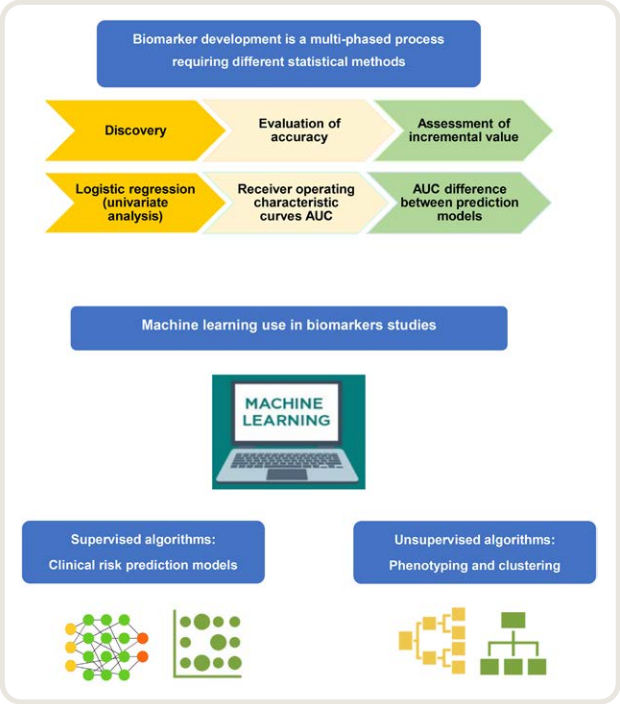
Evaluation of Biomarkers in Critical Care and Perioperative Medicine

A Clinician's Overview of Traditional Statistical Methods and Machine Learning Algorithms

Sabri Soussi, M.D., M.Sc., Gary S. Collins, Ph.D., Peter Jüni, M.D., Alexandre Mebazaa, M.D., Ph.D., Etienne Gayat, M.D., Ph.D., Yannick Le Manach, M.D., Ph.D.†

(Anesthesiology 2021; 134:15–25)



SUMMARY

Interest in developing and using novel biomarkers in critical care and perioperative medicine is increasing. Biomarkers studies are often presented with flaws in the statistical analysis that preclude them from providing a scientifically valid and clinically relevant message for clinicians. To improve scientific rigor, the proper application and reporting of traditional and emerging statistical methods (*e.g.*, machine learning) of biomarker studies is required. This Readers' Toolbox article aims to be a starting point to nonexpert readers and investigators to understand traditional and emerging research methods to assess biomarkers in critical care and perioperative medicine.

(Anesthesiology 2021; 134:15–25)

Biomarkers are increasingly used as personalized markers of diagnosis, in the assessment of disease severity or risk, and to prognosticate and guide clinical decisions.[1,2] Biomarkers exploring the cardiovascular system and kidneys, as well as inflammation, have proliferated in critical care and perioperative medicine. While existing guidelines are available to provide guidance on key information to report in a biomarker study, they do not explicitly provide guidance on appropriate statistical methods.[3–5] The use of inappropriate statistical methods for assessing the clinical value of biomarkers obfuscates any meaningful interpretation and usability of the study findings for clinicians.[1,2]

This article does not aim to be an exhaustive review of biostatistical and methodologic issues, but rather, to be a starting point for nonexpert readers and investigators to understand traditional and emerging research methods used to assess biomarkers in critical care and perioperative medicine. We provide toolboxes with reporting checklists to assist authors and readers in the use of these statistical methods.

Different Biomarker Development Phases

A biomarker may have several roles in clinical practice. It may provide a diagnosis, have a prognostic role, be used to

Image: Jorge A. Galvez, M.D., M.B.I., Terri Navarette, and Allan F. Simpao, M.D., M.B.I.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are available in both the HTML and PDF versions of this article. Links to the digital files are provided in the HTML text of this article on the Journal's Web site (www.anesthesiology.org).

†Deceased.

Submitted for publication March 23, 2019. Accepted for publication October 5, 2020. From the Department of Anesthesiology, Critical Care and Burn Center, Lariboisière - Saint-Louis Hospitals, University Medical Departments Parabol, Assistance Publique-Hôpitaux de Paris (AP-HP) Nord; Inserm Medical Research Unit 942, Cardiovascular Markers in Stress Conditions (MASCOT), University of Paris, France (S.S., A.M., E.G.); Interdepartmental Division of Critical Care, Keenan Research Centre for Biomedical Science and Institute of Medical Sciences, Faculty of Medicine, University of Toronto, Toronto, Canada (S.S.); Centre for Statistics in Medicine, University of Oxford, Oxford, United Kingdom (G.S.C.); Applied Health Research Centre, Li Ka Shing Knowledge Institute of St. Michael's Hospital, Department of Medicine and Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada (P.J.); and Departments of Anesthesia and Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University, Population Health Research Institute, Hamilton, Canada (Y.L.M.).

 Anesthesiology 2021; 134:15–25. DOI: 10.1097/ALN.0000000000003600

<zdoi;. DOI: 10.1097/ALN.0000000000003600>

Box 1. What to Look for in Research Using This Method

- Appropriate statistical analysis plan before biomarkers analysis
- Valid methods based on the clinical question/hypothesis, biomarker phase of development (*e.g.*, discovery, evaluation of accuracy and assessment of incremental value), and weaknesses of the statistical methods
- Avoidance of common pitfalls in biomarkers studies (*e.g.*, not considering properties of a biomarker assay, biomarker kinetics, imperfect accepted standard methods and different populations)
- Inappropriate use of machine learning algorithms that results in overfitting models, lack of independent validation, and lack of comparison with simpler modeling approaches

assess treatment responsiveness, or to guide the use of pharmaceuticals in treatment. Biomarkers were also proposed to identify critically ill patients' molecular subphenotypes, regardless of outcome. Examples of biomarkers with different roles used in critical care and perioperative medicine are presented in table 1.[6–14]

Biomarker development is a multiphased process requiring different statistical methods to accomplish the various objectives. The three phases of biomarker development, in chronological order, include (1) discovery; (2) evaluation of predictive (or diagnostic) accuracy; and (3) assessment of incremental value when added to existing clinical prediction (or diagnostic) tools.[15]

Statistical Methods to Evaluate Biomarkers

In the early phase of biomarker development, the association between a biomarker and the outcome is often assessed using regression models and the reporting of odds/hazard ratios or estimates of relative risks to quantify this association, preferably including an assessment of their value over established biomarkers or clinical characteristics. A prospective design is preferable as it facilitates clear inclusion criteria, data collection procedures (minimizing missing data), and standardization of measurements, and ensures all relevant clinical information is measured. Registering a protocol and prespecifying study objectives, biomarkers of interest, and statistical methods will reduce publication bias and selective reporting.[1]

A commonly used approach to estimate biomarker discrimination and the incremental value of a biomarker is to calculate the area under the receiver operating characteristic curve (AUC).[15] The receiver operating characteristic curve is formed by plotting false positive rates (1 − specificity) on the x-axis and the true positive rates (sensitivity) on the y-axis. The AUC quantifies the discriminative ability of the biomarker ranging from 0.5 (*i.e.*, no better than flipping a coin) to 1 (*i.e.*, perfect discrimination). Discrimination is the ability of the biomarker to differentiate those with and without the event (*e.g.*, quantifying whether those with the event tend to have higher biomarker values compared to those who do not).

So-called "optimal" biomarker thresholds are often determined based on maximizing the Youden index (maximum [sensitivity + specificity − 1]).[1] The Youden index is often used to determine the value of the biomarker that maximizes the sum of sensitivity and specificity. However, such an approach is problematic if the biomarker is used to either rule out (high sensitivity) or confirm (high specificity) a diagnosis when negative and positive likelihood ratios can be used to select thresholds. The 95% CI around the "optimal" cutoffs could be reported (*e.g.*, bootstrap resampling).[1,16] Furthermore, dichotomization (and indeed, categorization) of a biomarker is also biologically implausible, as no thresholds of a biomarker exist that cause a sudden change in risk (*e.g.*, there is typically no reason why a person's risk on either side of a cut-point will be dramatically different).

Table 1. Different Roles of a Biomarker in Clinical Practice or Research

Role	Description	Examples
Diagnosis of a disease	To provide a diagnosis more rapidly or more reliably than available diagnostic tools	High sensitivity troponin I for early diagnosis of myocardial infarction[6] Procalcitonin for diagnosis of bacterial infection[7]
Severity assessment and risk stratification	To identify subgroup of severe patients with worse outcome	Blood lactate identifies severe outcome in sepsis and trauma patients[8] Procalcitonin identifies severe outcome in respiratory tract infections[9]
Assessment of treatment effects	To identify the pharmacologic response to a treatment	Efficacy of low-molecular-weight heparins and clopidrogel[10,11]
Treatment monitoring	To assess the response to a therapeutic intervention	Procalcitonin may guide the duration of antibiotic therapy[12]
Patients clustering	To identify subphenotypes of patients with different molecular features independently of outcome (unsupervised approach)	Identification of a hyperinflammatory subphenotype in ARDS patients with a different response to a PEEP strategy[13] Identification of four different phenotypes (mainly based on markers of inflammation, coagulation, and renal injury) in sepsis with different responses to early goal-directed therapy[14]

ARDS, acute respiratory distress syndrome; PEEP, positive end-expiratory pressure.

Categorization (including dichotomization) of a continuous measurement (*e.g.*, biomarkers) should therefore be avoided during statistical analysis, as it will result in a loss of information and negatively impact predictive accuracy.[17–19] The statistical analysis should ideally retain continuous measurements on their original scale, allowing for nonlinear relationships to be considered (using restricted cubic splines or fractional polynomials).[20]

To assess the incremental value of a novel biomarker when added to a clinical model or a standard biomarker, the difference in AUC between two prediction models (improvement in discrimination) is often used.[21] Methods such as the DeLong nonparametric test and the Hanley and McNeil method are then used to compare AUCs of the biomarker under investigation against an already established biomarker or clinical model assessed in the same set of individuals.[22,23] The main limitation in comparing AUCs is that a relatively large "independent" association is needed to result in a meaningfully larger AUC for the new biomarker. In response to the insensitivity of comparing AUCs, reclassification methods (*e.g.*, net reclassification index, integrated discrimination index) have been proposed and are described in table 2.[8] However, despite their popularity, it has since been shown that these approaches offer little more than existing approaches and can be unreliable in certain situations.[24] Reclassification methods have been shown to have inflated false positive rates when testing the improved predictive performance of a novel biomarker.[25,26] Approaches based on net benefit using decision analytic methods are now widely recommended, as they allow for meaningful assessment of a new biomarker against an established biomarker or combination of biomarkers by comparing the benefits and risk of decisions (true positives) against their relative harms (false positives).[21,27–28] The comparison

Box 2. Where to Find More Information on This Topic

- Looney SW, Hagan JL. Analysis of biomarker data: A practical guide. Hoboken, New Jersey, John Wiley & Sons, Inc., 2015
  An introduction to biomarker analysis that includes the principles of good research study design; also contains SAS and R-based statistical packages
- Rabbee N. Biomarker analysis in clinical trials with R. New York, Chapman and Hall/CRC, 2020
  Describes the design and the statistical analysis plan of biomarker trials and covers the topic of combining multiple biomarkers to predict drug response/outcome using machine learning; reproducible codes and examples are provided in R

Table 2. Examples of Statistical Metrics and Methods to Evaluate Biomarkers

Statistical Metrics and Methods	Description
Discrimination/performance[19]	
Receiver operating characteristic curve	Visual description of discriminatory performance of a biomarker cut-off (true positive rate = sensitivity and false positive rate = 1 – specificity)
AUC	Summary measure of the receiver operating characteristic curve
Incremental value[49]	
AUC difference	Change in the AUC between two prediction models
Net reclassification index	Net reclassification index event (true positive rate variation) = change in the proportion of cases correctly identified Net reclassification index nonevent (false positive rate variation) = change in the proportion of controls incorrectly identified
Integrated discrimination index	Provides the difference in discrimination slopes
Regression models[36,49] (prediction models)	
Logistic regression, Cox regression	Logistic regression is mainly used for short-term binary outcomes, while survival methods (*e.g.*, Cox regression) are used for long-term (censored) outcomes
Regularized regression (ridge regression, least absolute shrinkage and selection operator, elastic net)	An extension of the classic regression methods with an imposed penalty to the fitted model
Supervised learning algorithms (predictive and prognostic models)[49]	
Support vector machines (polynomial, linear, radial basis kernel)	Represent the data in multidimensional feature space and fit a hyperplane (a subspace whose dimension is one less than that of its ambient space) that best isolates the data regarding the outcome
Tree-based (classification and regression trees, random forest, gradient boosted trees)	Decision trees partition the sample data by splitting the variables at discrete cut-points and presented graphically in the form of a tree (groups based on the relationship of the features with outcome)
Neural network	Nonlinear models that extract features from the data and create a set of combinations that best represent the underlying structure to predict an outcome
Unsupervised learning algorithms (clustering and phenotyping)[55,56]	
Latent class analysis	Identifies hidden subgroups (latent classes) in the dataset regardless of their outcome
Cluster analysis (fuzzy *c*-means, hierarchical cluster analysis)	Organizes the dataset into subgroups of maximum commonality based on distance between features

AUC, area under the receiving operating characteristics curve.

is made across all (or a range of) thresholds to evaluate whether the new biomarker has added clinical utility.

Clinical Risk Prediction Models Using Biomarkers

Clinical prediction models are typically developed using regression models (*e.g.*, logistic regression or Cox regression). Logistic regression is mainly used for short-term binary outcomes (*e.g.*, mortality, postoperative myocardial infarction), while survival methods (such as Cox regression) are used for time-to-event outcomes and allow for censoring. Methods for handling missing data should be considered before analysis (*e.g.*, multiple imputation).[29] Predictors with a high amount of missing data can be problematic, indicating the measurement is infrequently performed in daily practice and potentially limiting to a biomarker model's usefulness. The choice of which variables to include in a model needs consideration: variables should have clinical relevance and be readily available at the intended moment of use of the model. The functional form of any continuous variables (*e.g.*, biomarkers) should be appropriately investigated using fractional polynomials or restricted cubic splines to fully capture any nonlinearity in the association of the continuous variables with the outcome.[17,20] The number of candidate predictors to consider in multivariable modeling has historically been constrained relative to the number of outcome events to avoid overfitting, in a concept called events-per-variable that minimizes the risk of overfitting (a condition where a statistical model describes random variation in the data rather than the true underlying relationship).[30] It was widely recommended that studies should only be carried out when the events-per-variable exceeds 10. However, the events-per-variable concept has recently been refuted as having no strong scientific grounds.[31,32] More recently, sample size formulae have been developed that are context-specific to minimize the potential for overfitting; that depends not only on the number of events relative to the number of candidate predictors (*i.e.*, those considered for inclusion, not necessarily those that end up in the final model), but also on the total number of participants, the outcome proportion, and the expected predictive performance.[33]

The use of penalized regression methods (*e.g.*, least absolute shrinkage and selection operator, ridge regression, elastic net) can be considered since it facilitates the choice of variables to be included in the model while minimizing overfitting (table 2).[34–36] However, it was reported that penalized approaches do not necessarily solve problems associated with small sample size.[37] General and biomarker-specific considerations for a developing multivariable prediction models are summarized in Box 3.

More recently, machine learning methods have been gaining interest as an alternative approach to regression-based models in critical care and perioperative medicine.[38–40] Algorithms that improve the clinical use of biomarkers have been developed with machine learning.[41,42] A practical definition of machine learning is that it uses algorithms that automatically learn (*i.e.*, are trained) from data, contrary to clinical prediction models, which are based on prespecifying predictors and their functional forms. These algorithms are divided into two categories: supervised and unsupervised. Supervised machine learning algorithms are used to uncover the relationship between a set of clinical features and biomarkers and known outcomes (predictive and prognostic models; Supplemental Digital Content 1, http://links.lww.com/ALN/C503).[34] The main supervised learning algorithms (*e.g.*, artificial neural network, tree-based methods, support vector machines) are described in table 2. Supervised conventional statistical modeling (*e.g.*, logistic regression) and supervised machine learning should be considered complementary rather than mutually exclusive.[43,44] Marafino *et al.* used a set of vital signs and biologic data from the first 24 h of admission for more than 100,000 unique intensive care unit (ICU) patients in a supervised machine learning algorithm, incorporating measures of clinical trajectory to develop and validate ICU mortality prediction models. The developed

Box 3. General and Biomarker-specific Considerations for Developing Multivariable Prediction Models

- The biomarkers and other explanatory variables should not be highly correlated to avoid redundancy and collinearity (*e.g.*, blood urea nitrogen and creatinine levels in acute kidney injury).[49]
- The selection of predictor variables should depend on clinical relevance in addition to statistical results. Adjustment for covariates that influence the pharmacokinetics of a biomarker (*e.g.*, timing, age, renal function) should be performed.[49]
- Dichotomization or categorization of continuous biomarkers should be avoided. It is biologically implausible and will result in a loss of information and a negative impact on predictive accuracy.[18]
- Fractional polynomials or restricted cubic splines should be considered to account for nonlinearity when assessing continuous biomarkers.[17,20]
- The necessary sample size should be calculated *a priori*.[19]
- Penalized regression methods (*e.g.*, ridge regression, least absolute shrinkage and selection operator, elastic net) should be considered when developing prediction models for low dimensional data with few events to minimize overfitting.[35-37]
- Both discrimination and calibration and should be used to assess the accuracy of biomarkers regression models.[4]
- Methods for handling missing data should be considered (*e.g.*, multiple imputation).[29]
- Prediction models should be internally validated using cross-validation or bootstrapping.[4,49]
- External validation (*e.g.*, assessing model performance in other participant data than was used for the model development) is necessary for determining generalizability.[4,49] Interlaboratory biomarker assay reproducibility needs to be considered.

prediction model for mortality risk, leveraging serial data points for each predictor variable, exhibited discrimination comparable to classical mortality scores (*e.g.*, Simplified Acute Physiology Score III and Acute Physiologic Assessment and Chronic Health Evaluation IV scores).[41] In another example, Zhang *et al.* developed a prediction machine learning model that was used to differentiate between volume-responsive and volume-unresponsive acute kidney injury (AKI) in 6,682 critically ill patients. The extreme gradient boosting combined with a decision tree model was reported to outperform the traditional logistic regression model in differentiating the two groups.[42]

Machine learning is often claimed to have superior performance in high-dimensional settings (*i.e.*, with a large number of explanatory variables). However, there is limited evidence to support this claim in fair and meaningful comparisons with regression-based approaches, as observed in a recent systematic review that showed no performance benefit in clinical studies.[45] While machine learning algorithms are often declared to perform well, they require very large datasets, massive computations, and sufficient expertise.[46] As such, they should not be considered as an "easy path to perfect prediction." Limitations include overfitting, which captures random errors in the training dataset and makes the algorithm not generalizable to future predictions.[47] Approaches to control for overfitting should be adapted from the established clinical prediction model literature to provide an unbiased assessment of predictive accuracy. The other disadvantage of supervised machine learning algorithms is that the underlying association between covariates and outcome cannot be fully understood by clinicians ("black box" models).[48] Conversely, in logistic regression models, the regression coefficient of each covariate can be easily interpreted as the odds ratio (exponentiation of the regression coefficient), which reflects the magnitude of the association with the outcome. A causal interpretation of any association in a prediction model should be avoided, as the aim of a prediction model is to predict and not attribute causality.[49] The interpretation of a model that includes biomarkers reflecting distinct pathophysiological pathways (*e.g.*, myocardial injury, endothelial dysfunction) and their associations with outcome is more intuitive for clinicians when using classical regression models than machine learning algorithms.

Regardless of whether more traditional regression-based approaches or modern machine learning have been used to develop a prediction model, their predictive accuracy can be assessed with several metrics. The two widely recommended measures are calibration and discrimination.[4,49] Calibration assesses how well the risk predicted from the model agrees with the actual observed risk. Calibration can be assessed graphically by plotting the observed risk of outcome against the predicted risk (*e.g.*, mortality, postoperative AKI).[50] Discrimination is a measure of how well the biomarker model can discriminate those who have and those who do not have the outcome of interest (mainly evaluated by the AUC). Another measure of predictive accuracy is the Brier score (squared difference between patient outcome and predicted risk), which reflects the clinical utility of prediction models. However, it has been suggested that the Brier score does not appropriately evaluate the clinical utility of diagnostic tests or prediction models.[51] In practice, no one measure is enough, and the use of multiple metrics characterizing different components of prediction accuracy is required.[52]

Assessing model performance is an important and vital step. During the development of a prediction model, internal validation, using cross-validation or bootstrapping, that mimics the uncertainty in the building process and uses only the original study sample to assess model performance should be carried out.[4,49] The reason to carry out an internal validation is to obtain a bias-corrected estimate of model performance, and for regression-based models, the regression coefficients can be subsequently shrunk due to overfitting.[54] A stronger test of a model is to carry out an external validation, which consists of assessing the performance (discrimination and calibration) of the prediction model in different participant data than was used for the model development (typically collected from different institutions).[4,49] It is often expected that upon external validation, the calibration of the model will be poorer, and methods to recalibrate the model should be considered.[55]

Phenotyping and Clustering Using Biomarkers

Unsupervised machine learning algorithms are used to identify naturally occurring clusters or subphenotypes of patients who have similar clinical or biologic/molecular features without targeting a specific outcome (Supplemental Digital Content 2, http://links.lww.com/ALN/C504).[55,56] Several popular unsupervised learning algorithms (*e.g.*, latent class analysis, cluster analysis) are described in table 2.

An example of using this method in critical care is in personalized medicine research. Patients sharing the same clinical/biologic characteristics are more likely to respond to targeted treatments (*e.g.*, ventilation strategy, fluid administration strategy, statins).[13,14,57] For example, Calfee *et al.* identified two different subphenotypes in acute respiratory distress syndrome (ARDS) patients using latent class analysis (mainly based on clinical data and inflammatory biomarkers) with a different response to a positive end-expiratory pressure (PEEP) strategy.[13] The same group identified two different subphenotypes of ARDS in the Hydroxymethylglutaryl-CoA Reductase Inhibition with Simvastatin in Acute Lung Injury to Reduce Pulmonary Dysfunction cohort, with distinct clinical and biologic features (cytokines) and different clinical outcomes. The hyperinflammatory subphenotype had improved survival with simvastatin compared with placebo.[57] Finally, Seymour *et al.* retrospectively identified four different phenotypes (mainly based on markers of inflammation, coagulation, and renal injury) in sepsis with different responses to early goal-directed therapy.[14]

Challenges and Common Pitfalls in Studies Evaluating Biomarkers

Properties of Biomarker Assay

The precision of the measurement of a biomarker should be assessed. Along this line, the biologic assay and its measurement errors should be reported. The biomarker assay should be sensitive, detecting low concentrations of the biomarker, and specific, in that it is not affected by other molecules. Interlaboratory biomarker assay reproducibility should be considered when assessing a biomarker model performance in a cohort collected from a different institution (external validation).

Another potential issue is that the same biomarker can be produced by different cells with a different pathway mechanism. For example, urinary kidney injury molecule-1 (a biomarker of kidney injury) can also be produced by kidney cancer cells in the absence of kidney injury.[58,59] This point is difficult to control when analyzing data, as the physiology of a novel biomarker is often incompletely known.

Role of Time and Biomarker Kinetics

The timing of biomarker measurement is important to consider. For example, optimal information needed for the diagnosis of myocardial infarction in the postoperative period is obtained at the peak of troponin I concentrations (~24 h).

In major surgery and critical care, biomarkers of interest such as troponin T, N-terminal pro-B-type natriuretic peptide, and C-reactive protein may have completely different kinetics.[60] The main issue in these conditions is the timing of biomarker measurement, which has to take into consideration not only the biomarkers kinetics, but also the time of onset of various pathophysiological processes (*e.g.*, major surgery with a secondary onset of sepsis). Correlations between repeated measurements of the biomarker within an individual should also be considered during analysis. The use of mixed models instead of repeated measures analysis of variance offers distinct advantages in many instances.[61]

Another issue is that renal or hepatic function could influence the elimination of a biomarker and thus its diagnostic properties. This point is important to consider in elderly (with chronic organ dysfunctions), as well as major surgery and critical care, patients who are more likely to present with organ failure.

Along this line, the choice of the "optimal" biomarker measurement timing and adjustment for covariates (*e.g.*, age, renal function) are a real challenge when including them in regression models and machine learning algorithms with clinical parameters gathered in real time.

Imperfect Accepted Standard Methods

The choice of the reference test used to define diseased and nondiseased patients (*e.g.*, postoperative AKI, postoperative myocardial infarction) should be carefully considered. Novel biomarkers are frequently evaluated against accepted standards that are assumed to classify patients with perfect accuracy according to the presence or absence of disease. In practice, reference tests are rarely unerring predictors of disease and tend to misclassify patients. In the case of an imperfect accepted standard (*e.g.*, delayed increase in serum creatinine in the case of AKI[62]), patient misclassification introduces biases into the sensitivity and specificity estimates of the new biomarker. One of the main methods suggested to improve an "imperfect" reference standard is composite reference standards. The rationale is that combining results of different imperfect tests leads to a more accurate reference test. Nevertheless, the accuracy of this approach has been questioned.[63]

There are some situations in which the outcome is not dichotomous (diseased or nondiseased patients) but continuous (*e.g.*, creatinine level variation) or ordinal (*e.g.*, AKI network stages). In this case, a nonparametric estimator of the novel biomarker diagnostic accuracy with an interpretation analogous to the AUC can be applied.[64]

Different Populations

The studied population could greatly influence the diagnostic and prognostic performance of a test. For example, there are different cutoff points of cardiac troponin I to diagnose postoperative myocardial infarction in noncardiac *versus* cardiac surgery, or even in cardiac surgery patients with different procedures (coronary artery bypass graft *vs.* valve surgery).[65] Diagnostic test results may also vary in populations with different demographic characteristics and chronic illnesses (*e.g.*, age, chronic kidney disease). Therefore, authors should describe the exact studied population about which they want to make inference. Adjustment for covariates (external influences) is a major point when including biomarkers in regression models.

Associated Clinical Predictors or Multiple Biomarkers

To assess associated clinical predictors or multiple biomarkers regarding an outcome, a risk prediction model could be developed using logistic regression or Cox regression. Two models are then built and compared based on the difference in the AUC or the difference in the Harrell C-statistic, the first with usual predictors and the second with usual predictors and the novel biomarkers, respectively.[49,66] A multiple biomarker approach could also be applied. For example, stratification of long-term outcome is improved when adding several novel biomarkers of cardiac (N-terminal pro-B-type natriuretic peptide and soluble ST2) and vascular failure (bioactive adrenomedullin) to the multivariable clinical model.[67]

Conceptual issues related to the planning and analysis of biomarker performance are presented in Box 4. This methodologic approach could lead to a decrease in bias and thus obtain a pooled estimation of the biomarker performance. A summary of the most common avoidable pitfalls is presented in Box 5.

Box 4. Conceptual Issues Related to the Planning and Analysis of Biomarker Performance: A General Overview

Predefined endpoint

- Severity assessment and risk stratification
- Prediction of treatment effects or therapeutic monitoring[1]

Study design

- Ideally: Prospective, multicenter study
- Sample size consideration
- Population: Clear description, sufficient number of events[19]

Data analysis

- Receiver operating characteristic curve analysis (area under curve [CI 95%], sensitivity and specificity at multiple thresholds)[19]
- Comparison with established biomarkers or clinical parameters using decision-analytic methods (*e.g.*, net benefit)[21,27–28]
- If multivariable regression model, assessment of collinearity of factors and biomarkers[49]
- Describe any variable selection procedures

If multiple biomarkers

- Comparison of model performance[49,66]
- Detailed description of the combination of the biomarkers

Box 5. How to Avoid Common Pitfalls of the Evaluation of a Biomarker

Properties of a Biomarker Assay

- The precision of measurement of the biomarker should be assessed and reported.[1]
- The measurement of the biomarker should be sensitive (detects low concentration) and specific (without interferences with other molecules) and be reported.
- Interlaboratory biomarker assay reproducibility should be considered when assessing a biomarker model performance in a cohort collected from a different institution (external validation).[1]

Role of Time and Biomarker Kinetics

- The timing of biomarker measurement is important to consider. For example, in the postoperative period, the maximum amount of information to diagnose myocardial infarction is obtained at the peak of troponin I (~24 h).[60]
- Correlations between repeated measurements of the biomarker within an individual should also be considered during analysis. The use of mixed models instead of repeated-measures ANOVA offers distinct advantages in many instances.[61]
- Renal or hepatic dysfunction in critical patients could influence the elimination of a biomarker and thus its diagnostic properties. In this condition, adjustment for these covariates should be performed.[1]

Imperfect Accepted Standard Methods

- The choice of the reference test used to define diseased and nondiseased patients should be carefully considered.[62]
- In the case of an imperfect accepted standard (*e.g.*, delayed increase in serum creatinine in the case of acute kidney injury), the classification potential of the new biomarker could be falsely decreased.

Different Populations

- The studied population could greatly influence the diagnostic and prognostic performance of a test—for example, cardiac troponin I to diagnose postoperative myocardial infarction in noncardiac *versus* cardiac surgery.[65]
- Authors should describe the exact studied population about which they want to make inference.

Conclusions

Biomarker evaluations need a rigorously documented statistical analysis plan, which should be set up before analysis. Investigators need to choose methods based on the clinical question/hypothesis, biomarker phase of development (*i.e.*, discovery, evaluation of accuracy, assessment of incremental value), and weaknesses of the statistical methods. Biomarkers studies are often presented with statistical analyses pitfalls (*e.g.*, not considering properties of a biomarker assay, biomarker kinetics, imperfect accepted standard methods, and different populations) that preclude them from providing a pragmatic scientific message for anesthesiologists and intensivists. Therefore, the tables and toolboxes provided in this article could be used in addition to existing guidelines by investigators, editors, and reviewers to ensure the publication of high-quality biomarker studies for informed readers.

Furthermore, novel biostatistical techniques (*e.g.*, machine learning) are used more and more in critical care and perioperative medicine research. Machine learning is a promising tool to improve outcome prediction and patient subphenotyping to personalize treatments in critical patients. However, we believe that there is a real need for further research to better evaluate the role of machine learning to predict pathology or response to treatment. A direct implementation of machine learning in clinical decision making is as deleterious for patients as a poorly implemented statistical approach. Tables are provided in this article to help the reader to better understand machine learning techniques applied in health care and to avoid their misuse (*e.g.*, overfitting, lack of independent validation, lack of comparison with simpler modeling approaches).

Research Support

Support was provided solely from institutional and/or departmental sources.

Competing Interests

Dr. Jüni has received honoraria to the institution for participation in advisory boards from Amgen; has received research

grants to the institution from AstraZeneca (Cambridge, United Kingdom), Biotronik (Berlin, Germany), Biosensors International (Singapore), Eli Lilly (Indianapolis, Indiana), and The Medicines Company (Parsippany-Troy Hills, New Jersey); and serves as an unpaid member of the steering groups of trials funded by AstraZeneca (Cambridge, United Kingdom), Biotronik (Berlin, Germany), Biosensors (Singapore), St. Jude Medical (St. Paul, Minnesota), and The Medicines Company. Dr. Mebazaa has received speaker's honoraria from Abbott (Chicago, Illinois), Orion (Auckland, New Zealand), Roche (Basel, Switzerland), and Servier (Suresnes, France); and fees as a member of the advisory boards and/or steering committees and/or research grants from BMS (New York, New York), Adrenomed (Hennigsdorf, Germany), Neurotronik (Durham, North Carolina), Roche (Basel, Switzerland), Sanofi (Paris, France), Sphyngotec (Hennigsdorf, Germany), Novartis (Basel, Switzerland), Otsuka (Chiyoda City, Tokyo, Japan), Philips (Amsterdam, Netherlands) and 4TEEN4 (Hennigsdorf, Germany). Dr. Gayat received fees as a member of the advisory boards and/or steering committees and/or from research grants from Magnisense (Paris, France), Adrenomed (Hennigsdorf, Germany), and Deltex Medical (Chichester, United Kingdom). The remaining authors declare no competing interests.

Correspondence

Address correspondence to Dr. Soussi: Department of Anesthesiology and Critical Care and Burn Unit, St. Louis Hospital, 1 avenue Claude Vellefaux, 75010, Paris, France; and Interdepartmental Division of Critical Care, Keenan Research Centre for Biomedical Science and Institute of Medical Sciences, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada. sabri.soussi@uhn.ca.

References

1. Ray P, Le Manach Y, Riou B, Houle TT: Statistical evaluation of a biomarker. Anesthesiology 2010; 112:1023–40
2. Codorniu A, Lemasle L, Legrand M, Blet A, Mebazaa A, Gayat E: Methods used to assess the performance of biomarkers for the diagnosis of acute kidney injury: A systematic review and meta-analysis. Biomarkers 2018; 26:1–30
3. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HC, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF; STARD Group: STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. Clin Chem 2015; 61:1446–52
4. Collins GS, Reitsma JB, Altman DG, Moons KG: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Circulation 2015; 131:211–9
5. Altman DG, McShane LM, Sauerbrei W, Taube SE: Reporting recommendations for tumor marker prognostic studies (REMARK): Explanation and elaboration. BMC Med 2012; 10:51
6. Boeddinghaus J, Nestelberger T, Twerenbold R, Koechlin L, Meier M, Troester V, Wussler D, Badertscher P, Wildi K, Puelacher C, du Fay de Lavallaz J, Rubini Giménez M, Zimmermann T, Hafner B, Potlukova E, Miró Ò, Martin-Sanchez FJ, Keller DI, Reichlin T, Mueller C; APACE investigators: High-sensitivity cardiac troponin I assay for early diagnosis of acute myocardial infarction. Clin Chem 2019; 65:893–904
7. Thomas-Rüddel DO, Poidinger B, Kott M, Weiss M, Reinhart K, Bloos F; MEDUSA study group: Influence of pathogen and focus of infection on procalcitonin values in sepsis patients with bacteremia or candidemia. Crit Care 2018; 22:128
8. Raux M, Le Manach Y, Gauss T, Baumgarten R, Hamada S, Harrois A, Riou B, Duranteau J, Langeron O, Mantz J, Paugam-Burtz C, Vigue B; TRAUMABASE Group: Comparison of the prognostic significance of initial blood lactate and base deficit in trauma patients. Anesthesiology 2017; 126:522–33
9. Kutz A, Briel M, Christ-Crain M, Stolz D, Bouadma L, Wolff M, Kristoffersen KB, Wei L, Burkhardt O, Welte T, Schroeder S, Nobre V, Tamm M, Bhatnagar N, Bucher HC, Luyt CE, Chastre J, Tubach F, Mueller B, Schuetz P: Prognostic value of procalcitonin in respiratory tract infections across clinical settings. Crit Care 2015; 19:74
10. Golukhova EZ, Ryabinina MN, Bulaeva NI, Grigorian MV, Kubova MCh, Serebruany V: Clopidogrel response variability: Impact of genetic polymorphism and platelet biomarkers for predicting adverse outcomes post stenting. Am J Ther 2015; 22:222–30
11. Pannucci CJ, Fleming KI, Bertolaccini CB, Prazak AM, Huang LC, Pickron TB: Assessment of anti-factor Xa levels of patients undergoing colorectal surgery given once-daily enoxaparin prophylaxis: A clinical study examining enoxaparin pharmacokinetics. JAMA Surg 2019; 154:697–704
12. Hohn A, Balfer N, Heising B, Hertel S, Wiemer JC, Hochreiter M, Schröder S: Adherence to a procalcitonin guided antibiotic treatment protocol in patients with severe sepsis and septic shock. Ann Intensive Care 2018; 8:68
13. Calfee CS, Delucchi K, Parsons PE, Thompson BT, Ware LB, Matthay MA; NHLBI ARDS Network: Subphenotypes in acute respiratory distress syndrome: Latent class analysis of data from two randomised controlled trials. Lancet Respir Med 2014; 2:611–20

14. Seymour CW, Kennedy JN, Wang S, Chang CH, Elliott CF, Xu Z, Berry S, Clermont G, Cooper G, Gomez H, Huang DT, Kellum JA, Mi Q, Opal SM, Talisa V, van der Poll T, Visweswaran S, Vodovotz Y, Weiss JC, Yealy DM, Yende S, Angus DC: Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA 2019; 321:2003–17
15. Parikh CR, Thiessen-Philbrook H: Key concepts and limitations of statistical methods for evaluating biomarkers of kidney disease. J Am Soc Nephrol 2014; 25:1621–9
16. Ware LB, Zhao Z, Koyama T, Brown RM, Semler MW, Janz DR, May AK, Fremont RD, Matthay MA, Cohen MJ, Calfee CS: Derivation and validation of a two-biomarker panel for diagnosis of ARDS in patients with severe traumatic injuries. Trauma Surg Acute Care Open 2017; 2:e000121
17. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG: Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med 2016; 35:4124–35
18. Jenniskens K, Naaktgeboren CA, Reitsma JB, Hooft L, Moons KGM, van Smeden M: Forcing dichotomous disease classification from reference standards lead to bias in diagnostic accuracy estimates: A simulation study. J Clin Epidemiol 2019; 111:1–10
19. Le Manach Y, Collins G: Statistical Methods in Hemodynamic Research, Perioperative Hemodynamic Monitoring and Goal Directed Therapy, from Theory to Practice. Edited by Cannesson M, Pearse R. Cambridge, United Kingdom, University Printing House, 2014, pp 8–13
20. Zhang Z: Multivariable fractional polynomial method for regression model. Ann Transl Med 2016; 4:174
21. Cook NR: Quantifying the added value of new biomarkers: How and how not. Diagn Progn Res 2018; 2:14
22. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36
23. 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
24. Pepe MS: Problems with risk reclassification methods for evaluating prediction models. Am J Epidemiol 2011; 173:1327–35
25. Hilden J, Gerds TA: A note on the evaluation of novel biomarkers: Do not rely on integrated discrimination improvement and net reclassification index. Stat Med 2014; 33:3405–14
26. Burch PM, Glaab WE, Holder DJ, Phillips JA, Sauer JM, Walker EG: Net reclassification index and integrated discrimination index are not appropriate for testing whether a biomarker improves predictive performance. Toxicol Sci 2017; 156:11–3
27. Vickers AJ, Cronin AM, Elkin EB, Gonen M: Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak 2008; 8:53
28. Vickers AJ, Van Calster B, Steyerberg EW: Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016; 352:i6
29. Li T, Hutfless S, Scharfstein DO, Daniels MJ, Hogan JW, Little RJ, Roy JA, Law AH, Dickersin K: Standards should be applied in the prevention and handling of missing data for patient-centered outcomes research: A systematic review and expert consensus. J Clin Epidemiol 2014; 67:15–32
30. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49:1373–9
31. Ogundimu EO, Altman DG, Collins GS: Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol 2016; 76:175–82
32. van Smeden M, de Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ, Reitsma JB: No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol 2016; 16:163
33. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, Moons KGM, Collins G, van Smeden M: Calculating the sample size required for developing a clinical prediction model. BMJ 2020; 368:m441
34. Sanchez-Pinto LN, Luo Y, Churpek MM: Big data and data science in critical care. Chest 2018; 154:1239–48
35. Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ: Review and evaluation of penalised regression methods for risk prediction in low dimensional data with few events. Stat Med 2016; 35:1159–77
36. Ajana S, Acar N, Bretillon L, Hejblum BP, Jacqmin-Gadda H, Delcourt C; BLISAR Study Group: Benefits of dimension reduction in penalized regression methods for high-dimensional grouped data: A case study in low sample size. Bioinformatics 2019; 35:3628–34
37. Van Calster B, van Smeden M, De Cock B, Steyerberg EW: Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. Stat Methods Med Res 2020; 962280220921415
38. Lee CK, Hofer I, Gabel E, Baldi P, Cannesson M: Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. Anesthesiology 2018; 129:649–62
39. Kendale S, Kulkarni P, Rosenberg AD, Wang J: Supervised machine-learning predictive analytics for prediction of postinduction hypotension. Anesthesiology 2018; 129:675–88

40. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, Rinehart J, Cannesson M: Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. Anesthesiology 2018; 129:663–74
41. Marafino BJ, Park M, Davies JM, Thombley R, Luft HS, Sing DC, Kazi DS, DeJong C, Boscardin WJ, Dean ML, Dudley RA: Validation of prediction models for critical care outcomes using natural language processing of electronic Health Record Data. JAMA Netw Open 2018; 1:e185097
42. Zhang Z, Ho KM, Hong Y: Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. Crit Care 2019; 23:112
43. Beam AL, Kohane IS: Big data and machine learning in health care. JAMA 2018; 319:1317–8
44. Komorowski M: Artificial intelligence in intensive care: Are we there yet? Intensive Care Med 2019; 45:1298–300
45. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B: A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019; 110:12–22
46. van der Ploeg T, Austin PC, Steyerberg EW: Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014; 14:137
47. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS: Predictive analytics in health care: How can we know it works? J Am Med Inform Assoc 2019; 26:1651–4
48. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G: Artificial intelligence in anesthesiology: Current techniques, clinical applications, and limitations. Anesthesiology 2020; 132:379–94
49. Steyerberg E: Evaluation of Performance, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, Springer-Verlag, 2009, pp 255–79
50. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L: A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc 2020; 27:621–33
51. Assel M, Sjoberg DD, Vickers AJ: The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagn Progn Res 2017; 1:19
52. Le Manach Y, Collins G, Rodseth R, Le Bihan-Benjamin C, Biccard B, Riou B, Devereaux PJ, Landais P: Preoperative Score to Predict Postoperative Mortality (POSPOM): Derivation and validation. Anesthesiology 2016; 124:570–9
53. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG: Internal and external validation of predictive models: A simulation study of bias and precision in small samples. J Clin Epidemiol 2003; 56:441–7
54. Vergouwe Y, Nieboer D, Oostenbrink R, Debray TPA, Murray GD, Kattan MW, Koffijberg H, Moons KGM, Steyerberg EW: A closed testing procedure to select an appropriate method for updating prediction models. Stat Med 2017; 36:4529–39
55. Castela Forte J, Perner A, van der Horst ICC: The use of clustering algorithms in critical care research to unravel patient heterogeneity. Intensive Care Med 2019; 45:1025–8
56. Sammut C, Webb GI: Mixture model, Encyclopedia of Machine Learning. New York, Springer-Verlag, 2011, pp 680–3
57. Calfee CS, Delucchi KL, Sinha P, Matthay MA, Hackett J, Shankar-Hari M, McDowell C, Laffey JG, O'Kane CM, McAuley DF; Irish Critical Care Trials Group: Acute respiratory distress syndrome subphenotypes and differential response to simvastatin: Secondary analysis of a randomised controlled trial. Lancet Respir Med 2018; 6:691–8
58. Lei L, Li LP, Zeng Z, Mu JX, Yang X, Zhou C, Wang ZL, Zhang H: Value of urinary KIM1 and NGAL combined with serum Cys C for predicting acute kidney injury secondary to decompensated cirrhosis. Sci Rep 2018; 8:7962
59. Zhang PL, Mashni JW, Sabbisetti VS, Schworer CM, Wilson GD, Wolforth SC, Kernen KM, Seifman BD, Amin MB, Geddes TJ, Lin F, Bonventre JV, Hafron JM: Urine kidney injury molecule-1: A potential non-invasive biomarker for patients with renal cell carcinoma. Int Urol Nephrol 2014; 46:379–88
60. Ostermann M, Ayis S, Tuddenham E, Lo J, Lei K, Smith J, Sanderson B, Moran C, Collinson P, Peacock J, Rhodes A, Treacher D: Cardiac troponin release is associated with biomarkers of inflammation and ventricular dilatation during critical illness. Shock 2017; 47:702–8
61. Ma Y, Mazumdar M, Memtsoudis SG: Beyond repeated measures ANOVA: Advanced statistical methods for the analysis of longitudinal data in anesthesia research. Reg Anesth Pain Med 2012; 37:99–105
62. Waikar SS, Bonventre JV: Creatinine kinetics and the definition of acute kidney injury. J Am Soc Nephrol 2009; 20:672–9
63. Schiller I, van Smeden M, Hadgu A, Libman M, Reitsma JB, Dendukuri N: Bias due to composite reference standards in diagnostic accuracy studies. Stat Med 2016; 35:1454–70
64. Obuchowski NA: An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. Stat Med 2006; 25:481–93
65. Tevaearai Stahel HT, Do PD, Klaus JB, Gahl B, Locca D, Göber V, Carrel TP: Clinical relevance of troponin T profile following cardiac surgery. Front Cardiovasc Med 2018; 5:182

66. Pencina MJ, D'Agostino RB: Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation Stat Med 2004; 23:2109–23
67. Gayat E, Cariou A, Deye N, Vieillard-Baron A, Jaber S, Damoisel C, Lu Q, Monnet X, Rennuit I, Azoulay E, Léone M, Oueslati H, Guidet B, Friedman D, Tesnière A, Sonneville R, Montravers P, Pili-Floury S, Lefrant JY, Duranteau J, Laterre PF, Brechot N, Chevreul K, Michel M, Cholley B, Legrand M, Launay JM, Vicaut E, Singer M, Resche-Rigon M, Mebazaa A: Determinants of long-term outcome in ICU survivors: Results from the FROG-ICU study. Crit Care 2018; 22:8