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

Xavier Robin[a], Natacha Turck[a], Alexandre Hainard[a], Frédérique Lisacek[b], Jean-Charles Sanchez[a] and Markus Müller[☡, b]

[a]Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, University of Geneva, Switzerland
[b]Proteome Informatics Group, Swiss Institute of Bioinformatics, Geneva, Switzerland

December 2009

Abstract

A large number of biomarkers were discovered by proteomics techniques over the past few years. 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 in panels to improve clinical performances has emerged. We present here a discussion of the bioinformatics aspects of biomarker panels and concomitant challenges including high dimensionality, low patient number and reproducibility.

Citation: Robin X., Turck N., Hainard A., *et al.* (2009). Bioinformatics for protein biomarker panel classification: What is needed to bring biomarker panels into in vitro diagnostics? *Expert Review of Proteomics* 6 (6) p. 675–689. PMID: 19929612. DOI: 10.1586/EPR.09.83.

Abbreviations: AUC: area under the curve; ELISA: enzyme-linked immunosorbent assay; FDA: american food and drug administration; GSTP: Glutathione S-transferase P; HAT: human African trypanosomiasis; H-FABP: Fatty acid-binding protein, heart; iTRAQ: isobaric tag for relative and absolute quantitation; IVD: in vitro diagnostics; MALDI: matrix-assisted laser desorption/ionization; MS/MS: tandem mass spectrometry; ROC: receiver operating characteristic; SELDI: surface-enhanced laser desorption/ionization; SVM: support vector machines; WFNS: world federation of neurological surgeons committee.

Keywords: Panel, combination of biomarkers, classification, validation, proteomics, biomarkers, in-vitro diagnostics.

Submitted: June 1st, 2009; Revised: October 7th, 2009; Accepted: October 8th, 2009. Republished here with the permission of Future Science Group for Expert Reviews of Proteomics.

Introduction

As part of clinical practice, it is common to measure the concentration of a protein, called biomarker, in a biological sample either to diagnose a disease, to early predict the outcome or to 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 or C-reactive protein for assessment of inflammation. Proteomics techniques such as two-dimensional gel electrophoresis (Hanash 2003; Steel *et al.* 2005) and mass spectrometry (Domon & Aebersold 2006; Hanash *et al.* 2008; Panchaud *et al.* 2008; Steel *et al.* 2005) 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 (time,

☡ Corresponding author. Swiss Institute of Bioinformatics, CMU, 1 Michel-Servet, CH-1211 Geneva 4, Switzerland. Tel.: +41 22 379 5847; fax: +41 22 379 5858; e-mail address: Markus.Mueller@isb-sib.ch.

huge costs required to validate these molecules as well as the accuracy of assays not high enough to be translated directly into clinical practice) and biological (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 like age, sex, physiological constant or clinical scores with biomarkers (Knickerbocker *et al.* 2007). Like a single marker, a panel allows answering different clinical questions. Apart from increasing accuracy biomarker panels help studying different pathophysiological pathways and shedding light on a disease from different angles. For instance in the context of a brain damage condition (aneurysmal subarachnoid hemorrhage), Turck et al. (Turck *et al.*) recently demonstrated that a combination of brain parameters associated with a clinical score and also a cardiac biomarker could predict 6-months outcome better than the biomarkers taken individually. In the same manner, Hainard et al. (Hainard *et al.* 2009) proposed a combination of inflammatory cytokines and one brain damage marker. In both cases, the combination of different kind of biomarkers improves the classification.

In contrast to the traditional single analyte interpretation, several new challenges arise which could also explain why panels are not widespread yet. First, appropriate methods are required to combine information from multiple biomarkers. These methods must be efficient and yield correct patient classification, but they also have to be comprehensible to medical practitioners to gain acceptance. Secondly, the risk of overfitting the data is increased because of the higher dimensionality (Baggerly *et al.* 2004; Diamandis 2004; Feng & Yasui 2004). A careful validation is required to ensure that a panel truly performs better than individual biomarkers, hence avoiding raising false hopes. And finally, appropriate experimental design (Oberg & Vitek ; Stead *et al.* 2008) 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 see whether they addressed these new challenges and if so, how. We mainly focus on protein biomarker panels, but also include related work on analysing protein or gene expression microarray and protein mass spectrometry data if we deem it relevant for protein panels. We will also review methods that allow validating obtained models and their performance, as well as the strategies available to compare different panels and their combination. This review addresses clinical researchers who seek a basic introduction to statistical methods and pitfalls of biomarker panel research and statisticians who would like to learn more about recent work and clinical aspects of this subject.

From discovery to IVD

IVD encompasses any type of assays 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 that are performed by trained personnel in dedicated clinical chemistry labs.

Vitzthum et al. (Vitzthum *et al.* 2005) reviewed the needs in proteomics to push the discovered molecules into IVD. The crucial points are that the classification must be reliable and give information valuable for decision making; measurements must be both exact and robust, and the test accuracy must meet sufficient (positive or negative) predictive values.

Target performance of IVD tests have to be chosen according to the clinical question. As pointed out by Dodd and Pepe (Dodd & Pepe 2003), “large monetary costs result from high false-positive rates”. Similarly, failure to diagnose a disease can dramatically impact on patient’s health, which may even lead to death. Therefore, IVD (single biomarker or panel) as a helpful clinical practice must display a sufficient discrimination power and answer a well-defined question. In other words, one should focus on high sensitivity/specificity or high predictive values rather than global accuracy.

An IVD test aims at determining the state of the patient. Usually for biomarker tests a decision threshold (also called cut-off) is chosen. Any value below the cut-off will mean that the test result is

Word	Common abbreviation	Formula	Definition
Prevalence			frequency of the positive occurrence in the studied population
Rule-in (confirmatory)			a test done in an attempt to confirm the presence of a disease
Rule-out (screening)			a test done 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	$\frac{SE}{1-SE} \times \frac{SP}{1-SP}$	effect of a given increase of the studied marker.

Table 1: Clinical classification definitions.

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 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 and so 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 called a screening test. A negative result to a screening test implies a high probability for the patient 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 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 HAT field would select only patients with parasite in the brain (stage 2 of the disease), who will be subsequently subjected to a very toxic treatment. Patient without brain infection (stage 1) must be excluded because they could be potentially killed by the inappropriate medication (Hainard *et al.* 2009).

Predictive values (negative or positive) need to take the class prevalence into account since otherwise 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 for the patient to be truly positive or negative for a given group of patients.

Commercial panels

From a commercial point of view, McCormick (McCormick *et al.* 2007) 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. However, they pointed out that acceptance of biomarkers is hindered by the lack of data sharing (due to technical or strategic reasons) as well as insufficient validation and targeting.

In the USA, medical devices (including IVD) must get approval by the American food and drug administration (FDA). Hackett and Gutman (Hackett & Gutman 2005) highlighted the difficulties that are raised by the combination of several markers and the use of statistical models. FDA review procedures for device acceptation 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 measures simultaneously four markers (namely MMP-9, BNP, D-dimer and S100) and computes a *Multimarker Index* (MMX) using a proprietary algorithm. Two cut-offs are defined, associated with a high or low risk for the patient to have a stroke, while patients in the intermediate region need further investigation. It was accepted by the FDA for premarket approval application in 2005, was withdrawn by the manufacturer one year later to allow further clinical studies, but it was recently reintroduced. The Triage Stroke Panel was applied by Vanni et al. (Vanni *et al.*) in a neurological emergency service to discriminate patients with stroke or without among those having a suspicion of stroke. Sibon et al. (Sibon *et al.* 2009) compared it with an established neurological scale evaluated by nurses. Brouns et al. (Brouns *et al.* 2009) 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 MMX score.

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 CK-MB, Myoglobin, and Troponin I) (Apple *et al.* 1999). However, it cannot truly be called a panel as measurements are not combined into a single final score.

Applied Genomics sells several immuno-histochemistry 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 [19].

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 get an improved patient classification is not new. In psychiatry, Hoffer and Osmond (Hoffer & Osmond 1961) applied a combination of neuropsychiatric variables in the early 1960's to distinguish schizophrenic patients from normal individuals. 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 WFNS score (World Federation of Neurological Surgeons Committee 1988) was defined to assess patients' neurological status. It consists of the combination of three easy-to-assess clinical variables. Eye, verbal and motor responses are evaluated on a scale ranging from 1 to 4, 5 and 6 respectively. An intermediate score ranging from 3 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. (Woolas *et al.* 1993) showed the potential of using several serum markers together in 1993. 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 didn't use this observation to make a true statistical combination. In 2000, Hill et al. (Hill *et al.* 2000) were among the first to report the use of a panel of protein biomarkers. 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 section 2.3, panels can also combine biomarkers and clinical parameters. But 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.

Pre-processing

Normalization and 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 (sample container, time to freezing, storage temperature) and instrumental errors can lead to measurement variations (Ferguson *et al.* 2007; Rai & Vitzthum 2006). 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 (Baggerly *et al.* 2004). Furthermore, biological variability between different patients due to sex, age, treatment, lifestyle, 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 (Little *et al.* 2008). 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 0 and the variance to 1, but otherwise does not affect the shape of the distribution. Yeo et al. (Yeo & Johnson 2000) proposed the box-cox transformation family, which includes the logarithmic transformation, to obtain distributions closer to the normal one. Another normalization method is the Quantile Normalization, where all values are transformed into their corresponding normal quantiles (Gentleman *et al.* 2005). 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 (ANOVA) can calculate the bias and variance introduced by each processing step and lead to more accurate comparisons (Oberg & Vitek 2009).

Feature selection

Another important pre-processing step is feature selection that turns out to be crucial in high-dimensionality problems such as mass spectra or microarrays, but is less important for lower dimensional biomarker panels. It consists in choosing 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 (Hilario & Kalousis 2008). It is also important to note that data for feature selection must not include the test data otherwise test performance would be too optimistic. Saeys et al. (Saeys *et al.* 2007) classified the feature selection methods into three categories: filter methods, wrapper methods and embedded methods. Filter methods consider only the intrinsic properties of single features independently from classification. To the contrary, 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 (see below).

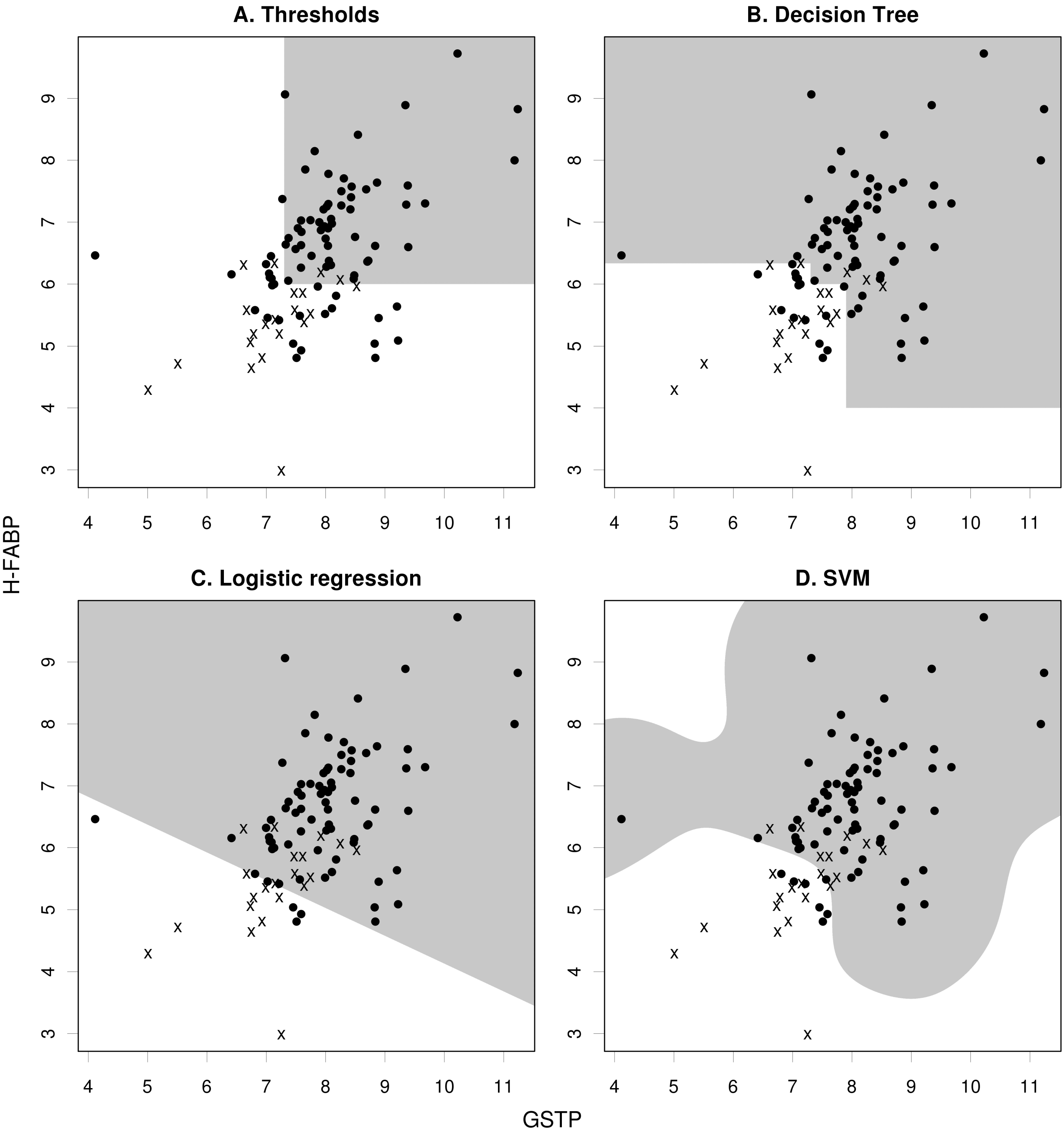

Figure 1: Classification by different methods. Data described by Hainard et al. (Hainard A *et al.* 2009) (GSTP and H-FABP concentrations illustrated in log scale). In grey, the region where the test would be considered positive by the method. Crosses and dots represent Stage 1, respectively Stage 2 HAT patients. A: Threshold-based method split the space into boxes; B: Decision trees can create more boxes; C: Logistic regression divide the data with a straight line; D: Support Vector Machines (SVM) can figure out complex separations but can also create linear partitions similarly to logistic regression (see Figure 4).

Several examples of feature selection are reported by Hilario and Kalousis (Hilario & Kalousis 2008). Baggerly et al. (Baggerly *et al.* 2003) used pre-processing, exhaustive search and genetic algorithms to reduce an initial 60 831 *m/z* values from mass spectrometry to filter 506 and then sets of 1 to 5 features, and then applied the feature sets to linear discriminant analysis. Petricoin et al. (Petricoin *et al.* 2002) also employed genetic algorithm with mass spectrometry, but in a wrapper method around a self-organizing map (SOM) algorithm.

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

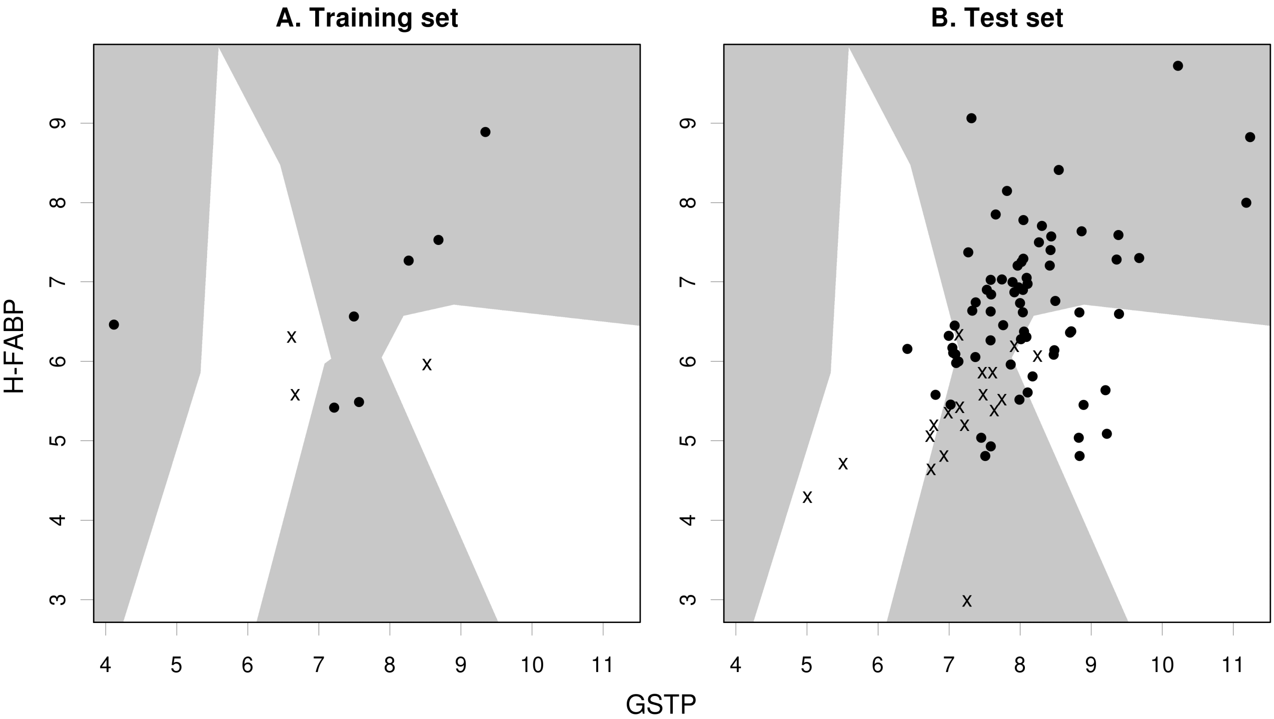

Figure 2: Overfitting with Nearest Neighbour algorithm (k=1). Same data (Hainard *et al.* 2009) as previously shown in Figure 1. In grey, the region where the test would be considered positive by the method. Crosses and dots represent Stage 1, respectively Stage 2 HAT patients. A: Training set of 10 patients determining the class regions (grey 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.

excellent textbooks such as that by Hastie et al. (Hastie *et al.* 2003). The classification task consists in 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, i.e. 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 disjoint test data.

This approach may seem fairly straight forward, but two main problems have to 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 to 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 poorly or not at all represented 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 maybe 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, which is actually linear, Figure 2). However, cross validation provides a means to at least partially mitigate this problem (see below). 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 this simple boundaries cannot reproduce the real ones 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 Support Vector Machines (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 (Faca *et al.* 2008; Hainard *et al.* 2009; Hill *et al.* 2000; Lejon *et al.* 2008; Montaner *et al.* 2008; Reynolds *et al.* 2003; Turck *et al.*) (Figure 1A), a set of thresholds, one for each biomarker, is selected usually in a univariate manner. Any value of a molecule below its respective threshold will mean 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 a minimal 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 mostly used 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 it is well adapted to biomarker panel data where class boundaries of a single marker can be often represented as single cut-off points.

Lejon et al. (Lejon *et al.* 2008) followed this approach to combine clinical and biochemical variables to predict trypanosomiasis treatment failure. Thresholds were chosen on univariate parameters to maximize the sum of sensitivity and specificity and two parameters were retained. On the same disease, Hainard et al. (Hainard *et al.* 2009) selected a panel of two cytokines and a brain-damage marker to assess the disease stage of 100 patients using a multivariate approach. The rationale is 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 into this direction (Reynolds *et al.* 2003). Vitzthum et al. also showed that different thresholds should be chosen for different clinical questions (Vitzthum *et al.* 2005). 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 patient rule Iinduction method (PRIM) (Hastie *et al.* 2003), 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. (Wang *et al.* 2004), but its usage seems scarce. Naïve Bayes is also a similar method where the thresholds are determined based on statistical criteria separately for every feature. Ralhan et al. (Ralhan *et al.* 2008) successfully applied it on proteins quantified by tandem mass spectrometry (MS/MS) after iTRAQ labelling on a small number of patients. It can be extended to deal with dependent data (Webb *et al.* 2005).

Decision trees

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

Classification and regression trees (CART) is one of the most popular tree based algorithms (Patz *et al.* 2007; Rosen *et al.* 1999; Seeber *et al.* 2008). Other methods are C4.5 decision trees (Reddy *et al.* 2008), J48 (Prados *et al.* 2004) or RPART. The latter allowed Ring et al. (Ring *et al.* 2006) to choose five out of several hundreds proteins and combine them into a decision tree able to classify 195 ER+ breast cancer patients into good, moderate or poor prognosis. However it seemed to be dependent on the cohorts on which the model was applied and was less predictive of outcome than other methods.

Trees perform well in combination with boosting algorithms (Wu *et al.* 2003), 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 (Breiman 2001) where classification is obtained from a combination of trees, each built from a small but random subset of the features.

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 to 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 (Han & Kamber 2001; Hastie *et al.* 2003).

Box 1: Decision trees.

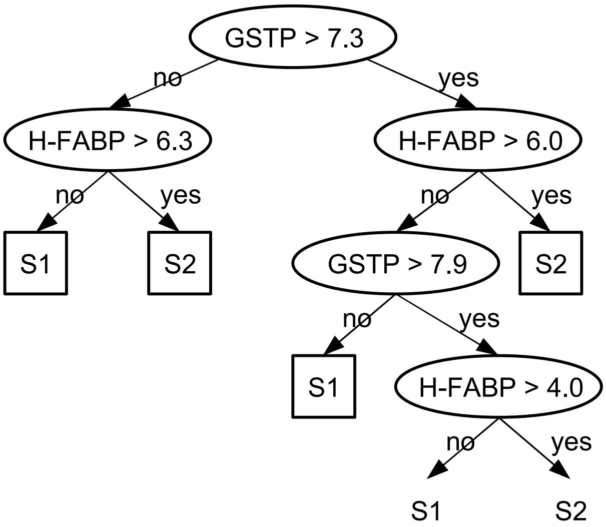

Figure 3: The decision tree corresponding to Figure 1B. It must be scanned from top to bottom. Each circle corresponds to a question. Depending on its answer, one follows the arrow to the left or right and goes to the next question, until a decision (square box) is reached. S1 and S2 mean Stage 1, respectively Stage 2 HAT classification.

A basic parallel or sequential AND/OR way of combining tests similar to decision trees has been proposed by Vitzthum et al. (Vitzthum *et al.* 2005). However, there is no evidence that it was applied in panels.

Logistic regression

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

Logistic regression can combine clinical or biomarker data, either continuous or categorical (Laskowitz *et al.* 2009; Lynch *et al.* 2004; Montaner *et al.* 2008; Reddy *et al.* 2008; Reynolds *et al.* 2003; Rosengart *et al.* 2007; Visintin *et al.* 2008; Zheng *et al.* 2007). For example, Visintin et al. (Visintin *et al.* 2008) trained several logistic regression models to screen ovarian cancer on several hundred patients and controls. Even though 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 (Welsh *et al.* 2009) or to combine clinical variables only (Wicki *et al.* 2001).

In its simplest form, logistic regression provides a linear separation of the feature space. It models the class probability $p(+|\mathbf{x})$, i.e. the probability that the n-dimensional feature vector x is classified positively, as a sigmoidal (s-shaped) function $f(z) = 1/(1+\exp(-z))$, where $z = \alpha_0 + \sum_{i=1}^{n} \alpha_i x_i$. The coefficients α_i have to be determined from the training sample by means of a maximum likelihood procedure, which usually converges to the unique global optimum (Hastie *et al.* 2003). If the different features x_i are properly normalized (same mean and standard deviation), the coefficients α_i give direct information about 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 $\mathbf{x} = (x_1, x_2)$ could be expanded to a higher dimensional vector $\mathbf{x}' = (x_1, x_1^2 + x_2^2, x_1x_2, x_2)$ or $x' = (x_1, x_1/x_2, x_2)$. The logistic regression is then applied to $\mathbf{x}'$ instead of $\mathbf{x}$.

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 (Pepe *et al.* 2004).

Box 2: Logistic regression.

SVM

Support vector machine (SVM) (Figure 1D, Box 3) is one of the most popular methods in machine learning. SVM has the advantage to provide a clear mathematical model with a globally optimal solution, to the contrary of 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 SVM can also find linear or polynomial separations similar to logistic regression.

SVM is preferred in higher dimensionality problems such as microarray (Schramm *et al.* 2005; Zervakis *et al.* 2009) or mass spectrometry (SELDI (Petricoin & Liotta 2004; Prados *et al.* 2004; Reddy *et al.* 2008) or MALDI (Ressom *et al.* 2007; Wu *et al.* 2003)) data analysis. Liu et al. (Liu *et al.* 2005) combined SVM with a genetic algorithm and obtained reproducible and fairly accurate results. It was also used by Wild et al. (Wild *et al.* 2008) 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.

Generalized additive models

Generalized additive models allowed Knickerbocker et al. (Knickerbocker *et al.* 2007) to combine protein microarray data with patient clinical information to predict survival after renal replacement. They added local polynomial functions (or splines) that allow defining non-linear relationships between the variables, as well as the 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 high risk by the clinical predictors.

Other methods

Several other methods were shown to perform well in proteomics. Gevaert et al. (Gevaert *et al.* 2006) applied a bayesian network on gene expression microarray data. This approach allows integrating clinical data in several manners: full integration, decision integration, 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.

Let us consider a 2-dimensional example where the 2 classes are completely separable by a straight line. It is easy to see that there are infinitely many straight lines that do the job, and the question is, which of these lines provides the best classification on a test sample. The support vector machine (SVM) (Cortes & Vapnik 1995) solves this problem by choosing the (usually unique) separating line that is farthest away from any data point. 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 $\mathbf{x}_i$ of class y_i (± 1) we have $\mathbf{w}\,\mathbf{x}_i + b \leq -1$ for $y_i = -1$ and $\mathbf{w}\,\mathbf{x}_i + b \geq 1$ for $y_i = 1$, where $\mathbf{w}$ is a vector orthogonal to the separating line. It can be shown (Cortes & Vapnik 1995) that the distance of the separating line to the next $\mathbf{x}_i$ is $1/|\mathbf{w}|$, therefore the SVM searches for the smallest $|\mathbf{w}|^2$, which satisfies the above inequalities. The lines $\mathbf{w}\,\mathbf{x}_i + b = -1$ for $y_i = -1$ and $\mathbf{w}\,\mathbf{x}_i + b = 1$ for $y_i = 1$ are called the margins, which lie parallel and at equal distance $1/|\mathbf{w}|$ 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. They introduced so-called slack variables $\xi_i \geq 0$ and reformulated the constraints as $\mathbf{w}\,\mathbf{x}_i + b \leq -1+\xi_i$ for $y_i = -1$ and $\mathbf{w}\,\mathbf{x}_i + b \geq 1-\xi_i$ for $y_i = 1$, i.e. for each $\mathbf{x}_i$ on the right side of its margin we have $\xi_i = 0$ and for each $\mathbf{x}_i$ on the wrong side of the margin $\xi_i > 0$, where $\xi_i\,/|\mathbf{w}|$ is the distance from the margin (Figure 4). Since we still would like to have a margin distance $2/|\mathbf{w}|$ as large as possible, but also as little miss-classification $\sum_{i=1}^{p} \zeta_i$ as possible, we search for a $\mathbf{w}$ satisfying the 'slack' inequalities above and minimizing $|w|^2 + C\sum_{i=1}^{p} \zeta_i$ where p is the number of samples and C a miss-classification weight. This is a quadratic programming problem, for which many efficient algorithms are available that 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 either lie on the margin or on the wrong side of it ($\mathbf{w}\,\mathbf{x}_i + b \geq -1$ for $y_i = -1$ and $\mathbf{w}\,\mathbf{x}_i + b \leq 1$ for $y_i = 1$), and $\alpha_i = 0$ for all other correctly classified vectors.

Cortes and Vapnik also showed that the SVM approach can be naturally extended to nonlinear separation (Cortes & Vapnik 1995). 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 centred at the support vectors (see (Cortes & Vapnik 1995) or (Hastie *et al.* 2003) for a detailed discussion of the kernel based formulation).

Box 3: Support Vector Machines.

Regularized discriminant analysis (RDA) (Hastie *et al.* 2003) is a classification method that can deal with strongly correlated data. It is based on linear discriminant analysis (Wu *et al.* 2003) or quadratic discriminant analysis. It can take into account the main effects of the markers as well as their interaction. Wild et al. (Wild *et al.* 2008) successfully used RDA to combine 2 to 3 molecules in patients with Rheumatoid arthritis. For prognostic purposes an attractive option is to analyze time series, if available. James and Hastie (James & Hastie 2001) proposed a classification based on spline regression of time series and linear discriminant analysis of the regression coefficients.

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 (Hammer & Bonates 2005). It was used by Reddy et al. (Reddy *et al.* 2008) to classify 48 ischemic stroke patients and 32 controls, and was applied on a validation set consisting of 60 patients. The methodology was also able to detect two outlier patients and showed good performance.

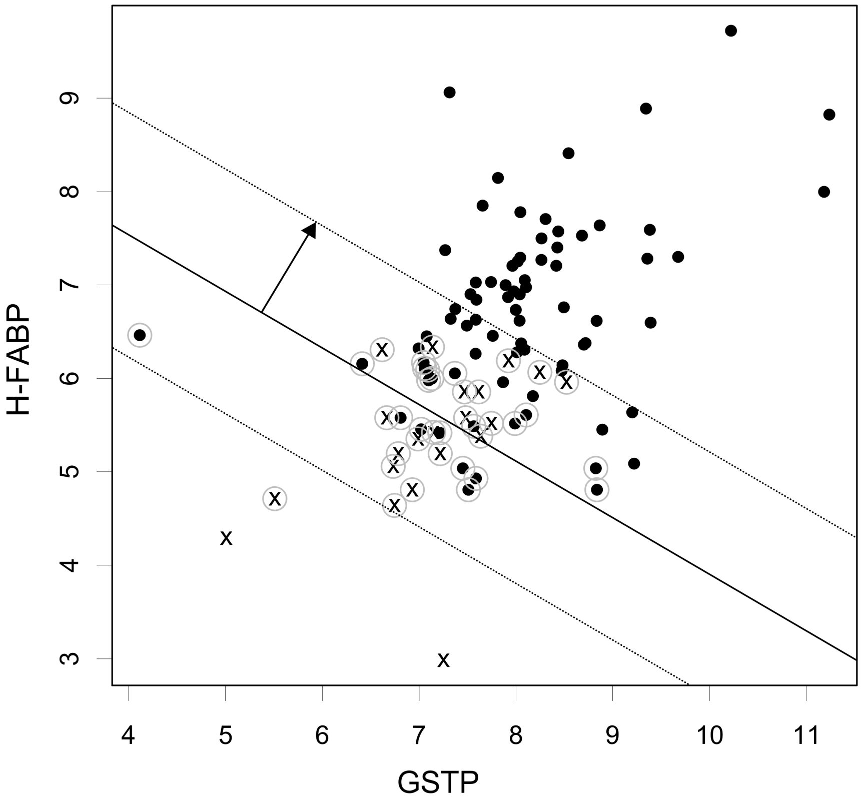

Figure 4: Support Vector Machines. Same data (Hainard *et al.* 2009) as previously shown in Figure 1 and Figure 2. Crosses and dots represent Stage 1, respectively Stage 2 HAT patients. Margins and the separation line are represented by dashed and a solid line respectively. Support vector observations are circled in grey. The arrow represents the vector $\mathbf{w}/|\mathbf{w}|^2$.

Reddy et al. (Reddy *et al.* 2008) and Prados et al. (Prados *et al.* 2004) applied multilayer perceptron, a type of linear neural network, and Cox proportional hazard models. The latter method was used in several other studies (Ishino *et al.* 2008; Ring *et al.* 2006; Rosengart *et al.* 2007; Ross *et al.* 2008).

Nearest neighbors (Wu *et al.* 2003) finds the *k* nearest samples and performs a majority vote to decide the classification. Linkov et al. (Linkov *et al.* 2007) defined a method they called ADE+PT, which is similar to a weighted nearest neighbor approach. 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 above, overfitting corresponds to the under-estimation of the classification error on the training set (Figure 2A) which cannot be validated on an independent test set (Figure 2B) (Hastie *et al.* 2003). High-dimensional data are especially prone to overfitting, as mentioned in Feng and Yasui (Feng & Yasui 2004) in the context of SELDI mass spectra, where a huge number of possible markers (peptide masses) are available. However, depending on the classifier, it can be a serious problem even for low dimensional data.

In the literature, Bhaskar et al. (Bhaskar *et al.* 2006) showed that validation is not done consistently in bioinformatics and Whiteley et al. (Whiteley *et al.* 2008) showed how even single biomarkers can be biased if its threshold is chosen on the same dataset. Several panel papers we previously mentioned did not perform any kind of validation of the accuracy of the reported classification (Faca *et al.* 2008; Hainard *et al.* 2009; Hill *et al.* 2000; Lejon *et al.* 2008; Montaner *et al.* 2008; Rosengart *et al.* 2007; Welsh *et al.* 2009) or simply mentioned that it would or should be done later. When this still is 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 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 lab. 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 (Patz *et al.* 2007; Reynolds *et al.* 2003; Visintin *et al.* 2008; Zheng *et al.* 2007). If not enough patients are available, randomization techniques such as permutation tests, cross-validation and bootstrapping (Feng & Yasui 2004) can help evaluate if the classification is significant or if it is only overfitting.

Permutation tests

Permutation tests (Hesterberg *et al.* 2005; Smit *et al.* 2007) allow determining if the classification result is significant. Patient labels are randomly permutated, and the problem is treated in the same way, giving information about 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 evaluating the robustness of a classification. In cross-validation, the data is split into k 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 (Hastie *et al.* 2003).

Typical values for k are 5 or 10 (Hastie *et al.* 2003). 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 overestimate 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 (Linkov *et al.* 2007; Reddy *et al.* 2008; Ring *et al.* 2006; Rosen *et al.* 1999; Visintin *et al.* 2008; Wild *et al.* 2008) for biomarker applications.

Several variants of cross-validation exist. When the data is not balanced, i.e. 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 width of a kernel, kernel-type, or number of principal components) are defined, and an external loop where the model is actually trained with these parameters and performance is evaluated (Smit *et al.* 2007).

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 (Hastie *et al.* 2003; Hesterberg *et al.* 2005).

In contrast to cross-validation sample size is not reduced but some data will be redundant. It is especially helpful to determine empirical confidence intervals (Carpenter & Bithell 2000). Several publications employed bootstrapping for validation (Knickerbocker *et al.* 2007; Lynch *et al.* 2004; Seeber *et al.* 2008). Similarly to double cross-validation, Feng et al. (Feng & Yasui 2004) proposed that cross validation should serve for model selection and bootstrap for estimation of the classification error.

Separate set validation

The ultimate validation is always to reproduce the experiment independently on different patients and within a different lab. However, mainly because of time and funding constraints, it cannot always be done, and one has to rely on previous investigations. For example, Whiteley et al. (Whiteley *et al.* 2008) showed that no publication using panels for the diagnosis of ischemic stroke validated its results on an independent patient cohort. They recommended independent validation as a good practice, also for other work dealing with patient classification. Reddy et al. (Reddy *et al.* 2008) and Gevaert et al. (Gevaert *et al.* 2006) for example rely on an independent cohort for validation.

Statistical method reporting

Proteomics is currently moving towards better reporting requirements, such as minimum information about a proteomics experiment (MIAPE) (Taylor 2006). A similar initiative exists in the medical community with the standards for reporting of diagnostic accuracy (STARD) (Bossuyt *et al.* 2003) that defines a checklist of 25 items to promote a coherent reporting of accuracies. But none of these initiatives fully covers the needs of panels. As a 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 labs, it is very important that the statistical analysis methods are discussed in detail and information about 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 (Hainard *et al.* 2009; Knickerbocker *et al.* 2007; Montaner *et al.* 2008). For cross-validation and bootstrapping, a graph such as that presented by Wild et al. (Wild *et al.* 2008) usually helps the reader understand how the performance test was applied and what the reported results really mean. 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 (Prados *et al.* 2004; Reddy *et al.* 2008; Wu *et al.* 2003). However there is no proper comparison. Reddy et al. (Reddy *et al.* 2008) states that “Logical analysis of data model has significantly better performance on the independent validation set compared to the other classification models.” However, there is no statistics to prove this difference and confidence intervals partially overlap. Prados et al. (Prados *et al.* 2004) used McNemar’s test for pairwise comparison of algorithms. Wu (Wu *et al.* 2003) eludes the problem by studying the stability of the model performance over several cross-validation or bootstrap replicates.

The most important point is that performance estimates should be compared on a dataset independent from the model definition (Hilario *et al.* 2006; LaBaer 2005). This can be done either with an independent validation cohort (separated or split), or performance can be estimated by means of cross-validation or bootstrapping.

ROC curves

Traditionally, performance of a test discriminating between two classes of patients is evaluated using a receiver operating characteristic (ROC) curve (Fawcett 2006). It 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 of the top right zone of the curve. To the contrary 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” one would reach the top left point corresponding to 100% sensitivity and 100% specificity. A major characteristic of a ROC curve is its area under the curve (AUC). The maximum AUC possible is 100%

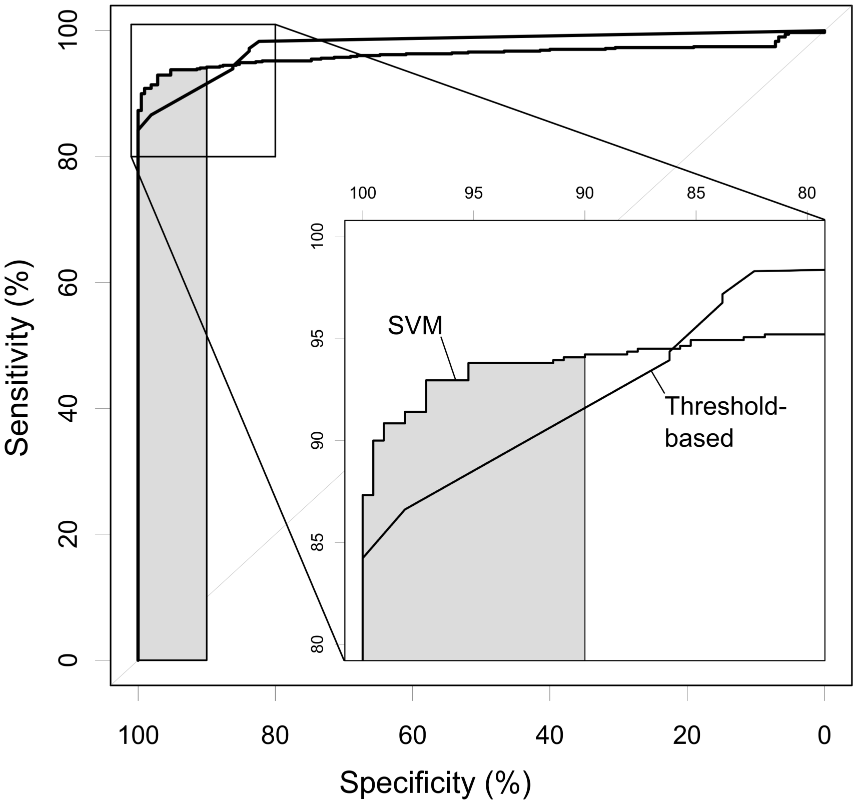

Figure 5: Two ROC curves for Support Vector Machines (SVM) and threshold-based classifiers. Cross-validation partial AUC (pAUC) between 100% and 90% specificity are shown in grey. Respective values of pAUC are 9.3% and 8.8% (perfect classification would correspond to a pAUC of 10%).

corresponding to a "perfect" classification. A non-discriminating ROC curve has an AUC of 50%. In 1989, McClish introduced the concept of partial area under the ROC curve (Dodd & Pepe 2003; McClish 1989; Thompson & Zucchini 1989). It consists in analyzing only a region of special interest of the ROC curve and allows selecting models with high specificity or sensitivity, rather than models with a better average performance, but potentially lower clinical value.

Hanley and McNeil (Hanley & McNeil 1983) and DeLong et al. (DeLong *et al.* 1988) proposed non-parametric 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 (McClish 1990). Baker (Baker 2000) proposed a method to select best thresholds from a multidimensional ROC curve.

An intrinsic property of ROC curves is that AUC of smooth curves tends to be greater than of trapezoidal or step ones (DeLong *et al.* 1988; Hanley & McNeil 1982). 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. (Gu *et al.* 2008) present a smoothing procedure based on bayesian bootstrap estimation.

Another option is to bootstrap and compute confidence intervals and see if the observed sample is compatible with the bootstrap distribution (Carpenter & Bithell 2000; Hesterberg *et al.* 2005). Reddy et al. (Reddy *et al.* 2008) adopted this solution.

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 (Lejon *et al.* 2008; Morais *et al.* 2008; Vasconcelos *et al.* 2006) can calculate the p-value that both classifiers are equally good (Salzberg 1997). Both tests are based on a 2x2 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 to calculate the 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 (Salzberg 1997) overview. 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 (Whiteley *et al.* 2008). Therefore, we recommend measuring all classifier performances with the same cross-validation method or on an independent test set.

Expert commentary

Interest in biomarker panels has been growing for the last few years. A number of publications 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 to understand for medical practitioners. The final choice of a method must be carefully validated.

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

The limited consensus about 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. A special care about 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 already 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 might have a great weight in a panel. Some progress was made towards this goal and with promising results (Gillette *et al.* 2005).

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.

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. SVM 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 done either by using a separate dataset or by means of cross-validation and/or bootstrap. A validation in an independent cohort measured by a different group is eventually required.
- Reporting detailed information about software and parameters set for preprocessing, classification, validation and comparison of methods should be seen as requirements. Reporting standards need to be developed.

Acknowledgements

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

JCS receives a grant from Proteome Sciences plc for brain marker discovery.

References

World Federation of Neurological Surgeons Committee (1988). Report of World Federation of Neurological Surgeons Committee on a Universal Subarachnoid Hemorrhage Grading Scale. *Journal of Neurosurgery* 68 (6), p. 985–986. PMID: 3131498. DOI: 10.3171/jns.1988.68.6.0985.

Apple F. S., Christenson R. H., Valdes R., *et al.* (1999). Simultaneous Rapid Measurement of Whole Blood Myoglobin, Creatine Kinase MB, and Cardiac Troponin I by the Triage Cardiac Panel for Detection of Myocardial Infarction. *Clinical Chemistry* 45 (2), p. 199–205. PMID: 9931041.

Baggerly K. A., Morris J. S., Wang J., *et al.* (2003). A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics* 3 (9), p. 1667–1672. PMID: 12973722. DOI: 10.1002/pmic.200300522.

Baggerly K. A., Morris J. S. and Coombes K. R. (2004). Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 20 (5), p. 777–785. PMID: 14751995. DOI: 10.1093/bioinformatics/btg484.

Baker S. G. (2000). Identifying Combinations of Cancer Markers for Further Study as Triggers of Early Intervention. *Biometrics* 56 (4), p. 1082–1087. PMID: 11129464. DOI: 10.1111/j.0006-341X.2000.01082.x.

Bhaskar H., Hoyle D. C. and Singh S. (2006). Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Computers in Biology and Medicine* 36 (10), p. 1104–1125. PMID: 16226240. DOI: 10.1016/j.compbiomed.2005.09.002.

Bossuyt P. M., Reitsma J. B., Bruns D. E., *et al.* (2003). The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. *Clinical Chemistry* 49 (1), p. 7–18. PMID: 12507954. DOI: 10.1373/49.1.7.

Breiman L. (2001). Random Forests. *Machine Learning* 45 (1), p. 5–32. DOI: 10.1023/A:1010933404324.

Brouns R., Van Den Bossche J., De Surgeloose D., *et al.* (2009). Clinical and biochemical diagnosis of small-vessel disease in acute ischemic stroke. *Journal of the Neurological Sciences* 285 (1-2), p. 185–190. PMID: 19619884. DOI: 10.1016/j.jns.2009.06.032.

Carpenter J. and Bithell J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* 19 (9), p. 1141–1164. PMID: 10797513. DOI: 10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F.

Cortes C. and Vapnik V. (1995). Support-vector networks. *Machine Learning* 20 (3), p. 273–297. DOI: 10.1023/A:1022627411411.

DeLong E. R., DeLong D. M. and Clarke-Pearson D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 44 (3), p. 837–845. PMID: 3203132.

Diamandis E. P. (2004). Analysis of Serum Proteomic Patterns for Early Cancer Diagnosis: Drawing Attention to Potential Problems. *Journal of the National Cancer Institute* 96 (5), p. 353–356. PMID: 14996856. DOI: 10.1093/jnci/djh056.

Dodd L. E. and Pepe M. S. (2003). Partial AUC Estimation and Regression. *Biometrics* 59 (3), p. 614–623. PMID: 14601762. DOI: 10.1111/1541-0420.00071.

Domon B. and Aebersold R. (2006). Mass Spectrometry and Protein Analysis. *Science* 312 (5771), p. 212–217. PMID: 16614208. DOI: 10.1126/science.1124619.

Faca V. M., Song K. S., Wang H., *et al.* (2008). A Mouse to Human Search for Plasma Proteome Changes Associated with Pancreatic Tumor Development. *PLoS Medicine* 5 (6), p. e123. PMID: 18547137. DOI: 10.1371/journal.pmed.0050123.

Fawcett T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27 (8), p. 861–874. DOI: DOI: 10.1016/j.patrec.2005.10.010.

Feng Z. and Yasui Y. (2004). Statistical considerations in combining biomarkers. *Disease Markers* 20 (2), p. 45–51. PMID: 15322313.

Ferguson R. E., Hochstrasser D. F. and Banks R. E. (2007). Impact of preanalytical variables on the analysis of biological fluids in proteomic studies. *Proteomics Clinical Applications* 1 (8), p. 739–746. PMID: 21136730. DOI: 10.1002/prca.200700380.

Gentleman R., Huber W., Carey V., *et al.*, Eds. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Statistics for Biology and Health. New York, Springer-Verlag.

Gevaert O., Smet F. D., Timmerman D., *et al.* (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* 22 (14), p. e184-190. PMID: 16873470. DOI: 10.1093/bioinformatics/btl230.

Gillette M. A., Mani D. R. and Carr S. A. (2005). Place of Pattern in Proteomic Biomarker Discovery. *Journal of Proteome Research* 4 (4), p. 1143–1154. PMID: 16083265. DOI: 10.1021/pr0500962.

Gu J., Ghosal S. and Roy A. (2008). Bayesian bootstrap estimation of ROC curve. *Statistics in Medicine* 27 (26), p. 5407–5420. PMID: 18613217. DOI: 10.1002/sim.3366.

Hackett J. L. and Gutman S. I. (2005). Introduction to the Food and Drug Administration (FDA) Regulatory Process. *Journal of Proteome Research* 4 (4), p. 1110–1113. PMID: 16083260. DOI: 10.1021/pr050059a.

Hainard A., Tiberti N., Robin X, *et al.* (2009). A combined CXCL10, CXCL8 and H-FABP panel for the staging of human African trypanosomiasis patients. *PLoS Neglected Tropical Diseases* 3 (6), p. e459. PMID: 19554086. DOI: 10.1371/journal.pntd.0000459.

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

Hammer P. L. and Bonates T. (2005) Logical Analysis of Data: From Combinatorial Optimization to Medical Applications. RUTCOR Research Report, rutcor.rutgers.edu/pub/rrr/reports2005/10_2005.pdf

Han J. and Kamber M. (2001). *Data mining*, Morgan Kaufmann.

Hanash S. (2003). Disease proteomics. *Nature* 422 (6928), p. 226–232. PMID: 12634796. DOI: 10.1038/nature01514.

Hanash S. M., Pitteri S. J. and Faca V. M. (2008). Mining the plasma proteome for cancer biomarkers. *Nature* 452 (7187), p. 571–579. PMID: 18385731. DOI: 10.1038/nature06916.

Hanley J. A. and McNeil B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143 (1), p. 29–36. PMID: 7063747.

Hanley J. A. and McNeil B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148 (3), p. 839–843. PMID: 6878708.

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

* *Reference book for statistical learning.*

Hesterberg T., Moore D. S., Monaghan S., *et al.* (2005). Bootstrap Methods and Permutation Tests. *Introduction to the Practice of Statistics*, W.H. Freeman & Company; 5th edition.

Hilario M., Kalousis A., Pellegrini C. and Müller M. (2006). Processing and classification of protein mass spectra. *Mass Spectrometry Reviews* 25 (3), p. 409–449. PMID: 16463283. DOI: 10.1002/mas.20072.

Hilario M. and Kalousis A. (2008). Approaches to dimensionality reduction in proteomic biomarker studies. *Briefings in Bioinformatics* 9 (2), p. 102–118. PMID: 18310106. DOI: 10.1093/bib/bbn005.

Hill M. D., Jackowski G., Bayer N., *et al.* (2000). Biochemical markers in acute ischemic stroke. *Canadian Medical Association Journal* 162 (8), p. 1139–1140-1139–1140. PMID: 10789628.

Hoffer A. and Osmond H. (1961). A card sorting test helpful in making psychiatric diagnosis. *Journal of Neuropsychiatry* 2, p. 306–330. PMID: 13714996.

Ishino M., Takeishi Y., Niizeki T., *et al.* (2008). Implications of BNP, H-FABP, and PTX3. *Circulation Journal* 72 (11), p. 1800–1805. PMID: 18832778. DOI: 10.1253/circj.CJ-08-0157.

James G. M. and Hastie T. J. (2001). Functional Linear Discriminant Analysis for Irregularly Sampled Curves. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 63 (3), p. 533–550. DOI: 10.1111/1467-9868.00297.

Knickerbocker T., Chen J. R., Thadhani R. and MacBeath G. (2007). An integrated approach to prognosis using protein microarrays and nonparametric methods. *Molecular Systems Biology* 3 (123), p. 1–8. PMID: 17593911. DOI: 10.1038/msb4100167.

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

LaBaer J. (2005). So, You Want to Look for Biomarkers (Introduction to the Special Biomarkers Issue). *Journal of Proteome Research* 4 (4), p. 1053–1059. PMID: 16083254. DOI: 10.1021/pr0501259.

Laskowitz D. T., Kasner S. E., Saver J., *et al.* (2009). Clinical Usefulness of a Biomarker-Based Diagnostic Test for Acute Stroke: The Biomarker Rapid Assessment in Ischemic Injury (BRAIN) Study. *Stroke* 40 (1), p. 77–85. PMID: 18948614. DOI: 10.1161/STROKEAHA.108.516377.

Lejon V., Roger I., Mumba Ngoyi D., *et al.* (2008). Novel Markers for Treatment Outcome in Late-Stage Trypanosoma brucei gambiense Trypanosomiasis. *Clinical Infectious Diseases|Clin. Infect. Dis.* 47 (1), p. 15–22. PMID: 18494605. DOI: 10.1086/588668.

Linkov F., Lisovich A., Yurkovetsky Z., *et al.* (2007). Early Detection of Head and Neck Cancer: Development of a Novel Screening Tool Using Multiplexed Immunobead-Based Biomarker Profiling. *Cancer Epidemiology Biomarkers & Prevention* 16 (1), p. 102–107. PMID: 17220337. DOI: 10.1158/1055-9965.EPI-06-0602.

Little R. R., Rohlfing C. L., Tennill A. L., *et al.* (2008). Standardization of C-peptide measurements. *Clinical Chemistry* 54 (6), p. 1023–1026. PMID: 18420730. DOI: 10.1373/clinchem.2007.101287.

Liu J. J., Cutler G., Li W., *et al.* (2005). Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics* 21 (11), p. 2691–2697. PMID: 15814557. DOI: 10.1093/bioinformatics/bti419.

Lynch J. R., Blessing R., White W. D., *et al.* (2004). Novel Diagnostic Test for Acute Stroke. *Stroke* 35 (1), p. 57–63. PMID: 14671250. DOI: 10.1161/01.STR.0000105927.62344.4C.

McClish D. K. (1989). Analyzing a Portion of the ROC Curve. *Medical Decision Making* 9 (3), p. 190–195. PMID: 2668680. DOI: 10.1177/0272989X8900900307.

McClish D. K. (1990). Determining a Range of False-positive Rates for Which ROC Curves Differ. *Medical Decision Making* 10 (4), p. 283–287. PMID: 2233158. DOI: 10.1177/0272989X9001000406.

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

Montaner J., Perea-Gainza M., Delgado P., *et al.* (2008). Etiologic Diagnosis of Ischemic Stroke Subtypes With Plasma Biomarkers. *Stroke* 39 (8), p. 2280–2287. PMID: 18535284. DOI: 10.1161/STROKEAHA.107.505354.

Morais D. F., Spotti A. R., Tognola W. A., *et al.* (2008). Clinical application of magnetic resonance in acute traumatic brain injury. *Arquivos de Neuro-Psiquiatria* 66 (1), p. 53–58. PMID: 18392415. DOI: 10.1590/S0004-282X2008000100013.

Oberg A. L. and Vitek O. (2009). Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments. *Journal of Proteome Research* 8 (5), p. 2144–2156. PMID: 19222236. DOI: 10.1021/pr8010099.

Panchaud A., Affolter M., Moreillon P. and Kussmann M. (2008). Experimental and computational approaches to quantitative proteomics: Status quo and outlook. *Journal of Proteomics* 71 (1), p. 19–33. PMID: 18541471. DOI: 10.1016/j.jprot.2007.12.001.

Patz E. F., Campa M. J., Gottlin E. B., *et al.* (2007). Panel of Serum Biomarkers for the Diagnosis of Lung Cancer. *Journal of Clinical Oncology* 25 (35), p. 5578–5583. PMID: 18065730. DOI: 10.1200/JCO.2007.13.5392.

Pepe M. S., Janes H., Longton G., *et al.* (2004). Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker. *American Journal of Epidemiology* 159 (9), p. 882–890. PMID: 15105181. DOI: 10.1093/aje/kwh101.

Petricoin E. F., Ardekani A. M., Hitt B. A., *et al.* (2002). Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* 359 (9306), p. 572–577. PMID: 11867112. DOI: 10.1016/S0140-6736(02)07746-2.

Petricoin E. F. and Liotta L. A. (2004). SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. *Current Opinion in Biotechnology* 15 (1), p. 24–30. PMID: 15102462. DOI: 10.1016/j.copbio.2004.01.005.

Prados J., Kalousis A., Sanchez J.-C., *et al.* (2004). Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics* 4 (8), p. 2320–2332. PMID: 15274126. DOI: 10.1002/pmic.200400857.

Rai A. J. and Vitzthum F. (2006). Effects of preanalytical variables on peptide and protein measurements in human serum and plasma: implications for clinical proteomics. *Expert Review of Proteomics* 3 (4), p. 409–426. PMID: 16901200. DOI: 10.1586/14789450.3.4.409.

Ralhan R., DeSouza L. V., Matta A., *et al.* (2008). Discovery and Verification of Head-and-neck Cancer Biomarkers by Differential Protein Expression Analysis Using iTRAQ Labeling, Multidimensional Liquid Chromatography, and Tandem Mass Spectrometry. *Molecular & Cellular Proteomics* 7 (6), p. 1162–1173. PMID: 18339795. DOI: 10.1074/mcp.M700500-MCP200 .

Reddy A., Wang H., Yu H., *et al.* (2008). Logical Analysis of Data (LAD) model for the early diagnosis of acute ischemic stroke. *BMC Medical Informatics and Decision Making* 8, p. 30–30. PMID: 18616825. DOI: 10.1186/1472-6947-8-30.

> ⊛ *Comparison of five methods (LAD, SVM, Decision tree, Logistic regression and Multilayer perceptron) on SELDI data. Exemplary validation with 10-fold cross-validation and an independent validation set.*

Ressom H. W., Varghese R. S., Drake S. K., *et al.* (2007). Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics* 23 (5), p. 619–626. PMID: 17237065. DOI: 10.1093/bioinformatics/btl678.

Reynolds M. A., Kirchick H. J., Dahlen J. R., *et al.* (2003). Early Biomarkers of Stroke. *Clinical Chemistry* 49 (10), p. 1733–1739. PMID: 14500614. DOI: 10.1373/49.10.1733.

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

Ring B. Z., Seitz R. S., Beck R., *et al.* (2006). Novel Prognostic Immunohistochemical Biomarker Panel for Estrogen Receptor-Positive Breast Cancer. *Journal of Clinical Oncology* 24 (19), p. 3039–3047. PMID: 16809728. DOI: 10.1200/JCO.2006.05.6564.

Rosen R. C., Cappelleri J. C., Smith M. D., *et al.* (1999). 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. *International Journal of Impotence Research* 11 (6), p. 319–326. PMID: 10637462.

Rosengart A. J., Schultheiss K. E., Tolentino J. and Macdonald R. L. (2007). Prognostic Factors for Outcome in Patients With Aneurysmal Subarachnoid Hemorrhage. *Stroke* 38 (8), p. 2315–2321. PMID: 17569871. DOI: 10.1161/STROKEAHA.107.484360.

Ross D. T., Kim C.-Y., Tang G., *et al.* (2008). Chemosensitivity and stratification by a five monoclonal antibody immunohistochemistry test in the NSABP B14 and B20 trials. *Clinical Cancer Research* 14 (20), p. 6602–6609. PMID: 18927301. DOI: 10.1158/1078-0432.CCR-08-0647.

Saeys Y., Inza I. and Larranaga P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23 (19), p. 2507–2517. PMID: 17720704. DOI: 10.1093/bioinformatics/btm344.

Salzberg S. (1997). On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. *Data Mining and Knowledge Discovery* 1 (3), p. 317–328. DOI: 10.1023/A:1009752403260.

Schramm A., Schulte J. H., Klein-Hitpass L., *et al.* (2005). Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling. *Oncogene* 24 (53), p. 7902–7912. PMID: 16103881. DOI: 10.1038/sj.onc.1208936.

Seeber B., Sammel M. D., Fan X., *et al.* (2008). Panel of markers can accurately predict endometriosis in a subset of patients. *Fertility and Sterility* 89 (5), p. 1073–1081. PMID: 17706208. DOI: 10.1016/j.fertnstert.2007.05.014.

Sibon I., Rouanet F., Meissner W. and Orgogozo J. M. (2009). Use of the Triage Stroke Panel in a neurologic emergency service. *The American Journal of Emergency Medicine* 27 (5), p. 558–562. PMID: 19497461. DOI: 10.1016/j.ajem.2008.05.001.

Smit S., van Breemen M. J., Hoefsloot H. C. J., *et al.* (2007). Assessing the statistical validity of proteomics based biomarkers. *Analytica Chimica Acta* 592 (2), p. 210–217. PMID: 17512828. DOI: 10.1016/j.aca.2007.04.043.

Stead D. A., Paton N. W., Missier P., *et al.* (2008). Information quality in proteomics. *Briefings in Bioinformatics* 9 (2), p. 174–188. PMID: 18281347. DOI: 10.1093/bib/bbn004.

Steel L. F., Haab B. B. and Hanash S. M. (2005). Methods of comparative proteomic profiling for disease diagnostics. *Journal of Chromatography B* 815 (1-2), p. 275–284. PMID: 15652816. DOI: 10.1016/j.jchromb.2004.10.072.

Taylor C. F. (2006). Minimum Reporting Requirements for Proteomics: A MIAPE Primer. *Proteomics* 6 (S2), p. 39–44. PMID: 17031795. DOI: 10.1002/pmic.200600549.

Thompson M. L. and Zucchini W. (1989). On the statistical analysis of ROC curves. *Statistics in Medicine* 8 (10), p. 1277–1290. PMID: 2814075. DOI: 10.1002/sim.4780081011.

Turck N., Vutskits L., Sanchez-Pena P., *et al.* A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage. *Intensive Care Medicine* 36 (1), p. 107–115. PMID: 19760205. DOI: 10.1007/s00134-009-1641-y.

Vanni S., Polidori G., Pepe G., *et al.* (2009). Use of Biomarkers in Triage of Patients with Suspected Stroke. *Journal of Emergency Medicine* Epub ahead of print. PMID: 19217237. DOI: 10.1016/j.jemermed.2008.09.028.

Vasconcelos O. M., Jr., Prokhorenko O. A., Kelley K. F., *et al.* (2006). A comparison of fatigue scales in postpoliomyelitis syndrome. *Archives of Physical Medicine and Rehabilitation* 87 (9), p. 1213–1217. PMID: 16935057. DOI: 10.1016/j.apmr.2006.06.009.

Visintin I., Feng Z., Longton G., *et al.* (2008). Diagnostic Markers for Early Detection of Ovarian Cancer. *Clinical Cancer Research* 14 (4), p. 1065–1072. PMID: 18258665. DOI: 10.1158/1078-0432.CCR-07-1569.

> ⊛ *Several Logistic regression models fitted with a very clear validation. Data was acquired in a multiplex bead assay.*

Vitzthum F., Behrens F., Anderson N. L. and Shaw J. H. (2005). Proteomics: from basic research to diagnostic application. A review of requirements & needs. *Journal of Proteome Research* 4 (4), p. 1086–1097. PMID: 16083257. DOI: 10.1021/pr050080b.

Wang P., Kim Y., Pollack J. and Tibshirani R. (2004). Boosted PRIM with Application to Searching for Oncogenic Pathway of Lung Cancer. *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, p. 604–609. DOI: 10.1109/CSB.2004.1332514.

Webb G. I., Boughton J. R. and Wang Z. (2005). Not So Naive Bayes: Aggregating One-Dependence Estimators. *Machine Learning* 58 (1), p. 5–24. DOI: 10.1007/s10994-005-4258-6.

Welsh P., Barber M., Langhorne P., *et al.* (2009). Associations of inflammatory and haemostatic biomarkers with poor outcome in acute ischaemic stroke. *Cerebrovascular Diseases* 27 (3), p. 247–253. PMID: 19176958. DOI: 10.1159/000196823.

Whiteley W., Tseng M.-C. and Sandercock P. (2008). Blood Biomarkers in the Diagnosis of Ischemic Stroke: A Systematic Review. *Stroke* 39 (10), p. 2902–2909. PMID: 18658039. DOI: 10.1161/STROKEAHA.107.511261.

Wicki J., Perneger T. V., Junod A. F., *et al.* (2001). Assessing Clinical Probability of Pulmonary Embolism in the Emergency Ward: A Simple Score. *Archives of Internal Medicine* 161 (1), p. 92–97. PMID: 11146703.

Wild N., Karl J., Grunert V. P., *et al.* (2008). Diagnosis of rheumatoid arthritis: multivariate analysis of biomarkers. *Biomarkers* 13 (1), p. 88–105. PMID: 18188726.

Woolas R. P., Xu F.-J., Jacabs I. J., *et al.* (1993). Elevation of Multiple Serum Markers in Patients With Stage I Ovarian Cancer. *Journal of the National Cancer Institute* 85 (21), p. 1748–1751. PMID: 8411259. DOI: 10.1093/jnci/85.21.1748.

Wu B., Abbott T., Fishman D., *et al.* (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 19 (13), p. 1636–1643. PMID: 12967959. DOI: 10.1093/bioinformatics/btg210.

Yeo I.-K. and Johnson R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* 87 (4), p. 954–959. DOI: 10.1093/biomet/87.4.954.

Zervakis M., Blazadonakis M. E., Tsiliki G., *et al.* (2009). Outcome prediction based on microarray analysis: a critical perspective on methods. *BMC Bioinformatics* 10, p. 53. PMID: 19200394. DOI: 10.1186/1471-2105-10-53.

Zheng Y., Katsaros D., Shan S. J. C., *et al.* (2007). A Multiparametric Panel for Ovarian Cancer Diagnosis, Prognosis, and Response to Chemotherapy. *Clinical Cancer Research* 13 (23), p. 6984–6992. PMID: 18056174. DOI: 10.1158/1078-0432.CCR-07-1409.