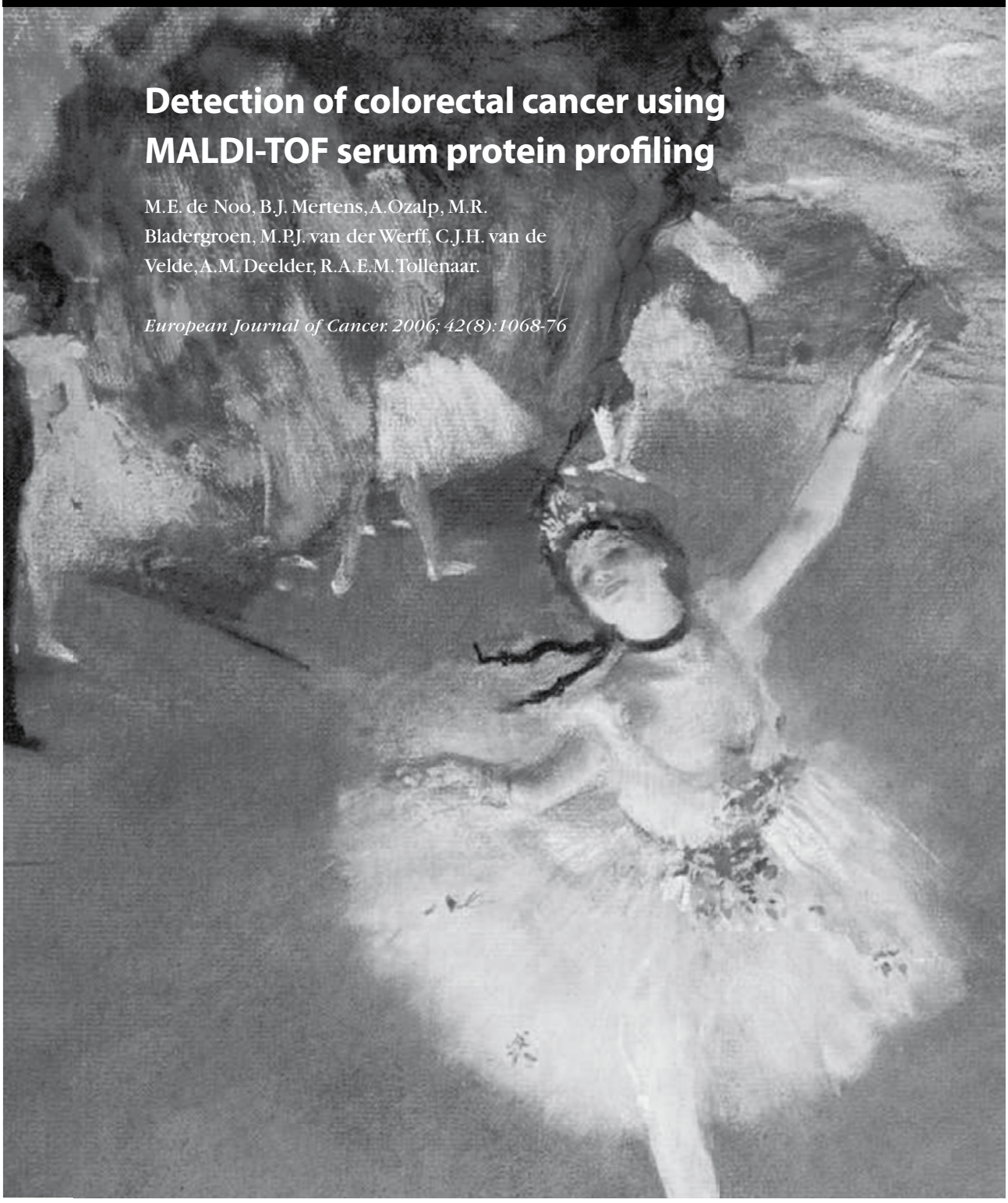
# Clinical proteomics in oncology : a passionate dance between science and clinic

Noo, M.E. de

# Chapter 4



## Detection of colorectal cancer using MALDI-TOF serum protein profiling

M.E. de Noo, B.J. Mertens, A. Ozalp, M.R. Bladergroen, M.P.J. van der Werff, C.J.H. van de Velde, A.M. Deelder, R.A.E.M. Tollenaar.

*European Journal of Cancer. 2006; 42(8):1068-76*

## ABSTRACT

*Purpose*

Serum protein profiling is a promising approach for classification of cancer versus non-cancer samples. The objective of our study was to assess the feasibility of mass spectrometry based protein profiling for the discrimination of colorectal cancer patients from healthy individuals.

*Experimental design*

In a randomised block design pre-operative serum samples obtained from 66 colorectal cancer patients and 50 controls, were used to generate MALDI-TOF protein profiles. After pre-processing of the spectra, linear discriminant analysis with double cross-validation was used to classify the protein profiles.

*Results*

A total recognition rate of 92.6%, a sensitivity of 95.2% and a specificity of 90.0% for the detection of CRC were shown. The area under the curve of the classifier was 97.3%, which demonstrates the high, significant separation power of the classifier.

*Conclusions*

Double cross-validation shows that classification can be attributed to information in the protein profile. Although preliminary, the high sensitivity and specificity indicate the potential usefulness of serum protein profiles for the detection of colorectal cancer.

## INTRODUCTION

Colorectal cancer (CRC) is among the most common malignancies and remains a leading cause of cancer-related morbidity and mortality. It is well recognised that CRC arises from a multistep sequence of genetic alterations that result in the transformation of normal mucosa to a precursor adenoma and ultimately to carcinoma. Given the natural history of CRC, early diagnosis appears to be the most appropriate tool to reduce disease-related mortality.[1;2] Currently, there is no early diagnostic test with high sensitivity, specificity and positive predictive value, which can be used as a routine screening tool. Therefore, there is a need for new biomarkers for colorectal cancer that can improve early diagnosis, monitoring of disease progression and therapeutic response and detect disease recurrence. Furthermore, these markers may give indications for targets for novel therapeutic strategies.

Proteomic expression profiles generated with mass spectrometry have been suggested as potential tools for the early diagnosis of cancer and other diseases. Different protein profiles may be associated with varying responses to therapeutics. It has been postulated that on the basis of the presence/absence of multiple low-molecular-weight serum proteins using time-of-flight (TOF) mass spectrometry technologies, such as SELDI-TOF and MALDI-TOF, biomarkers can be identified.[3-6] Although the data from these studies are encouraging, critical notes have been made on both study design and experimental procedures for proteomic profiling.[7-9] In addition, the importance of avoiding confounding biological variables, as well as technological factors that may bias the results, have previously been stressed by several authors.[10;11] Another recurrent topic for debate is the use of independent validation sets for the classification of diseased versus healthy individuals. A specific problem in the discovery-based research field of clinical proteomics is overfitting. Overfitting may occur in the analysis of large datasets when multivariate models show apparent discrimination that is actually caused by data over-interpretation, and hence give rise to results that are not reproducible.[9;12;13] The chance of overfitting, however, can be reduced by appropriate application of validatory estimation and assessment, such as through application of double cross-validation, when properly implemented.

The objective of this study was to assess the feasibility of mass spectrometry based protein profiling for the discrimination of colorectal cancer patients from healthy individuals. In addition to standardizing technical factors and biological variations, we performed blinded tests and employed a randomised block design experimentation to minimize impact of potential confounding factors and to avoid bias. To minimize danger of overfitting, among other reasons, we used a fairly inflexible classification method based on first-and-second order statistics only. Specifically, Fisher linear discriminant analysis was employed with double cross-validatory integrated estima-

tion and validation of error rate on the entire dataset to calculate an unbiased error rate assessment.

## MATERIAL AND METHODS

### Subjects

Serum samples were obtained from a total of 66 colorectal cancer patients one day before surgery. All patients with stage IV disease had synchronous metastatic disease confined to the liver. Colorectal cancer was histological confirmed on surgical specimens and preoperatively assessed with abdominal CT scan and carcinoembryonic antigen (CEA) levels. The extent of tumour spread was assessed by TNM classification based on histological examination of the resected specimen. All stages of colorectal cancer were represented in the patient group. The median age of the patient group was 62.8 years (range 32.6-90.3) and the male to female ratio was 31/35. Patients were included from October 2002 till December 2004 in our Center. The control group consisted of 50 healthy volunteers. The median age of the healthy symptom-free control group was 49.7 years (range 25.9-76.6) and the male to female ratio was 21/29. The controls were included from October till December 2004 (Table 1).



**Table** 1. Patient characteristics.

| | CRC patients | | Controls |
|---|---|---|---|
| | inclusion | results | |
| n = | 66 | 63 | 50 |
| Age (mean) | 62.6 | 62.2 | 49.7 |
| Age (range) | (32.6-90.3) | (32.6-90.3) | (25.9-76.6) |
| Male/female ratio | 34/32 | 31/32 | 21/29 |

### Study design

Having identified plate-to-plate and day-today variation as important potential batch effects, we used a randomised blocked design.[14;15] All the available 116 samples from both groups (controls and colorectal cancer) were randomly distributed across 3 plates in roughly equal proportions (Table 2). For colon cancer, the distribution of stadia across plates was again in random fashion and in approximately equal proportions (Table 3). The position on the plates of samples allocated to each plate was randomised as well. Each plate was then assigned to a distinct day, which completes the design. Analysis was carried out on 3 consecutive days, Tuesday to Thursday,

processing a single plate each day. A duplicate of this randomised blocked study was performed in the following week.

**Table** 2. Distribution and randomisation of serum samples of colorectal cancer patients with different TNM stage before and after the MALDI-TOF experiment. The distribution of stadia across plates was performed randomly and in approximately equal proportions.

| | Plate 1 | Plate 2 | Plate 3 | Total |
|---|---|---|---|---|
| Colorectal cancer | 22 | 22 | 19 | 63 |
| Controls | 17 | 17 | 16 | 50 |
| Total | 39 | 39 | 35 | 113 |

**Table** 3. Distribution and randomisation of serum samples of different groups over the three MS target plates.

| | TNM stage | Plate 1 | Plate 2 | Plate 3 |
|---|---|---|---|---|
| Inclusion | I | 4 | 4 | 3 |
| | II | 10 | 10 | 8 |
| | III | 4 | 4 | 4 |
| | IV | 4 | 4 | 4 |
| | 0 | 4 | 3 | 3 |
| | Total | 26 | 25 | 22 |
| Exclusion | I | 0 | 0 | 1 |
| | II | 0 | 0 | 1 |
| | III | 0 | 0 | 1 |
| | IV | 0 | 0 | 0 |
| | 0 | 4 | 3 | 3 |
| | Total | 4 | 3 | 6 |

### Serum samples

Informed consent was obtained from all patients and the Medical Ethical Committee approved the study. All blood samples were drawn while the patients or healthy controls were seated and non-fasting. The samples were collected in a 10 cc Serum Separator Vacutainer Tube and centrifuged 30 min later at 3000 rpm for 10 minutes. The serum samples were distributed into 1 ml aliquots and stored at -70 ºC until the experiment.[16]

### Isolation of peptides

The isolation of peptides from serum was performed using the magnetic beads, based hydrophobic interaction chromatography (MB-HIC) kit from Bruker, mainly according to the manufacturers instructions, adapted for automation on an 8-channel Hamilton STAR® pipetting robot (Hamilton, Martinsried, Germany). Magnetic beads with C8- functionality (MB-HIC8) were divided in 5-µl aliquots in a 96-well microtiter plate, which was placed on the magnetic beads separation device (MPC®-auto96, Dynal, Oslo, Norway), with the magnet down. Ten µl MB-HIC binding solution and 5-µl serum sample were added to the beads and carefully mixed using the mixing feature of the robot. The sample was incubated for 30 sec and the magnet was lifted, followed by a 30 sec waiting interval to settle the magnetic beads. The supernatant was removed and the magnet was lowered again. The magnetic beads were washed three times with MB-HIC washing solution (also provided with the kit) lifting and lowering the magnet as needed. The peptides were eluted from the beads using 10-µl 50% acetonitrile and 2-µl of this eluate was transferred to a fresh 384-well microtiter plate (Greiner). Most of the remaining eluate (6-µl) was transferred to an auto sampler vial containing 54-µl water and stored for later use. 15-µl $\alpha$-cyano-4-hydroxycinnamic acid (0.3 mg/l in ethanol: acetone 2:1) was added to the 1-µl eluate in the 384-well microtiter plate and mixed carefully. 1-µl of this mixture was spotted in quadruplicate on a MALDI AnchorChip™ (Bruker Daltonics, Bremen, Germany).

### Protein profiling

Matrix Assisted Laser Desorption Ionisation Time-Of-Flight (MALDI-TOF) mass spectrometry measurements were performed using an Ultraflex TOF/TOF instrument (Bruker Daltonics, Bremen, Germany) equipped with a SCOUT ion source, operating in linear mode. Ions formed with a N2 pulse laser beam (337 nm) were accelerated to 25 kV. With this specific serum preparation peptide/protein peaks in the m/z range of 960 to 11,169 Dalton were measured. An independent mass spectrometer operator performed the experiments at 3 consecutive days after cleaning of the instrument. One week later the experiment was duplicated in exactly same order. Hereafter the entire process of capturing and concentrating serum proteins using C8 magnetic beads including the generation of readouts of the MALDI-TOF spectra will be designated as the protein profiling procedure.

### Data processing

All unprocessed spectra were exported from the Ultraflex in standard 8-bit binary ASCII format. They consisted of approximately 45,000 mass-to-charge ratio (m/z) values, covering a domain of 1160-11,600 Dalton. To increase robustness, the average of four spots was used to represent one serum sample. Subsequently, we lightly

smoothed the spectra using the Whittaker[17] smoother. Due to the quadratic nature of the TOF-equation, the high-resolution spectra were binned using a linear scaling at the time scale, resulting in bin widths of approximately 1 Dalton at the beginning of the spectrum and 3 Dalton at the end at the mass/charge scale. The resulting spectra generally showed strong baseline effects. These were removed using an asymmetric least squares algorithm. To normalize the spectra, we calculated the median intensity of every spectrum and subtracted it from the original spectrum. Each of the thus normalised spectra was then also divided by the interquartile range of intensity within that spectrum. We consider this more robust than normalization of the spectra on the average, as it is less sensitive to the most extreme intensities. Finally, prior to classification and evaluation of error rate, the logarithm was taken of all intensity measurements (predominantly to ensure numerical stability of computations).

### Statistical data-analysis

Fully validated classification error rates were estimated based on a classical Fisher linear discriminant analysis through complete double cross-validatory joint estimation and assessment of class predictions, as is further explained in appendix 1.[18-20] Instead of ordinary leave-one-out cross-validatory choice of $k$, we employ double cross-validation. This is an extension of leave-one-out cross-validation which combines validatory 'choice of model' (the parameter $k$ in this case) with 'predictive assessment' (of the same model, through use of error rate or other suitable summary statistic). The reason for this additional "technical complication" is that we do not wish to incur the bias inherent in the assessment, which would normally result from a model choice based on ordinary leave-one-out validation only. In a double cross-validatory evaluation, we remove each individual in turn from the data (just as in ordinary leave-one-out cross-validation), after which the discriminant rule is fully recalibrated and optimised for prediction on the leftover data (now of size $n-1$, where $n$ is the total initial sample size) and using the same procedure in each case. The choice of the calibration rule (i.e. choice of $k$ in this case) to classify the left-out observation is then again based on a leave-one-out cross-validatory estimation (hence the name 'double-cross') within the leftover set of size $n-1$. The resulting classification rule is then applied to the left-out datum to obtain an unbiased allocation for this sample. This procedure is then repeated across all individuals and for each person separately, after which we can calculate a truly unbiased estimate of the misclassification rates on the basis of the thus validated (and calibrated) classifications. In other words, 'double-cross' is actually 'leave-one-out cross-validation within leave-one-out cross-validation' and it is precisely because of this that we can avoid bias in error rate estimation that an ordinary application of standard leave-one-out choice would imply.

## RESULTS

In the first week three different randomised target plates were successfully measured on three consecutive days in the middle of the week. A duplicate experiment

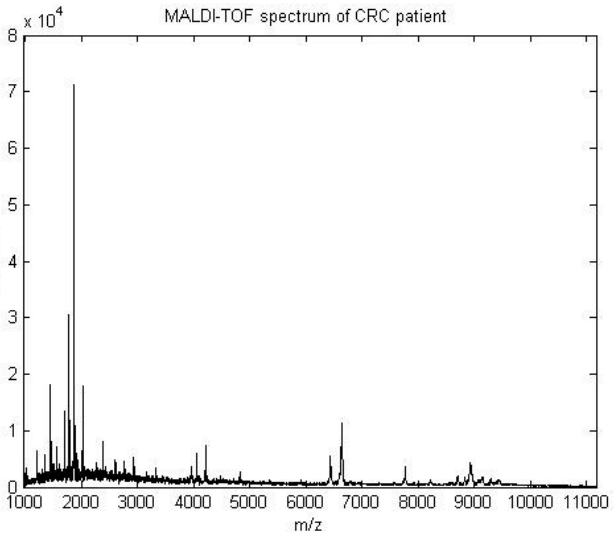



**Figure** 1a

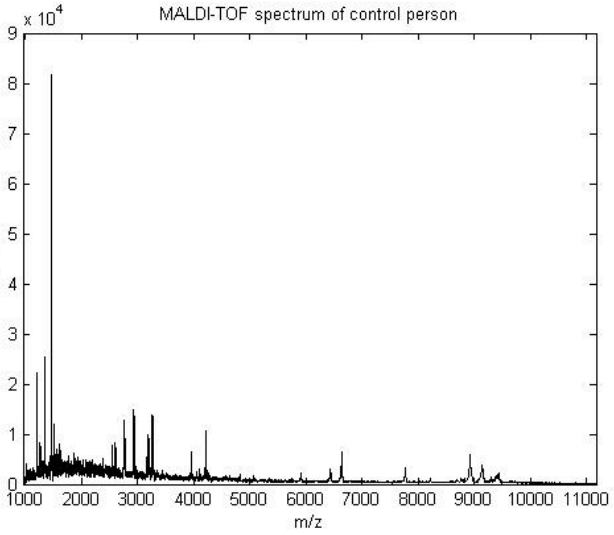



**Figure** 1b

**Figure** 1. MALDI-TOF spectrum of a colorectal cancer patient (1a) and a healthy subject (1b) after peptide isolation with C8 magnetic beads. On the Y-axis the relative intensity is shown. The mass to charge ration (m/z) is demonstrated on the X-axis in Dalton.

was performed in the second week on the same days. Figure 1 shows a raw data spectrum, directly obtained from the MALDI-TOF mass spectrometer. Before pre-processing and further analysis a mean spectrum of each sample was calculated over all four spots that were measured for each sample. In case all four spots from one sample showed spectra of poor quality due to a technical problem, the sample was left out of the analysis. This was the case for 3 CRC patients' samples. The above-described pre-processing steps resulted in a sequence of 4483 normalised m/z values ranging from 1160 to 11,600 Dalton, for each individual.

### Detection of colorectal cancer

Double cross-validatory analysis and evaluation carried out on the protein spectra measured in week 1, correctly classified 45 of the 50 controls as not cancer. Sixty of the 63 cancer samples were correctly classified as malignant, including 9 of 10 TNM stage I patients (Table 4). The remaining 2 misclassified patients had stage II disease. All patients with stage III and IV disease were correctly recognised as malignant within the double cross-validatory evaluation. These validated results thus yield a total recognition rate of 92.6%, a sensitivity of 95.2% and a specificity of 90.0% for the detection of CRC (Table 5). To analyze the actual discriminative power of the classifier, we produced an ROC-curve (again based on the double cross-validatory classification probabilities), visualizing the performance of the two-class classifier in figure 2. The AUC of the classifier was 97.6%.

**Table** 4. Double cross-validatory classification of serum samples. A positive test results assigns subjects to the CRC group and a negative to the controls. In the horizontal plane the actual histologically confirmed diagnosis is stated.

| | Test results for detection of CRC | | |
|---|---|---|---|
| | Neg | Pos | Total |
| Controls | 45 | 5 | 50 |
| CRC patients | 3 | 60 | 63 |
| | 48 | 65 | 113 |

**Table** 5. Cross-validated classification results for the detection of CRC. TRR is the total recognition rate; Sens and Spec are sensitivity and specificity respectively. AUC is the estimated area under the ROC curve.

| Method | First week | | | | Second week | | | |
|---|---|---|---|---|---|---|---|---|
| | TRR | Sens | Spec | AUC | TRR | Sens | Spec | AUC |
| PCA selection | 92.6 | 95.2 | 90.0 | 97.3 | 88.8 | 80.6 | 97.1 | 96.8 |

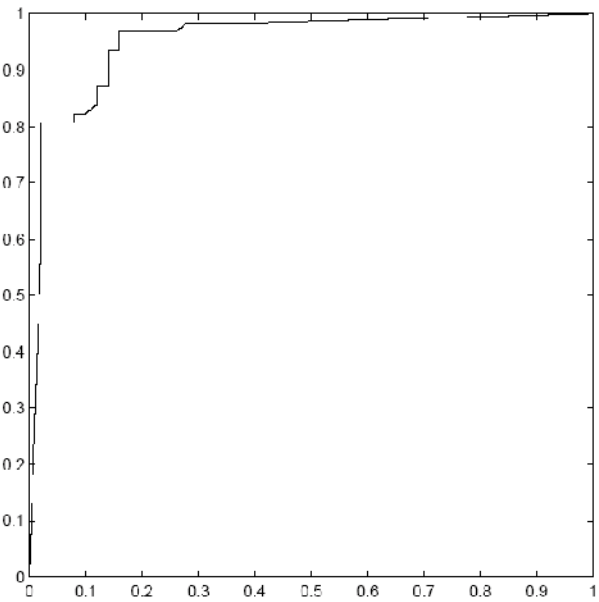



**Figure** 2. ROC-curve for the double cross-validated two-group classifier. The true positive recognition rate (sensitivity) is demonstrated on the y-axis against the false negative recognition rate (1-specificity) on the x-axis of the classifier.

We repeated the entire double cross-validatory evaluation executed with the week 1 data using the duplicate measured spectra from week 2. This procedure was identical to that carried out in week 1 and used the same calibration spectra. However, prior to classifying each left-out datum in the outer "shell" of the double cross-validatory procedure, we substituted the week 1 data with the corresponding measured spectra from the same sample in week 2. In this manner, we could calculate a double cross-validatory error rate, which takes the effect of replicate measurement of the spectrum (and thus also recalibration of the equipment) into account. The effect of classifying the remaining replicate data was that the recognition rate dropped to 88.8%. The sensitivity and specificity for the detection of CRC for the second week data was 80.6% and 97.1% respectively (Table 4). The associated AUC of this repeat double cross-validatory estimation on week 2 was 96.8%.

It is of interest to evaluate bias of the double cross-validatory calculations. Hence, we performed a permutation exercise, which randomly permutes and reassigns the class labels across subjects and then repeats the entire double cross-validation procedure. Carrying out this procedure more than 600 times resulted in a median recognition rate of 50.0% (95% confidence interval is [36.3, 72.7]). The median AUC was 49.4% with confidence interval of [24.8, 64.2]. As both median recognition rates

and AUC's equal 50%, there is thus no substantial evidence of bias remaining within the cross-validatory calculation.

Having executed the above-described validatory evaluation, we can explore the nature of the classification through a post hoc analysis. We found that the first two principal components provide most of the between-group separation. Figure 3 shows a plot of the correlation coefficients, with the class indicator, which can be calculated from the linear discriminant weightings in the region between 1160 and 11,600 Dalton.[20;21] The remainder of the plot is not shown, as the coefficients are effectively zero in that range. As can be seen, the classification is achieved primarily through a contrast in peak intensities between the first and second principal component. This can also be seen from the scatter plot shown in figure 4: low intensities at the first peak for cases separates cases from controls. Likewise, a small contribution for controls at the second peak separates controls from patients. To illustrate these results further, we can simply calculate the contrast between the two peak intensities directly across all subjects and construct a simple one-dimensional summary of the data, as shown in the histogram displayed in figure 5, which shows overlapping histograms of this (ad hoc) contrast for each group separately. The separation is clearly visible. We may also quantify the significance of this difference by performing a two-sample Student t-test on this contrast, which is $t=14.0$ ($p<0.0001$).

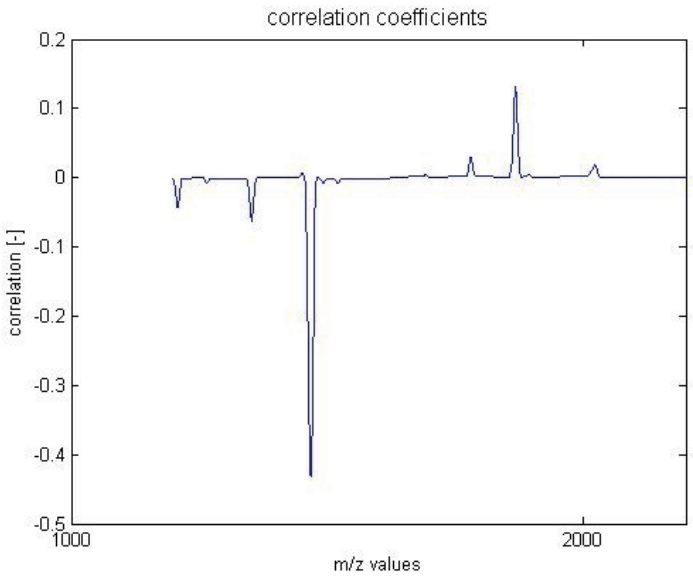



**Figure** 3. Correlation coefficients of two first principal components with the class indicator. The correlation coefficients were calculated from the linear discriminant weightings. The negative correlation of the first peak is an indicator for the control group and the positive correlation of the second peak points out the cases.

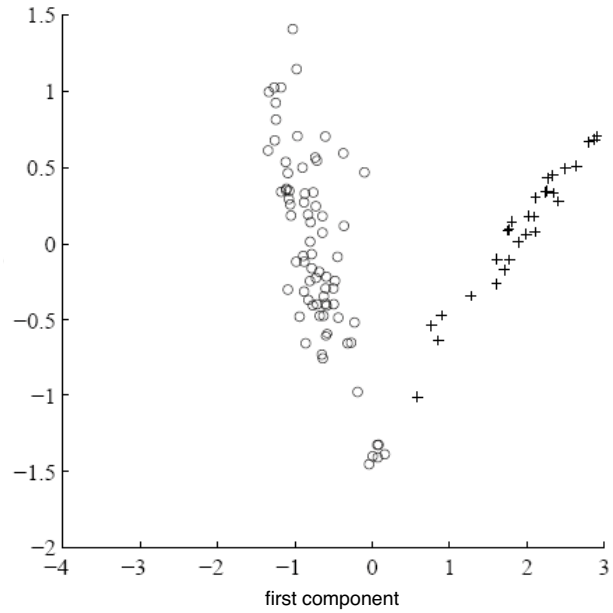



**Figure** 4. Scatter plot of the first two principle components on basis of which the classification patient-control group was made.

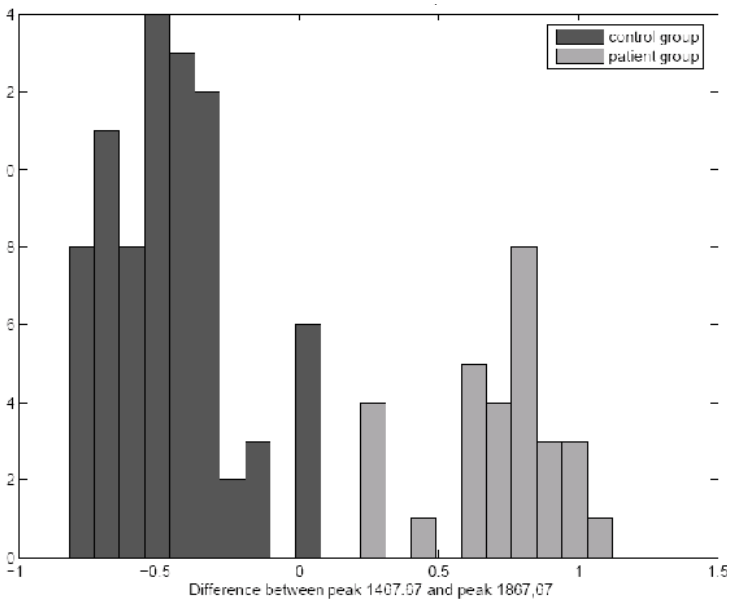



**Figure** 5. Histogram showing the difference between the normalized intensities of the two most discriminating "peaks" (bins). The X-axis shows the difference between the normalized intensities of the peaks. On the Y-axis the number of subjects is displayed.

## DISCUSSION

Our study supports the hypothesis that serum protein profiles can discriminate a normal from a malignant state of organs, in our case of the colon. Here we show that, based upon information in MALDI-TOF serum spectra, a classifier could be constructed for the detection of CRC. This classifier, calibrated and validated on spectra of week one demonstrated a sensitivity and specificity of 95.2% and 90.0% respectively. Thirty-four patients out of thirty-seven with early stage disease (stage 1 and 2) and all patients with stage 3 or 4 disease were correctly classified as having cancer. For the misclassified control subjects it was not possible to retrieve the current physical state as it concerned anonymous healthy controls.

Sensitivity and specificity of 80.6% and 97.1% respectively was achieved when the entire double cross-validatory evaluation was repeated for the data of week 2. The latter evaluation, through use of replicate measurements within the double cross-validation, is likely to provide the more realistic assessment of true error rates and appears to better represent possible diagnostic potential as will be discussed further in this paper.

Although previous studies have reported similar high classification results for various solid tumours, we prefer evaluation through a thorough study design and double cross-validation of classification as proposed in this study.[3-6;12;22;23] As a great variety of different discriminating peaks for the same malignancy have been described,[3;4;24] caution with proteomic data has been stressed before.[7;8] The discrepancies in discriminating protein profiles, found by different research groups, lead to serious concerns regarding biological variations and technological reproducibility issues. Therefore, we used a standardised and well-documented sample collection and a thorough study design, matching biological variables and pre-analytical conditions.[16] Still, patient samples from all stages of CRC were equally distributed over the different target plates, as was the male/female ratio between the two groups, excluding these factors as a discriminator in the detection classifier. Unfortunately there was significant difference in age; the control group being younger than the CRC patients. Ideally, the control group should consist of age-matched symptom-free individuals undergoing a colonoscopy showing no aberrations. However, due to the nature of the intervention, ethical legislation and the increasing disease burden with ageing this is difficult to realize in clinical practice. Notwithstanding, we performed an analysis to examine the differences in intensity of most discriminating peaks based on age, gender and sample age. In the present study there was no significant contribution of one of these factors on the most discriminating peaks of our classification model (data not shown).

A source of bias may be the presence of batch effects, such as day-to-day variation or plate-to-plate variation. The presence of batch effects is unavoidable and – rather than to eliminate them from the design – a better approach is to account for and accommodate these effects, in such a way that they do not lead to errors of artificially induced group separation. Consequently, we randomly distributed the available samples from each group across the batches such that proportions were equal across batches within group. The so-called randomised block design ensured that the batch effect – if it materialised – would not induce an artificial between-group effect.[14;15]

A crucial point of discussion in the evolving field of clinical proteomics is validation of classification.[9;25] Given the sample size achievable within the experiment, use of a separate (possibly set-aside) validation set was precluded. The other problem is 'predictive optimisation'. However, as evaluation of predictive performance of the classifier is our primary focus, it is crucial that calibration is not carried out on the same data used for validation, which in turn would require an additional tuning set. Again, this would greatly increase the burden of collecting sufficient samples. For these reasons, other studies often carry out predictive optimisation on the full data in practice - which results in optimistically biased error rate evaluations, particularly with high-dimensional data such as in mass spectrometry proteomics.[26] As we have already suggested, another option is to reduce the available calibration data prior to optimisation, so as to set aside data, both for a training and validation set. However, this 'solution' is not as innocent as would appear at first sight, since it typically reduces the calibration set beyond the point of what is needed for reasonable calibration. Once more, this is particularly the case in high-dimensional cases such as clinical proteomics, where samples of malignancies are relatively difficult to obtain. Both problems may be avoided by carrying out a double-cross-validatory approach, which avoids the need for separate test and validation sets to yield unbiased error rate estimates. The double validatory aspect of the procedure results from the fact that the discriminant rule constructed to classify the left-out data was optimised through a secondary cross-validatory evaluation within the first cross-validatory layer (i.e. full cross-validation again on each 'leftover' set after removal of an observation). In this manner, we are able to integrate predictive optimisation and predictive unbiased validation in the same procedure, without loss of data – which is a crucial requirement to get realistic estimates of error rate with high-dimensional data while reducing the risk of overfitting.[27] Although the principle is sound and understood, this procedure has until recently not been applied in practice due to the considerable computational cost and (algebraic) complexity of the method.

Our classifier is based on Fisher linear discrimination, which has been derived and may be justified based on a variety of principles of inference, such as maximization

of the between-group separation relative to within-group error in the two-group case or the likelihood principle for normally distributed within-group populations. The methodology has been amply studied and has been established as reliable and robust form of classification and discriminant analysis. Furthermore, Fisher discrimination does not require an assumption of within-group normal dispersion.[21;28;29] Hastie et al. contains an up-to-date account of many new applications that demonstrate the continuing success of the approach.[18;21;28-30] Much similar and confirmatory experience has accumulated in related fields of application, which identifies this classification method as most reliable in high-dimensional analysis.[19;31] For proteomic mass spectra, principal components are attractive as it provides a means of non-parametrically smoothing and pooling information across peaks.

The controversy about the use of protein profiles as a pattern diagnostic without analysis of the diagnostic biomarkers remains to be solved for its clinical application. Identification and functional analysis of these discriminating proteins/peptides might render new insights on tumour development and environmental responsiveness, which could eventually be translated in new diagnostic and prognostic insights for the clinician. Unfortunately, little success has been booked so far in assigning reproducible discriminating biomarkers.[12;25] Though this study showed two most discriminating mass values of MALDI-TOF based protein profiling analysis to be low molecular weight fragments, we have not identified these potential biomarkers yet.

In the present study we used patterns of proteomic signatures from high dimensional mass spectrometry data to generate a diagnostic classifier for the detection of CRC. To our knowledge, this is the first double cross-validatory study in a randomised block design in this field of research. Although independent validation would strengthen the observations and follow up studies are now underway, we obtained maximal reliability in classification in this study while maintaining protection against overfitting. Due to the relatively small sample size we have chosen to use our entire dataset for a within-study validation to avoid optimistic biased (error) misclassification rates. To assess the performance of our classifier a further independent validation study will be necessary. In addition, in future studies the specificity of discriminating protein profiles for colorectal cancer have to be assessed in comparison with other cancer types. Nevertheless, we are currently able to detect CRC accurately on the basis of differences in actual information in the serum protein profiles with a rigorously standardised approach and exclusion of batch effects. Thus, although introduction in a routine clinical setting may take longer than originally hoped for, this study is an initial proof for a successful evolution of the potentially great use of discriminating protein profiles in the detection of CRC.

Fisher linear discriminant analysis may be defined as assigning an observation to the group for which the smallest within-group distance $D_g(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_g)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_g)^T$ is found for the corresponding observed feature vector $\boldsymbol{x} = (x_1, ..., x_p)$ with respect to the $g^{th}$ group ($g$=1,2 here, for either cases or controls), where $p$ is the dimensionality of the problem, $\boldsymbol{\mu}_g$ denotes the population within-group sample mean for the $g^{th}$ group and $\boldsymbol{\Sigma}$ is the (common) within-group dispersion matrix. We may estimate the population means through the within-group sample means. When the dimensionality of the problem is greater than the sample size, as is the case in this problem, the observed within-group pooled covariance matrix $\boldsymbol{S}$ will typically not be of full rank and hence special measures are called for before we can apply the above paradigm in this context. This can be achieved through an initial principal components decomposition of the observed within-group pooled covariance matrix $\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$, where $\boldsymbol{Q}$ and $\boldsymbol{\Lambda} = diag(\lambda_1, ..., \lambda_r)$ are the matrices of principal component weightings and variances respectively ($r$ is the rank of the pooled covariance matrix). We then re-estimate the within-group covariance matrix by only retaining the first $k$ components only: $\boldsymbol{S}_{(k)} = \boldsymbol{Q}_{(k)}\boldsymbol{\Lambda}_{(k)}\boldsymbol{Q}_{(k)}{}^T$, which account for most of the variation in the spectra. The discriminant rule may now be expressed as assigning an observation to the group for which we observe the smallest sample estimate $D_g(\boldsymbol{x}) = (\boldsymbol{x} - \overline{\boldsymbol{x}}_g)\boldsymbol{S}_{(k)}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}_g)^T$.

In the two-group case, this is also equivalent to least-squares regression analysis using the Moore-Penrose inverse of the pooled covariance matrix when $k$=$r$ (all components kept, also known as shortest least squares regression), or else is equivalent to so-called shrunken least-squares regression.[20,21] When choosing k<r, the choice may be made through appeal to a (cross-) validatory evaluation of the performance of the respective possible choices for the parameter $k$. The above methodology has been described and compared to other methods in the recent paper by Mertens[18], which shows this method to be competitive in the closely related high-dimensional setting for classification with microarrays. Much similar and confirmatory experience has accumulated in related fields of application, which identifies this classification method as reliable and stable in high-dimensional analysis, as has been described by Stone and Jonathan, among others.[19,31]

**Appendix** 1

## REFERENCES


1. Ruo,L., Gougoutas,C., Paty,P.B., Guillem,J.G., Cohen,A.M., and Wong,W.D. (2003) Elective bowel resection for incurable stage IV colorectal cancer: prognostic variables for asymptomatic patients. *J.Am.Coll.Surg.*, 196, 722-728.
2. Gill,S. and Sinicrope,F.A. (2005) Colorectal cancer prevention: is an ounce of prevention worth a pound of cure? *Semin.Oncol.*, 32, 24-34.
3. Adam,B.L., Qu,Y., Davis,J.W., Ward,M.D., Clements,M.A., Cazares,L.H., Semmes,O.J., Schellhammer,P.F., Yasui,Y., Feng,Z., and Wright,G.L., Jr. (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.*, 62, 3609-3614.
4. Petricoin,E.F., III, Ornstein,D.K., Paweletz,C.P., Ardekani,A., Hackett,P.S., Hitt,B.A., Velassco,A., Trucco,C., Wiegand,L., Wood,K., Simone,C.B., Levine,P.J., Linehan,W.M., Emmert-Buck,M.R., Steinberg,S.M., Kohn,E.C., and Liotta,L.A. (2002) Serum proteomic patterns for detection of prostate cancer. *J.Natl.Cancer Inst.*, 94, 1576-1578.
5. Rai,A.J., Zhang,Z., Rosenzweig,J., Shih,I., Pham,T., Fung,E.T., Sokoll,L.J., and Chan,D.W. (2002) Proteomic approaches to tumor marker discovery. *Arch.Pathol.Lab Med.*, 126, 1518-1526.
6. Yanagisawa,K., Shyr,Y., Xu,B.J., Massion,P.P., Larsen,P.H., White,B.C., Roberts,J.R., Edgerton,M., Gonzalez,A., Nadaf,S., Moore,J.H., Caprioli,R.M., and Carbone,D.P. (2003) Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet*, 362, 433-439.
7. Hu,J., Coombes,K.R., Morris,J.S., and Baggerly,K.A. (2005) The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief.Funct.Genomic.Proteomic.*, 3, 322-331.
8. Coombes,K.R., Morris,J.S., Hu,J., Edmonson,S.R., and Baggerly,K.A. (2005) Serum proteomics profiling-a young technology begins to mature. *Nat.Biotechnol.*, 23, 291-292.
9. Ransohoff,D.F. (2004) Rules of evidence for cancer molecular-marker discovery and validation. *Nat.Rev.Cancer*, 4, 309-314.
10. Boguski,M.S. and McIntosh,M.W. (2003) Biomedical informatics for proteomics. *Nature*, 422, 233-237.
11. Villanueva,J., Philip,J., Entenberg,D., Chaparro,C.A., Tanwar,M.K., Holland,E.C., and Tempst,P. (2004) Serum Peptide profiling by magnetic particle-assisted, automated sample processing and maldi-tof mass spectrometry. *Anal.Chem.*, 76, 1560-1570.
12. Diamandis,E.P. (2004) Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *J.Natl.Cancer Inst.*, 96, 353-356.
13. 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, 777-785.
14. Box,G.E.P., Hunter W.G., and Hunter J.S. (1978) *Statistics for experimenters*. John Wiley & Sons, Inc..
15. Cox D.R. and Reid N. (2000) *The theory of the design of experiments*. Chapmann/Hall CRC.
16. de Noo,M.E., Tollenaar,R.A.E.M., Ozalp,A., Kuppen,P.J.K., Bladergroen,M.R., and Deelder A.M. (2005) Reliability of human serum protein profiles generated with C8 magnetic beads assisted MALDI-TOF mass spectrometry. *Anal.Chem.*, 77, 7232-7241.
17. Whittaker,E.T. (2005) *On a new method of graduation*.
18. Mertens,B.J.A. (2003) Microarrays, pattern recognition and exploratory data analysis. *Statistics in Medicine*, 22, 1879-1899.
19. Mervyn Stone and Philip Jonathan (1994) Statistical thinking and technique for QSAR and related studies. Part II: Specific methods. *Journal of Chemometrics*, 8, 1-20.
20. Ripley,B.D. (1996) *Pattern recognition and neural networks*. Cambridge University Press.

21. Seber,G.A.F. (2005) *Multivariate Observations*. John Wiley & Sons Inc.
22. Yu,J.K., Chen,Y.D., and Zheng,S. (2004) An integrated approach to the detection of colorectal cancer utilizing proteomics and bioinformatics. *World J.Gastroenterol.*, 10, 3127-3131.
23. Diamandis,E.P. (2003) Point: Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? *Clin.Chem.*, 49, 1272-1275.
24. Qu,Y., Adam,B.L., Yasui,Y., Ward,M.D., Cazares,L.H., Schellhammer,P.F., Feng,Z., Semmes,O.J., and Wright,G.L., Jr. (2002) Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin.Chem.*, 48, 1835-1843.
25. Somorjai,R.L., Dolenko,B., and Baumgartner,R. (2003) Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics.*, 19, 1484-1491.
26. Baggerly,K.A., Morris,J.S., Wang,J., Gold,D., Xiao,L.C., and Coombes,K.R. (2003) A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics.*, 3, 1667-1672.
27. Stone,M. (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36, 111-147.
28. McLachlan (2004) *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons Inc.
29. Hand,D.J. (1997) *Construction and assessment of classification rules*. John Wiley and Sons; Inc.
30. Hastie,T., ., Tibshirani,R., and and Friedman,J. (2001) *The elements of statistical learning*. Springer-verlag.
31. Stone M. and Jonathan P. (1993) Statistical Thinking and technique for QSAR and related studies. *Journal of Chemometrics*, 7, 455-475.