of all estimates of the measure (*e.g.*, sensitivity, area under the curve) would decrease if the study was repeated over and over again. When LHRs are reported, CIs that include 1 indicate that the study has not shown convincing evidence of any diagnostic value of the investigated biomarker. Therefore, the reader does not know whether a test with a positive LHR of 20 but a 95% CI of 0.7–43 is useful. A study reporting a positive LHR of 5.1 with a 95% CI of 4.0–6.0 provides more precise evidence than another study arriving at a positive LHR of 9.7 with a 95% CI of 2.3–17. Usually the sample size in critical care medicine studies is small, leading to wide CIs. Likewise, too often, studies concerning diagnostic tests are underpowered to allow statistically sound inferences about the differences in test accuracy.

The reporting of CIs enables the researcher and the reader to effectively communicate or understand the values presented, taking into account the uncertainty inherent with any sample size. This is particularly important because most of these variables are calculated using only a fraction of the whole population studied: for example, an interesting sensitivity of 0.90 in a large population of 500 patients (but only 10 presenting the disease) may not seem so interesting when considering its 95% CI: 0.60–0.98. Likelihood and diagnostic ratios are ratios of probabilities but should also be reported with their CIs. Moreover, CI enables the reader to directly interpret statistical inference.[58]

Role of Time

In most clinical situations, the issue of the time of biomarker measurement is of limited interest, mainly because the time of onset of the pathologic process and or disease is unknown. However, in other situations, the time of onset can be readily determined. This is the case for acute chest pain and for the appearance in the blood of a biomarker for myocardial infarction. In that example, although troponin is recognized as an ideal biomarker (both very sensitive and very specific), it needs more time to be detected than myoglobin, which is considered as a poorer diagnostic biomarker but one that appears earliest (fig. 5).[59] The importance of timing can be crucial in perioperative medicine, particularly in the postoperative period, because timing of the insult (anesthesia/surgery) is precisely known. For example, in cardiac surgery, Fellahi *et al.*[11] suggested that troponin should be measured 24 h after cardiopulmonary bypass to gain the maximum information. In contrast, the time profile of another biomar-

[‖] Morrissey J, London A, Lambert M, Luo J, Kharasch E: Specificity of the urinary biomarkers KIM-1 or NGAL to detect perioperative kidney injury (abstract A1623). Paper presented at the Annual Meeting of the American Society of Anesthesiologists, New Orleans, Louisiana, October 17–21, 2009.

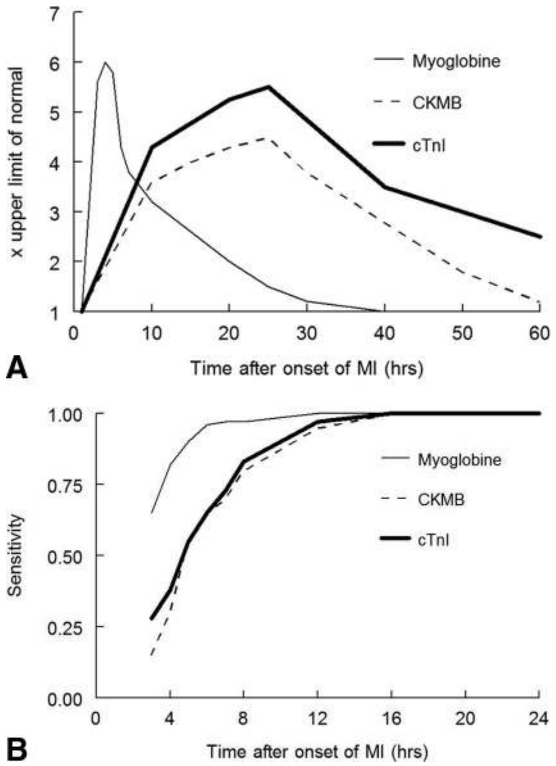

Fig. 5. Effect of time. (*A*) Schematic evolution of blood concentrations of main biomarkers of acute myocardial infarction (MI): myoglobin, creatine kinase MB (CKMB), and cardiac troponin I (cTnI) after the onset of chest pain. (*B*) Schematic evolution of their respective sensitivity. Adapted from data from De Winter *et al.*[59]

ker such as BNP may be completely different in that surgical setting.[60]

The issue of time of measurement may also be crucial when considering the pathophysiologic process assessed by biomarkers, which supposes a clear understanding of these processes, which is not always the case. In sepsis, the simultaneous occurrence of proinflammatory processes (tumor necrosis factor-α and interleukin-6) and antiinflammatory processes (interleukin-1), and the complex interaction of therapeutics, may render difficult the analysis of biomarkers, particularly when the onset of the various infection processes (onset of infection *vs.* onset of severe infection *vs.* onset of shock) remains vague.

Different Populations

Diagnostic tests may substantially vary when measured in different patient populations, particularly when studied populations are defined by characteristics such as demographic features (age and sex) and spectrum of the disease (severity, acute *vs.* chronic illness, pathologic location of form).[61] Moreover, the diagnostic test may work well in a global population but not in a given subgroup. For example, procalcitonin may not be a good biomarker of infection in pyelonephritis[28] or intraabdominal abcess.[62] Procalcitonin is not a good biomarker for infection in a population exposed to heatstroke even though half of them are truly infected simply because heatstroke itself increases procalcitonin.[63] In the perioperative period, the type of surgery might be an important cause of variation. The properties of cardiac troponin I to diagnose postoperative myocardial infarction are fundamentally different in noncardiac *versus* cardiac surgery, just because cardiac surgery alone is responsible for important postoperative release of cardiac troponin, which has multiple causes: surgical cardiac trauma, extracorporeal circulation, and defibrillation.[60] Even when considering cardiac surgery, different cutoff points of cardiac troponin to predict major postoperative cardiac events are observed when comparing cardiopulmonary bypass, valve, or combined surgery.[49]

All these important issues are usually summarized as spectrum biases. Therefore, precise information concerning the population studied and its case mix are important to be provided by researchers and to be understood by readers (table 4).

The issue of different populations could be more widely analyzed as an issue of external influences (covariates). This might be the case when factors other than the disease affect the biomarker including factors that affect the test procedure (apparatus and centers), the value of the biomarker itself (see later for the influence on kinetics), or the relation of the biomarker to the outcome. Thus, adjustment for covariates may be an important component of evaluation of biomarkers.[64] When the covariate does not modify the ROC performance, the covariate-adjusted ROC curve is an appropriate tool to assess the classification accuracy and is analogous to the adjusted odds ratio in an association study. In contrast, when a covariate affects the ROC performance, the ROC curves for specific covariate groups should be used. Covariate adjustment may also be important when comparing biomarkers, even under a paired design, because unadjusted comparisons could be biased.[64]

Importance of the Biomarker Kinetics

A biomarker has its own kinetics implying metabolism and elimination. This important issue has been poorly recognized at least partly because the kinetics of biomarkers is often poorly investigated. Just as renal or liver insufficiency may influence the pharmacokinetics of drugs, they also could influence the kinetics of a biomarker and interfere with their diagnostic properties. For example, procalcitonin has been shown recently to be increased in patients with renal function who undergo vascular surgery. This increase was observed both in infected and noninfected patients (fig. 6) and interferes with the cutoff point chosen but not with the diagnostic performance.[65] This effect could be of paramount importance in the postoperative period or in the intensive care unit because these patients are more likely to present organ failures. When comparing two biologic forms derived from BNP, the active form and its prometabolite N-terminal prohormone brain natriuretic peptide, Ray *et al.*[66] observed that the diagnostic properties of N-terminal prohormone brain natriuretic peptide were decreased compared with BNP, probably because of the differential impact of renal function on these two biomarkers in an elderly population.

Table 4. The Standards for Reporting of Diagnostic Accuracy (STARD) Checklist for Reporting Diagnostic Studies*

Section and Topic	Item	Description
Title, abstract, keywords	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading "sensitivity and specificity"
Introduction	2	State the research questions or aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups
Methods, participants	3	Describe the study population: the inclusion and exclusion criteria and the settings and locations where the data were collected
	4	Describe participant recruitment: was this based on presenting symptoms, results from previous tests, or the fact that the participants had received the index test or the reference standard
	5	Describe participant sampling: was this a consecutive series of participants defined by selection criteria in items 3 and 4? If not specify how participants were further selected
	6	Describe data collection: was data collection planned before the index tests and reference standard were performed (prospective study) or after (retrospective study)?
Test methods	7	Describe the reference standard and its rationale
	8	Describe technical specifications of materials and methods involved, including how and when measurements were taken, or cite references for the index test or reference standard, or both
	9	Describe the definition of and rationale for the units, cut-off points, or categories of the results of the index tests and the reference standard
	10	Describe the number, training, and expertise of the persons executing and reading the index tests and the reference standard
	11	Where the readers of the index tests and the reference standard blind (masked) to the results of the other test? Describe any other clinical information available to the readers
Statistical methods	12	Describe the methods for calculating or comparing measures of diagnostic accuracy and the statistical methods used to quantify uncertainty (*e.g.*, 95% confidence interval)
Results, participants	13	Describe methods to quantify test reproducibility
	14	Report when study was done, including beginning and ending dates of recruitment
	15	Report clinical and demographic characteristics (*e.g.*, age, sex, spectrum of presenting symptoms, comorbidity, current treatments, and recruitment centers)
	16	Report how many participants satisfying the criteria for inclusion did or did not undergo the index tests or the reference standard, or both; describe why participants failed to receive either test (a flow diagram is strongly recommended)
Test results	17	Report time interval from index tests to reference standard, and any treatment administered between
	18	Report distribution of severity of disease (defined criteria) in those with the target condition and other diagnoses in participants without the target conditions
	19	Report a cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, report the distribution of the test results by the results of the reference standard
	20	Report any adverse events from performing the index test or the reference standard
Estimates	21	Report estimates of diagnostic accuracy and measured of statistical accuracy (*e.g.*, 95% confidence interval)
	22	Report how indeterminate results, missing results, and outliers of index tests were handled
	23	Report estimates of variability of diagnostic accuracy between readers, centers, or subgroup of participants if done
	24	Report estimates of test reproducibility, if done
Discussion	25	Discuss the clinical applicability of the study findings

* Table available at http://www.stard-statement.org/. Accessed December 2, 2009.

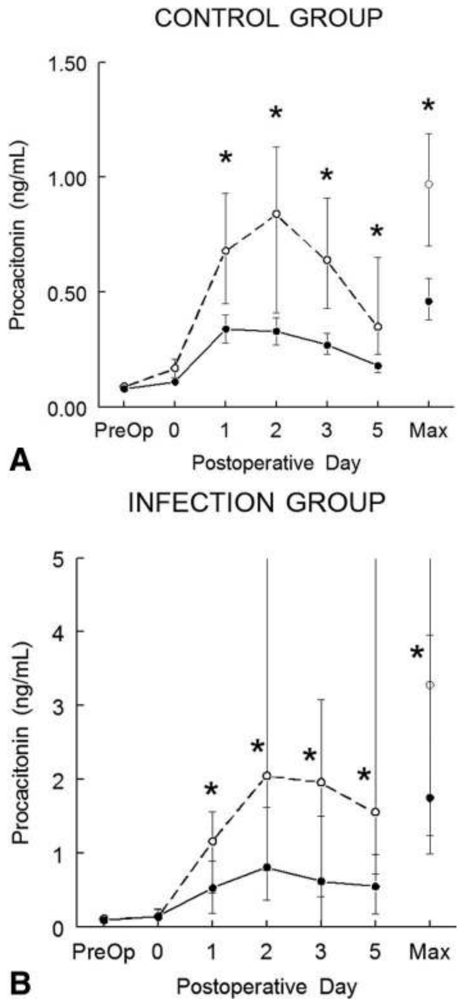

Fig. 6. Influence on renal function on a biomarker in the postoperative period after major vascular surgery. Comparison of procalcitonin in patients without (*full circles* and *full line*, n = 201) and with postoperative renal dysfunction (*open circle* and *dotted line*, n = 75) in the control group (*A*) and the infection group (*B*). Data are median (95% confidence interval). * $P < 0.05$ (between group comparisons). Reproduced with permission from Amour *et al*.[65]

One of the important variables associated with a decrease in organ function is age.[10] Because we are more frequently caring for elderly patients, it is important that biomarkers be tested not only in a middle-aged population but also in an elderly population.

Other Bias

The range of values reported for the sensitivity and specificity in different studies of any biomarker are often very wide. This variability is uncovered by most meta-analyses performed on biomarker studies.[67] Apart from differences concerning the cutoff point, which should be considered as a definition issue, one of the most important reason for this wide variation is that diagnostic studies are plagued by numerous biases[68]:

The main problem resides in the reference test used. In many clinical situations, the definition of case and controls do not rely on a perfect "gold standard" reference test (see infra), and often patients with ambiguous classification are ignored in the analysis. This leads to a case-control design that overestimates the diagnostic accuracy.[3,69] This bias (also called spectrum bias) may be associated with the largest bias effect.[3]

Selection bias occurs when nonconsecutive patients or not randomly selected patients are included.

The lack of blinding for the biomarker tested can also introduce bias, which usually overestimates diagnostic accuracy, although this effect seems to be relatively small.[3]

The verification bias is caused by the selection of a population of patients who actually receive the reference test, thus ignoring unverified patients, or when not all patients are subjected to the reference test, or when different reference tests are used (called partial verification bias).[3] This bias might be particularly confounding when the decision to perform the reference test is based on the result of the studied test.

The test may produce an uninterpretable result. Although this problem is frequently not formally reported, for these observations are removed from the analysis, this practice can introduce bias. For subjectively interpreted tests (which might be particularly the case for biomarkers measured with rapid bedside technique), interobserver variation can have a silent but important impact that could be neither estimated nor reported.

The accuracy of a biomarker may improve over time because of either improvement in the skills of the biologist or reader or improvement in technology. The measurement of cardiac troponin is a good illustration of progressive improvement in technology over the recent decade, leading to marked and progressive decrease of the cutoff determining normality.[70]

Finally, as for clinical trials, a publication bias might occur because studies showing encouraging results have a higher likelihood to be published. This bias is important to consider for meta-analysis. In the absence of registration of diagnostic studies, it is difficult to estimate the impact of this source of bias.

Statistical Power Issue

Although frequently overlooked,[71] statistical power considerations are as important for studies examining the diagnostic performance of a biomarker as in other types of research (*e.g.*, clinical trials). Thus, all studies on biomarkers should have included an *a priori* calculation of the number of patients needed to be included. The exact statistical power considerations that are relevant for interpreting a biomarker study are dependent on the nature or purpose of the study but generally focus on demonstrating that the sensitivity and/or specificity of a biomarker is superior to some stated value (*e.g.*, sensitivity >0.75). It is of note that this focus on sensitivity and specificity is the case even if the predictive values are of greater interest (as they often are), because predictive values are also dependent on prevalence of the underlying disease (fig. 2).

Calculating the required sample size to provide some level of desired statistical power ($1 - \beta$) is analogous to a traditional difference from a theoretical proportion (*i.e.*, one-sample difference in proportions) using a one-sided CI. For the calculation, the sensitivity or specificity values of the test are treated as a proportion and compared with a minimally acceptable value. For this calculation, a null hypothesis is posited that the sensitivity or specificity of the test is equal to a minimally acceptable value, with the alternative hypothesis that the test value is greater than this minimal value. To test this hypothesis, a type-I error rate must be specified (usually as $\alpha = 0.05$) to construct a one-sided CI ($1 - \alpha$). Further, because of the nature of the inference being conducted, the desired statistical power is conservatively set to 95% ($1 - \beta = 0.95$). Finally, the expected sensitivity or specificity of the biomarker must be anticipated such that the difference in the two proportions can be used in the calculation.

Although this process can seem daunting, there are several resources available to assist researchers and consumers of biomarker research. The calculation is now routinely available on most statistical software applications. The assumptions used in the calculation must still be provided by the user, but elegant algorithms can actually perform the calculation. Second, Flahault *et al.*[72] have recently provided an extensive overview of the process and have even provided tables of values that are routinely encountered. Finally, a growing list of internet sites host statistical power calculators for a variety of applications. Although many of these sites are not formally vetted, several are hosted by Universities and are, thus, quite useful.

Nevertheless, as we advocate going beyond sensitivity and specificity in this review, it should be emphasized that calculation of the required sample size should now be done considering either the sensitivity at a particular false-positive rate,[73] the AUC_{ROC},[73–75] including partial AUC_{ROC},[73,76] or the reclassification indices.[54] Moreover, the objective of a study could also be to determine the value of a cutoff or to compare two or more biomarkers. In fact, diagnosis assessments of biomarkers include numerous forms of statistical analyses, but we have to take into consideration sample size calculations, which is ever feasible even with some difficulties. Thus, clinicians have first to define the aim of the considered research and second to evaluate its ability to conduct this power calculation. In fact, these techniques are not yet available in most of the usual statistical software applications, and more advanced statistical software# might be dissuasive for a punctual use by clinical researchers. Thus, advice of a biostatistician may be very helpful.

Imperfect Reference Test

In a diagnostic study, the reference test should be a gold standard, but in many clinical situations this is not possible. A universally recognized standard may not exist (*e.g.*, cardiac failure), may not have been performed in many patients (*e.g.*, autopsy), or logistically could not be concurrently performed. For example, when evaluating BNP, echocardiography for heart failure is not always performed in the emergency department but is usually performed later during hospitalization.[34] Moreover, in many situations, biomarkers are compared with derived scores from several clinical metrics that have unknown reliability in place of a confirmed diagnosis. This practice was seen in the Framingham score for heart failure and use of the biomarker BNP, criteria of systemic inflammatory response syndrome and sepsis and use of procalcitonin,[77] and Risk, Injury, Failure, Loss, and End-stage Kidney (RIFLE) score for acute renal failure.[78]

When an imperfect reference test must be used, it should be recognized that measures of test performance can be distorted.[79] Glueck *et al.*[80] showed that when inappropriate reference standards are used, the observed AUC_{ROC} can be greater than the true area, with the typical direction of the bias being a strong inflation in sensitivity with a slight deflation of specificity. Taken together, this information warrants the use of reliable reference standards that are not prone to such bias.

There are several options available to improve a reference standard when a gold standard does not exist or cannot be used. First, expert consensus can be used to define the diagnosis. For this task, at least three experts are needed with a majority rule.[81] These experts should have complete access to all available information, except that concerning the biomarker test, to which they should be blinded. The statistical agreement between experts should be quantified and reported. A second option is to assign a probability value (*i.e.*, 0–1) that corresponds to a subjective or derived (logistic regression using dedicated variables) probability that a patient has the disease. Third, one can use covariance information to estimate a model of the multivariate normal distributions of disease-positive and disease-negative patients when several accurate tests are being compared. Finally, one can transform the diagnostic problem into a clinical outcome problem.[82]

There are also some situations in which the reference test outcome is not binary (yes or no) but ordinal or continuous. Obuchowski *et al.*[83] proposed a ROC type nonparametric measure of diagnostic accuracy. This is a discrimination test in which a diagnostic test is compared with a continuous reference test to determine how well it distinguishes outcome of the reference test.

New biomarkers do not only modify our diagnostic process but also change the definition of a disease.[84] For example, cardiac troponin has progressively modified the definition of the diagnosis of myocardial infarction.[70] Glasziou *et al.*[84] have proposed three main principles that may assist the replacement of a current reference test: the consequences of the new test can be understood through disagreements be-

# For example, R software. Available at: http://cran.r-project.org/. Accessed December 2, 2009.

tween the reference and the new test; resolving disagreements between new and references test requires a fair, but not necessarily perfect, “umpire” test; possible umpire tests include causal exposures, concurrent testing, prognosis, or the response to treatment. A fair umpire test means that it does not favor either the reference or the new test and, thus, is considered as unbiased.

STARD Statement for Diagnosis Studies

The STARD initiative was published recently to improve the quality of reporting for studies of diagnostic accuracy.[4] Complete and accurate reporting of biomarker studies allow the reader to detect potential bias and judge the clinical applicability and generalization of results. The STARD recommendations follow the template of the Consolidated Standards of Reporting Trials statement for the reporting of randomized controlled trials (RCTs).[5] The STARD guideline attempts to improve the reporting of several factors that may threaten the internal or external validity of the results of a study, including design deficiencies, selection of the patients, execution of the index test, selection of the reference standard, and analysis of the data. That these reporting improvements are needed is evidenced by a survey of diagnostic accuracy studies published in major medical journals between 1978 and 1993 that found generally poor methodologic quality and underreporting of key methodologic elements.[85] Similar shortcomings were observed in most specialized journals.[86]

The STARD guideline (table 4) provides a checklist of 25 items to verify that relevant information is provided. Similar to the reporting of RCTs, a flow diagram is strongly recommended, with an item advocating the extensive use of CIs. Although the STARD initiative is a crucial step for improving reporting, because of the heterogeneity of available methods, only a few general recommendations concerning the statistical methods were offered.

Associated Clinical Predictors and/or Multiple Biomarkers

Pretest risk for all patients of a population is rarely equal, and clinical predictors, such as age, are most often present. The clinical question remains how a new biomarker improves the risk stratification determined by the classic predictors (clinical and biologic). The risk prediction obtained with a new biomarker alone, even if excellent, may have no clinical application if it does not improve the risk stratification obtained with the usual predictors. The use of a risk prediction model (a statistical model that combines information from several predictors) is the most frequent approach. The purpose of a risk prediction model is to accurately stratify individuals into clinically relevant risk categories. The common types of models include logistic regression, Cox proportional hazard, and classification trees. Two nested models are then constructed and compared, one with usual predictors and the other with usual predictors and the new biomarkers.

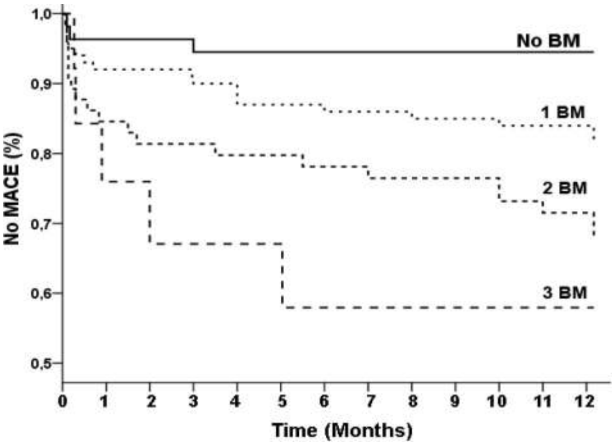

Fig. 7. Multiple biomarkers in cardiac surgery: cumulative postoperative survival at mean of covariates (European System for Cardiac Operative Risk Evaluation) without major cardiac events (MACE) according to elevation of cardiac troponin I (>3.5 ng/ml), B-type natriuretic peptide (>880 pg/ml), and C-reactive protein (>180 mg/l). Patients were categorized according to elevation of no biomarker (BM) (n = 58; survival rate at 1 yr, 95%), only one BM (n = 98; survival rate at 1 yr, 82%), two BMs (n = 56; survival rate at 1 yr, 63%), or three BMs (n = 12; survival rate at 1 yr, 58%). All survival curves significantly differ from each other ($P < 0.05$). Reproduced with permission from Fellahi *et al.*[60]

In most clinical situations, clinicians want to apply more than one biomarker. A multiple biomarker approach is more widely used in several domains of modern medicine. For example, stratification of the cardiovascular risk in the general population is improved when considering several biomarkers such as C-reactive protein, troponin, and BNP.[87] After cardiac surgery, a multiple biomarker approach has been shown to improve the prediction of poor long-term outcome when compared with the classic clinical Euroscore (fig. 7).[60] There are two main approaches: (1) several biomarkers testing the same pathophysiologic process; (2) different biomarkers testing different pathologic processes. For example, C-reactive protein may assess the postoperative inflammatory response, BNP the cardiac strain, and troponin myocardial any ischemic damage, all of them influencing final outcome in cardiac surgery.[11]

However, the results from different tests are usually not independent of each other, even if they assess different pathophysiologic mechanisms, indicating that sequential Bayesian calculations may not be appropriate. Other statistical methods, which take into account colinearity and interdependence, should be considered. In the case of a logistic regression, the regression coefficients can be used to calculate the probability of disease presence, and a simplified score can then be derived from these coefficients. Biases associated with the multivariate analyses are common and are largely able to impact the replication of these scores. The best way to limit this is the use of both internal and external validations of the models.[88]

Meta-analysis of Diagnostic Studies

Systematic reviews are conducted to help gain insight on the available evidence for a research topic. For a meta-analysis,

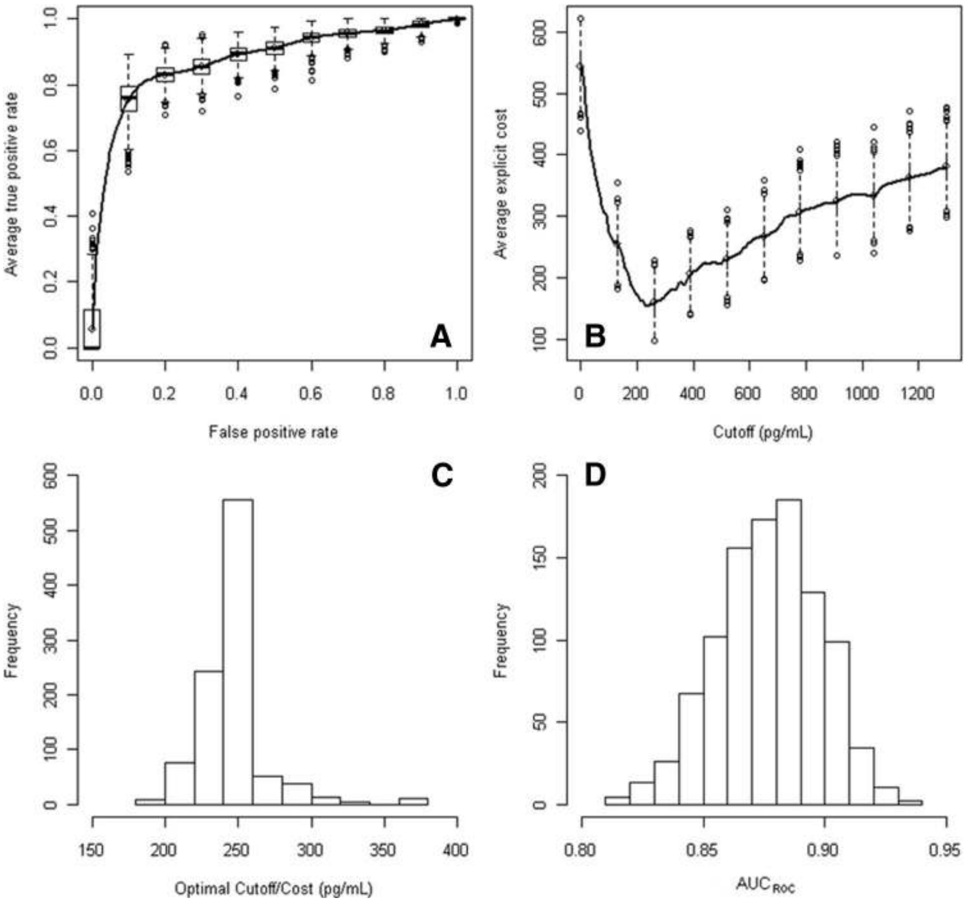

Fig. 8. Reanalysis of the predictive value of brain natriuretic peptide for cardiogenic pulmonary edema in elderly patients (>65 yr), patients admitted to the emergency department for acute dyspnea. (*A*) A bootstrap analysis (1,000 random samples) was performed to obtain a *box plot* in the receiver-operating characteristics (ROC) curve. (*B*) A cost-benefit analysis was performed to choose the best cutoff point. (*C*) The bootstrap analysis also allowed the determination of the best cutoff using the Youden method; this could also provide another definition of the gray zone or a 95% confidence interval for the cutoff point. (*D*) The bootstrap analysis shows the distribution of the area under the receiver-operating characteristic curve (AUC_{ROC}). Adapted from data from Ray *et al.*[31]

this is conducted in a two-stage process where summary statistics are first computed for individually considered studies, and then a weighted average is computed across studies.[89] In this regard, meta-analyses of biomarker diagnostic studies are similar to other types of meta-analyses.[67] For that reason, the reporting of meta-analyses of biomarker diagnostic studies should generally follow existing guidelines for meta-analysis such as Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA).[90]

Despite general similarities, the conduct of meta-analysis for biomarker studies differs from meta-analyses of RCT in several important ways. First, the assessment of study quality for diagnostic studies varies considerably from RCTs. Individual studies on the same biomarker can vary considerably on the choice of threshold used, the population under study, and even the measurement of the biomarker or reference standard. The choice of patient recruitment strategy can impact the assessment as well, with one study finding that recruiting patients and controls separately can lead to an overestimation of the test's diagnostic accuracy.[3] To assist in the evaluation of study quality, several specialized tools have been created, such as STARD guidelines (table 4), and the quality assessment of studies of diagnostic accuracy (QUADAS) has been included in systematic reviews.[91] The accurate characterization of a biomarker's performance in a particular setting for a specific population is dependent on sorting through the available evidence to primarily focus on only relevant studies of high quality.

The statistical techniques used to aggregate the results of biomarker studies also differ from the meta-analyses of RCTs. The meta-analysis of diagnostic studies requires the consideration of two index measures (*e.g.*, sensitivity and specificity), as opposed to a single index in the meta-analysis of an RCT.[92] It is also expected that heterogeneity in the indices will be observed from several different sources,[3] and this heterogeneity must be considered in the statistical model used to pool the estimates.[93] The choice of which type of model and estimation strategy to use is not trivial, with several novel techniques such as the hierarchical summary ROC[94] and multivariate random-effects meta-analysis[95] offering distinct advantages over traditional approaches. For the interested reader, Deeks *et al.*[92] offers an informative illustration of the meta-analytical process.

Conclusions

The studies of biomarkers, either as diagnostic or prognostic variables, are often presented with poor biostatistics and methodologic flaws that often preclude them from providing a reliable and reproducible scientific message. This practice dramatically contrasts with bioengineering companies' prolific development of new biomarkers exploring the cardiovascular system, kidney, central nervous system, inflammation, and sepsis. To address this gap in methodologic quality, some recommendations have been produced recently, but even they did not cover the entire biomarker domain.[4] Two main reasons may explain this situation. First, there is a widely recognized delay between the development of biostatistical techniques and their implementation in medical journals. Second, even in biostatistics, this domain has not been thoroughly explored and developed. Thus, there is an urgent need to accelerate the improvement of our methods in analyzing biomarkers, particularly concerning the use of the ROC curve, choice of cutoff point, including the definition of a gray zone, appropriate *a priori* calculation of the number of patients to include, and the extensive use of validation techniques. Admittedly, if we retrospectively look at one of our recent studies,[66] we now realize that considerable improvement in information could have been provided (fig. 8), that should be considered as a promising and encouraging signal.[96]

There are important potential shortcomings in biomarker studies. Investigators should be aware of this when designing their studies, editors and reviewers when analyzing a manuscript, and readers when interpreting results.

References

1. Baker M: In biomarker we trust? Nature Biotechnol 2005; 23:297–304
2. Riou B: Troponin: Important in severe trauma and a first step in the biological marker revolution. ANESTHESIOLOGY 2004; 101:1259–60
3. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM: Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282:1061–6
4. Bossuyt PM, Reitsma JR, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, Lijmer JG: The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. Ann Intern Med 2003; 138:W1–12
5. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche PC, Lang T; CONSORT GROUP (Consolidated Standards of Reporting Trials): The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. Ann Intern Med 2001; 134:663–94
6. Le Manach Y, Perel A, Coriat P, Godet G, Bertrand M, Riou B: Early and delayed myocardial infarction after abdominal aortic surgery. ANESTHESIOLOGY 2005; 102:885–91
7. Jebali MA, Hausfater P, Abbes Z, Aouni Z, Riou B, Ferjani M: Assessment of the accuracy of procalcitonin to diagnose postoperative infection after cardiac surgery. ANESTHESIOLOGY 2007; 107:232–8
8. Hausfater P, Juillien G, Madonna-Py B, Haroche J, Bernard M, Riou B: Serum procalcitonin measurement as diagnostic and prognostic marker in febrile adult patients presenting to the emergency department. Crit Care 2007; 11:R60
9. Konstantinides S, Geibel A, Olschewski M, Kasper W, Hruska N, Jackle S, Binder L: Importance of cardiac troponins I and T in risk stratification of patients with acute pulmonary embolism. Circulation 2002; 106:1263–8
10. Rivera R, Antognini J: Perioperative drug therapy in elderly patients. ANESTHESIOLOGY 2009; 110:1176–81
11. Fellahi JL, Hanouz JL, Gué X, Guillou L, Riou B: Kinetics analysis of cardiac troponin I release is no more accurate than a single 24 h measurement to predict adverse outcome after conventional cardiac surgery. Eur J Anaesthesiol 2008; 25:490–7
12. Simon T, Verstuyft C, Mary-Krause M, Quteineh L, Drouet E, Méneveau N, Steg G, Ferrières J, Danchin N, Becquemont L; French Registry of acute ST-elevation and non-St-elevation myocardial infarction (FAST-MI) investigators: Genetic determinants of response to clopidogrel and cardiovascular events. N Engl J Med 2009; 360:363–75
13. Nobre V, Harbarth S, Graf JD, Rohner P, Pugin J: Use of procalcitonin to shorten antibiotic treatment duration in septic patients: A randomized trial. Am J Respir Crit Care Med 2008; 177:498–505
14. Howell MD, Donnino M, Clardy P, Talmor D, Shapiro NI: Occult hypoperfuision and mortality in patients with suspected infection. Intensive Care Med 2007; 33:1892–9
15. Zweig MH, Campbell G: Receiver-operating characteristics (ROC) plots: A fundamental evaluation tool in clinical medicine. Clin Chem 1993; 39:561–77
16. Christ-Crain M, Jaccard-Stoltz D, Bingisser R, Gencay MM, Huber PR, Tamm M, Muller B: Effect of procalcitonin-guided treatment on antibiotic use and outcome in lower respiratory tract infections: Cluster-randomized, single-blinded intervention trial. Lancet 2004; 363:600–7
17. Schneider HG, Lam L, Lokuge A, Krum H, Naughton MT, De Villiers Smit P, Bystrzycki A, Eccleston D, Federman J, Flannery G, Cameron P: B-type natriuretic peptide testing, clinical outcomes, and health services use in emergency department patients with dyspnea: A randomized trial. Ann Intern Med 2009; 150:365–71
18. Marshall JC, Reinhardt K; for the International Sepsis Forum: Biomarkers of sepsis. Crit Care Med 2009; 37:2290–8
19. Parmigiani G: Modeling in Medical Decision Making: A Bayesian Approach. New York, John Wiley and Sons, 2002
20. Foxcroft DR, Kypri K, Simonite V: Bayes' Theorem to estimate population prevalence from Alcohol Use Disorders Identification Test (AUDIT) scores. Addiction 2009; 104:1132–7
21. Fagan TJ: Nomogram for Bayes theorem. N Engl J Med 1975; 293:257
22. Alberg AJ, Park JW, Hager BW, Brock MV, Diener-West M: The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. JGIM 2004; 19:460–5
23. Youden WJ: Index for rating diagnostic tests. Cancer 1950; 3:32–5
24. Hilden J, Glasziou P: Regret graphs, diagnostic uncertainty and Youden's index. Stat Med 1996; 15:969–86
25. Falcoz PE, Laluc F, Toubin MM, Puyraveau M, Clement F, Mercier M, Chocron S, Etievent JP: Usefulness of procalcitonin in the early detection of infection after thoracic surgery. Eur J Cardiothorac Surg 2005; 27:1074–8
26. Brenner H, Gellefer O: Variation of sensitivity, specificity, and likelihood ratios and predictive values with disease prevalence. Stat Med 1997; 16:981–91
27. Cook NR: Use and misuse of the receiver operating characteristic curve in risk stratification. Circulation 2007; 115:928–35
28. Lemiale V, Renaud B, Moutereau S, N'Gako A, Salloum M, Calmettes MG, Hervé J, Boraud C, Santin A, Grégo JC, Braconnier F, Roupie E: A single procalcitonin level does not predict adverse outcomes of women with pyelonephritis. Eur Urol 2007; 51:1394–401
29. Gibot S, Kolopp-Sarda MN, Béné MC, Cravoisy A, Levy B, Faure GC, Bollaert PE: Plasma level of a triggering receptor expressed on myeloid cells-1: Its diagnostic accuracy in

patients with suspected sepsis. Ann Intern Med 2004; 141:9–15

30. Sweets JA: Measuring the accuracy of diagnostic systems. Science 1988; 240:1285–93
31. Ray P, Arthaud M, Lefort Y, Birolleau S, Beigelman C, Riou B; the EPIDASA Group: Usefulness of B-type natriuretic peptide in elderly patients with acute dyspnea. Intensive Care Med 2004; 30:2230–6
32. McClish DK: Analyzing a portion of the ROC curve. Med Decis Making 1989; 9:190–5
33. Hanley JA, McNeil BJ: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148:839–43
34. DeLong ER, DeLong DM, Clarke-Pearson DL: Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1988; 44:837–45
35. Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP: A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. Med Decis Making 1997; 17:94–102
36. Zhang DD, Zhou XH, Freeman DH Jr, Freeman JL: A nonparametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. Stat Med 2002; 21:701–15
37. Maisel AS, Krishnaswamy P, Nowak RM, McCord J, Hollander JE, Duc P, Omland T, Storrow AB, Abraham W, Wu A, Clopton P, Steg G, Westheim A, Wold Knudsen C, Perez A, Kazanegra R, Herrmann H, McCullough P: Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. N Engl J Med 2002; 347:161–7
38. Schisterman EF, Perkins NJ, Bondell H: Optimal cut-points and its corresponding Youden index to discriminate individuals using pooled blood samples. Epidemiology 2005; 16:73–81
39. Perkins NJ, Schisterman EF: The inconsistency of "optimal-"cutpoints obtained using two criteria based on the receiver operating characteristics curve. Am J Epidemiol 2006; 163:670–5
40. Hausfater P, Fillet AM, Rozenberg F, Arthaud M, Trystram D, Huraux JM, Lebon P, Riou B: Prevalence of viral infection markers by polymerase chain raction amplification and interferon-alpha measurement among patients undergoing lumbar puncture in an emergency department. J Med Virol 2004; 73:137–46
41. McNeil BJ, Keeler E, Adelstein SJ: Primer on certain elements of medical decision making. N Engl J Med 1975; 293:211–5
42. Metz CE: Basic principles of ROC analysis. Sem Nucl Med 1978; 8:283–8
43. Cantor SB, Sun CC, Tortolero-Luna G, Richards-Kortum R, Follen M: A comparison of C/B ratios from studies using receiver operating characteristic curve analysis. J Clin Epidemiol 1999; 52:885–92
44. Ewald B: Post hoc choice of cut points introduced bias to diagnostic research. J Clin Epidemiol 2006; 59:798–801
45. Beck JR, Shultz EK: The use of relative operating characteristics (ROC) curve in test performance evaluation. Arch Pathol Lab Med 1986; 110:13–20
46. Hilgers RA: Distribution-free confidence bounds for ROC curves. Methods Inf Med 1991; 30:96–101
47. Molinaro AM, Simon R, Pfeiffer RM: Prediction error estimation: A comparison of resampling method. Bioinformatics 2005; 21:3301–7
48. Carpenter J, Bithell J: Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. Stat Med 2000; 19:1141–64
49. Fellahi JL, Hedoire F, Le Manach Y, Monier E, Guillou L, Riou B: Determination of the threshold of cardiac troponin I associated with a adverse postoperative outcome after cardiac surgery: A comparative study between coronary artery bypass graft, valve surgery, and combined surgery. Crit Care 2007; 11:R106
50. Brown MD, Reeves MJ: Interval likelihood ratios: Another advantage for the evidence-based diagnostician. Ann Emerg Med 2003; 42:292–7
51. Ware JH: The limitations of risk factors as prognostic tools. N Engl J Med 2006; 355:2615–7
52. 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
53. Cook N, Ridker PM: Advances in measuring the effect of individual predictors of cardiovascular risk: The role of classification measures. Ann Intern Med 2009; 150:795–802
54. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS: Evaluating the added predictive ability of a new marker: From are under the ROC curve to reclassification and beyond. Stat Med 2008; 27:157–72
55. Hausfater P, Megarbane B, Dautheville S, Patzak A, Andronikof A, Santin A, André S, Korchia L, Terbaoui N, Kierzek G, Doumenc B, Leroy C, Riou B: Prognostic factors in non-exertionnal heatstroke. Intensive Care Med 2009; 36:272–80
56. Greenland S: The need for reorientation toward cost-effective prediction: Comments on "Evaluating the added predictive ability of a new marker: From area under ROC curve to reclassification and beyond" by M.J. Pencina et al, Statistics in Medicine. Stat Med 2008; 27:199–206
57. Saah AJ, Hoover DR: "Sensitivity" and "specificity" reconsidered: The meaning of the terms in analytical and diagnostic settings. Ann Intern Med 1997; 126:91–4
58. Altman DG: Diagnostic tests, In: Statistics with Confidence, 2nd edition. Edited by Altman DG, Machin D, Bryant TN, Gardner MJ. Bristol, United Kingdom, BMJ Books, 2000, pp 105–9
59. De Winter RJ, Koster RW, Sturk A, Sanders GT: Value of myoglobin, troponin T, and CM-MBmass in ruling out an acute myocardial infarction in the emergency room. Circulation 1995; 92:3401–7
60. Fellahi J-L, Hanouz J-L, Le Manach Y, Gué X, Monier E, Guillou L, Riou B: Simultaneous measurement of cardiac troponin I, B-type natriuretic peptide, and C-reactive protein for the prediction of long-term cardiac outcome after cardiac surgery. ANESTHESIOLOGY 2009; 111:250–7
61. Mower WR: Evaluating bias and variability in diagnostic test. Ann Emerg Med 1999; 33:85–91
62. Hausfater P: Le dosage de la procalcitonine en pratique clinique chez l'adulte. Rev Med Intern 2007; 28:296–305
63. Hausfater P, Hurtado M, Pease S, Juillien G, Lvovschi V, Salehabadi S, Lidove O, Wolf M, Bernard M, Chollet-Martin S, Riou B: Is procalcitonin a marker of critical illness in heatstroke? Intensive Care Med 2008; 34:1377–83
64. Janes H, Pepe MS: Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: An old concept in a new setting. Am J Epidemiol 2008; 168:89–97
65. Amour J, Birenbaum A, Langeron O, Le Manach Y, Bertrand M, Coriat P, Riou B, Bernard M, Hausfater P: Influence of renal dysfunction on the accuracy of procalcitonin to diagnose postoperative infection after vascular surgery. Crit Care Med 2008; 36:1147–54
66. Ray P, Arthaud M, Birolleau S, Isnard R, Lefort Y, Boddaert J, Riou B; the EPIDASA Group: Comparison of brain natriuretic peptid and probrain natriuretic peptid in the diagnosis of cardiogenic pulmonary edema in patients older than 65 years. J Am Geriatr Soc 2005; 53:643–8
67. Ryding ADS, Kumar S, Worthington AM, Burgess D: Prognostic value of Brain natriuretic peptide in noncardiac surgery. A Meta-analysis. ANESTHESIOLOGY 2009; 111:311–9
68. Begg CB: Biases in the assessment of diagnostic tests. Stat Med 1987; 6:411–23
69. Fischer JE, Bachmann LM, Jaeschke R: A reader's guide to the interpretation of diagnostic test properties: Clinical example of sepsis. Intensive Care Med 2003; 29:1043–51
70. Archan S, Fleisher LA: From creatine kinase-MB to troponin: The adoption of a new standard. ANESTHESIOLOGY 2010; 112: 1005–12

71. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM: Sample sizes of studies on diagnostic accuracy: Literature survey. BMJ 2006; 332:1127–9
72. Flahault A, Cadilhac M, Thomas G: Sample size calculation should be performed for design accuracy in diagnostic test studies. J Clin Epidemiol 2005; 58:859–62
73. Obuchowski NA: Sample size calculations in studies of test accuracy. Stat Methods Med Res 1998; 7:371–92
74. Liu JP, Ma MC, Wu CY, Tai JY: Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves. Sat Med 2006; 25:1219–38
75. Obuchowski NA, McClish DK: Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. Sat Med 1997; 16:1529–42
76. Li CR, Liao CT, Liu JP: A non-inferiority test for diagnostic accuracy based on the paired partial areas under ROC curves. Stat Med 2008; 27:1762–76
77. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, Cohen J, Opal SM, Vincent JL, Ramsay G; SCCM/ESICM/ACCP/ATS/SIS: 2001 SCCM/ESICM/ACCP/ATS/SIS International sepsis definitions conference. Crit Care Med 2003; 31:1250–6
78. Abosaif NY, Tolba YA, Heap M, Russel J, El Nahas AM: The outcome of renal failure in the intensive care unit according to RIFLE: Model application, sensitivity, and predictability. Am J Kidney Dis 2005; 46:1038–48
79. Valenstein PN: Evaluating diagnostic tests with imperfect standards. Am J Cin Pathol 1990; 93:252–8
80. Glueck DH, Lamb MM, O'Donnell CI, Ringham BM, Brinton JT, Muller KE, Lewin JM, Alonzo TA, Pisano ED: Bias in trials comparing paired continuous tests can cause researchers to choose the wrong screening modality. BMC Med Res Methodol 2009; 9:4
81. Ray P, Birolleau S, Lefort Y, Becquemin MH, Beigelman C, Isnard R, Teixeira A, Arthaud M, Riou B, Boddaert J: Acute respiratory failure in the elderly: Etiology, emergency diagnosis and prognosis. Crit Care 2006; 10:R82
82. Henckelman RM, Kay I, Bronakill MJ: Receiver operating characteristic (ROC) analysis without truth. Med Decis Making 1990; 10:24–9
83. Obuchowski NA: An RO: C-type measure of diagnostic accuracy when the gold standard is continuous-scale. Stat Med 2006; 25:481–93
84. Glasziou P, Irwig L, Dekks JJ: When should a new test become the current reference standard? Ann Intern Med 2008; 149:816–22
85. Reid MC, Lachs MS, Feinstein AR: Use of methodological standards in diagnostic tests research. Getting better but still not good. JAMA 1995; 274:645–51
86. Obuchowski NA, Lieber ML, Wians FH: ROC curves in clinical chemistry: Uses, misuses, and possible solutions. Clin Chem 2004; 50:1118–25
87. Zethelius B, Berglund L, Sundström J, Ingelsson E, Basu S, Larsson A, Venge P, Arnlöv J: Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. N Engl J Med 2008; 358:2107–16
88. Katz MH: Multivariable analysis: A primer for readers of medical research. Ann Intern Med 2003; 138:644–50
89. Deeks JJ, Altman DG, Bradburn MJ: Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis, Systematic Reviews in Health Care: Meta-analysis in Context, 2nd edition. Edited by Egger M, Davey Smith G, Altman DG. London, BMJ Books, 2001
90. Moher D, Liberati A, Tetzlaff J, Altman DG; The PRISMA Group: Preferred reporting items for systematic reviews and meta-analysis: The PRISMA statement. PLOS Med 2009; 6:e100097
91. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J: The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2009; 6:9
92. Deeks JJ: Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001; 323:157–62
93. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA: A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 2007; 8:239–51
94. Rutter CM, Gatsonis CA: A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001; 20:2865–84
95. Hamza TH, Arends LR, van Houwelingen HC, Stijnen T: Multivariate random effects of meta-analysis of diagnostic test with multiple thresholds. BMC Med Res Methodol 2009; 9:73
96. Sing T, Sander O, Beerenwinkel N, Lengauer T: ROCR: Visualizing classifier performance in R. Bioinformatics 2005; 21:3940–1