UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Article scientifique | Revue de la littérature | 2009

Published version | Open Access

This is the published version of the publication, made available in accordance with the publisher's policy.

Bioinformatics for protein biomarker panel classification: what is needed to bring biomarker panels into in vitro diagnostics?

Robin, Xavier Arnaud; Turck, Natacha; Hainard, Alexandre; Lisacek, Frédérique; Sanchez, Jean-Charles; Muller, Markus Johann

How to cite

ROBIN, Xavier Arnaud et al. Bioinformatics for protein biomarker panel classification: what is needed to bring biomarker panels into in vitro diagnostics? In: Expert review of proteomics, 2009, vol. 6, n° 6, p. 675–689. doi: 10.1586/epr.09.83

This publication URL: https://archive-ouverte.unige.ch/unige:19574
Publication DOI: 10.1586/epr.09.83

© This document is protected by copyright. Please refer to copyright holder(s) for terms of use.

For reprint orders, please contact reprints@expert-reviews.com

Expert Reviews

Bioinformatics for protein biomarker panel classification: what is needed to bring biomarker panels into *in vitro* diagnostics?

Expert Rev. Proteomics 6(6), 675–689 (2009)

Xavier Robin, Natacha Turck, Alexandre Hainard, Frédérique Lisacek, Jean-Charles Sanchez† and Markus Müller

†*Author for correspondence*
Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, Geneva, Switzerland
jean-charles.sanchez@unige.ch

A large number of biomarkers have been discovered over the past few years using proteomics techniques. Unfortunately, most of them are neither specific nor sensitive enough to be translated into *in vitro* diagnostics and routine clinical practice. From this observation, the idea of combining several markers into panels to improve clinical performances has emerged. In this article, we present a discussion of the bioinformatics aspects of biomarker panels and the concomitant challenges, including high dimensionality and low patient number and reproducibility.

Keywords: biomarker • classification • combination of biomarkers • *in vitro* diagnostics • panel • proteomics • validation

Background

As part of clinical practice, it is common to measure the concentration of a protein, known as a biomarker, in a biological sample to diagnose a disease, predict the outcome early or monitor a therapy. Examples of commonly accepted biomarkers include troponin I for detecting acute myocardial infarction, prostate-specific antigen for the screening of prostate cancer, glycated hemoglobin for the control of long-term glycemia and C-reactive protein for assessment of inflammation. Proteomics techniques, such as 2DGE [1,2] and mass spectrometry [2–5], have led to the discovery of numerous biomarkers, most of which are not currently available to medical practitioners. Possible explanations for this gap between proteomics research and routine practice are technical (e.g., the time and huge costs required to validate these molecules, as well as the accuracy of assays not being high enough to be translated directly into clinical practice) and biological (e.g., inter- and intra-individual variability).

When several biomarkers are measured, they are often considered separately, irrespective of the additional information contained in their joined interpretation. Combining several biomarkers into a single classification rule helps to improve their classification accuracy and, therefore, their clinical usefulness. Hereafter, we will call such a combination a panel. Potentially, a panel could even combine clinical parameters, such as age, sex, physiological constants or clinical scores, with biomarkers [6]. Similar to a single marker, a panel allows us to answer different clinical questions. Apart from increasing accuracy, biomarker panels help in the study of different pathophysiological pathways and shed light on diseases from different angles. For instance, in the context of a brain damage condition (e.g., aneurysmal subarachnoid hemorrhage), Turck *et al.* recently demonstrated that a combination of brain parameters associated with a clinical score and a cardiac biomarker could predict 6-month outcomes better than the biomarkers taken individually could [7]. In the same manner, Hainard *et al.* proposed a combination of inflammatory cytokines and one brain damage marker [8]. In both cases, the combination of different kinds of biomarkers improved the classification.

In contrast to the traditional single-analyte interpretation, several new challenges arise, which could also explain why panels are not yet

 10.1586/EPR.09.83 © 2009 Expert Reviews Ltd ISSN 1478-9450

widespread. First, appropriate methods are required to combine information from multiple biomarkers. These methods must be efficient and yield correct patient classification, but they must also be comprehensible to medical practitioners in order to gain acceptance. Second, the risk of overfitting the data is increased because of the higher dimensionality [9–11]. A careful validation is required to ensure that a panel truly performs better than individual biomarkers, hence avoiding raising false hopes. Finally, appropriate experimental design [12,13] and validation are the most important factors for ensuring the quality of the results.

After a short overview of *in vitro* diagnostics (IVD), we will review recent papers that describe combinations of biomarkers and/or clinical parameters in panels, and examine whether they addressed these new challenges and, if so, how. We mainly focus on protein biomarker panels but also include related work on analyzing protein or gene expression microarray and protein mass spectrometry data where we deem it relevant for protein panels. We will also review methods that allow the validation of obtained models and their performance, as well as the strategies available to compare different panels and their combinations. This review addresses clinical researchers who seek a basic introduction to statistical methods and the pitfalls of biomarker-panel research and statisticians who would like to learn more about recent work and the clinical aspects of this subject.

From discovery to IVD

In vitro diagnostics encompasses any type of assay performed on a patient sample in a controlled environment to answer a clinical question, including diagnostic, prognostic or monitoring tests. It typically includes point-of-care tests, which are quick and simple assays performed beside the patient with portative equipment, and laboratory tests, which are performed by trained personnel in dedicated clinical chemistry laboratories.

Vitzthum *et al.* reviewed the need in proteomics to push discovered molecules into IVD [14]. The crucial points are that the classification must be reliable and provide information that is valuable for decision making; measurements must be both exact and robust, and the test accuracy must meet sufficient (positive or negative) predictive values.

The target performance of IVD tests must be chosen according to the clinical question. As pointed out by Dodd and Pepe, "large monetary costs result from high false-positive rates" [15]. Similarly, failure to diagnose a disease can dramatically impact on a patient's health, which may even lead to death. Therefore, IVD (single biomarker or panel) as a helpful clinical practice must display sufficient discriminative power and answer a well-defined question. In other words, one should focus on high sensitivity and/or specificity or high predictive values rather than global accuracy.

An IVD test aims to determine the state of the patient. For biomarker tests, a decision threshold (also known as a cut-off) is usually chosen. Any value below the cut-off will indicate that the test result is negative, while a value above the threshold will be deemed as a positive result. The test result, together with the observed true outcome, will define the sensitivity and specificity (see Table 1 for definitions).

Predictive tests can be split into two categories: 'rule-out' and 'rule-in' tests. Rule-out tests reject negative patients while avoiding false negatives. In these tests, the sensitivity is of prime importance, as is the negative predictive value. However, the level of false positive must be kept low enough in order to preserve both specificity and positive predictive value at acceptable levels. When the test is applied to asymptomatic patients, it is termed a screening test. A negative result to a screening test implies that the patient is highly likely to be healthy, while a positive result only means that more investigations are required. For example, in the context of human African trypanosomiasis (HAT), a potential rule-out test would be applied to exclude the patients not infected by the parasite. All patients with a negative test would then be classed as free of the parasite, with a very high confidence. Similarly, rule-in tests (also called confirmatory tests) try to include only positive patients and generate as few false positives as possible. The specificity and positive predictive values must be very high. A rule-in test applied in the HAT field would select only patients with parasites in the brain (stage 2 of the disease), who would be subsequently subjected to a very toxic treatment. Patients without brain infection (stage 1) must be excluded, because they could potentially be killed by the inappropriate medication [8].

Predictive values (negative or positive) need to take the class prevalence into account since even a test with a very high specificity could have a low positive predictive value. If the prevalence of the disease is very low, there would be a larger number of false positive, only because of the larger number of controls. This property makes predictive values more difficult to compute than specificity or sensitivity. Despite this complication, predictive values are usually more valuable because they express the probability of the patient being truly positive or negative for a given group of patients.

Commercial panels

From a commercial point of view, McCormick showed how both pharmaceutical companies and medical practitioners could profit from biomarkers and biomarker panels to predict the safety of a treatment, identify risk and responder candidates, and monitor therapies [101]. However, they pointed out that the acceptance of biomarkers is hindered by the lack of data sharing (owing to technical or strategic reasons), as well as insufficient validation and targeting.

In the USA, medical devices (including IVD) must obtain approval by the US FDA. Hackett and Gutman highlighted the difficulties that are raised by the combination of several markers and the use of statistical models [16]. FDA review procedures for device acceptance focus on the test result, and a simple model can be accepted at the condition that it is independently validated.

To our knowledge, only the Biosite® company sells panels of protein biomarkers for blood samples. The Triage® Stroke Panel simultaneously measures four markers (namely matrix metalloproteinase 9, brain natriuretic peptide, D-dimer and S100β) and computes a multimarker index using a proprietary algorithm. Two cut-offs are defined, associated with a high or low risk for the patient having a stroke, while patients in the intermediate region require further investigation. It was accepted by the FDA for premarket

Table 1. Clinical classification definitions.

Word	Common abbreviation	Formula	Definition
Prevalence			Frequency of the positive occurrence in the studied population
Rule in (confirmatory)			A test performed in an attempt to confirm the presence of a disease
Rule out (screening)			A test performed in an attempt to exclude the presence of a disease
True negatives	TN		Negative patients correctly classified as negatives
True positives	TP		Positive patients correctly classified as positives
False negatives	FN		Positive patients incorrectly classified as negatives
False positives	FP		Negative patients incorrectly classified as positives
Sensitivity	SE	TP/(TP+FN)	Proportion of positive patients correctly detected by the test
Specificity	SP	TN/(TN+FP)	Proportion of negative patients correctly rejected by the test
Positive predictive value	PPV	TP/(TP+FP)	Proportion of positive tests that correctly indicate positive patients
Negative predictive value	NPV	TN/(TN+FN)	Proportion of negative tests that correctly indicate negative patients
Odds ratio	OR	(SE/(1-SE))(SP/(1-SP))	Effect of a given increase of the studied marker

approval application in 2005, was withdrawn by the manufacturer 1 year later to allow further clinical studies, but it was recently reintroduced. The Triage Stroke Panel was applied by Vanni *et al.* in a neurological emergency service to discriminate patients with or without stroke among those suspected of having stroke [17]. Sibon *et al.* compared it with an established neurological scale evaluated by nurses [18]. Brouns *et al.* analyzed only the D-dimer measurement to compare it with the assay of another manufacturer to discriminate small-artery and large-artery acute ischemic stroke, but they did not make use of the multimarker index score [19].

The same company previously marketed in 1999 a Triage® Cardiac Panel for the diagnosis of cardiovascular diseases. This test measures three proteins, known as cardiac markers (namely creatine kinase MB, myoglobin and troponin I) [20]. However, it cannot truly be called a panel as the measurements are not combined into a single final score.

Applied Genomics sells several immunohistochemistry panels. Tissue arrays are stained, and each staining is assessed in a binary manner. The results are then combined with a Cox proportional hazards model into a single score stratifying patients into low, medium or high risk. One of the available panels provides prognostic information for breast cancer outcome [18].

Tools for panels

History

In terms of biomarkers, a panel is the combination of more than one variable into a single classification rule. The idea of combining several medical parameters to obtain an improved patient classification is not new. In psychiatry, Hoffer and Osmond applied a combination of neuropsychiatric variables in the early 1960s to distinguish schizophrenic patients from normal individuals [21]. They defined 145 questions that could be answered by true or false, covering perceptions, thoughts and feelings. Complex algorithms would then compute several scores. However, the set of questions and the scoring algorithms were not justified. Later, in 1988, the World Federation of Neurological Surgeons (WFNS) score was defined to assess patients' neurological status [22]. It consists of the combination of three easy-to-assess clinical variables. Eye, verbal and motor responses are evaluated on a scale ranging from one to four, five and six, respectively. An intermediate score ranging from three to 15 is computed, and the final score depends on the range of this intermediate score and the presence of a motor deficit.

In the field of biomarkers, Woolas *et al.* showed the potential of using several serum markers together in 1993 [23]. They observed that most of their patients with stage 1 ovarian cancer were positive for at least one of the three markers they tested. However, they did not use this observation to make a true statistical combination. In 2000, Hill *et al.* were among the first to report the use of a panel of protein biomarkers [24]. They tested four biomarkers, and they observed that 93% of their acute ischemic stroke patients were positive for at least one of the four markers of the panel.

As detailed in the later section 'Classification using panels', panels can also combine biomarkers and clinical parameters. However, prior to discussing the various approaches for panel classification, we briefly review some important data preprocessing and data-normalization steps, which are performed prior to classification.

Preprocessing

Normalization & reproducibility

Several types of errors can disturb the results of biomarker concentration measurements and mitigate reproducibility. It has been shown that sample collection from different centers and by different nurses as well as sample handling variability (i.e., sample container, time to freezing and storage temperature) and instrumental errors can lead to measurement variations [25,26]. When dealing with high-dimensional mass spectra, reproducibility of the experiments becomes a problem, and it has been shown that proper sample and data processing, as well as feature selection,

are of major importance [9]. Furthermore, biological variability between different patients owing to sex, age, treatment, lifestyle and chronic diseases, or even within a single patient taken at different times, can confuse the analysis. All these sources of variation make it difficult to compare the results of different experiments and to draw conclusions.

On the experimental side, normalization methods often require a 'calibration' sample, which has constant values over all the experiments [27]. Using calibration curves, concentration measurements of biomarkers can be adjusted for each patient and systematic offsets in the measurements reduced. However, only instrumental offsets can be reduced in this way and other offsets due to sample acquisition and treatment need further bioinformatics normalization.

This computational normalization equalizes the mean and variance of distributions of different biomarker measurements, making them more comparable. A very simple normalization method consists of the z-score transformation, which sets the mean to zero and the variance to one, but otherwise does not affect the shape of the distribution. Yeo *et al.* proposed the box–cox transformation family, which includes the logarithmic transformation, to obtain distributions closer to the normal one [28]. Another normalization method is the quantile normalization technique, where all values are transformed into their corresponding normal quantiles [29]. However, this is an extreme normalization and the structure of the data can be lost in the process. Based on technical and biological replicates, analysis of variance can calculate the bias and variance introduced by each processing step and lead to more accurate comparisons [13].

Feature selection

Another important preprocessing step is feature selection, which is crucial in high-dimensionality problems, such as mass spectra or microarrays, but is less important for lower dimensional biomarker panels. It consists of selecting the biomarkers and patient parameters that will be included in the panel. The choice of the feature-selection method strongly depends on the classification algorithm and the data [30]. It is also important to note that data for feature selection must not include the test data; otherwise, the test performance would be too optimistic. Saeys *et al.* classified the feature-selection methods into three categories: filter methods, wrapper methods and embedded methods [31]. Filter methods consider only the intrinsic properties of single features independently from classification. Conversely, wrapper and embedded methods perform the feature search at the same time as the classifier model is trained. In wrapper methods, the search for optimal features is performed by an optimization procedure, which evaluates the performance of a given classifier on different feature subsets. Embedded techniques can include or eliminate features during the classifier-training procedure. Such embedded techniques can be implemented, for example, in logistic regression, random forests, neural networks or support vector machines (SVMs; see later).

Several examples of feature selection are reported by Hilario and Kalousis [30]. Baggerly *et al.* used preprocessing and exhaustive search and genetic algorithms to reduce an initial 60 831 *m/z* value from mass spectrometry to filter 506 and then sets of one to five features, and then applied the feature sets to linear discriminant analysis [32]. Petricoin *et al.* also employed genetic algorithm with mass spectrometry, but in a wrapper method around a self-organizing map algorithm [33].

Classification using panels

Biomarker panels rely on a well-established field of statistics, known as multivariate classification or supervised learning. There is a vast amount of literature available, and much of it is summarized in excellent textbooks, such as that by Hastie *et al.* [34]. The classification task consists of attributing a class label to every patient by means of the vector of biomarker concentrations and clinical scores. In the case of two classes, this corresponds to dividing the space of all possible panel vectors into two distinct regions, one region for every class (Figure 1). The way the classifier determines these regions depends on the method used. In all cases, the algorithms learn these boundaries from training data, that is, a set of panel vectors known to belong to a diseased or healthy patient. Once the region boundaries are fixed, the performance can be evaluated on equally annotated but disjointed test data.

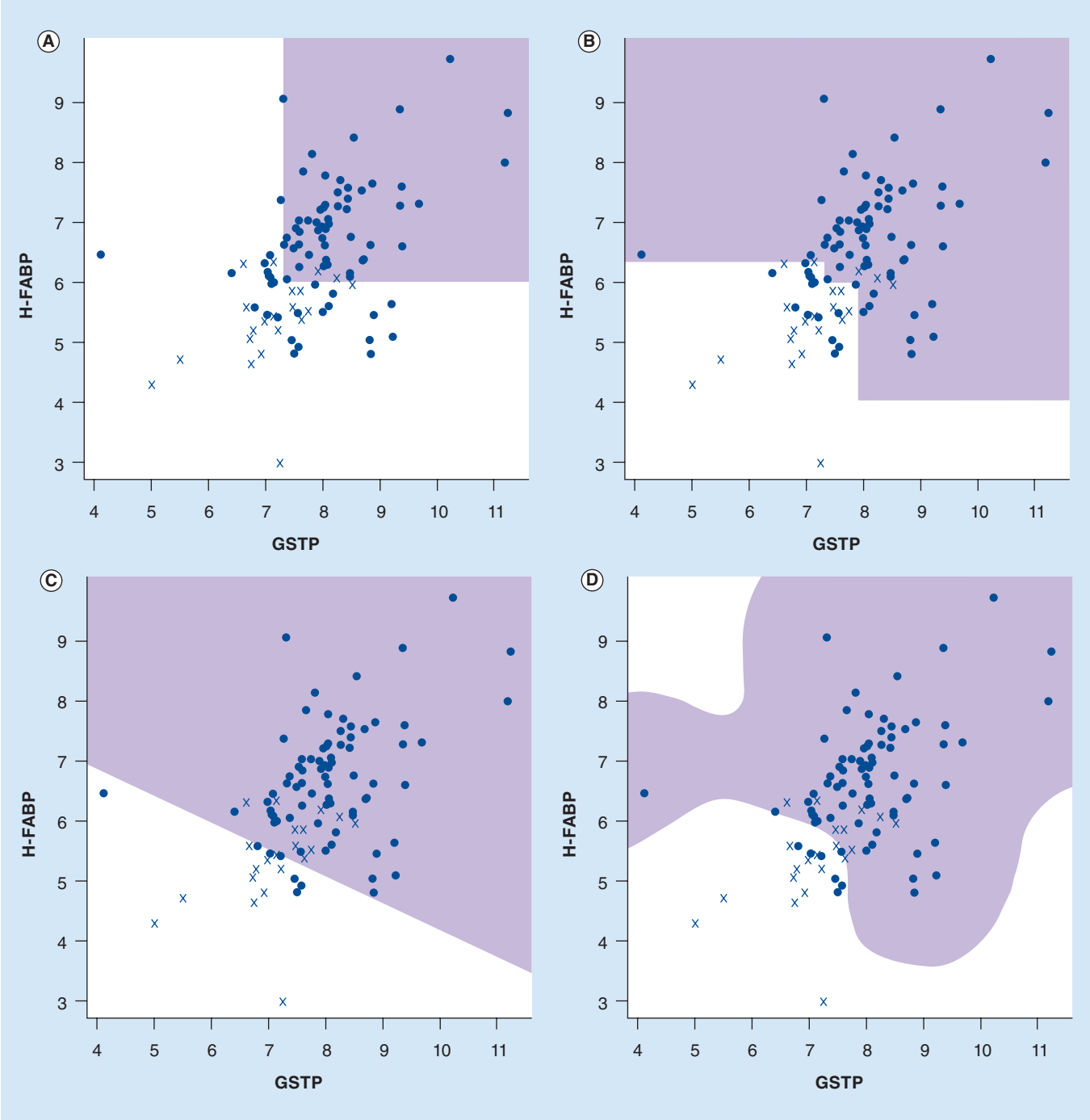
This approach may seem fairly straightforward, but two main problems must be dealt with: the low number of samples in the training set and overfitting the data. The former problem is paramount in many biomarker projects, since the number of patients is usually small (from a few to several hundred patients) compared with the number of markers. The patients are then only sparsely distributed in the panel vector space, and many parts of the class regions are only represented poorly or not at all in the training set, which makes it more difficult for the classifier to find the correct regions. Figure 1 illustrates this problem since neither the upper left nor the lower right corners contain any data points and, considering only these training data, it is impossible to predict the classifier results in these regions. The latter problem is perhaps less severe but equally important. Since the shape and smoothness of the boundaries between the class regions is not known (linear or curved), the regions obtained from the training data might be wrong even if they fit the training data very well, because the model defined in the classifier is wrong (i.e., the classifier might yield an arbitrarily curved boundary that is actually linear Figure 2). However, cross validation provides a means to at least partially mitigate this problem (see later). As a rule of thumb, the fewer patients there are in the training and test sets, the simpler the class boundaries should be to avoid overfitting, even if these simple boundaries cannot reproduce the true boundaries correctly.

We now discuss the main methods applied to define biomarker panels. Threshold-based methods and logistic regression are probably the most popular ones. Tree-based methods are also widely used, whereas SVM is a method of choice for many high-dimensional problems. We will now detail some methods and show how they are applied.

Threshold-based

In threshold-based methods, a set of thresholds, one for each biomarker, is selected, usually in a univariate manner (Figure 1A) [7,8,24,35–38]. Any value of a molecule below its

respective threshold will indicate that the test result is negative, while a value above the threshold will be deemed a positive result. In some rare cases, it can be necessary to reverse the order and to consider values below the threshold as positive results. The score of the test for a patient corresponds to the number of biomarker molecules, whose concentration value exceeds (or is below for negative biomarkers) the threshold. Similar to a majority voting, a patient is classified positively if this score is higher than



Figures 1. Classification by different methods. (GSTP and H-FABP concentrations illustrated in log scale). The gray area shows the region where the test would be considered positive by the method. Crosses and dots represent Stage 1 and Stage 2 human African trypanosomiasis patients, respectively. **(A)** Threshold-based methods split the space into boxes. **(B)** Decision trees can create more boxes. **(C)** Logistic regression divides the data with a straight line. **(D)** Support vector machines can compute complex separations but can also create linear partitions similarly to logistic regression (see Figure 4).
GSTP: Glutathione *S*-transferase Pi; H-FABP: Heart-type fatty acid-binding protein.
Redrawn from from [8].

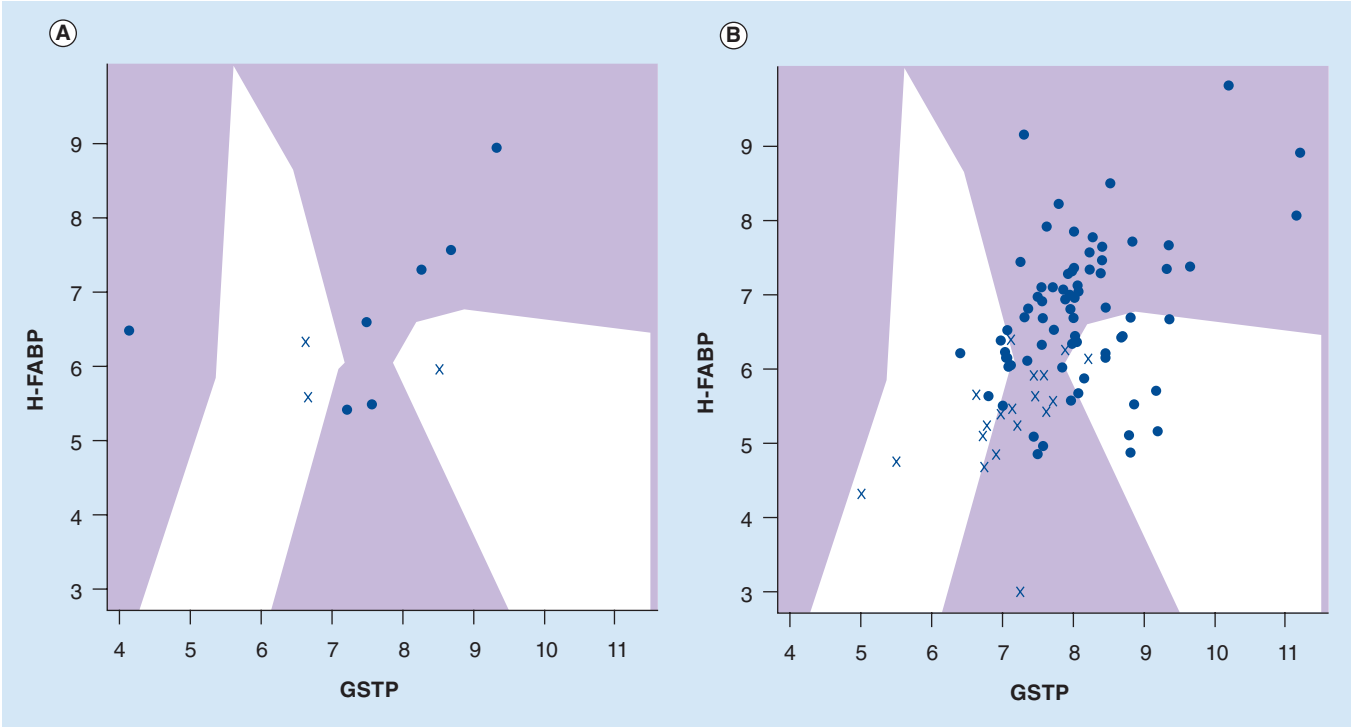


Figure 2. Overfitting with the nearest-neighbour algorithm (*k* = 1). The gray area shows the region where the test would be considered positive by the method. Crosses and dots represent Stage 1 and Stage 2 human African trypanosomiasis patients, respectively. **(A)** Training set of ten patients determining the class regions (gray or white background). **(B)** The pattern defined in **(A)** is applied to a test set of 90 different patients. Most Stage 1 patients and many Stage 2 patients are misclassified in the test set. The choice of the training and test set is purely for illustrational purposes.
GSTP: Glutathione *S*-transferase Pi; H-FABP: Heart-type fatty acid-binding protein.
Redrawn from [8].

a minimum number. To take a purely theoretical example, one could set a minimum of two out of five parameters, where any two positive molecules of the panel would raise a positive test, but if only one is positive, the panel result would be negative. The minimal number can be chosen based on several criteria, usually depending on the targeted sensitivity or specificity or by cross validation. It is used mostly for ELISA and clinical data but not in higher-dimensional problems. The threshold method has the major advantage that results are easy to interpret. Additionally, its simple boundary structure reduces the possibility of overfitting the training data. In our view, the threshold method is well adapted to biomarker panel data, where class boundaries of a single marker can often be represented as single cut-off points.

Lejon *et al.* followed this approach to combine clinical and biochemical variables to predict trypanosomiasis treatment failure [38]. Thresholds were chosen on univariate parameters to maximize the sum of sensitivity and specificity, and two parameters were retained. For the same disease, Hainard *et al.* selected a panel of two cytokines and a brain-damage marker to assess the disease stage of 100 patients using a multivariate approach [8]. The rationale was that interactions between molecules in a panel can be complex and good univariate thresholds are not necessarily the best thresholds in a panel. Other attempts have been made in this direction [36]. Vitzthum *et al.* also showed that different thresholds should be chosen for different clinical questions [14]. This means that if a threshold discriminates well between classes for one question, it may not automatically be accurate in other problems.

A similar technique is the patient rule-induction method [34], where two thresholds (lower and upper) are chosen, and a patient is positive only if the biomarker value is included in the range. This can bring out patients with particularly low values, but the clinical and biological relevance of such a criterion is not obvious. It was applied by Wang *et al.*, but its usage seems scarce [39]. Naive Bayes is another similar method, in which the thresholds are separately determined based on statistical criteria for every feature. Ralhan *et al.* successfully applied it to proteins quantified by MS/MS after isobaric tag for relative and absolute quantitation labeling on a small number of patients [40]. It can be extended to deal with dependent data [41].

Decision trees

Decision trees are similar to threshold-based methods, but they can find more complex boundaries (Figures 1B & 3 & Box 1). Different tree methods exist and vary in the construction of the tree from the training set, that is, the selection of a feature and a threshold for each node, and in the pruning strategy.

Classification and regression trees (CARTs) are one of the most popular tree-based algorithms [42–44]. Other methods are C4.5 decision trees [45], J48 [46], or recursive partitioning and regression trees (RPART). The latter allowed Ring *et al.* to select five proteins

out of several hundreds and combine them into a decision tree that was able to classify 195 estrogen receptor-positive breast cancer patients into good, moderate or poor prognosis [47]. However, it seemed to be dependent on the cohorts to which the model was applied and was less predictive of outcome than other methods.

Trees perform well in combination with boosting algorithms [48], which can strongly improve the classification results. The idea is to boost the classification performance of a simple classifier (e.g., a strongly pruned tree) by iteratively applying it to modified versions of the data, where the weight of the misclassified training observations is increased. Each successive tree classifier is then forced to focus on those misclassified observations, and the final classification is calculated as the weighted average over all tree classifiers. Trees also form the basis of the random forest algorithm [49], where classification is obtained from a combination of trees, each built from a small but random subset of the features.

A basic parallel or sequential AND/OR way of combining tests similar to decision trees has been proposed by Vitzthum *et al.* [14]. However, there is no evidence that it was applied in panels.

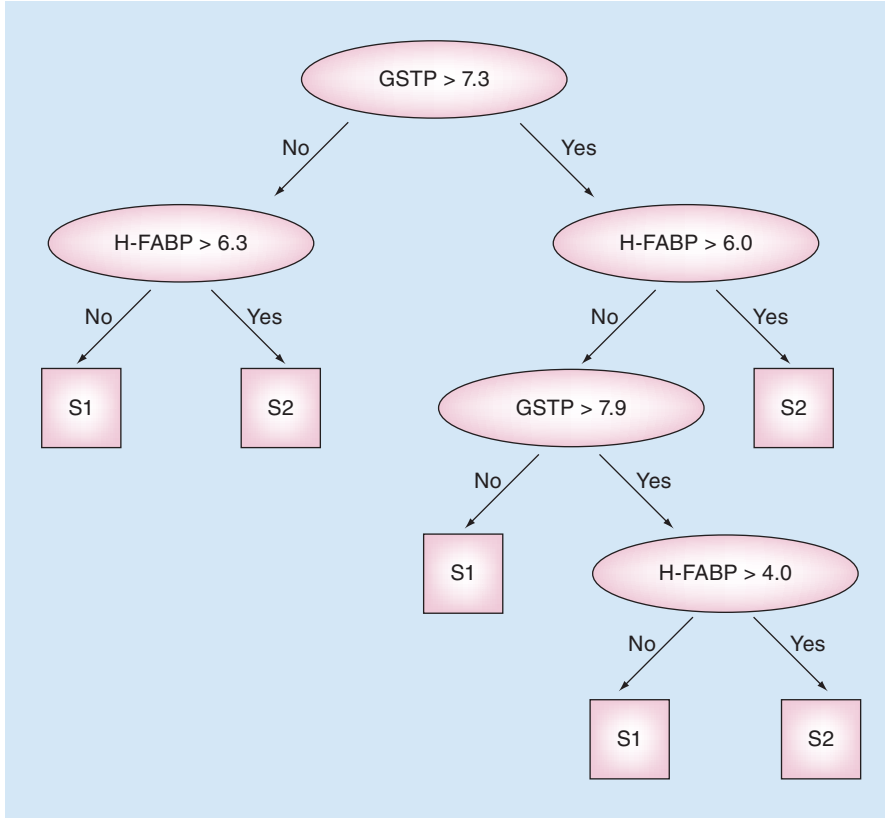


Figure 3. The decision tree corresponding to Figure 1B. Each circle corresponds to a question. Depending on its answer, one follows the arrow to the left or right and goes to the next question, until a decision (square box) is reached.
GSTP: Glutathione *S*-transferase Pi; H-FABP: Heart-type fatty acid-binding protein; S1: Stage 1 human African trypanosomiasis classification; S2: Stage 2 human African trypanosomiasis classification.

Logistic regression

Logistic regression is a very popular linear regression method in the medical field, where the simplicity and robustness of the models produced is appreciated (Figure 1C & Box 2). The method is based on a clear mathematical formulation and yields a globally optimal solution. Interaction terms can be entered to model nonlinear class boundaries, but this requires *a priori* information regarding the structure of the data and, therefore, is not commonly used.

Logistic regression can combine clinical or biomarker data, either continuous or categorical [36,37,45,50–54]. For example, Visintin *et al.* trained several logistic regression models to screen ovarian cancer on several hundred patients and controls [50]. Although some individual biomarkers displayed a significantly lower performance in the test set, regression models were stable, denoting the robustness of the technique. Logistic regression was also applied to combine protein markers with clinical parameters [55] or to combine clinical variables only [56].

Support vector machines

Support vector machines (Figures 1D & 4 & Box 3) are one of the most popular methods in machine learning. SVMs have the advantage of being able to provide a clear mathematical model with a globally optimal solution, contrary to neural networks or others learning methods that can get trapped in a local optimum. It performs well in a large variety of tasks, and it was applied in very different fields, ranging from text pattern recognition to analysis of gene expression microarrays. However the underlying concepts are more difficult to grasp for non-mathematicians. Figure 1D shows the result of classification with a radial basis kernel, but SVMs can also find linear or polynomial separations similar to logistic regression.

The SVM is preferred in higher dimensionality problems, such as microarray [57,58] or mass spectrometry (SELDI [45,46,59] or MALDI [48,60]) data analysis. Liu *et al.* combined use of an SVM with a genetic algorithm and obtained reproducible and fairly accurate results [61]. It was also used by Wild *et al.* to classify ELISA data for patients suffering from rheumatoid arthritis, but only to challenge the regularized discriminant analysis and confirm the results generated by the latter technique [62].

Generalized additive models

Generalized additive models allowed Knickerbocker *et al.* to combine protein microarray data with patient clinical information to predict survival after renal replacement [6]. They added local polynomial functions (or splines) that allow defining nonlinear relationships between the variables, as well as the

Box 1. Decision trees.

- Decision trees are simple but powerful methods that split the feature space into a set of boxes and attribute a class (or a probability) to each one. Figure 3 displays a typical representation of the decision tree corresponding to Figure 1B
- To build a decision tree, a series of binary splits based on a threshold of one of the variables is performed. For each step, the variable that yields the best split is selected. Every outcome of a test (positive or negative) creates a branch, which either leads to a new test or to a terminal leaf, corresponding to a box in the feature space. Each of the boxes is defined by the unique path leading to it, and it is possible to calculate a class probability or binary outcome within the box. The tree is then pruned and the less informative decision branches are removed to simplify the tree and avoid overfitting. The number of splits and the minimal number of observations allowed in each terminal leaf must be carefully investigated, for example, by cross validation [34,90]

detection of inflexion points. They trained two models separately, one for clinical parameters and one for protein biomarkers, and showed that the experimental predictors could only add information for patients detected as being at high risk by the clinical predictors.

Other methods

Several other methods were shown to perform well in proteomics. Gevaert *et al.* applied a Bayesian network on gene expression microarray data [63]. This approach allows the integration of clinical data in several manners: full, decision and partial integration. In full integration, the clinical and microarray datasets are merged and handled as a single dataset. In decision integration, two models are trained, one clinical and one with microarray data, and the final decision is generated as a combination of the weighted probability of the clinical panel with the microarray one. Finally, in partial integration, the network structures are determined separately for each dataset and joined into one single structure before performing the learning step for the merged clinical and microarray datasets.

Regularized discriminant analysis is a classification method that can deal with strongly correlated data [34]. It is based on linear discriminant analysis [48] or quadratic discriminant analysis. It can take into account the main effects of the markers as well as their interaction. Wild *et al.* successfully used regularized discriminant analysis to combine two to three molecules in patients with rheumatoid arthritis [62]. For prognostic purposes, an attractive option is to analyze the time series, if available. James and Hastie proposed a classification based on spline regression of time series and linear discriminant analysis of the regression coefficients [64].

Logical analysis of data is a method that finds approximations of subsets of observations by combinatorics and optimization. Its application in the medical field had been reviewed previously [102]. It was used by Reddy *et al.* to classify 48 ischemic stroke patients and 32 controls, and was applied on a validation set consisting of 60 patients [45]. The methodology was also able to detect two outlier patients and showed good performance.

Reddy *et al.* [45] and Prados *et al.* [46] applied multilayer perceptron, a type of linear neural network, and Cox proportional hazard models. The latter method was used in several other studies [47,52,65,66].

'Nearest neighbors' finds the *k* nearest samples and performs a majority vote to decide the classification [48]. Linkov *et al.* defined a method they called ADE+PT, which is similar to a weighted nearest-neighbor approach [67]. There is no evidence of its application in any other published study.

Performance validation

Why?

Once a panel is defined, its performance must be evaluated. As stated earlier, overfitting corresponds to the underestimation of the classification error on the training set (Figure 2A), which cannot be validated on an independent test set (Figure 2B) [34]. High-dimensional data are especially prone to overfitting, as mentioned in Feng and Yasui in the context of SELDI mass spectra, where a huge number of possible markers (peptide masses) are available [11]. However, depending on the classifier, it can be a serious problem even for low dimensional data.

Box 2. Logistic regression.

- In its simplest form, logistic regression provides a linear separation of the feature space. It models the class probability p(+|*x*), that is, the probability that the *n*-dimensional feature vector, *x*, is classified positively, as a sigmoidal (s-shaped) function:

$$f(z) = \frac{1}{(1+e^{-z})}$$

where

$$z = \alpha_0 + \sum_{i=1}^{n} \alpha_i x_i$$

The coefficients α_i must be determined from the training sample by means of a maximum likelihood procedure, which usually converges to the unique global optimum [34]. If the different features x_i are properly normalized (same mean and standard deviation), the coefficients α_i provide direct information regarding the importance of a feature for the correct classification in the logistic regression model. It is also possible to expand the features by explicitly including interaction and nonlinear terms. For example, the feature vector

$$x = (x_1, x_2)$$

could be expanded to a higher dimensional vector

$$x' = (x_1,\ x_1^2\ +\ x_2^2, x_1 x_2, x_2)$$

or

$$x' = (x_1,\ x_1/x_2, x_2)$$

The logistic regression is then applied to *x'* instead of *x*

- Odds ratios measure the effect of a given increase of the studied marker. They are frequently used in relation to logistic regression. However, their use as a measure of performance is difficult [91]

Box 3. Support vector machines.

- Let us consider a 2D example where the two classes are completely separable by a straight line. It is easy to see that there are many straight lines that do the job; the question is which of these lines provides the best classification on a test sample? The support vector machine solves this problem by selecting the (usually unique) separating line that is farthest away from any data point [92]. It can be shown that this line often yields better classification performance on a test set since it is as far away as possible from the critical points, which lie close to the class boundary. Mathematically, the linear separation can be formulated as follows: for each feature vector x_i of class y_i (±1) we have we have $w\,x_i + b \leq -1$ for $y_i = -1$ and $w\,x_i + b \geq 1$ for $y_i = 1$ where w is a vector orthogonal to the separating line. It can be shown [92] that the distance of the separating line to the next x_i is $1/|w|$; therefore, the support vector machine searches for the smallest $|w|^2$, which satisfies the aforementioned inequalities. The lines $w\,x_i + b = -1$ for $y_i = -1$ and $w\,x_i + b = 1$ for $y_i = 1$ are termed the margins, which lie parallel and at equal distance $1/|w|$ to the separating line and touch one or more data points of the corresponding class
- In almost all real-life applications, classes are not linearly separable. Cortes and Vapnik, however, showed that a similar approach still works in these cases [92]. They introduced so-called slack variables $\xi_i \geq 0$ and reformulated the constraints as $w\,x_i + b \leq -1+\xi i$ for $y_i = -1$ and $w\,x_i + b \geq 1-\xi i$ for $y_i = 1$, that is, for each x_i on the right side of its margin, we have $\xi_i = 0$, and for each x_i on the wrong side of the margin, $\xi_i > 0$, where $\xi_i/|w|$ is the distance from the margin (Figure 4). Since we still would like to have a margin distance $2/|w|$ as large as possible, but also as little misclassification

$$\sum_{i=1}^{p} \xi_i$$

as possible, we search for a w value that satisfies the 'slack' inequalities mentioned and minimizing

$$|w|^2 + C\sum_{i=1}^{p} \xi_i$$

where p is the number of samples and C a misclassification weight. This is a quadratic programming problem, for which many efficient algorithms are available, which usually converge to a unique solution. It can be shown that

$$w = \sum_{i=1}^{p} \alpha_i y_i x_i$$

where $\alpha_i > 0$ for those sample vectors (so-called support vectors), which lie either on the margin or on the wrong side of it ($w\,x_i + b \geq -1$ for $y_i = -1$ and $w\,x_i + b \leq 1$ for $y_i = 1$), and $\alpha_i = 0$ for all other correctly classified vectors
- Cortes and Vapnik also showed that the support vector machine approach can be naturally extended to nonlinear separation [92]. In Figure 1D for example, we used a radial basis kernel, which yields the class indicator function as a sum over radial basis functions, which are centered at the support vectors (see [34,92] for a detailed discussion of the kernel-based formulation)

In the literature, Bhaskar *et al.* showed that validation is not performed consistently in bioinformatics [68] and Whiteley *et al.* showed how even single biomarkers can be biased if its threshold is chosen on the same dataset [69]. Several panel papers that we previously mentioned did not perform any kind of validation of the accuracy of the reported classification [8,24,35,37,38,52,55] or simply mentioned that it would or should be done later. While this is still acceptable for single biomarkers, doing so with panels could lead to false hopes and should be avoided in the future. Therefore, it is crucial to have a separate dataset that includes patient data, which is independent from the model definition, to test that model. Ideally, the dataset should originate from a separate cohort of patients with biomarker concentration measured in a different laboratory. However, such validation data is often unavailable, and the number of patients is often too small to split the data into independent training and test sets of the same size.

How?

Apart from using an independent validation dataset, which is not always possible, several computational methods can overcome this issue. If the number of patients is sufficient, a subset of the sample population can be left aside for the training process and kept as validation set, which was done by several groups [36,44,50,53]. If not enough patients are available, randomization techniques, such as permutation tests, cross validation and bootstrapping, can help with evaluation if the classification is significant or if it is only overfitting [11].

Permutation tests

Permutation tests allow the determination of whether the classification result is significant [70,71]. Patient labels are randomly permutated, and the problem is treated in the same way, providing information concerning the classification error under the random hypothesis. If the efficiency of the classification of random patients is comparable to that of real patients, it is a strong indication that the method is overfitting the training data.

Cross validation

Cross validation is a purely computational method that allows evaluation of the robustness of a classification. In cross validation, the data are split into a number, *k*, of equal-sized parts. Sequentially, *k*-1 parts are used to train the classifier model, and the remaining one is kept to test the performance of the model. When all parts have been used as test sets, performance is averaged [34].

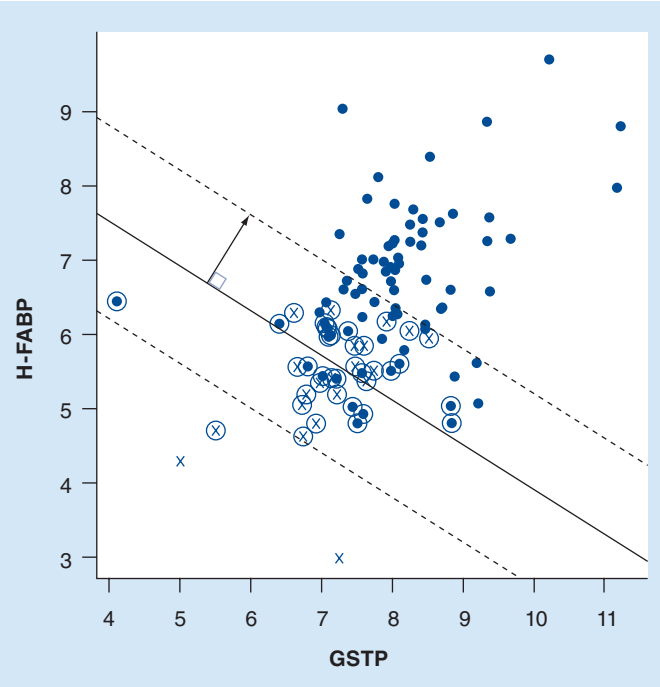


Figure 4. Support vector machines. Crosses and dots represent Stage 1 and Stage 2 human African trypanosomiasis patients, respectively. Margins and the separation line are represented by dashed and solid lines, respectively. Support vector observations are circled in gray. The arrow represents the vector $w/|w|^2$.
GSTP: Glutathione *S*-transferase Pi; H-FABP: Heart-type fatty acid-binding protein.
Redrawn from [8].

Typical values for *k* are five or ten [34]. If *k* is equal to the sample size, it is a leave-one-out cross validation. The problem with cross validation is that the training sample size is smaller, which can lead to overestimation of the prediction error. For biologists and clinicians, another problem is that each round of cross validation can choose a different model. Therefore, it must be made clear we evaluate the error of the method, not of the model itself. Several groups applied cross validation for biomarker applications [43,45,47,50,62,67].

Several variants of cross validation exist. When the data are not balanced, that is, one class has a much smaller patient number than the other, a stratified cross validation can be performed, where both classes are represented in the same proportion in each *k* fold than in the whole set. Another variant is double cross validation, which combines an internal loop where the model meta-parameters (such as the width of a kernel, the kernel-type or the number of principal components) are defined, and an external loop where the model is actually trained with these parameters and performance is evaluated [70].

Bootstrapping

Bootstrapping involves randomly selecting items with replacement in order to obtain a new sample of the same size as the original one. Approximately 37% of the original sample will not be selected and can be used as a test set. This procedure can be repeated a large number of times to get a good approximation [34,72].

In contrast to cross validation, sample size is not reduced but some data will be redundant. It is particularly helpful for determining empirical confidence intervals [73]. Several publications employed bootstrapping for validation [6,42,51]. Similarly to double cross validation, Feng *et al.* proposed that cross validation should serve as model selection and bootstrap as estimation of the classification error [11].

Separate set validation

The ultimate validation is always to reproduce the experiment independently on different patients and within a different laboratory. However, mainly because of time and funding constraints, it cannot always be carried out, and one must rely on previous investigations. For example, Whiteley *et al.* showed that no publication using panels for the diagnosis of ischemic stroke validated its results on an independent patient cohort [70]. They recommended independent validation as a good practice, also for other work dealing with patient classification. Reddy *et al.* [45] and Gevaert *et al.* [63], for example, rely on an independent cohort for validation.

Statistical method reporting

Proteomics is currently moving towards better reporting requirements, such as the 'Minimum Information about a Proteomics Experiment' model [73]. A similar initiative exists in the medical community with the Standards for Reporting of Diagnostic Accuracy that defines a checklist of 25 items to promote a coherent reporting of accuracies [74]. However, none of these initiatives fully covers the needs of panels. As good reporting of panel performance is absolutely required to gain medical community acceptance, we believe that reporting standards will be needed for panels. Detailing what this standard would be is out of the scope of this review, but we can highlight a few points of major importance.

In order to allow the ultimate independent validation by different laboratories, it is very important that statistical analysis methods are discussed in detail and information regarding the software and corresponding parameters is provided. Stating which software was used is important, since default parameters may differ in distinct implementations of the same method. Most studies do not follow this advice, with few exceptions [6,8,37]. For cross validation and bootstrapping, a graph such as that presented by Wild *et al.* usually helps the reader understand how the performance test was applied and what the reported results really mean [62]. Other requirements will need to be discussed by the panel community.

Comparison of methods

As mentioned earlier, several models can be generated from one dataset. Therefore, model comparison is crucial in order to optimize the final selection.

Several papers analyze datasets with more than one method [45,46,48]. However, there is no proper comparison. Reddy *et al.* states that "logical analysis of data model has significantly better performance on the independent validation set compared with the other classification models" [45]. However, there are no statistics to prove this difference, and confidence intervals

partially overlap. Prados *et al.* used McNemar's test for pairwise comparison of algorithms [46]. Wu eludes the problem by studying the stability of the model performance over several cross validation or bootstrap replicates [48].

The most important point is that performance estimates should be compared on a dataset that is independent from the model definition [75,76]. This can be carried out either with an independent validation cohort (separated or split), or by estimating performance via cross validation or bootstrapping.

Receiver-operating characteristic curves

Traditionally, performance of a test discriminating between two classes of patients is evaluated using a receiver-operating characteristic (ROC) curve [77]. This shows the variation of sensitivity and specificity of a test as the decision threshold changes. When the decision threshold is low, sensitivity is high and specificity is low, thus corresponding to the top right zone of the curve. Conversely, when the decision threshold is high, specificity is high and sensitivity is low, which corresponds to the bottom left part of the curve (Figure 5, see also Table 1).

A biomarker with no discrimination power would be characterized by a diagonal line, while a 'perfect' biomarker would reach the top left point corresponding to 100% sensitivity and 100% specificity. A major characteristic of a ROC curve is its AUC. The maximum AUC possible is 100%, corresponding to a 'perfect' classification. A nondiscriminating ROC curve has an AUC of 50%. In 1989, McClish introduced the concept of partial area under the ROC curve [15,78,79]. It consists of analyzing only a region of special interest of the ROC curve and allows the selection of models with high specificity or sensitivity, rather than models with a better average performance but potentially lower clinical value.

Hanley and McNeil [80] and DeLong *et al.* [81] proposed nonparametric methods to compare ROC curves derived from the same sample. McClish described a method to find a specific region within a ROC curve that is different [82]. Baker proposed a method to select best thresholds from a multidimensional ROC curve [83].

An intrinsic property of ROC curves is that the AUC of smooth curves tend to be greater than those of trapezoidal or step graphs [81,84]. Therefore, classification methods or predictors that can take only a few values (such as clinical scores) will not work as well as continuous predictors (such as biomarkers). Several smoothing procedures can be applied to reduce this problem. For example, logistic or other regression techniques will produce smooth estimates of the class probabilities. Gu *et al.* present a smoothing procedure based on Bayesian bootstrap estimation [85].

Another option is to bootstrap and compute confidence intervals and see if the observed sample is compatible with the bootstrap distribution [71,72]. Reddy *et al.* adopted this solution [45].

Classifications

Statistical tests should also be applied in order to judge the significance of differences between classifiers. If only two classifiers are compared, a simple binomial or McNemar test [38,86,87] can calculate the p-value to show that both classifiers are equally good [88]. Both tests are based on a 2 × 2 table, where the diagonal elements count the number of patients where both classifiers agree (either correctly or erroneously), and the off-diagonal elements indicate the number of patients where only one of the classifiers produces the right prediction. The off-diagonal elements are then compared with the calculated p-values. The number of patients where both classifiers agree does not enter into these calculations, which can cause a problem if the number of ties is much larger than the number of discrepancies, and these tests will overestimate the difference between the classifiers. Other, more sophisticated and general tests and methods for testing multiple classifiers are also described in Salzberg's overview [89]. Often, several parameterizations of the same classifiers are tested and the best one is retained. This can lead to overly optimistic results if the p-values are not adjusted for multiple testing. For example, if 20 independent parameterizations are tested at a 5% significance level, one of these parameterizations may exceed the significance level just by chance.

A panel should perform better than each of its individual markers. When comparing the performance of a panel with that of an individual marker, it is important to be as fair as possible. In most publications, the predictions of individual markers are not evaluated by cross validation, which may lead to overly optimistic results [69]. Therefore, we recommend measuring all classifier performances with the same cross validation method or on an independent test set.

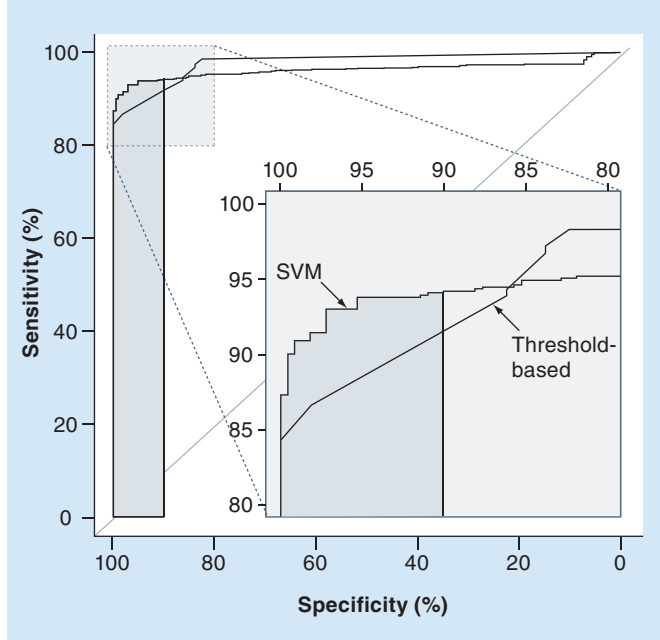


Figure 5. Two receiver-operating characteristic curves for support vector machines and threshold-based classifiers. Cross validation of partial AUC between 100 and 90% specificity are shown in gray. Respective values of partial AUC are 9.3 and 8.8% (perfect classification would correspond to a partial AUC of 10%).
SVM: Support vector machine.

Expert commentary

Interest in biomarker panels has been growing over the last few years. A number of publications have demonstrated that the approach has a big potential and could be suitable for various clinical applications. They applied many different methods, based on thresholds, decision trees, logistic regression, SVM and several other techniques. None of these methods is clearly superior. SVMs are well studied and tend to work well, even for high-dimensional data, whereas threshold-based methods are easy to implement and understand for medical practitioners. The final choice of a method must be carefully validated.

New markers, although they do not individually perform better than the current ones, could bring useful complementary pieces of information to a panel if they allow evaluation of the state of different pathways. However, such a relation must be sought during the discovery phase, which is made difficult by the very low sample size commonly used.

The limited consensus regarding accepted statistical methods and tools hamper their adoption, and could explain why the number of panels available in clinical practice is still limited. We predict that such standardized methods and tools will soon be made available and that the field will continue to grow despite these current limitations. Validation and comparison are of major importance in the evaluation of panels. It is not always possible to obtain an independent validation cohort, but in this case, the model must be evaluated by cross validation or bootstrap. Here again, the lack of clear guidelines and standards makes it difficult to compare different methods and impedes the credibility of the published results.

Five-year view

To gain a broad acceptance, future panel studies will need to define and follow reporting standards. Special care regarding validation will be required. Robust statistical methods of comparison must still be defined and are a crucial step. There is clearly a critical need for standardized methodologies and reporting standards to gain the medical practitioner's confidence. It is not unreasonable to say that in the absence of a strict enforcement of guidelines, most authors will not comply with better validation and reporting.

In the future, proteomics researchers willing to work with panels will need to think about combinations during the discovery process. Standard feature-selection techniques that select only a few of the best individual markers might reject proteins that are less efficient individually but that might have a greater weight in a panel. Some progress has been made towards this goal, with promising results [89].

We can imagine that proteomics biomarkers, which are still not commonly used in clinical practice, and panels, might contribute to new and more efficient IVD tools. However, given that the field is only in its first stages, it will probably take more than five years to see protein panels used in large scale clinical practice.

Acknowledgements

The authors would like to thank Proteome Sciences Plc for their keen support.

Financial disclosures

Jean-Charles Sanchez receives a grant from Proteome Sciences Plc for brain marker discovery. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Key issues

- A panel is the combination of information from several molecules into one predictor.
- Several methods can be applied. None of them is clearly superior. Support vector machines are usually preferred for high-dimensional data, such as mass spectra, while logistic regression or threshold-based methods are commonly preferred with ELISA-measured biomarkers.
- Methods are difficult to compare, and no efficient comparison tool is available yet.
- An especially careful validation is required in order not to overestimate the performance. It can be achieved either by using a separate dataset or by means of cross validation and/or bootstrap. Validation in an independent cohort measured by a different group is eventually required.
- Reporting detailed information regarding software and parameters set for preprocessing, classification, validation and comparison of methods should be seen as requirements. Reporting standards need to be developed.

References

Papers of special note have been highlighted as:
• of interest
•• of considerable interest

1 Hanash S. Disease proteomics. *Nature* 422(6928), 226–232 (2003).

2 Steel LF, Haab BB, Hanash SM. Methods of comparative proteomic profiling for disease diagnostics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 815(1–2), 275–284 (2005).

3 Domon B, Aebersold R. Mass spectrometry and protein analysis. *Science* 312(5771), 212–217 (2006).

4 Hanash SM, Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. *Nature* 452(7187), 571–579 (2008).

5 Panchaud A, Affolter M, Moreillon P, Kussmann M. Experimental and computational approaches to quantitative proteomics: status quo and outlook. *J. Proteomics* 71(1), 19–33 (2008).

6 Knickerbocker T, Chen JR, Thadhani R, MacBeath G. An integrated approach to prognosis using protein microarrays and nonparametric methods. *Mol. Syst. Biol.* 3(123), 1–8 (2007).

• **A complete study with a multiplex protein array approach and a thorough statistical analysis.**

7 Turck N, Vutskits L, Sanchez-Pena P *et al.* A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage. *Intensive Care Med.* DOI: 10.1007/s00134-009-1641-y (2009) (Epub ahead of print).

8 Hainard A, Tiberti N, Robin X *et al.* A combined CXCL10, CXCL8 and H-FABP panel for the staging of human African trypanosomiasis patients. *PLoS Negl. Trop. Dis.* 3(6), E459 (2009).

- **Impressive results on a neglected tropical disease and a potential clinical application in the short to medium term.**

9 Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 20(5), 777–785 (2004).

10 Diamandis EP. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *J. Natl Cancer Inst.* 96(5), 353–356 (2004).

11 Feng Z, Yasui Y. Statistical considerations in combining biomarkers. *Dis. Markers* 20(2), 45–51 (2004).

12 Stead DA, Paton NW, Missier P *et al.* Information quality in proteomics. *Brief Bioinform.* 9(2), 174–188 (2008).

13 Oberg AL, Vitek O. Statistical design of quantitative mass spectrometry-based proteomic experiments. *J. Proteome Res.* 8(5), 2144–2156 (2009).

14 Vitzthum F, Behrens F, Anderson NL, Shaw JH. Proteomics: from basic research to diagnostic application. A review of requirements and needs. *J. Proteome Res.* 4(4), 1086–1097 (2005).

15 Dodd LE, Pepe MS. Partial AUC estimation and regression. *Biometrics* 59(3), 614–623 (2003).

16 Hackett JL, Gutman SI. Introduction to the Food and Drug Administration (FDA) regulatory process. *J. Proteome Res.* 4(4), 1110–1113 (2005).

17 Vanni S, Polidori G, Pepe G *et al.* Use of biomarkers in triage of patients with suspected stroke. *J. Emerg. Med.* DOI: 10.1016/j.jemermed.2008.09.028 (2009) (Epub ahead of print).

18 Sibon I, Rouanet F, Meissner W, Orgogozo JM. Use of the Triage Stroke Panel in a neurologic emergency service. *Am. J. Emerg. Med.* 27(5), 558–562 (2009).

19 Brouns R, Van Den Bossche J, De Surgeloose D, Sheorajpanday R, De Deyn PP. Clinical and biochemical diagnosis of small-vessel disease in acute ischemic stroke. *J. Neurol. Sci.* 285(1–2), 185–190 (2009).

20 Apple FS, Christenson RH, Valdes R *et al.* Simultaneous rapid measurement of whole blood myoglobin, creatine kinase MB, and cardiac troponin I by the triage cardiac panel for detection of myocardial infarction. *Clin. Chem.* 45(2), 199–205 (1999).

21 Hoffer A, Osmond H. A card sorting test helpful in making psychiatric diagnosis. *J. Neuropsychiatr.* 2, 306–330 (1961).

22 Report of World Federation of Neurological Surgeons Committee on a Universal Subarachnoid Hemorrhage Grading Scale. *J. Neurosurg.* 68(6), 985–986 (1988).

23 Woolas RP, Xu F-J, Jacabs IJ *et al.* Elevation of multiple serum markers in patients with stage I ovarian cancer. *J. Natl Cancer Inst.* 85(21), 1748–1751 (1993).

24 Hill MD, Jackowski G, Bayer N, Lawrence M, Jaeschke R. Biochemical markers in acute ischemic stroke. *Can. Med. Assoc. J.* 162(8), 1139–1140 (2000).

25 Ferguson RE, Hochstrasser DF, Banks RE. Impact of preanalytical variables on the analysis of biological fluids in proteomic studies. *Proteomics Clin. Appl.* 1(8), 739–746 (2007).

26 Rai AJ, Vitzthum F. Effects of preanalytical variables on peptide and protein measurements in human serum and plasma: implications for clinical proteomics. *Expert Rev. Proteomics* 3(4), 409–426 (2006).

27 Little RR, Rohlfing CL, Tennill AL *et al.* Standardization of C-peptide measurements. *Clin. Chem.* 54(6), 1023–1026 (2008).

28 Yeo I-K, Johnson RA. A new family of power transformations to improve normality or symmetry. *Biometrika* 87(4), 954–959 (2000).

29 Gentleman R, Huber W, Carey V, Irizarry RA, Dudoit S. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Gentleman R, Huber W, Carey V, Irizarry RA, Dudoit S (Eds). Springer-Verlag, NY, USA (2005).

30 Hilario M, Kalousis A. Approaches to dimensionality reduction in proteomic biomarker studies. *Brief Bioinform.* 9(2), 102–118 (2008).

31 Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19), 2507–2517 (2007).

32 Baggerly KA, Morris JS, Wang J, Gold D, Xiao L-C, Coombes KR. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics* 3(9), 1667–1672 (2003).

33 Petricoin EF, Ardekani AM, Hitt BA *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359(9306), 572–577 (2002).

34 Hastie T, Tibshirani R, Friedman J. *Elements of Statistical Learning: data mining, inference, and prediction.* Springer-Verlag, NY, USA (2003).

- **Reference book for statistical learning.**

35 Faca VM, Song KS, Wang H *et al.* A mouse to human search for plasma proteome changes associated with pancreatic tumor development. *PLoS Med.* 5(6), E123 (2008).

36 Reynolds MA, Kirchick HJ, Dahlen JR *et al.* Early biomarkers of stroke. *Clin. Chem.* 49(10), 1733–1739 (2003).

- **First attempt to select a panel with thresholds defined in a multivariate manner.**

37 Montaner J, Perea-Gainza M, Delgado P *et al.* Etiologic diagnosis of ischemic stroke subtypes with plasma biomarkers. *Stroke* 39(8), 2280–2287 (2008).

38 Lejon V, Roger I, Mumba Ngoyi D *et al.* Novel markers for treatment outcome in late-stage *Trypanosoma brucei gambiense* trypanosomiasis. *Clin. Infect. Dis.* 47(1), 15–22 (2008).

39 Wang P, Kim Y, Pollack J, Tibshirani R. Boosted PRIM with application to searching for oncogenic pathway of lung cancer. *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference.* Stanford, CA, USA, 16–19 August 2004.

40 Ralhan R, DeSouza LV, Matta A *et al.* Discovery and verification of head-and-neck cancer biomarkers by differential protein expression analysis using iTRAQ labeling, multidimensional liquid chromatography, and tandem mass spectrometry. *Mol. Cell. Proteomics* 7(6), 1162–1173 (2008).

41 Webb GI, Boughton JR, Wang Z. Not so naive Bayes: aggregating one-dependence estimators. *Machine Learn.* 58(1), 5–24 (2005).

42 Seeber B, Sammel MD, Fan X *et al.* Panel of markers can accurately predict endometriosis in a subset of patients. *Fertil. Steril.* 89(5), 1073–1081 (2008).

43 Rosen RC, Cappelleri JC, Smith MD, Lipsky J, Peña BM. Development and evaluation of an abridged 5-item version of the International Index of Erectile Function (IIEF-5) as a diagnostic tool for erectile dysfunction. *Int. J. Impot. Res.* 11(6), 319–326 (1999).

44 Patz EF, Campa MJ, Gottlin EB, Kusmartseva I, Guan XR, Herndon JE. Panel of serum biomarkers for the diagnosis of lung cancer. *J. Clin. Oncol.* 25(35), 5578–5583 (2007).

45 Reddy A, Wang H, Yu H *et al.* Logical Analysis of Data (LAD) model for the early diagnosis of acute ischemic stroke. *BMC Med. Inform. Decis. Mak.* 8, 30 (2008).

•• **Comparison of five methods (Logical Analysis of Data, support vector machine, decision tree, logistic regression and multilayer perceptron) on SELDI data. Exemplary validation with tenfold cross-validation and an independent validation set.**

46 Prados J, Kalousis A, Sanchez J-C, Allard L, Carrette O, Hilario M. Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics* 4(8), 2320–2332 (2004).

47 Ring BZ, Seitz RS, Beck R *et al.* Novel prognostic immunohistochemical biomarker panel for estrogen receptor-positive breast cancer. *J. Clin. Oncol.* 24(19), 3039–3047 (2006).

48 Wu B, Abbott T, Fishman D *et al.* Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 19(13), 1636–1643 (2003).

49 Breiman L. Random forests. *Machine Learn.* 45(1), 5–32 (2001).

50 Visintin I, Feng Z, Longton G *et al.* Diagnostic markers for early detection of ovarian cancer. *Clin. Cancer Res.* 14(4), 1065–1072 (2008).

•• **Describes several logistic regression models fitted with a very clear validation. Data were acquired in a multiplex bead assay.**

51 Lynch JR, Blessing R, White WD, Grocott HP, Newman MF, Laskowitz DT. Novel diagnostic test for acute stroke. *Stroke* 35(1), 57–63 (2004).

52 Rosengart AJ, Schultheiss KE, Tolentino J, Macdonald RL. Prognostic factors for outcome in patients with aneurysmal subarachnoid hemorrhage. *Stroke* 38(8), 2315–2321 (2007).

53 Zheng Y, Katsaros D, Shan SJC *et al.* A multiparametric panel for ovarian cancer diagnosis, prognosis, and response to chemotherapy. *Clin. Cancer Res.* 13(23), 6984–6992 (2007).

54 Laskowitz DT, Kasner SE, Saver J, Remmel KS, Jauch EC; BRAIN Study Group. Clinical usefulness of a biomarker-based diagnostic test for acute stroke: the Biomarker Rapid Assessment in Ischemic Injury (BRAIN) study. *Stroke* 40(1), 77–85 (2009).

55 Welsh P, Barber M, Langhorne P, Rumley A, Lowe GDO, Stott DJ. Associations of inflammatory and haemostatic biomarkers with poor outcome in acute ischaemic stroke. *Cerebrovasc. Dis.* 27(3), 247–253 (2009).

56 Wicki J, Perneger TV, Junod AF, Bounameaux H, Perrier A. Assessing clinical probability of pulmonary embolism in the emergency ward: a simple score. *Arch. Intern. Med.* 161(1), 92–97 (2001).

57 Zervakis M, Blazadonakis ME, Tsiliki G, Danilatou V, Tsiknakis M, Kafetzopoulos D. Outcome prediction based on microarray analysis: a critical perspective on methods. *BMC Bioinformatics* 10, 53 (2009).

58 Schramm A, Schulte JH, Klein-Hitpass L *et al.* Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling. *Oncogene* 24(53), 7902–7912 (2005).

59 Petricoin EF, Liotta LA. SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. *Curr. Opin. Biotechnol.* 15(1), 24–30 (2004).

60 Ressom HW, Varghese RS, Drake SK *et al.* Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics* 23(5), 619–626 (2007).

61 Liu JJ, Cutler G, Li W *et al.* Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics* 21(11), 2691–2697 (2005).

62 Wild N, Karl J, Grunert VP *et al.* Diagnosis of rheumatoid arthritis: multivariate analysis of biomarkers. *Biomarkers* 13(1), 88–105 (2008).

63 Gevaert O, Smet FD, Timmerman D, Moreau Y, Moor BD. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* 22(14), E184–E190 (2006).

64 James GM, Hastie TJ. functional linear discriminant analysis for irregularly sampled curves. *J. R. Statist. Soc. B* 63(3), 533–550 (2001).

65 Ishino M, Takeishi Y, Niizeki T *et al.* Implications of BNP, H-FABP, and PTX3. *Circ. J.* 72(11), 1800–1805 (2008).

66 Ross DT, Kim C-Y, Tang G *et al.* Chemosensitivity and stratification by a five monoclonal antibody immunohistochemistry test in the NSABP B14 and B20 trials. *Clin. Cancer Res.* 14(20), 6602–6609 (2008).

67 Linkov F, Lisovich A, Yurkovetsky Z *et al.* Early detection of head and neck cancer: development of a novel screening tool using multiplexed immunobead-based biomarker profiling. *Cancer Epidemiol. Biomarkers Prev.* 16(1), 102–107 (2007).

68 Bhaskar H, Hoyle DC, Singh S. Machine learning in bioinformatics: a brief survey and recommendations for practitioners. *Comput. Biol. Med.* 36(10), 1104–1125 (2006).

69 Whiteley W, Tseng M-C, Sandercock P. Blood biomarkers in the diagnosis of ischemic stroke: a systematic review. *Stroke* 39(10), 2902–2909 (2008).

70 Smit S, van Breemen MJ, Hoefsloot HCJ, Smilde AK, Aerts JMFG, de Koster CG. Assessing the statistical validity of proteomics based biomarkers. *Anal. Chim. Acta* 592(2), 210–217 (2007).

71 Hesterberg T, Moore DS, Monaghan S, Clipson A, Epstein R. Bootstrap methods and permutation tests. In: *Introduction to the Practice of Statistics (5th Edition).* WH Freeman & Company, NY, USA (2005).

72 Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.* 19(9), 1141–1164 (2000).

73 Taylor CF. Minimum reporting requirements for proteomics: a MIAPE primer. *Proteomics* 6(Suppl. 2), 39–44 (2006).

74 Bossuyt PM, Reitsma JB, Bruns DE *et al.* The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin. Chem.* 49(1), 7–18 (2003).

75 LaBaer J. So, you want to look for biomarkers (introduction to the special biomarkers issue). *J. Proteome Res.* 4(4), 1053–1059 (2005).

76 Hilario M, Kalousis A, Pellegrini C, Müller M. Processing and classification of protein mass spectra. *Mass Spectrom. Rev.* 25(3), 409–449 (2006).

77 Fawcett T. An introduction to ROC analysis. *Pattern Recogn. Lett.* 27(8), 861–874 (2006).

78 McClish DK. Analyzing a portion of the ROC Curve. *Med. Decis. Making* 9(3), 190–195 (1989).

79 Thompson ML, Zucchini W. On the statistical analysis of ROC curves. *Stat. Med.* 8(10), 1277–1290 (1989).

80 Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148(3), 839–843 (1983).

81 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44(3), 837–845 (1988).

82 McClish DK. Determining a range of false-positive rates for which ROC curves differ. *Med. Decis. Making* 10(4), 283–287 (1990).

83 Baker SG. Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics* 56(4), 1082–1087 (2000).

84 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1), 29–36 (1982).

85 Gu J, Ghosal S, Roy A. Bayesian bootstrap estimation of ROC curve. *Stat. Med.* 27(26), 5407–5420 (2008).

86 Morais DF, Spotti AR, Tognola WA, Gaia FFP, Andrade AF. Clinical application of magnetic resonance in acute traumatic brain injury. *Arquivos de Neuro-Psiquiatria* 66(1), 53–58 (2008).

87 Vasconcelos OM Jr, Prokhorenko OA, Kelley KF *et al.* A comparison of fatigue scales in postpoliomyelitis syndrome. *Arch. Phys. Med. Rehabil.* 87(9), 1213–1217 (2006).

88 Salzberg S. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Min. Knowl. Discov.* 1(3), 317–328 (1997).

89 Gillette MA, Mani DR, Carr SA. Place of pattern in proteomic biomarker discovery. *J. Proteome Res.* 4(4), 1143–1154 (2005).

90 Han J, Kamber M. *Data Mining.* Morgan Kaufmann, CA, USA (2001).

91 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.* 159(9), 882–890 (2004).

92 Cortes C, Vapnik V. Support-vector networks. *Machine Learn.* 20(3), 273–297 (1995).

Websites

101 McCormick T, Martin K, Hehenberger M. The evolving role of biomarkers – Focusing on patients from research to clinical practice. *IBM Global Business Services* (2007) www.03.ibm.com/industries/healthcare/doc/content/resource/insight/2799019105.html

102 Hammer PL, Bonates T. Logical analysis of data: from combinatorial optimization to medical applications. RUTCOR Research Report (10–2005) (2005) http://rutcor.rutgers.edu/pub/rrr/reports2005/10_2005.pdf

Affiliations

- Xavier Robin
  Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, Geneva, Switzerland
- Natacha Turck
  Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, Geneva, Switzerland
- Alexandre Hainard
  Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, Geneva, Switzerland
- Frédérique Lisacek
  Swiss Institute of Bioinformatics, Medical University Centre, Geneva, Switzerland
- Jean-Charles Sanchez
  Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, Medical University Centre, Geneva, Switzerland
  jean-charles.sanchez@unige.ch
- Markus Müller
  Swiss Institute of Bioinformatics, Medical University Centre, Geneva, Switzerland