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

Noo, M.E. de

Clinical Proteomics in oncology:

A passionate dance between science and clinic

Cover: **L'etoile [La danseuse sur la scene]** 1871,
by Edgar Degas, Musee d'Orsay, Paris

ISBN 978-90-9022-020-8

Printed by Optima Grafische Communicatie B.V. Rotterdam

This thesis is financially supported by BIOMET, Farwick Groenspecialisten,Greiner, KCI Medical, LCS Systemen, Nycomed, Roche, Sanofi Aventis, Smith & Nephew Hoofddorp, Stichting ter bevordering van de Chirurgische Wetenschappen te Leiden, Tyco Healthcare

Clinical Proteomics in oncology:

A passionate dance between science and clinic

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van de rector Magnificus prof. mr. P.F. van der Heijden,

volgens besluit van het College voor Promoties

te verdedigen op dinsdag 9 oktober 2007

klokke 16.15 uur

door

Mirre E. de Noo

geboren te Enschede in 1978.

PROMOTIECOMMISSIE:

Promotores :	Prof. Dr. A.M. Deelder
	Prof. Dr. R.A.E.M. Tollenaar
Referent:	Dr. L. van 't Veer *(NKI-AVL, Amsterdam)*
Overige Leden:	Prof. Dr. I.H.M. Borel Rinkes *(UMC, Utrecht)*
	Prof. M.J. Duffy *(St. Vincent's University Hospital, Dublin)*
	Dr. M. Frohlich
	Prof. Dr. D.W. Hommes
	Dr. B.J.A. Mertens
	Prof. Dr. C.J.H. van de Velde

Voor Han

Surgeons must be very careful
When they take the knife
Underneath their fine incisions
Stirs the culprit–Life!

Emily Dickinson

CONTENTS

Chapter 1

Introduction

COLORECTAL CANCER

Colorectal adenocarcinoma is the third most common cancer and the fourth most frequent cause of death due to cancer worldwide. Yearly almost one million new cases occur global, with 492000 related deaths.[1] In developed countries it is the second most common tumour, with a lifetime risk of 5%, but its incidence and mortality are currently decreasing.[2;3] Surgery is the cornerstone of therapy when the disease is confined to the bowel wall. This results in 70 to 80% of patients who can be resected with curative intent.[4] After curative surgery the five-year survival rate for patients with localised disease is 90%, decreasing to 65% in case of metastised disease in the lymph nodes. Adjuvant radiation therapy, chemotherapy, or both are beneficial in selected patients. For colorectal cancer the TNM staging system remains the gold standard for prognostication of the disease relying entirely on morphological and histopathological appearance of the tumour. Classification of tumours into these TNM stages with distinct clinical courses enables clinicians to define treatment. However, tumours with similar histopathological characteristics may have different clinical outcome and responsiveness to therapy.[5] Therefore, a detailed diagnosis would allow a more individualised treatment that may avoid unnecessary morbidity and increase survival. Despite these optimised treatment strategies for colorectal cancer patients, early detection of colorectal cancer will increase survival most. Colorectal cancer is optimal to employ early detection, as precancerous and early cancerous lesions are well defined in a multistep sequence of genetic alterations that result in the transformation of normal mucosa to a precursor adenoma and ultimately to carcinoma. Thus, given the natural history of the malignancy, early diagnosis appears to be the most appropriate tool to reduce disease-related mortality.[6-8]

BIOMARKERS IN COLORECTAL CANCER

Biomarkers are molecules that indicate the presence of cancer in the body. Most biomarkers are based on abnormal presence, absence or alterations in genes, RNA, proteins and metabolites. Since the molecular changes that occur during tumour development can take place over a number of years, some biomarkers may be used to detect colorectal cancer early. Furthermore, they might be used to predict prognosis, monitor disease progression and therapeutic response. Gion et al. classified different circulating biomarkers according to their clinical application.[9] These candidate biomarkers however, are frequently found in relatively low concentrations amid a sea of other biomolecules, so biomarker research and possible diagnostic tests depend critically on the ability to make high sensitive and accurate biochemical measure-

ments. Ideally, biomarkers should be specific for the disease and easy accessible, such as serum, plasma or urine, to increase their clinical applicability.

Carcinoembryonic antigen (CEA) is the best-characterised serologic tumour marker for monitoring colorectal cancer. However, its use as a population based screening tool for early detection and diagnosis of the disease is hindered by its low sensitivity and specificity. Fletcher showed that for screening purposes in a normal population, a cut-off concentration of 2.5 µg/L CEA would yield a sensitivity of 30-40%. Based on these data he calculated that there would be 250 false positive tests for every true positive test, i.e. a patient with cancer. Furthermore, 60% of the cancers would not be detected.[10-13]

Faecal occult-blood testing (FOBT) is another biomarker for which clinical trials have shown evidence of a decreased risk of death correlated with increased detection of the disease. This approach is a non-invasive option that limits the need for follow-up colonoscopy to patients with evidence of bleeding. Neoplasms bleed intermittently, however, allowing some to escape detection with faecal occult-blood testing. Annual retesting is therefore necessary but is still insufficient, detecting only 25 to 50% of colorectal cancers and 10% of adenomas. The specificity of FOBT is also limited by frequent false positive reactions to dietary compounds, medications, and gastrointestinal bleeding from causes other than colorectal cancer.[14-17] However, population screening for colorectal cancer based on FOBT is already implemented in several countries, including a trial in the Netherlands. The expectation is that even though the techniques still has its flaws, a population screening for colorectal cancer will decrease mortality with 15-20%.[14] This can be attributed to the fact that colorectal cancer develops as a multistep sequence of precancerous and early cancerous lesions over a relatively long period of time. However, these early stages of the disease can only be detected by screening. Furthermore, scientific evidence clearly shows that, in the case of CRC, early detection and treatment leads to more benefit than treatment that has started later. These reasons among other Wilson and Jungner's criteria, that also apply for colorectal cancer, are explain that a population based screening trial has started in the Netherlands, although the technique still has some limitations.[18]

BREAST CANCER

With over 1 million new cases in the world each year, breast cancer is the commonest malignancy in women and comprises 18% of all female cancers.[19] In 2005, breast cancer caused 502,000 deaths (7% of cancer deaths; almost 1% of all deaths) worldwide. The most recent data from the Surveillance, Epidemiology, and End

Results (SEER) program of the National Cancer Institute indicate that the lifetime probability of developing invasive breast cancer is one in nine.[20] Despite increasing incidence rates, annual mortality rates from breast cancer have decreased over the last decade (2.3% per year from 1990 to 2002).[21] The effect of reduction due to early diagnosis of breast cancer has been outlined with patients' data by the Surveillance, Epidemiology, and End Results program in a competing-risk analysis calculating probabilities of death from breast cancer and other causes according tot stage, race and age at diagnosis.[22] Reasons for the decline in mortality rates in western Europe, Australia and the Americas include widespread mammography screening, precise diagnosis, and increased number of women receiving tailor made treatment- including extensive use of tamoxifen and the use of chemotherapy.[23] There are many risk factors for breast cancer, including age and gender, race, lifestyle and dietary factors, reproductive and hormonal factors, family history and genetic factors, exposure to ionizing radiation and environmental factors. Although many epidemiological risk factors have been identified, the cause of any individual breast cancer is often unknown. In other words, epidemiological research informs the patterns of breast cancer incidence across certain populations, but often not in a given individual.

Once the diagnosis of breast cancer is established, the choice of initial treatment depends upon the stage or extent of disease. Although initial treatment decisions are made on the size and appearance of the primary tumour and the presence of palpable axillary nodes, the surgical and pathological findings are used to determine the pathologic disease stage, which dictates the prognosis and need for adjuvant systemic therapy. The most important are the status of the draining axillary lymph nodes, tumour size, whether the tumour expresses hormone receptors and/or the protein HER2, and a woman's age or menopausal status. Up to one-third of women with non-palpable axillary lymph nodes will be found to harbour metastases, while one-third of those with palpable nodes will be pathologically free of nodal involvement. In women with breast cancer who are younger than 50 years of age, chemotherapy increases their 15-year survival rate by 10%; in older women the increase is 3%.[24] However, chemotherapy has a wide range of acute and long-term side effects that substantially affect the patient's quality of life.[25] As it is not possible to accurately predict the risk of metastasis development in individual patients, nowadays more than 80% of them receive adjuvant chemotherapy, although only approximately 40% of the patients relapse and ultimately die of metastatic breast cancer. Therefore, many women who would be cured by local treatment alone, which includes surgery and radiotherapy, will be 'over-treated' and suffer the toxic side effects of chemotherapy needlessly.[26]

Women who have oestrogen sensitive (ER positive) tumours receive some form of hormonal therapy to block the cancer-promoting effect of oestrogen.[27] The use of tamoxifen was shown to significantly reduce the risk of recurrence and increase ten year survival in women with ER positive and ER unknown status tumours and its gradual widespread use is one of the main factors associated with the dramatic fall in mortality during the late twentieth century.[28;29] Most postmenopausal women receive tamoxifen for five years. Trials are ongoing to establish even more effective drugs and regimens for pre- and postmenopausal patients, taking into account side-effects as well as survival times. The ATAC trial recently reported its early results comparing anastrazole alone, anastrazole plus tamoxifen, and tamoxifen alone for postmenopausal women and has shown the benefits of anastrozole over tamoxifen in disease-free survival in early breast cancer.[30] In premenopausal women oestrogen production may be stopped by surgery (removing the ovaries), radiotherapy or drugs that reversibly suppress the ovaries (LHRH analogues).

In a recent meta-analysis, a mortality reduction of 38% (age <50 years) and 20% (age 50-69 years) with chemotherapy is shown, followed by a further reduction of 31% from tamoxifen. When combined together, the final mortality reductions would be 57% and 45%, respectively 57% reduction for women younger than 50 years of age and for those of age 50–69 years.[24] Moreover, breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome.

DIAGNOSIS AND BIOMARKERS IN BREAST CANCER

The procedures most commonly used in breast-cancer diagnosis is mammography, and to a lesser extent ultrasonography, MRI, and PET. In addition, physical examination remains important because a certain proportion (11%) of breast cancers is not seen on mammography.[31] Mammography remains the most important diagnostic tool in women with breast tissue that is not dense and is used in many countries as a population based screening in woman older than 50 years. The effect of breast screening in terms of breast cancer mortality reduction persists after long-term follow-up. A recent meta-analysis of seven randomised trials – concluded that there was a 15-20% reduction in risk of death from breast cancer in women attending mammography.[32] The effect of mammography screening is age-dependent and the highest effect is seen in women aged 55-69 years. This effect was not seen in woman under the age of 50, probably because of the higher density of the breast tissue.[33] Thus, after menopause, mammography is generally the best method to discover tiny, non-palpable lesions. By contrast, ultrasonography is the most effective procedure to diagnose small tumours in women with dense breast and to differentiate

solid lesions from cystic lesions.[34] Although mammography can identify suspicious micro calcifications, it is not good at distinguishing between breast densities and has difficulty in identifying certain lobular invasive carcinomas, Paget's disease of the nipple, inflammatory carcinoma, and particularly peripheral, small carcinomas.[35] MRI is mainly used as a problem-solving method after conventional diagnostic procedures. The technique is highly sensitive and mainly used for the screening of high-risk, *BRCA*-positive patients. It is also useful for identification of primary foci in non-palpable lesions and axillary metastases with no evidence of a primary focus, and for assessment of response to neoadjuvant chemotherapy.[36] Although MRI has good diagnostic accuracy, the rate of false-positive cases is still high and MRI findings cannot be a sole indication for breast surgery.[37] PET is presently used to discover undetected metastatic foci in any distant organ and can assess the status of axillary nodes in the preoperative staging process.[38]

Currently, mammography remains the most important diagnostic tool since serum tumour markers play no role of importance in the diagnosis of breast cancer due to a lack of sensitivity and specificity. Consequently, a major focus of present research is the identification of new biomarkers and drug targets to improve (early) detection and treatment; since early detection and more individualised treatment would benefit the individual patient and avoid unnecessary morbidity.

A NEW DIAGNOSTIC PARADIGM: CLINICAL PROTEOMICS

Proteomics is the large-scale study of proteins, particularly their presence, structure and functions. The term 'proteomics' was coined to make an analogy with genomics, the study of the genes. Proteomics is often considered the next step in the study of biological systems, after genomics. It is more complicated than genomics, mostly because while an organism's genome is rather constant, a proteome differs from cell to cell and constantly changes through its biochemical interactions with the genome and the environment. However, this functional state of a cell is very interesting for research goals and especially in oncogenesis. Clinical proteomics is referred to as mass spectrometry based proteomics using easy accessible body fluids.

Mass spectrometry is an analytical technique used to measure the mass-to-charge ratio of ions. It is most generally used to find the composition of a physical sample by generating a mass spectrum representing the masses of sample components. The mass spectrum is measured by a mass spectrometer. Matrix-assisted laser desorption/ionisation (MALDI) is a soft ionisation technique used in mass spectrometry, allowing the analysis of biomolecules which tend to be fragile and fragment when ionised by more conventional ionisation methods. The ionisation is triggered by a laser beam

(normally a nitrogen laser). A matrix is used to protect the biomolecule from being destroyed by direct laser beam and to facilitate vaporization and ionisation. The type of a mass spectrometer most widely used with MALDI is the TOF (time-of-flight mass spectrometer), mainly due to its large mass range. These mass spectrometers use an electric field to accelerate the ions through the same potential, and then measures the time they take to reach the detector. If the particles all have the same charge, then their kinetic energies will be identical, and their velocities will depend only on their masses. Lighter ions will reach the detector first, as shown in figure 1. The TOF measurement procedure is also ideally suited to the MALDI ionisation process since the pulsed laser takes individual 'shots' rather than working in continuous operation. MALDI-TOF instruments are typically equipped with an "ion mirror", deflecting ions with an electric field, thereby doubling the ion flight path and increasing the resolution. First, a sample has to be introduced into the ionisation source of the instrument. Once inside the ionisation source, the sample molecules are ionised, because ions are easier to manipulate than neutral molecules. These ions are extracted into the analyzer region of the mass spectrometer where they are separated according to their mass-to-charge ratios (m/z). The separated ions are detected and this signal is sent to a data system where the m/z ratios are stored together with their relative abundance for presentation in the format of an m/z spectrum.

Proteomic pattern diagnostics is a recent and potentially revolutionary approach for early disease detection, prognostication, and monitoring in oncology. The use of proteomic technologies might benefit biomarker discovery and treatment modalities: serum protein profiling for early disease detection and molecular signal mapping to instigate pharmocoproteomic therapeutic interventions.[39] Thus, several authors hypothesised that proteomic patterns generated with mass spectrometry are correlated to biological events occurring in the entire organism and are likely to change in the

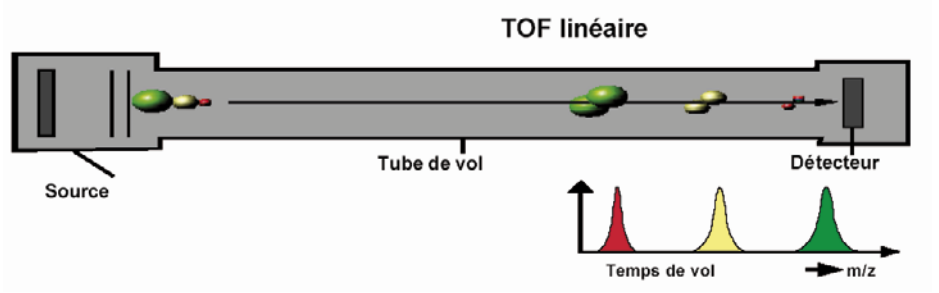


Figure 1. Schematic version of MALDI-TOF mass spectrometry principle.
Sample molecules are ionised with a laser source. Then an electric field is used to accelerate the ions in a flight tube. The detector measures their flight time. If the particles all have the same charge, then their kinetic energies will be identical, and their velocities will depend only on their masses. The smaller ones (red) will reach the detector earlier than the heavier ones (green).

presence of disease. New types of bioinformatic pattern recognition algorithms were used to identify patterns of protein changes in order to discriminate cancer patients from healthy individuals with promising results.

Petricoin and his co-workers were the first to state that finding a single disease-related biomarker is like searching for a needle in a haystack; each entity has to be separated and identified individually.[40;41] Moreover, they postulated that the blood proteome constantly changes as a consequence of the perfusion of the diseased organ adding, subtracting, or modifying the circulating proteome. These differences might be the result of proteins being abnormally produced or shed and added to the serum proteome, clipped or modified as a consequence of the disease process, or subtracted from the proteome owing to disease-related proteolytic degradation pathways. Therefore, protein pattern diagnostics would provide an easy and reliable tool for detection of cancer. The advantages of the proteomic pattern approach were stressed in several papers. In addition to the high sensitivity and specificity, cost-effectiveness, easy accessibility of body fluid and especially the high-throughput, ultimately allowing application in future screening studies, were mentioned.[42;43] Next to these hopeful voices, soon critical notes were made on analytical reproducibility and the use of the so-called black box approach, lacking identification of discriminating proteins.[44]

CLINICAL PROTEOMICS IN ONCOLOGY

Cancer is known to be the consequence of genetic alterations. A gene, however, is only potential information that must be put into a functional form. The DNA is transcribed into RNA before translation into protein, the functional manifestation of the genetic code. During the transformation of a healthy cell into a neoplastic cell, including alterations in expression, activity, localization and differential protein modification, changes also occur in the protein level. Identifying and understanding these changes is the underlying theme in cancer proteomics.[45]

In 2002 several studies discriminated patients with various cancers from healthy subjects on the basis of presence/absence of multiple low-molecular-weight serum proteins using SELDI-TOF mass spectrometry technologies.[42;46-48] The authors hypothesised that proteomic patterns are correlated with biological events occurring in the entire organism and are likely to change in the presence of disease. New types of bioinformatic pattern recognition algorithms were used to identify patterns of protein changes in order to discriminate cancer patients from healthy individuals with promising results. Several studies have shown that biomarkers can be identified on the basis of the presence/absence of multiple low-molecular-weight serum

proteins.[41;42;46-49] Furthermore, different profiles may be associated with varying responses to therapeutics and other clinically relevant parameters and may also serve as prediction for treatment outcome.

Although serum protein patterns showed high sensitivity and specificity as an early diagnostic tool in several studies, critical notes have been made on biological variation, pre-analytical conditions and analytical reproducibility of serum protein profiles that would make it difficult to differentiate a normal from a pathological and/or malignant status.[50] In addition, the reproducibility of serum protein profiles has been questioned, which however relates more to the bioinformatical analysis of the measured protein profiles than the capturing and measuring techniques itself.[51-53] Thus, if proteomics spectra are ultimately to be applied in a routine clinical setting, collection and processing of the data will need to be subject to stringent quality control procedures.[54]

OUTLINE OF THIS THESIS

Given the natural history of colorectal and breast cancer, early diagnosis appears to be the most appropriate tool to reduce disease-related mortality.[6;7] 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 both types of cancer that can improve early diagnosis, monitoring of disease progression and therapeutic response and detect disease recurrence. Proteomic expression profiles generated with mass spectrometry have been suggested as potential tools for the early diagnosis of cancer and other diseases. Because it is still in its infancy, many problems have to be overcome before clinical proteomics can be transferred form bench to bedside. **Chapter 2** gives an insight in the different fields of translational research in colorectal cancer by our group. In **chapter 3** reliability of human serum protein profiling using MALDI-TOF mass spectrometry is analysed. We present a pipeline for pre-processing, statistical data analysis and presentation of MALDI-TOF spectra. This novel analysis method was used to assess the effect of variable pre-analytical conditions on human serum protein profiles, and their effect on reproducibility. In line with the logistic conditions in a routine clinical setting, the effects of sample handling and storage, and also circadian rhythm factors on the serum protein profiles were analysed. In **chapter 4 and 5** the feasibility of mass spectrometry based protein profiling for the discrimination of colorectal cancer patients from healthy individuals was assessed. 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. Especially, validation of our classifier, as a possible pitfall, was given much attention. Therefore, we performed a linear discriminant analysis with double cross-validation to separate cancer patients from healthy subjects. **Chapter 6** reports on results from an identical designed protein profiling study for the detection of breast cancer. In **chapter 7** a first validated study on the detection of breast cancer based on mass spectrometry generated protein profiles is described. In this study the same randomised blocked design and double cross validation is used, however the classifier was validated in an independent set of new patients and controls. Finally, the results and conclusions of all above mentioned studies and especially the current status of clinical proteomics in cancer are discussed in **chapter 8**.

A Dutch summary of this thesis is written in **chapter 9**.

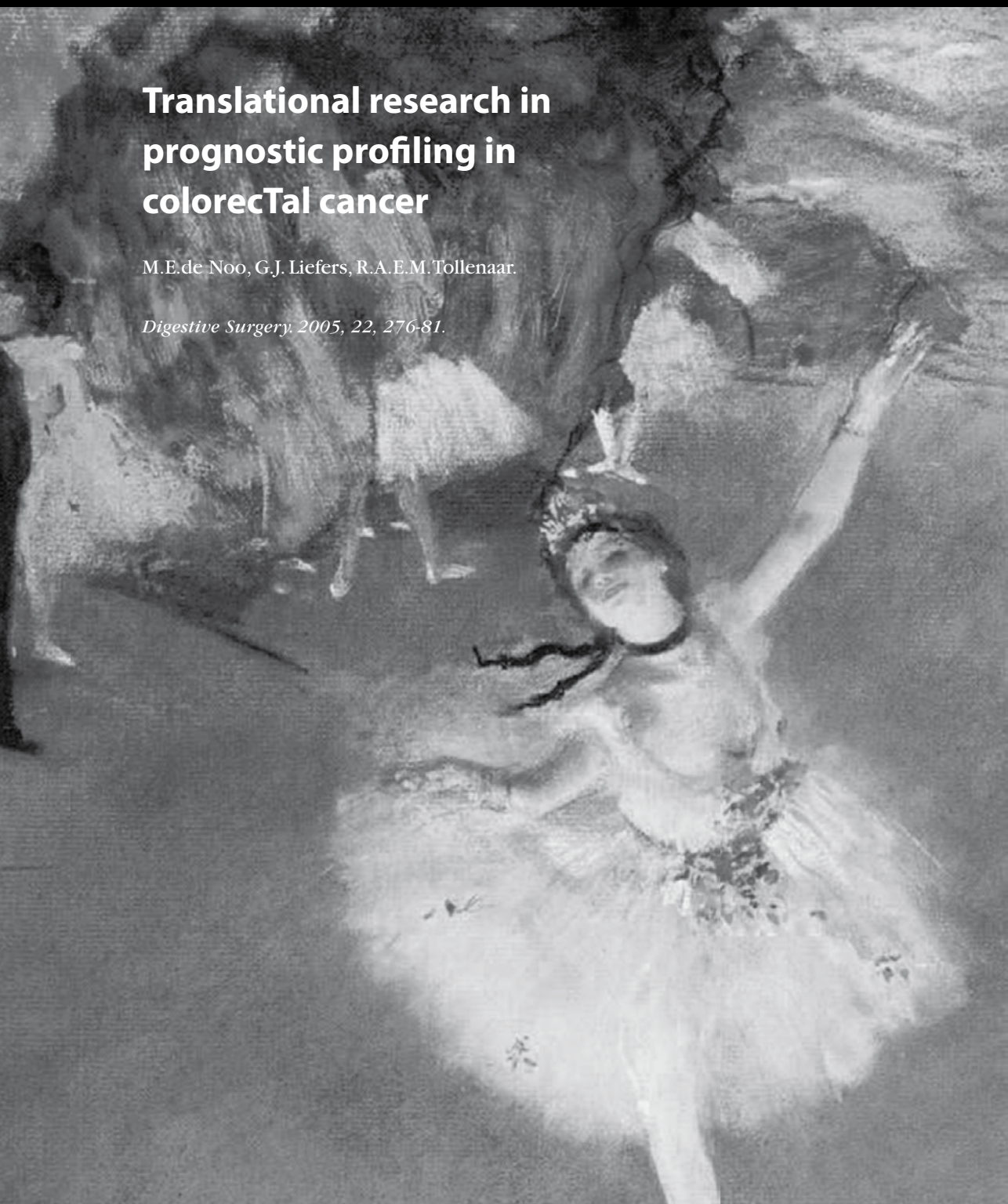
REFERENCES

1. Weitz,J., Koch,M., Debus,J., Hohler,T., Galle,P.R., and Buchler,M.W. (2005) Colorectal cancer. *Lancet*, 365, 153-165.
2. Russo,M.W., Wei,J.T., Thiny,M.T., Gangarosa,L.M., Brown,A., Ringel,Y., Shaheen,N.J., and Sandler,R.S. (2004) Digestive and liver diseases statistics, 2004. *Gastroenterology*, 126, 1448-1453.
3. Jemal,A., Tiwari,R.C., Murray,T., Ghafoor,A., Samuels,A., Ward,E., Feuer,E.J., and Thun,M.J. (2004) Cancer statistics, 2004. *CA Cancer J Clin*, 54, 8-29.
4. Pfister,D.G., Benson,A.B., III, and Somerfield,M.R. (2004) Clinical practice. Surveillance strategies after curative treatment of colorectal cancer. *N.Engl.J Med.*, 350, 2375-2382.
5. Liefers,G.J. and Tollenaar,R.A. (2002) Cancer genetics and their application to individualised medicine. *Eur.J.Cancer*, 38, 872-879.
6. 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.
7. 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.
8. Hawk,E.T. and Levin,B. (2005) Colorectal cancer prevention. *J Clin Oncol*, 23, 378-391.
9. Gion,M. and Daidone,M.G. (2004) Circulating biomarkers from tumour bulk to tumour machinery: promises and pitfalls. *Eur.J Cancer*, 40, 2613-2622.
10. Duffy,M.J., van Dalen,A., Haglund,C., Hansson,L., Klapdor,R., Lamerz,R., Nilsson,O., Sturgeon,C., and Topolcan,O. (2003) Clinical utility of biochemical markers in colorectal cancer: European Group on Tumour Markers (EGTM) guidelines. *Eur.J.Cancer*, 39, 718-727.
11. Fletcher,R.H. (2002) Rationale for combining different screening strategies. *Gastrointest.Endosc.Clin.N.Am.*, 12, 53-63.
12. Winawer,S., Fletcher,R., Rex,D., Bond,J., Burt,R., Ferrucci,J., Ganiats,T., Levin,T., Woolf,S., Johnson,D., Kirk,L., Litin,S., and Simmang,C. (2003) Colorectal cancer screening and surveillance: clinical guidelines and rationale-Update based on new evidence. *Gastroenterology*, 124, 544-560.
13. Ouyang,D.L., Chen,J.J., Getzenberg,R.H., and Schoen,R.E. (2005) Noninvasive testing for colorectal cancer: a review. *Am.J.Gastroenterol.*, 100, 1393-1403.
14. Pignone,M., Rich,M., Teutsch,S.M., Berg,A.O., and Lohr,K.N. (2002) Screening for colorectal cancer in adults at average risk: a summary of the evidence for the U.S. Preventive Services Task Force. *Ann.Intern.Med.*, 137, 132-141.
15. Ransohoff,D.F. and Lang,C.A. (1997) Screening for colorectal cancer with the fecal occult blood test: a background paper. American College of Physicians. *Ann.Intern.Med.*, 126, 811-822.
16. Ransohoff,D.F. (2005) Colon cancer screening in 2005: status and challenges. *Gastroenterology*, 128, 1685-1695.
17. Mandel,J.S., Bond,J.H., Church,T.R., Snover,D.C., Bradley,G.M., Schuman,L.M., and Ederer,F. (1993) Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. *N.Engl.J.Med.*, 328, 1365-1371.
18. de Visser,M., van Ballegooijen,M., Bloemers,S.M., van Deventer,S.J., Jansen,J.B., Jespersen,J., Kluft,C., Meijer,G.A., Stoker,J., de Valk,G.A., Verweij,M.F., and Vlems,F.A. (2005) Report on the Dutch consensus development meeting for implementation and further development of population screening for colorectal cancer based on FOBT. *Cell Oncol.*, 27, 17-29.
19. Cancer Facts and Figures 2007. American Cancer Society. 2007.
20. Jemal,A., Siegel,R., Ward,E., Murray,T., Xu,J., Smigal,C., and Thun,M.J. (2006) Cancer statistics, 2006. *CA Cancer J.Clin.*, 56, 106-130.

21. Edwards,B.K., Brown,M.L., Wingo,P.A., Howe,H.L., Ward,E., Ries,L.A., Schrag,D., Jamison,P.M., Jemal,A., Wu,X.C., Friedman,C., Harlan,L., Warren,J., Anderson,R.N., and Pickle,L.W. (2005) Annual report to the nation on the status of cancer, 1975-2002, featuring population-based trends in cancer treatment. *J.Natl.Cancer Inst.*, 97, 1407-1427.
22. Schairer,C., Mink,P.J., Carroll,L., and Devesa,S.S. (2004) Probabilities of death from breast cancer and other causes among female breast cancer patients. *J.Natl.Cancer Inst.*, 96, 1311-1321.
23. Peto,R., Boreham,J., Clarke,M., Davies,C., and Beral,V. (2000) UK and USA breast cancer deaths down 25% in year 2000 at ages 20-69 years. *Lancet*, 355, 1822.
24. (2005) Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet*, 365, 1687-1717.
25. Eifel,P., Axelson,J.A., Costa,J., Crowley,J., Curran,W.J., Jr., Deshler,A., Fulton,S., Hendricks,C.B., Kemeny,M., Kornblith,A.B., Louis,T.A., Markman,M., Mayer,R., and Roter,D. (2001) National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer, November 1-3, 2000. *J.Natl.Cancer Inst.*, 93, 979-989.
26. Weigelt,B., Peterse,J.L., and 't Veer,L.J. (2005) Breast cancer metastasis: markers and models. *Nat.Rev.Cancer*, 5, 591-602.
27. Wishart,G.C., Gaston,M., Poultsidis,A.A., and Purushotham,A.D. (2002) Hormone receptor status in primary breast cancer--time for a consensus? *Eur.J.Cancer*, 38, 1201-1203.
28. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. (1992) 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. Early Breast Cancer Trialists' Collaborative Group. *Lancet*, 339, 1-15.
29. Tamoxifen for early breast cancer: an overview of the randomised trials. (1998) Early Breast Cancer Trialists' Collaborative Group. *Lancet*, 351, 1451-1467.
30. Baum,M., Budzar,A.U., Cuzick,J., Forbes,J., Houghton,J.H., Klijn,J.G., and Sahmoud,T. (2002) Anastrozole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer: first results of the ATAC randomised trial. *Lancet*, 359, 2131-2139.
31. Benson,S.R., Blue,J., Judd,K., and Harman,J.E. (2004) Ultrasound is now better than mammography for the detection of invasive breast cancer. *Am.J.Surg.*, 188, 381-385.
32. Gotzsche,P.C. and Nielsen,M. (2006) Screening for breast cancer with mammography. *Cochrane.Database.Syst.Rev.*,CD001877.
33. Nystrom,L., Andersson,I., Bjurstam,N., Frisell,J., Nordenskjold,B., and Rutqvist,L.E. (2002) Long-term effects of mammography screening: updated overview of the Swedish randomised trials. *Lancet*, 359, 909-919.
34. Helvie,M.A., Chan,H.P., Adler,D.D., and Boyd,P.G. (1994) Breast thickness in routine mammograms: effect on image quality and radiation dose. *AJR Am.J.Roentgenol.*, 163, 1371-1374.
35. Gordon,P.B. and Goldenberg,S.L. (1995) Malignant breast masses detected only by ultrasound. A retrospective review. *Cancer*, 76, 626-630.
36. Kneeshaw,P.J., Turnbull,L.W., and Drew,P.J. (2003) Current applications and future direction of MR mammography. *Br.J.Cancer*, 88, 4-10.
37. Szabo,B.K., Aspelin,P., Wiberg,M.K., and Bone,B. (2003) Dynamic MR imaging of the breast. Analysis of kinetic and morphologic diagnostic criteria. *Acta Radiol.*, 44, 379-386.
38. Wahl,R.L., Siegel,B.A., Coleman,R.E., and Gatsonis,C.G. (2004) Prospective multicenter study of axillary nodal staging by positron emission tomography in breast cancer: a report of the staging breast cancer with PET Study Group. *J.Clin.Oncol.*, 22, 277-285.
39. Posadas,E.M., Simpkins,F., Liotta,L.A., Macdonald,C., and Kohn,E.C. (2005) Proteomic analysis for the early detection and rational treatment of cancer--realistic hope? *Ann.Oncol.*, 16, 16-22.

40. Petricoin,E.F. and Liotta,L.A. (2002) Proteomic analysis at the bedside: early detection of cancer. *Trends Biotechnol.*, 20, S30-S34.
41. Wulfkuhle,J.D., Liotta,L.A., and Petricoin,E.F. (2003) Proteomic applications for the early detection of cancer. *Nat.Rev.Cancer*, 3, 267-275.
42. Petricoin,E.F., Ardekani,A.M., Hitt,B.A., Levine,P.J., Fusaro,V.A., Steinberg,S.M., Mills,G.B., Simone,C., Fishman,D.A., Kohn,E.C., and Liotta,L.A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359, 572-577.
43. Petricoin,E.E., Paweletz,C.P., and Liotta,L.A. (2002) Clinical applications of proteomics: proteomic pattern diagnostics. *J.Mammary.Gland.Biol.Neoplasia.*, 7, 433-440.
44. 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.
45. Srinivas,P.R., Verma,M., Zhao,Y., and Srivastava,S. (2002) Proteomics for cancer biomarker discovery. *Clin.Chem.*, 48, 1160-1169.
46. 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.
47. 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.
48. 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.
49. 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.
50. Boguski,M.S. and McIntosh,M.W. (2003) Biomedical informatics for proteomics. *Nature*, 422, 233-237.
51. 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.
52. Yasui,Y., Pepe,M., Thompson,M.L., Adam,B.L., Wright,G.L., Jr., Qu,Y., Potter,J.D., Winget,M., Thornquist,M., and Feng,Z. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics.*, 4, 449-463.
53. 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.
54. Coombes,K.R., Fritsche,H.A., Jr., Clarke,C., Chen,J.N., Baggerly,K.A., Morris,J.S., Xiao,L.C., Hung,M.C., and Kuerer,H.M. (2003) Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin. Chem.*, 49, 1615-1623.

Chapter 2



Translational research in prognostic profiling in colorecTal cancer

M.E.de Noo, G.J. Liefers, R.A.E.M.Tollenaar.

Digestive Surgery. 2005, 22, 276-81.

ABSTRACT

There is a widening gap between basic research and clinical practice, particularly for colorectal cancer. In recent years, many have expressed concerns regarding the disconnection between the promises of basic science and the delivery of better individual health. In this paper we describe some of our research in serum proteomics, microarrays and minimal residual disease dedicated to this field and discuss some of the roadblocks ahead in translational research. We conclude that translational medicine should be a collective effort for the medical community as a whole with adequate financial support and sound, measurable outcome. Since extensive validation of the above mentioned research fields is necessary, adequate funding is required. This may require some adjustments in the current funding policy because it involves non-innovative studies. Furthermore, the pool of researchers/clinicians capable of performing translational research must be increased. Additionally, there should be an enhanced participation of patients in clinical trials and an optimization of the efficiency of these trials using validated surrogate markers. Only when these conditions are fulfilled will the 'post-genomic' era of biomedical research have unprecedented opportunities to innovate and improve therapy for cancer.

INTRODUCTION

There is a widening gap between basic research and clinical practice particularly for colorectal cancer. Over the last decades our molecular knowledge on the genesis of colorectal cancer has increased dramatically. Despite this increase our treatment of patients remains largely the same: 'en-bloc' removal of the primary tumour and draining lymph nodes when possible, staging according to standard Dukes' or TNM classification systems and adjuvant treatment with cytotoxic drugs and/or radiation therapy. Despite mounting evidence of abundant heterogeneity of both clinical course of disease and responsiveness to therapy, 'tailor-made' medicine is an item in review papers and editorials instead of every-day-practice.

The paradigm for the translation of new information has been conceptualised by some as a highway. A 'translational highway' running from basic biomedical research to individualised patient care with improved health as a result (Figure 1). In recent years many have expressed concerns regarding the disconnection between the promise of basic science and the delivery of better health.[1] In a special communication for the JAMA, Donald Berwick addresses the problem of disseminating

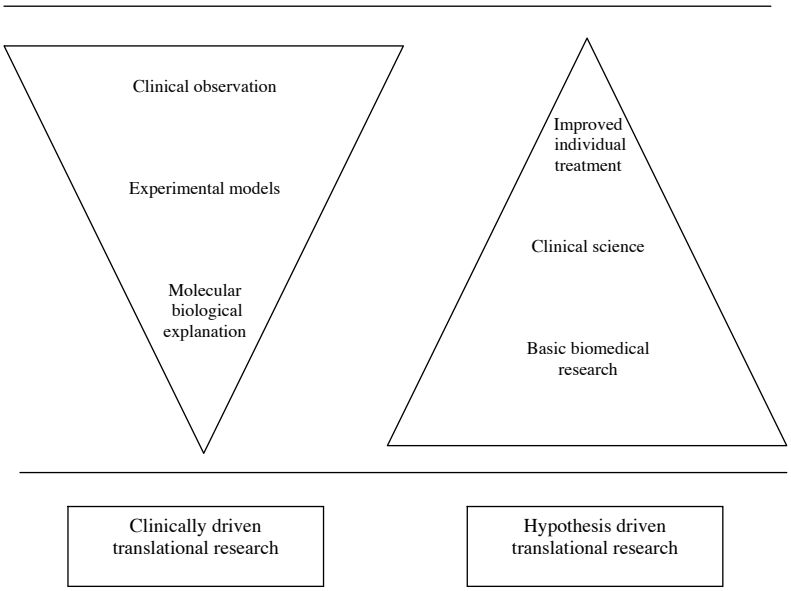


Figure 1. Current biomedical research places high priority on defining molecular mechanisms of disease with the ultimate aim of improving health of the individual patient (hypothesis driven). However, part of the failure to translate hypotheses derived from complex experimental models into improved patient care can be explained by the fact that many of these hypotheses do not translate to human pathology. It is therefore pivotal for successful translational medicine to promote research based on clinical observations and corresponding molecular biological explanations (clinically driven).

Table 1. *Factors associated with perceptions of an innovation*

Perception of an innovation that influences the rate of spread
Perceived benefit of the change
Compatibility with beliefs and needs of potential adopters
Complexity of the proposed innovation
Trialability (testing the change on a small scale)
Observability (watching others try the change first)

Adapted from: E.M. Rogers, Diffusion if innovations, 4th ed. New York, NY: Free Press 1995

innovations in health care as postulated by Rogers.[2] One of the obstacles he mentions, is that apart from lack of knowledge about the expected consequences of innovations or the perceived benefit, these innovations must resonate with currently felt needs and beliefs. Other factors associated with perceptions of an innovation are the complexity the proposed innovation, trial ability (testing the change on a small scale) and observability (watching others trying the change first), as shown in table 1. For colorectal cancer the 'needs and beliefs' are evident. First, the prognostic information from our standard classification system needs refinement. This is exemplified by the fact that despite of lack of evidence of residual disease in Dukes' B colorectal cancer patients, 30% die of recurrent disease within five years. Second, there is a need for tools that allows us to predict or monitor therapy response to avoid unnecessary morbidity. And third there is a need for new molecular targets that allows the development of cancer specific drugs that lack the side effects of current cytotoxic chemotherapeutics.

In this paper we would like to describe some of our research dedicated to this field and discuss some of the roadblocks ahead.

PROTEOMICS

Cancer is often described as a genetic disease. A gene alone, however, is only potential information that must be put into a functional form. The DNA is transcribed into RNA before translation into protein, the manifestation of the genetic code. During the transformation of a healthy cell into a neoplastic cell, including alterations in expression, activity and localization and differential protein modification, changes occur in the protein level. Identifying and understanding these changes is the underlying theme in cancer proteomics.[3]

Proteomic pattern diagnostics is a recent and potentially revolutionary approach for early disease detection, prognostication, and monitoring in oncology. The use of proteomic technologies might benefit biomarker discovery and treatment modalities: serum protein profiling for early disease detection and molecular signal mapping to instigate pharmocoproteomic therapeutic interventions.[4]

Several studies have shown that biomarkers can be identified on the basis of the presence/absence of multiple low-molecular-weight serum proteins using mass spectrometry technologies such as SELDI-TOF and MALDI-TOF.[5-9] Patterns of these peptides can be correlated to biological events occurring in the entire organism and are likely to change in the presence of disease. In oncology new types of bioinformatic pattern recognition algorithms have been used to identify patterns of protein changes in order to discriminate cancer patients from healthy individuals.[10] Furthermore, different profiles may be associated with varying responses to therapeutics and other clinically relevant parameters and may also serve as prediction for treatment outcome. Although serum protein patterns showed high sensitivity and specificity as an early diagnostic tool in several studies, critical notes have been made on biological variation, pre-analytical conditions and analytical reproducibility of serum protein profiles that would make it difficult to differentiate a normal from a pathological and/or malignant status.[11] In addition, the reproducibility of serum protein profiles has been questioned, which however relates more to the bioinformatical analysis of the measured protein profiles than the capturing and measuring techniques itself.[12-14] Thus, if proteomics spectra are ultimately to be applied in a routine clinical setting, collection and processing of the data will need to be subject to stringent quality control procedures.[15]

In a recently submitted study we assessed the reproducibility of our MALDI-TOF protein profiling procedure after capture and elution of serum peptides with C8 magnetic beads. Corresponding to the logistical conditions in a routine clinical setting, the effects of sample handling and storage, and also individual factors on the serum protein profiles were analysed. The reproducibility of the used capturing technique with C8 magnetic beads and MALDI-TOF analysis is acceptable and satisfactory for large discriminating studies. The time of blood collection and the number of freeze-and-thaw cycles had no influence on serum protein profiles. However, sample handling prior to serum centrifugation did have considerable effect on serum protein profiles. All together, we have shown in this study that effects of handling and storage procedures on serum protein profiles lie within acceptable limits. To prevent bias in classification studies we stress the importance of a standardised collection of all blood samples, from the point of sample handling and storage until freezing the samples. Although the importance of homogeneity and uniformity within sample groups must be stressed, variation of such factors cannot totally be

excluded in a clinical setting. The most important issues for discriminating studies at this moment are a standardised and well-documented sample collection and a thorough study design. Further research for the statistical data-analysis is needed. Due to the lack of discriminating profiles, serum protein profiling is not ready for introduction in a routine clinical setting. Nevertheless, based on the present data and these of Villanueva et al. [16], we feel that the methodology can be standardised to a level which allows application as a diagnostic and prognostic tool. Therefore, we are now in the process of carrying out a study to determine whether serum protein profiles can differentiate colorectal cancer patients from individuals with benign bowel disorders and healthy subjects. Further, identification and functional analysis of these discriminating proteins will render new insights on tumour development and environmental responsiveness.

MICROARRAYS (PROGNOSTIC FACTORS)

Over the last decades, numerous molecular factors with prognostic and predictive value have been described. Specifically loss of heterozygousity (LOH) of chromosome 18q and microsatellite instability (MSI) have been repeatedly implicated both as prognosticators as well as predictive for 5-fluorouracil based chemotherapy.[17] Despite multiple studies with large number of patients and unequivocal outcome data these markers have not yet found their way into routine treatment planning for patients with colorectal cancer. One of the reasons for this may be that the observed differences are studied retrospectively, which diminishes the expected benefit of using these markers in clinical decision making. Furthermore, with respect to the triability, it takes a tremendous amount of work for potential adopters to prospectively validate these markers. Also, the use of a single marker disregards the biological complexity of tumour development. New techniques, such as cDNA microarray analysis enable the parallel monitoring of expression levels of thousands of genes. Current cDNA microarray protocols are based on the Southern blot technique in which labelled nucleic acid molecules are hybridised to complementary nuclear acid molecules attached to a solid surface such as glass. Technical innovations such as miniaturization and fluorescence-based detection greatly enhance the throughput. A microarray consists of thousands of small spots of multiple copies of amplified cDNA spotted on a glass microscopic slide. Each spot represents a unique sequence from a named gene or expressed sequence tag (EST) and one slide can hold up to 10,000 probes. As a target for analysis, total RNA or mRNA from two cell populations is used (e.g. cell lines, clinical samples and animal models). Fluorescent marker dyes such as Cy3 and Cy5 are incorporated into target cDNA. The labelled cDNA from the

two cell populations of interest are mixed with a labelled control sample and hybridised to the probes on the glass slides. The array is scanned using confocal laser microscopy. After excitation and emission of fluorescence, signals can be measured and displayed. This results in a matrix of thousands of green, red and yellow spots. When, for example, a gene is equally expressed in test and control samples, both the red and green fluorescent signals will be equally strong and will be visualised as a yellow dot. Consequently, in the case of differential expression, the red to green ratio will shift. Following hybridisation and scanning, large amounts of data are available for processing. A variety of software tools are available which can help to measure fluorescent signal ratios, exclude artefacts and normalize data.

In a small, unpublished series of rectal cancer patients we have tested the hypothesis that microarray analysis could distinguish between patients with and without liver metastases. In collaboration with the Institute of Medical Sciences, University of Tokyo, Japan, we analysed tumour RNA from 20 rectal cancer patients; 12 patients with liver metastases and 8 patients without. RNA was extracted from fresh frozen tissue samples using laser capture micro dissection (LCM). After amplification and labelling, probes were hybridised to a microarray consisting of 9,216 genes. After scanning, the differential expression ratio for each gene was determined.

Data were analysed according to the 'leave-one-out' methodology as described. The resulting set of 30 genes could correctly predict the presence of liver metastases in 10 out of 12 patients. These data are currently being validated in a larger series. However these preliminary data show that, as in many other tumours, cDNA microarrays are promising new tools for the prognostication of patients with colorectal cancer.

For the translation of these experimental techniques into standard care, some of the roadblocks ahead can be easily envisioned. First the proposed superiority over our standard classification system must be (repeatedly) demonstrated in large groups of patients. To achieve this, tissue banks with fresh frozen tissues and serum must be established for validation studies. With adequate funding, these tissues can be collected from patients who are randomised in clinical trials and made available to the research community. International initiatives from the NCI and EORTC underline this view.

Secondly, the introduction and acceptance of prognostic gene sets would be more anticipated when experiments show a causal role of each of the genes in the clinical course of the disease. Microarray data are therefore by no means endpoints. Rather, they are hypothesis driven starting points for the development of new therapeutic strategies.

MINIMAL RESIDUAL DISEASE

Detection of metastatic cells by molecular techniques has been reported to increase the sensitivity over standard pathological examination. Metastatic cells can be found in histopathological negative lymph nodes, bone marrow and blood of colorectal cancer patients. Many of the published papers indicate a poor prognosis in patients with molecular detected metastases in all of the mentioned sites. Despite this, molecular techniques are not routinely used in the staging of patients.

The prognosis for colorectal cancer patients whose lymph nodes are not involved (stage II) is good. Five-year survival rates approximate 70%. In the Netherlands, adjuvant chemotherapy, therefore, is not considered standard care. Our group studied 26 stage II patients to detect micro metastases in lymph nodes by reverse-transcriptase-PCR of carcinoembryonic antigen (CEA) mRNA in microscopic negative lymph nodes. Overall, micro metastases could be detected in one or more lymph nodes from 14 patients (54%). These patients fared significantly worse than the patients without micrometastases. In this study, survival dropped from 75% to 36% based on the presence of micrometastatic disease. When only cancer-related deaths were considered, survival dropped from 91% to 50% respectively. The relative risk for cancer related death associated with the presence of micrometastases was 11.7 (95% C.I.: 1.2-106.9; P=0.03).[18]

This study is one of the first to relate micrometastatic disease to patient outcome and provides a rationale for the selection of patients who might benefit from adjuvant therapy. Since our publication others have confirmed these findings but there has been no massive introduction or these techniques into daily practice. The reason for this is that the pivotal question whether the prognosis of patients 'upstaged' by molecular techniques improves after adjuvant treatment remains unanswered. The perceived benefit for this innovation therefore may be low and is subject of ongoing investigation by our group and others. A second reason for the lack of adoption of these techniques is that they are complex and time consuming. Sentinel node (SN) biopsy has been introduced to minimize the extent of surgery and to enable assessment of minimal residual disease (MRD) without compromising accurate staging or survival.[19] For colorectal cancer the SN concept could be used to limit the number of nodes amenable for detailed molecular analysis. We are currently in the process of evaluation of micrometastases in sentinel nodes from colorectal cancer patients.

Another area of research is MRD detection in bone marrow. Viable cancer cells can be found in bone marrow from 20-40% of patients with colorectal cancer. This phenomenon correlates with an adverse prognosis. We have tested different methods for MRD detection, including automated microscopy and RT-PCR and preliminary results indicate prognostic relevance of these tests for different stages of colorectal cancer.

[20] Not all cancer cells that can be found in bone marrow are clinically relevant since they are present even in patients that never relapse. Experimental studies in breast cancer show that tumours consist of heterogeneous populations of cells with distinct tumorigenic potential.[21;22]

Minimal residual disease may arise from tumorigenic or non-tumorigenic cancer cells. Only when tumorigenic cancer cells metastasize, clinically relevant metastasis will occur.[23] Support for this theory comes from observations that disseminated minimal residual cancer cells from patients with and without overt distant metastasis are genotypically different.[24] Therefore the development of diagnostic tools that allow for the prospective identification of tumorigenic minimal residual cells may have therapeutic significance for patients with solid tumours. This will be one of the research goals for our group in the coming years.

CONCLUSION

In the 'post-genomic era' of biomedical research there are unprecedented opportunities to innovate and improve therapy for cancer. These opportunities are limited by today's clinical infrastructure. Efforts to validate and implement novel therapies are characterised by lack of funding and fragmentation. For a successful translation of novel biomedical discoveries to improved, individual health there are several issues to be addressed. First of all, translational medicine should be a collective effort for the medical community as a whole with adequate financial support and sound, measurable outcome. As extensive validation of the above mentioned research fields is necessary, adequate funding is required. This may require some adjustments in the current funding policy as it involves non-innovative studies. Secondly, the pool of researchers/clinicians capable of performing translational research must be increased. Thirdly, there must be an enhanced participation of patients in clinical trials and we have to optimize the efficiency of these trials using validated surrogate markers. Especially when we move towards 'tailor-made' medicine, evidence from large randomised trials (with inherently large groups of uniformly treated patients) will be more difficult to obtain. Current clinical trials must be appended with basic biomedical science studies, with collection of tissues for retrospective analysis. Last, we have to deal with regulatory and cultural aspects of the implementation of health innovations.

For the coming years it is the goal of our group to integrate three lines of research; MRD detection, cDNA microarray analysis and proteomics (Figure 2). We believe that integrating these techniques will improve the detection and staging of colorectal cancer and allow more precise prediction and monitoring therapy responses of individual patients.

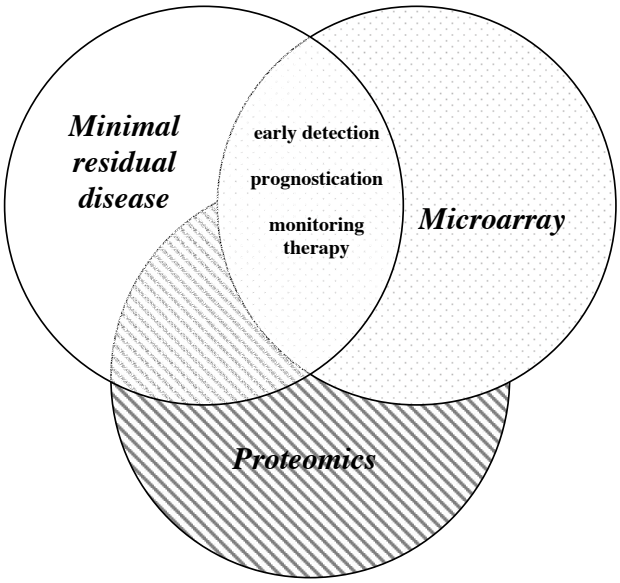


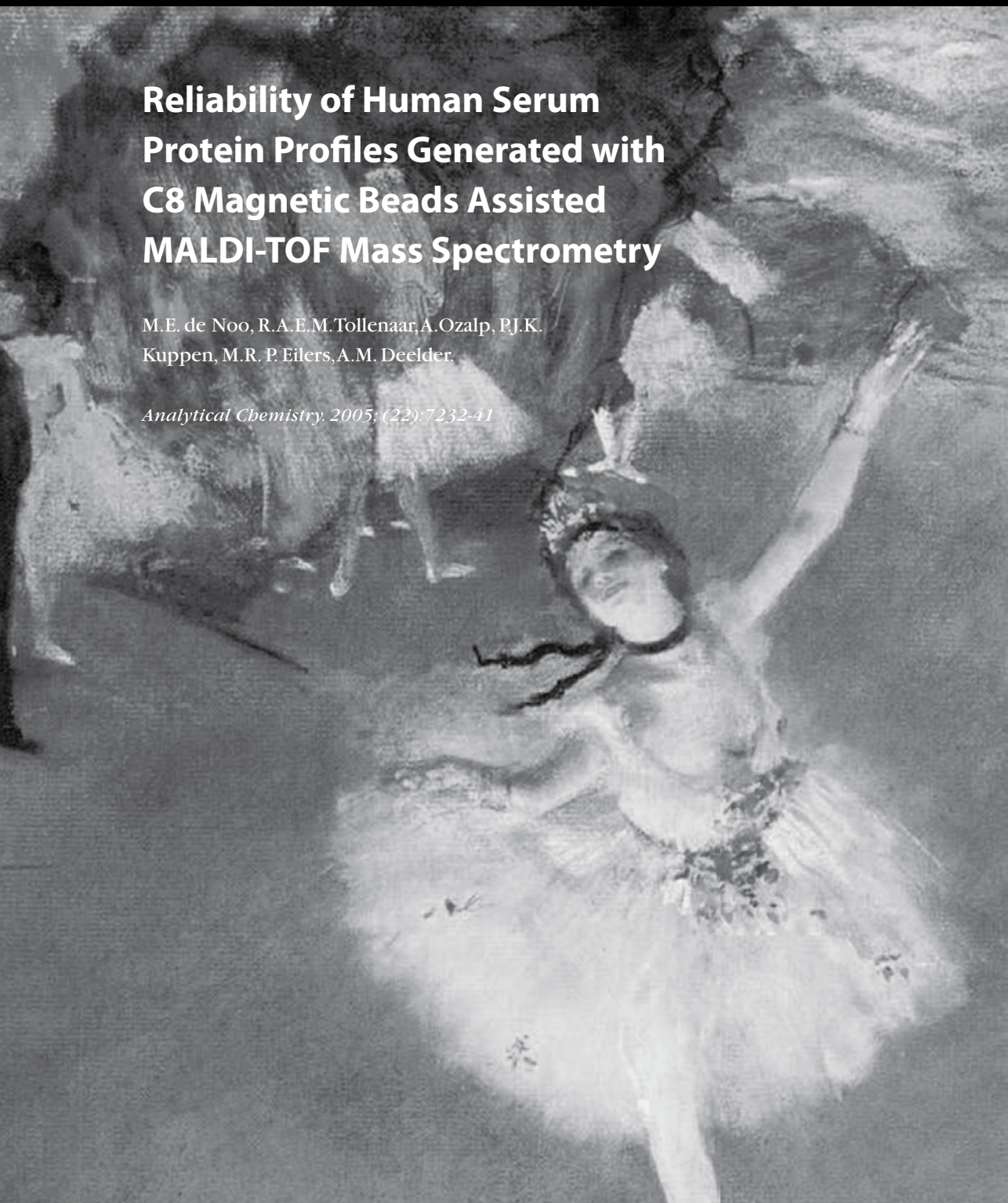
Figure 2. Integrating the three different research techniques will result in a better understanding of the molecular mechanisms of colorectal cancer and will facilitate translating hypotheses derived from basic science into improved patient care. The combination of the different research techniques may result in earlier detection, prognostication and treatment monitoring of colorectal cancer.

REFERENCES

1. Sung,N.S., Crowley,W.F., Jr., Genel,M., Salber,P., Sandy,L., Sherwood,L.M., Johnson,S.B., Catanese,V., Tilson,H., Getz,K., Larson,E.L., Scheinberg,D., Reece,E.A., Slavkin,H., Dobs,A., Grebb,J., Martinez,R.A., Korn,A., and Rimoin,D. (2003) Central challenges facing the national clinical research enterprise. *JAMA*, 289, 1278-1287.
2. Berwick,D.M. (2003) Disseminating innovations in health care. *JAMA*, 289, 1969-1975.
3. Srinivas,P.R., Verma,M., Zhao,Y., and Srivastava,S. (2002) Proteomics for cancer biomarker discovery. *Clin.Chem.*, 48, 1160-1169.
4. Posadas,E.M., Simpkins,F., Liotta,L.A., Macdonald,C., and Kohn,E.C. (2005) Proteomic analysis for the early detection and rational treatment of cancer--realistic hope? *Ann.Oncol.*, 16, 16-22.
5. 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.
6. 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.
7. 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.
8. 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.
9. Petricoin,E.F., Ardekani,A.M., Hitt,B.A., Levine,P.J., Fusaro,V.A., Steinberg,S.M., Mills,G.B., Simone,C., Fishman,D.A., Kohn,E.C., and Liotta,L.A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359, 572-577.
10. Wulfkuhle,J.D., Liotta,L.A., and Petricoin,E.F. (2003) Proteomic applications for the early detection of cancer. *Nat.Rev.Cancer*, 3, 267-275.
11. Boguski,M.S. and McIntosh,M.W. (2003) Biomedical informatics for proteomics. *Nature*, 422, 233-237.
12. 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.
13. Yasui,Y., Pepe,M., Thompson,M.L., Adam,B.L., Wright,G.L., Jr., Qu,Y., Potter,J.D., Winget,M., Thornquist,M., and Feng,Z. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics.*, 4, 449-463.
14. 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.
15. Coombes,K.R., Fritsche,H.A., Jr., Clarke,C., Chen,J.N., Baggerly,K.A., Morris,J.S., Xiao,L.C., Hung,M.C., and Kuerer,H.M. (2003) Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin. Chem.*, 49, 1615-1623.
16. 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.

17. Liefers,G.J. and Tollenaar,R.A. (2002) Cancer genetics and their application to individualised medicine. *Eur.J.Cancer*, 38, 872-879.
18. Liefers,G.J., Cleton-Jansen,A.M., van de Velde,C.J., Hermans,J., van Krieken,J.H., Cornelisse,C.J., and Tollenaar,R.A. (1998) Micrometastases and survival in stage II colorectal cancer. *N.Engl.J.Med.*, 339, 223-228.
19. Doekhie,F.S., Peeters,K.C., Tollenaar,R.A., and van de Velde,C.J. (2004) Minimal residual disease assessment in sentinel nodes of breast and gastrointestinal cancer: a plea for standardization. *Ann.Surg.Oncol.*, 11, 236S-241S.
20. Mesker,W.E., Doekhie,F.S., Vrolijk,H., Keyzer,R., Sloos,W.C., Morreau,H., O'Kelly,P.S., de Bock,G.H., Tollenaar,R.A., and Tanke,H.J. (2003) Automated analysis of multiple sections for the detection of occult cells in lymph nodes. *Clin.Cancer Res.*, 9, 4826-4834.
21. Klein,C.A., Blankenstein,T.J., Schmidt-Kittler,O., Petronio,M., Polzer,B., Stoecklein,N.H., and Riethmuller,G. (2002) Genetic heterogeneity of single disseminated tumour cells in minimal residual cancer. *Lancet*, 360, 683-689.
22. Solakoglu,O., Maierhofer,C., Lahr,G., Breit,E., Scheunemann,P., Heumos,I., Pichlmeier,U., Schlimok,G., Oberneder,R., Kollermann,M.W., Kollermann,J., Speicher,M.R., and Pantel,K. (2002) Heterogeneous proliferative potential of occult metastatic cells in bone marrow of patients with solid epithelial tumors. *Proc.Natl.Acad.Sci.U.S.A*, 99, 2246-2251.
23. Al Hajj,M., Wicha,M.S., Benito-Hernandez,A., Morrison,S.J., and Clarke,M.F. (2003) Prospective identification of tumorigenic breast cancer cells. *Proc.Natl.Acad.Sci.U.S.A*, 100, 3983-3988.
24. Schmidt-Kittler,O., Ragg,T., Daskalakis,A., Granzow,M., Ahr,A., Blankenstein,T.J., Kaufmann,M., Diebold,J., Arnholdt,H., Muller,P., Bischoff,J., Harich,D., Schlimok,G., Riethmuller,G., Eils,R., and Klein,C.A. (2003) From latent disseminated cells to overt metastasis: genetic analysis of systemic breast cancer progression. *Proc.Natl.Acad.Sci.U.S.A*, 100, 7737-7742.

Chapter 3



Reliability of Human Serum Protein Profiles Generated with C8 Magnetic Beads Assisted MALDI-TOF Mass Spectrometry

M.E. de Noo, R.A.E.M. Tollenaar, A. Ozalp, P.J.K. Kuppen, M.R. P. Eilers, A.M. Deelder.

Analytical Chemistry. 2005; (22):7232-41

ABSTRACT

Protein profiling with mass spectrometry is a promising approach for classification and identification of biomarkers. However, there is debate about measurement quality and reliability. Here we present a pipeline for pre-processing, statistical data analysis and presentation. Serum samples of sixteen healthy individuals are used to generate protein profiles with a high-resolution MALDI-TOF after isolation of peptides with C8 magnetic beads. Analysis of variance (ANOVA) was performed after binning, normalization and baseline correction of the mean spectra. Relative variations in the spectra are expressed as coefficient of variation (CV), which depending on the respective preanalytical variation parameter investigated, was found to range between 0.15 and 0.67 in this study. With this novel method the reproducibility of our protein profiling procedure could be quantified. We showed that circadian rhythm and the number of freeze-thaw cycles had relatively limited influence on serum protein profiles, whereas the period between collection and serum centrifugation had a more pronounced effect.

INTRODUCTION

Proteomic pattern diagnostics is a recent and potentially revolutionary technology and approach for early disease detection, surveillance, and monitoring in oncology.[1] In proteomics proteins and functional protein networks as well as their dynamic alteration during physiological and pathological processes are characterised. It is a potential powerful tool in the discovery of disease biomarkers, as the proteome reflects both the intrinsic genetic program of an organism and the impact of its immediate environment.[2] Human serum contains thousands of peptides, most of which are thought to be fragments of larger proteins, but their precise nature remains largely undetermined. High throughput mass spectrometry can generate a proteome/peptidomic fingerprint of a given body fluid, such as serum. Patterns of these peptides can be correlated to biological events occurring in the entire organism and are likely to change in the presence of disease. In oncology new types of bioinformatic pattern recognition algorithms have been used to identify patterns of protein changes in order to discriminate cancer patients from healthy individuals.[3] Furthermore, different profiles may be associated with varying responses to therapeutics and other clinically relevant parameters and may also serve as prediction for treatment outcome. Several studies have shown that biomarkers can be identified on the basis of the presence/absence of multiple low-molecular-weight serum components using time-of-flight (TOF) mass spectrometry technologies such as SELDI-TOF and MALDI-TOF.[4-7] In general, although most studies measure serum components in a range in which primarily peptides and protein degradation products as well as small proteins are detected, the term protein profiling is generally accepted to describe this approach. Although essentially imprecise, this term will also be used in this study. Petricoin et al. showed that patterns of low-molecular-weight serum proteins reflect the pathological state of organs. In addition, these disease-related protein patterns could be useful in the early detection of ovarian cancer.[8] Based on discriminating serological protein profiles that study showed a sensitivity of 100%, specificity of 95% and a positive predictive value of 94% for the detection of ovarian cancer.

Although serum protein patterns have shown high sensitivity and specificity as an early diagnostic tool in several studies, critical notes have been made on biological variation, pre-analytical conditions and analytical reproducibility of serum protein profiles, which would make it difficult to differentiate a normal from a pathological and/or malignant status.[9] In addition, the reproducibility of serum protein profiles has been questioned, however more with respect to the bioinformatical analysis of the measured protein profiles than to the capturing and measuring techniques itself.[10-12] Thus, if proteomics spectra are ultimately to be applied in a routine clinical setting, collection and processing of the data will need to be subject to stringent

quality control procedures.[13] In fact, some critics argue that discriminating protein profiles are so far based more on experimental artefacts than on real biological differences.[14]

There are many factors that are thought to have an influence on serum protein profiles, complicating clear and unambiguous study findings. These factors include environmental and individual factors such as race, age, diet, smoking, stress, general physical condition and use of drugs, which all may influence serum protein profiles. Pre-analytical conditions of human serum also appear to influence protein pattern outcomes. So far, only a few studies have reported on the effects of different serum sample preparations and the use of a magnetic beads based approach to capture and concentrate serum proteins for MALDI-TOF mass spectrometry.[15-17] Since data processing and statistical analysis of protein spectra are essential elements in clinical proteomics, the objective of this study was to quantify the relative contributions of sources of variability on the protein spectra. To this end we developed a novel data processing pipeline, which was performed with an analysis of variance (ANOVA) of the spectra, after the spectra had been made comparable, reduced to common mass channels and the noise had been filtered. Strong baselines were always present in the spectra and had to be removed. This novel analysis method was used to assess the effect of variable pre-analytical conditions on human serum protein profiles, and their effect on reproducibility. In contrast to the above-mentioned study, we have chosen to primarily focus on assaying serum with C8 magnetic beads with hydrophobic functionality, followed by MALDI-TOF analysis. In line with the logistic conditions in a routine clinical setting, the effects of sample handling and storage, and also circadian rhythm factors on the serum protein profiles were analysed.

MATERIAL AND METHODS

Serum samples

Blood was collected from 16 healthy adult volunteers, 8 men and 8 women, by antecubital venipuncture. All blood samples were drawn from the left arm while the volunteers were seated. Approximately 10 ml venous blood was collected in a 10 cc Serum Separator Vacutainer Tube (BD Vacutainer Systems, Preanalytical Solutions, Plymouth, UK) at three different time points throughout the day. The first sample was drawn between 8 and 9 a.m. when all individuals had been fasting since midnight. The second specimen was obtained half an hour after lunch, between 1 and 2 p.m. and the last sample between 5 and 6 p.m. Thirty minutes after collection serum was separated by centrifugation at 3,000rpm for 10 minutes, divided into aliquots (Greiner) and stored at -70ºC. The serum procurement, data management and blood

collection protocol were according to the guidelines of the Medical Ethical Committee of the Leiden University Medical Center. Informed consent was obtained from all subjects.

Protein profiling

To enhance signal quality magnetic beads based on hydrophobic interaction chromatography (MB-HIC kit, Bruker Daltonics, Leipzig, Germany) were used for sample preparation prior to MALDI-TOF mass spectrometry analysis. Five µl of serum was diluted with 10 µl binding solution and 5 µl magnetic beads were added. The solution was mixed by carefully pipetting five times. After 30 seconds supernatant was separated from the magnetic beads in a magnetic beads separator (MBS, Bruker) and discarded. This was followed by three washing steps with 100 µl wash solution (MB-HIC kit, Bruker) and supernatant was discarded each time. After 1 minute in 10µl e lution solution (50% Acetonitrile) the magnetic beads were separated in the MBS from the elution solution. An amount of 1 µl of this eluate, containing the captured peptides/proteins, was mixed with 10 µl matrix solution and 1 µl of this mixture was transferred to an Anchor Chip target plate ™ (Bruker Daltonics, Bremen, Germany) and allowed to dry before introduction into the mass spectrometer. Alpha-cyano-4-hydroxycinnamicacid (HCCA) was used as matrix (0.3 mg/ml in Ethanol: Acetone 2:1). Each sample was deposited onto four spots of the target plate. 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 the employed serum preparation peptide/protein peaks in the m/z range of 1500 to 10,000 were measured. An independent mass spectrometer operator performed all measurements with blinded samples. 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 spectra were compiled, and qualified mass peaks with mass-to-charge ratios (m/z) between 1500 and 10,000 were auto-detected. Each mass spectrum, as exported in an ASCII file, consisted of approximately 45,000 pairs of mass-to-charge values (Dalton) and ion counts. As we preferred to analyze the data using the intensity of the mass spectra per bin, the first processing step was to collect and average the data in bins of 1 Dalton wide. To reduce noise the Whittaker smoother was applied, using second differences, $\lambda = 100$ and weights proportional to the number of raw data points per

bin.[18] The resulting spectra generally showed strong baselines, which had to be removed before further processing. We used the asymmetric least squares algorithm as described in the appendix of Eilers 2004 [19]; figure 1 shows a typical example. The intensity scale of the baseline-corrected spectra was un-calibrated. To normalize the spectra we divided each mass spectrum by the median of the intensities. We consider this to be more robust than normalization on the average (or equivalently, the area under the curve), as the median is less sensitive to spurious large peaks. While this is an ad-hoc solution, we hope to find relatively stable regions in the spectra, so that we can normalize on medians over these regions in further research.

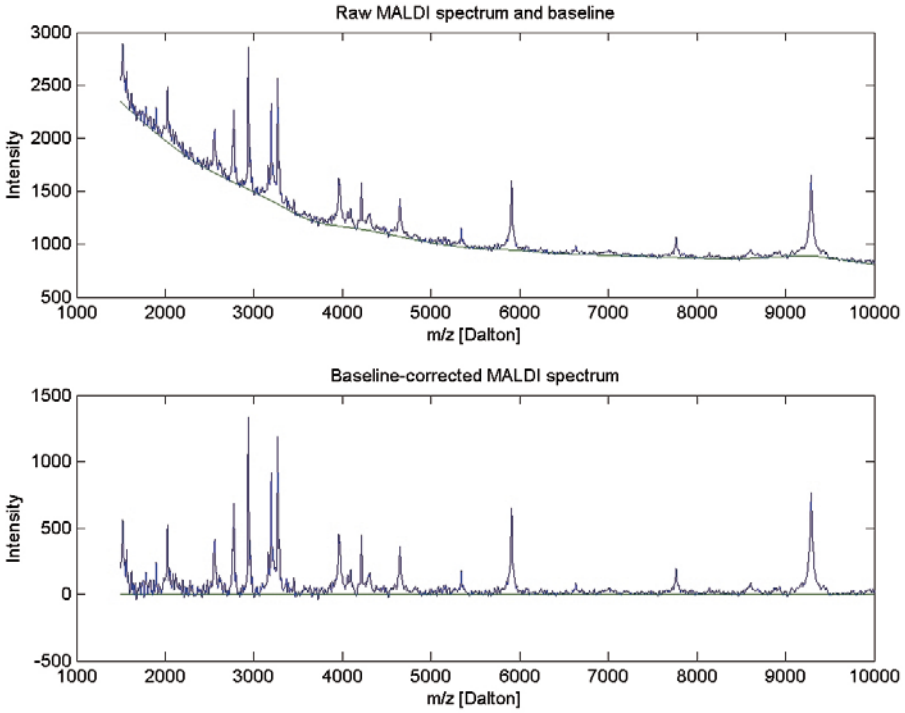


Figure 1. MALDI-TOF spectrum before and after baseline correction.

Statistical data analysis

To quantify the effects of experimental conditions, variability between individual persons and noise, we applied analysis of variance (ANOVA). Consider, as an example, an experiment in which we have P subjects, T storage times and C storage temperatures points. For each combination of subject and time we have measured a spectrum. First, we concentrate on only one, arbitrary, mass channel. We have PTC measurements, which we indicate by y_{ptc}. The ANOVA model assumes that $y_{ptc} = \mu+\alpha_p+\beta_t+\gamma_c+e_{ptc}$. Here μ is the overall mean, α_p is the effect of person P, β_t the ef-

fect of storage time t, γ_c is the effect of storage temperature s and e_{ptc} is random variation. The values of μ and the vectors α, β and γ that minimize the sum of the squares of the elements of e are the so-called least squares estimates and the standard result of ANOVA. If all combinations of persons, storage times and temperatures are present, they are the averages per person (storage time, temperature) over all spectra. When some combinations are missing, a somewhat more complicated regression approach has to be used.

The ANOVA was performed for each bin on the mass axis. This results in 1) one spectrum for μ, the average spectrum; 2) P spectra of person effects; 3) T spectra of time effects; 4) C spectra of storage temperature effects and a spectrum of s, the standard deviation of the noise for each sample. The single spectra of μ and s are easy to present and study, but the multiple spectra of the effects can be voluminous. We summarised them by computing standard deviations of α, β and γ per mass bin. The final results are a plot of five spectra for each of the performed experiments, but only shown in figure 3. The plot shows that, generally, the standard deviations increase when the overall mean increases. A simple measure of this relationship would be the coefficient of variation, like s/μ or s_α/μ, where s_α indicates the standard deviation of α. Unfortunately, this can provide wildly fluctuating results when μ is near zero. Therefore, we computed $cv = \Sigma s_i \mu_i / \Sigma_i \mu_i^2$, which is the slope of a regression line through the origin in a scatter plot of s vs. μ. The summation can be over the whole mass range; this result is reported as a number in the title of each graph of standard deviations for all experiments. In addition, we graphically present CV as computed in m/z windows 500 Dalton wide.

To investigate the influence of the effective bin width on computed CVs, we varied the smoothing parameter λ over a large range, artificially increasing peak width up to five times.

EXPERIMENTS

Reproducibility

In a first set of experiments both the reproducibility of repeated measurements of the same eluate and the reproducibility of repeated analysis of the same samples on four different days were determined. Serum samples of 8 randomly chosen individuals, drawn at one time point during the day were used. Each of these serum samples was processed only once, and measured 8 times with MALDI-TOF according to the standard protocol. Additionally, to determine the inter-measurement variation, protein profiling from 4 of the 8 serum samples was performed on 4 consecutive days. In all experiments samples were preparated just before each MALDI-TOF measurement.

Sample handling

To simulate 'realistic' logistical factors, the effects of sample handling and storage temperature prior to serum centrifugation on serum protein profiles were studied. Serum samples of 4 out of the 16 randomly chosen individuals, all drawn at the same time point were used for this experiment. From each individual 7 aliquots were stored at both room temperature and the same number at 4 °C. After a period of 30 minutes, 1 hr, 2, 4, 8, 24 and 48 hours serum samples were processed according to the standard protocol and protein profiles of all samples were compared for each individual.

Freeze-thaw cycles

To determine the utility of (archival) serum banking, effects of multiple freeze-thaw cycles on serum protein profiles were determined. Serum, drawn at one time point, of 8 randomly chosen individuals was used. Serum of each individual was divided into 11 primary aliquots. From each serum sample one aliquot was measured within 30 minutes after blood collection. The remaining ten sets were immediately frozen at -70 °C. Four hours after the initial freezing, all aliquots were removed from the freezer. Two aliquots of each sample were left at room temperature and the rest on ice for approximately 2 hours until completely thawed. Following the first freeze-thaw cycle, two samples, one thawed on ice and one at room temperature, were assayed. The remaining sets of aliquots were refrozen at –70 °C for 4 hours. Again one sample of each individual was allowed to stand at room temperature and the rest on ice for 2 hours until completely thawed. Subsequently, two samples were processed and the rest refrozen. This was repeated after respectively three and four freeze-thaw cycles, but all samples were thawed on ice.

Circadian rhythm

In a last set of experiments, effects of at which moment of the day blood was drawn on serum protein profiles were studied by analyzing serum samples of 16 individuals, drawn at three different times over the day. All samples were frozen and thawed once and assayed on one day according to the standard protocol.

RESULTS

The data processing pipeline described above was applied to all our experiments. In a preliminary step the influence of effective bin width was studied. We found that stronger filtering, which corresponds to increasing the effective bin width, broadens peaks in both mean and standard deviation spectra, but that the CV did not change much (less than 20%). Therefore, the subsequent experiments were analysed with bins of 1 m/z.

Reproducibility

A test concerning intra-measurement reproducibility was performed by determining the coefficient of variation (CV) over 8 MALDI-TOF spectra for each subject, as shown in figure 2. The CV of the reproducibility within one measurement was less than 20% for 6 out of 8 subjects. Subject D2 and D5 showed slightly higher CV's of 22% and 29%, respectively.

The inter-measurement reproducibility of 4 serum samples performed on 4 different days is shown in table 1. The range in CV between the spectra within one individual (14-23%) is similar to the CV between the consecutive days after correction for differences between individuals (17-26%). However, the variation in spectra between the 4 consecutive days was minor, with an increase in CV on day 4 - 26% (Figure 3).

Table 1. Coefficient of variation (CV) for inter-measurement reproducibility. The CV was determined over 4 MALDI-TOF spectra of each individual, all measured at consecutive days.

Subject	CV (in %)
F	23
G	20
M	22
R	14

Sample handling

To establish the effects of serum sample handling, an ANOVA was performed for effects of persons, time and temperature and residual variation (Figure 4). After correction for inter-individual differences and residual standard deviations with ANOVA, CV between storages at room temperature or at 4 °C was calculated to be 45 and 50%, respectively. The CV of the samples stored for different periods of time before centrifugation ranged from 42% to 67% (Table 2). There was no correlation between the storage time and the coefficient of variation.

Freeze-thaw cycles

The effects of multiple freeze-thaw cycles on serum protein profiles were determined for 10 sets with various storage circumstances, as set 5 had to be left out of the analysis due to technical problems. Table 3 shows the coefficient of variation between persons for different freeze-thaw cycles. In fresh serum samples (set 1) the CV was highest with 64%. With the growing number of hours that serum samples were stored in the fridge at 4 °C, the CV decreased to a minimum of 24% after 8

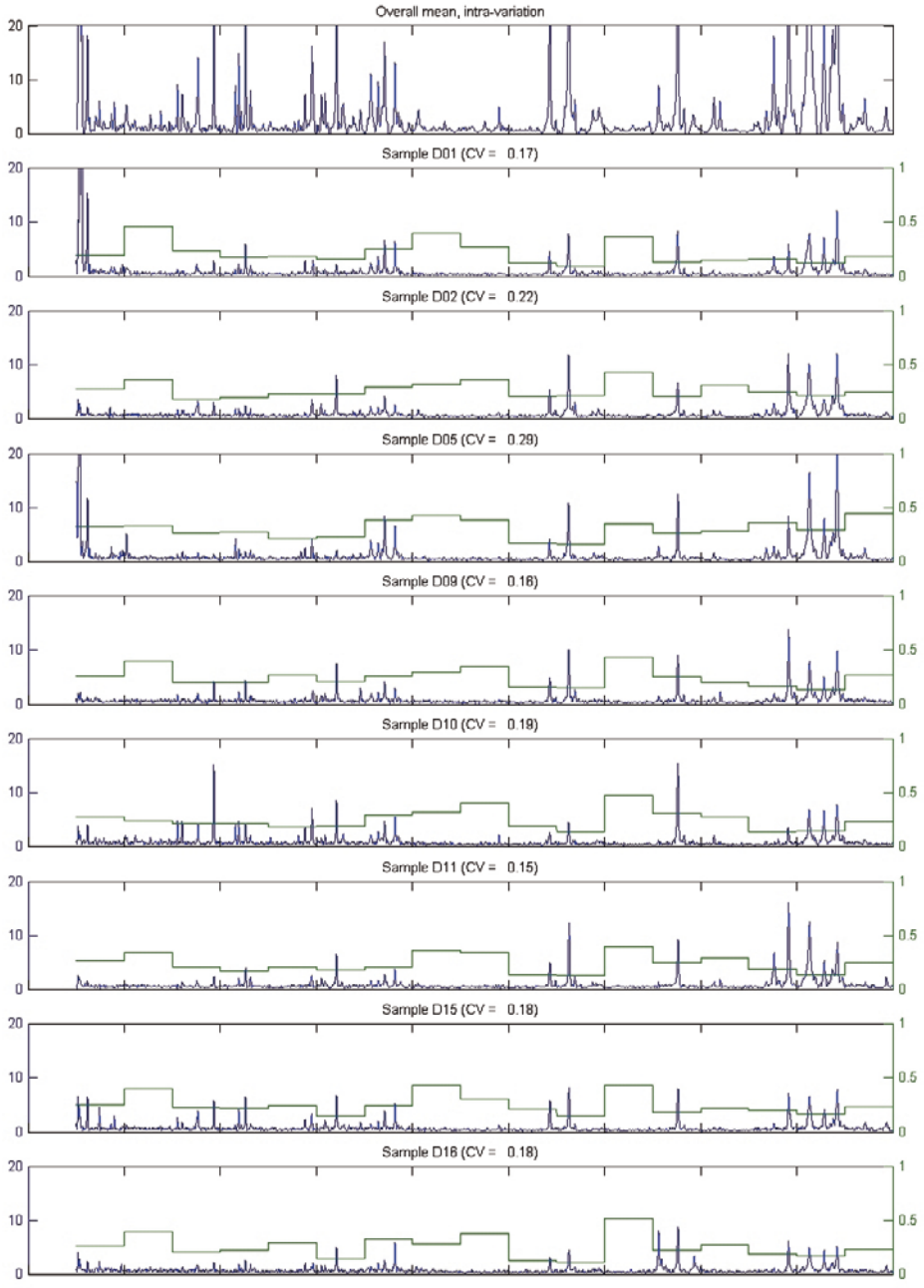


Figure 2. Intra-measurement reproducibility. The CV was determined over 8 MALDI spectra of each individual, all processed in one run.

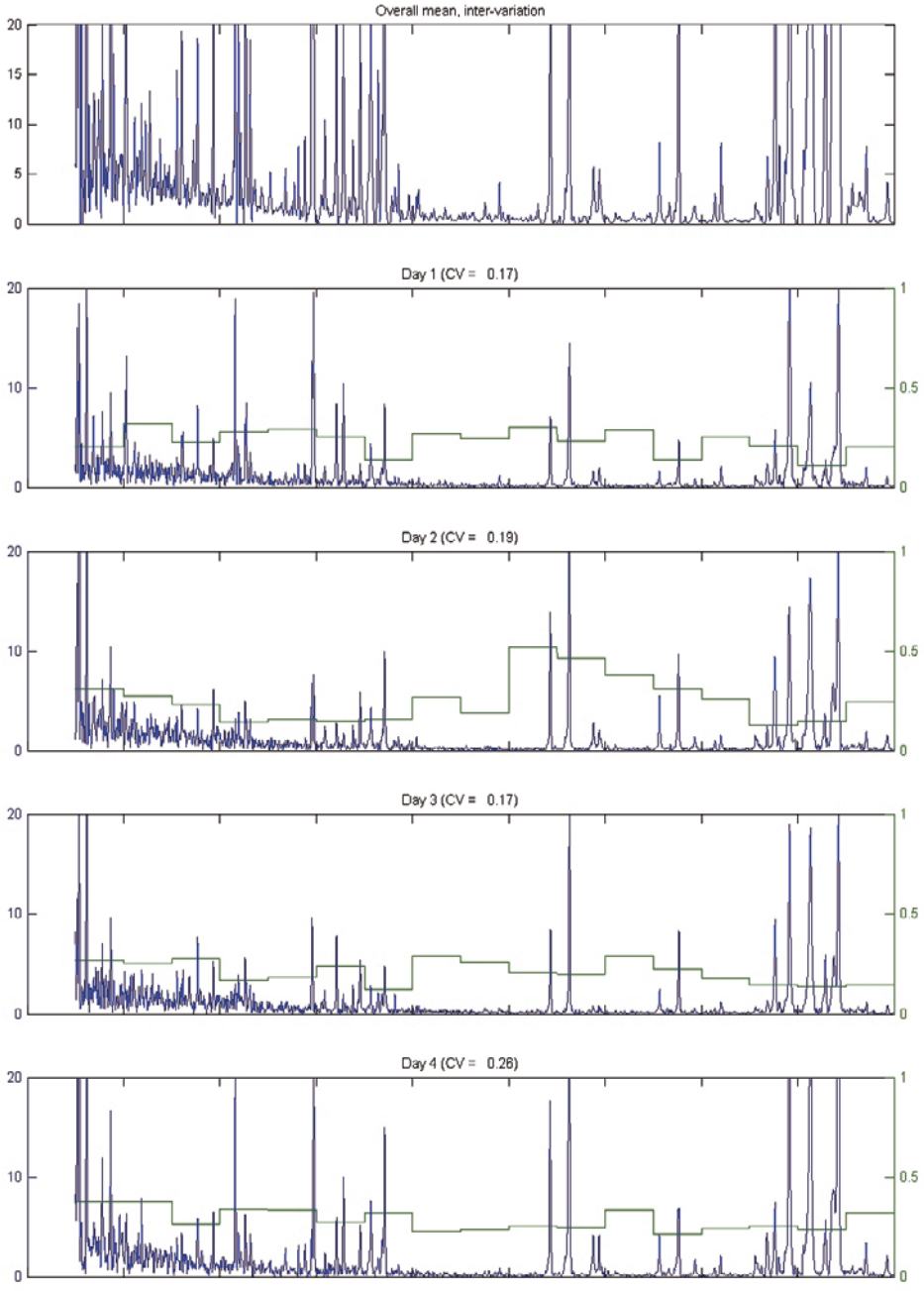


Figure 3. ANOVA for inter-measurement reproducibility. The CV is calculated for spectra that are measured on the same day, after correction for inter individual differences.

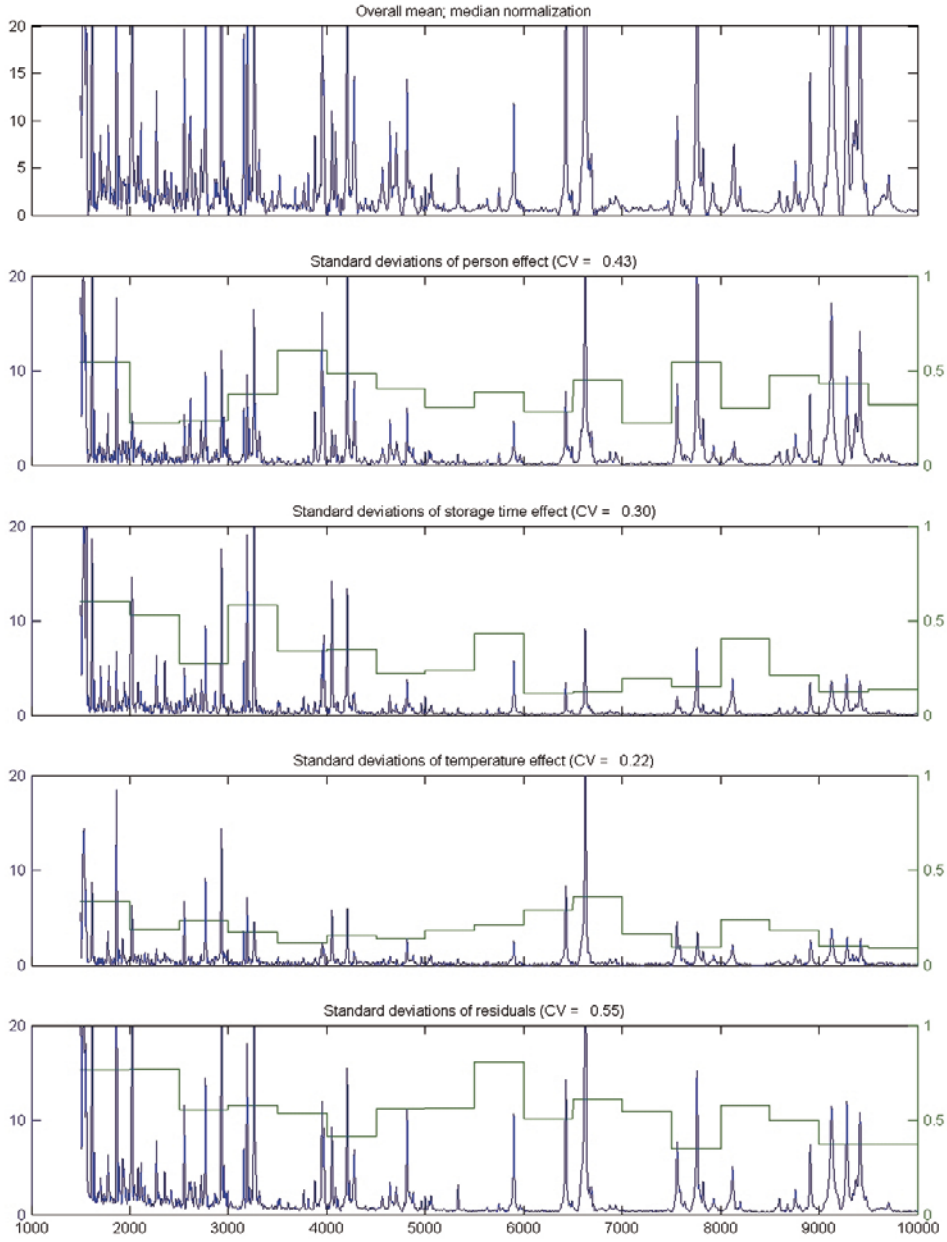


Figure 4. ANOVA of sample handling. From top to bottom: the average spectrum; the variation in spectra due to person's effect; the variation in spectra due to time effects is shown. Finally, the effects of storage temperature variation and the standard deviation of the noise for each sample are presented.

Table 2. Coefficient of variation (CV) between storage times of venous blood before serum centrifugation, regardless the temperature of storage

Storage time	CV (in %)
30 min	52
1 hr	42
2 hrs	63
4 hrs	53
8 hrs	49
24 hrs	52
48 hrs	67

Table 3. Coefficient of variation (CV) between persons per freeze-thaw set. Each set consisted of serum samples of 8 subjects. Each set was stored under different circumstances, namely after none or 1 to 4 freeze-thaw cycles. Sets 1 to 4 were not frozen at all, but stored at 4 ° C during different periods of time.

Set	No freeze-thaw cycles	Temp	CV (in %)
1	0	21 ° C	64
2	0 (2 hrs)	4 ° C	40
3	0 (4 hrs)	4 ° C	39
4	0 (8 hrs)	4 ° C	24
6	1	on ice	26
7	1	21 ° C	39
8	2	on ice	28
9	2	21 ° C	25
10	3	on ice	28
11	4	on ice	28

hours. The number of freeze-thaw cycles had no influence on the CV. All CV of sets 6 to 11 were smaller than 28%, with exception of set 7 (thawed at room temperature after one cycle) with a CV of 39%.

To get an impression of the patterns of change with freeze-thaw cycles, we applied the following procedure to each of the 8 subjects: 1) selected the 8 spectra of the reference set spectra; 2) subtracted these reference spectra from the individual spectra of all sets; 3) regressed (per spectrum) the absolute value of the corresponding reference spectrum, to calculate a CV. The so computed CVs are presented in figure

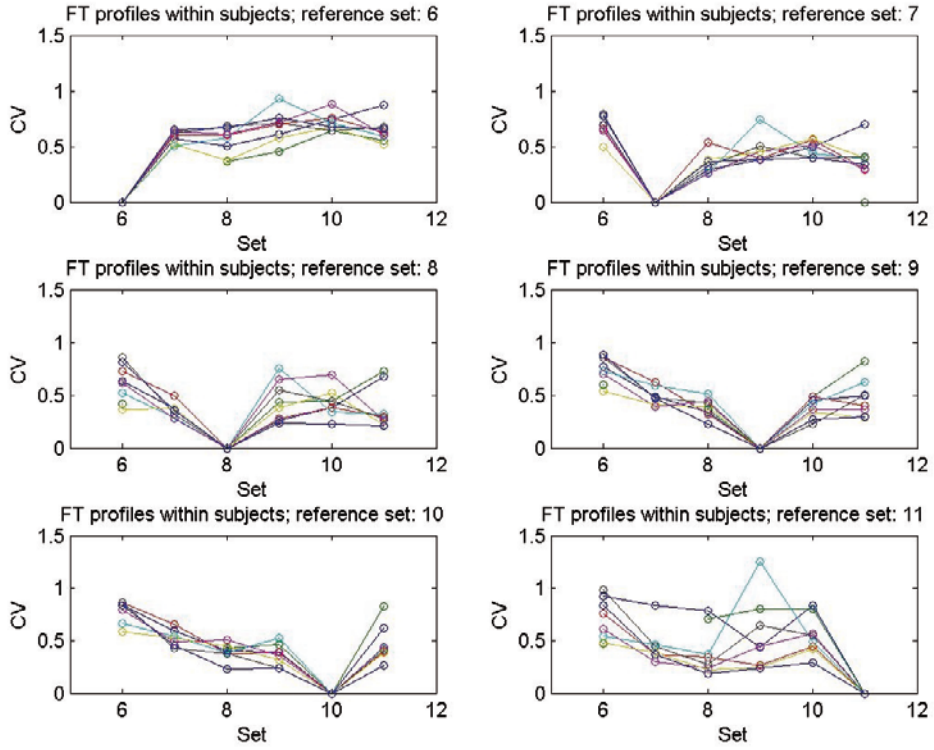


Figure 5. Coefficient of variation between the samples of one reference set and set 6 to 11. On the Y-axis the CV is stated . The sets of the freeze-thaw experiment, as described in table 3, are represented on the X-axis.

5. Generally, all 8 subjects showed the same patterns per reference set, with one or two outliers. In contrast to the other reference sets, 6 showed a continuous increase in CV per extra set. Between the reference sets 7 and 10, the patterns became more identical and the CV was decreased over all sets. In set 11, the variation between the subjects increased, became more variable.

Circadian rhythm

The effect of time variation of blood drawing on serum protein profiles is shown in table 4. Spectra in set 1, collected at 8 a.m. when subjects were fasting, showed a CV between the 8 individuals of 51%. Set 2, collected half an hour after lunch, and set 3, drawn at the end of the afternoon, non-fasting, resulted in 44% and 55%, respectively. No large difference in CV was found between the three sets.

Table 4. Coefficient of variation (CV) between individuals per time point of blood collection. Each set consists of serum samples of 16 subjects, all drawn at time point as indicated in the table.

Set	Time of blood drawing	Fasting	CV
1	8-9 a.m.	Yes	0.51
2	13-14 p.m.	No	0.44
3	18-19 p.m.	No	0.55

DISCUSSION

So far, in a limited number of studies, proteomics-based approaches have shown promising results for the generation of diagnostic profiles in serum. Substantial attention was given to analyze low molecular weight protein patterns from easy-accessible body fluids. To qualify as a future diagnostic test, the entire procedure of protein profiling should be easy to use, robust, reproducible and affordable.[15] High-throughput will also be essential for embedding protein profiling in the clinical setting. The use of fractionation protocols, such as reversed phase magnetic beads, to reduce the complexity of biological samples in MALDI-TOF is needed to avoid signal suppression effects.[20] Therefore, direct analysis of serum is not feasible. In this study we have chosen to use the increasingly accepted C8 magnetic bead capturing technique, taking into consideration that only a small fraction of proteins, from the potential ten thousands of proteins and peptides in human serum can be analysed with this approach. In future studies we will evaluate capturing techniques with different functionalities. In our MALDI-TOF experiments we obtained 'rich' mass spectra, containing many peaks and showing much detail. Our novel data processing pipeline proved to be an effective tool for quality assessment. Baseline correction, binning and filtering provided uniformly structured data in which most typical artefacts had been removed. The ANOVA algorithm separates the sources of variation and provides easily understood numerical summaries of their relative strength.

There is much room for further improvement and refinement. Calibration of the spectra is now based on the median over the domain of interest (1500 to 10,000 Dalton). This is a natural, but rather arbitrary choice. It would be attractive if stable areas in spectra could be located on which to base calibration, or if a reliable spiking procedure was available.

The ANOVA assumed an additive model for the spectral intensities, which is acceptable to compare the relative influence of logistical factors. However, one could argue that a multiplicative model might hold as well, or perhaps even better. It is not possible to simply take logarithms and replicate the ANOVA, as many mass channels

contain negative numbers after baseline correction, caused by noise. A threshold may solve this problem, but overall coefficients of variation depend on the level of this threshold.

We have analysed the data in the form of binned spectra. An alternative approach is to detect individual peaks and analyze peak lists.[21;22] For our purpose, quantification of the reproducibility and of the effects of logistical factors, this would offer no advantages. The experiments with increased smoothing showed only a small influence of the effective bin width on the CV. In a peak list, each peak acts like one 'bin' representing a group of highly correlated intensities around it. Whether we compute a local coefficient of variation by averaging over these individual intensities or over a smaller set of representative peak heights makes little difference. A disadvantage of peak lists is the need for finding complete lists for all spectra, because missing peaks complicate the ANOVA. Furthermore, we used the Whittaker smoother to remove noise and in baseline removal.[18] Compared to wavelets, it has the following advantages: one has continuous control over smoothness and one very short Matlab function does all the work, eliminating any need for toolboxes.[22]

With the employed statistical data analysis the intra-measurement experiments showed a good reproducibility. It is generally accepted that factors like matrix composition and ionisation suppression influence the quality of the MALDI spectra, which in turn will always result in a certain degree of variance in intensity of the generated spectra. This phenomenon can be seen in spectra of subject D5. All spectra of this individual were of inferior quality, possibly due to ionisation suppression or poor matrix solvent composition.[23] Ion suppression results from the presence of less volatile compounds that can change the efficiency of droplet formation or droplet evaporation. This in turn affects the amount of charged ion in the gas phase that ultimately reaches the detector and may result in lower quality spectra.[24;25] To minimize these influences, we used HCCA as a matrix and each sample was spotted four times. However, differences in ionisation rate and thus in peak intensity are intrinsic to the technique and have to be accounted for in the statistical analysis.

The inter-measurement reproducibility within one individual corresponded to the intra-measurement reproducibility for all 4 individuals. However, there seems to be a very small but acceptable day-to-day variation between the different experiments. Therefore, we recommend performing all experiments on one day or to correct for day-to-day variation. To further enhance intra and inter-measurement reproducibility application of robotics for sample processing is recommended. Indeed, implementation of an automated procedure on an 8-channel Hamilton STAR® pipetting robot (Hamilton, Martinsried, Germany) did result in a further reduction of the CV (data not shown).

Ultimately, it might be advisable to include a synthetic peptide mix in the generated spectra for external calibration. In larger profiling studies, batch effects should be taken into account in the design of the study.

Moreover, it is interesting to speculate on the potential discriminating power of MALDI spectra. In the reproducibility experiment we found the overall CV of the error to be 0.18 and that of the person effect 0.33. This can be expressed as a reliability coefficient $r = 0.33^2/(0.33^2 + 0.18^2) = 77\%$. This indicates that nearly 80% of variation in spectra is related to differences between persons. This is not a percentage that indicates that on the basis of whole spectra discrimination between individual spectra will be possible. The graphs of CV as computed for windows of 500 Dalton show strong variations, suggesting that better discrimination could in principle be achieved by using selected parts of the m/z domain. Of course, our data were generated from healthy volunteers, so it remains to be determined how much spectra will differ between healthy and diseased persons.

In this study the largest effect was observed for sample handling conditions. There was no correlation between the increasing number of hours before centrifugation and the variation between the serum protein profiles, but the overall variation was larger. This would already justify acceptance of a certain time range after blood collection and before centrifugation. Furthermore it is unlikely that in a hospital's daily practice this factor could be rigorously standardised. Thus, although a standard time period would be ideal, we accept a delay of 0-4 hours between the moment of blood collection and serum centrifugation. In view of the fact that there was no large difference between the storage temperatures and logistical factors, leaving all blood samples to stand at room temperature before centrifugation seems justified.

The effect of increasing numbers of freeze-thaw cycles was small and consistent, with the exception of set 7, in which serum samples were thawed only once at room temperature. The coefficient of variation in this set was larger than in all other sets, as shown in table 3. This might be explained by the fact that protein degradation occurs sooner at room temperature, as also demonstrated in sets 2-4. This phenomenon might be explained by proteolytic activity and the fact that hydrophobic interactions are strengthened, while with increasing temperature the hydrogen bonding is weakened and the electrostatic interactions are not changed due to its entropic origin.[26] Whereas the range in coefficients of variation between increasing numbers of freezes and thaw cycles is small, fresh serum samples provided the largest variation between persons, almost double in comparison to other sets. Furthermore, in fresh serum samples the number of peaks observed was less than 50, as also reported by other groups.[15;27] We suggest that in this early stage of defining optimal parameters/conditions for serum pattern diagnostics the use of fresh serum samples is better avoided. This seems contradictory, as proteolytic activity after thawing implicates a

loss of proteins and peptides and thus of information. However, on the condition that all samples are treated according to a standard protocol, this would not be critical for a black box approach. Thus it would seem that the use of archival material is safe with respect to the effect of freezing and thawing; nevertheless it remains of paramount importance that the entire sample handling and storage procedure is standardised. Based on the fact that the coefficient of inter-group variation in reference set 8 is lower than in the other sets (Figure 5), we prefer to use serum samples for further studies, which have undergone two freeze-thaw cycles. Moreover, our choice is mainly rooted in practical and logistical reasons, as in many large hospitals; sample collection is centralised in the clinical chemical laboratory.

With only minimal variation observed between protein profiles from samples collected at three different time points over the day, circadian rhythm seems to have limited effect on individual serum protein profiles. This is an encouraging fact, as blood samples can be collected all over the day, which increases the future applicability of serum protein profiling in the clinic. Furthermore, there is no indication that fasting has any influence on serum protein profiles, which also facilitates future clinical use.

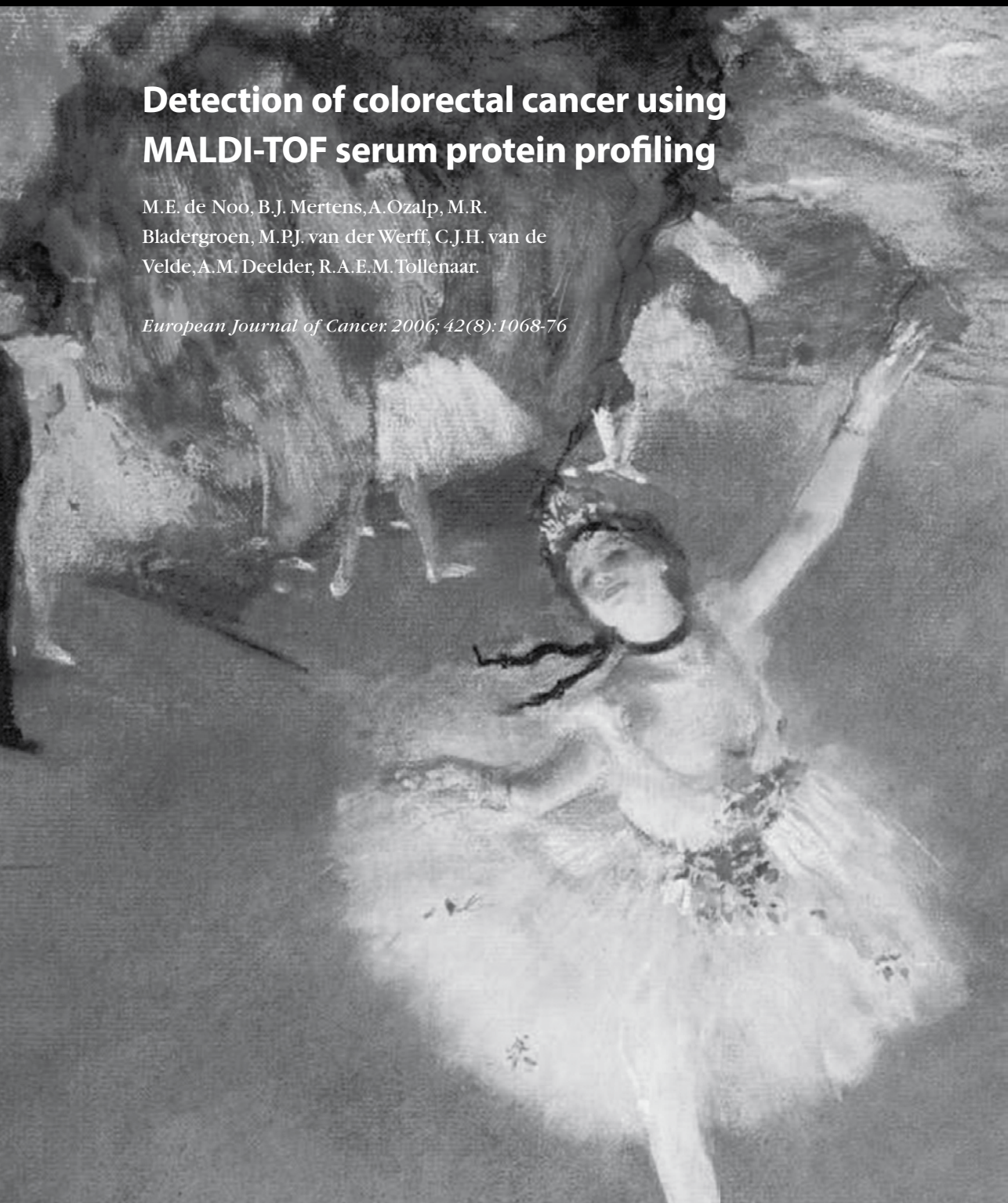
All together, we have presented a method to assess the reproducibility of a protein profiling procedure using a high-end MALDI-TOF. Our appliance of ANOVA over the mean spectra allowed analysis of the effects of handling and storage procedures on serum protein profiles. The results from this study stress the importance of a standardised collection of all blood samples, from the moment of sample handling and storage until freezing the samples in order to prevent bias in classification studies. Although the importance of homogeneity and uniformity within sample groups must be stressed, variation of such factors can not totally be excluded in a clinical setting. The most important issues for discriminating studies at this moment are a standardised and well-documented sample collection and a thorough study design. Based on the present data and those of Villanueva et al.[15], we feel that the methodology can be standardised to a level which allows application as a tool in biomarker discovery. Although it remains to be seen whether actual biomarkers can reliably be identified with the current technique, we are now in the process of carrying out a study to determine whether serum protein profiles can differentiate colorectal cancer patients from individuals with benign bowel disorders and healthy subjects. To this end and to facilitate high-throughput studies, we developed an automated platform for our capturing technique with C8 magnetic beads with reverse-phase based functionality and we used the MS instrument's AutoXecute function to further enhance reproducibility (data not shown). In addition to large clinical studies as mentioned above, such a platform would also be valuable for more large-scale studies as e.g. inter group variance (cases versus controls) under different experimental setups.

REFERENCES

1. Poon,T.C. and Johnson,P.J. (2001) Proteome analysis and its impact on the discovery of serological tumor markers. *Clin.Chim.Acta*, 313, 231-239.
2. Srinivas,P.R., Srivastava,S., Hanash,S., and Wright,G.L., Jr. (2001) Proteomics in early detection of cancer. *Clin.Chem.*, 47, 1901-1911.
3. Wulfkuhle,J.D., Liotta,L.A., and Petricoin,E.F. (2003) Proteomic applications for the early detection of cancer. *Nat.Rev.Cancer*, 3, 267-275.
4. 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.
5. 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.
6. 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.
7. 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.
8. Petricoin,E.F., Ardekani,A.M., Hitt,B.A., Levine,P.J., Fusaro,V.A., Steinberg,S.M., Mills,G.B., Simone,C., Fishman,D.A., Kohn,E.C., and Liotta,L.A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359, 572-577.
9. Boguski,M.S. and McIntosh,M.W. (2003) Biomedical informatics for proteomics. *Nature*, 422, 233-237.
10. 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.
11. Yasui,Y., Pepe,M., Thompson,M.L., Adam,B.L., Wright,G.L., Jr., Qu,Y., Potter,J.D., Winget,M., Thornquist,M., and Feng,Z. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics.*, 4, 449-463.
12. 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.
13. Coombes,K.R., Fritsche,H.A., Jr., Clarke,C., Chen,J.N., Baggerly,K.A., Morris,J.S., Xiao,L.C., Hung,M.C., and Kuerer,H.M. (2003) Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin. Chem.*, 49, 1615-1623.
14. Sorace,J.M. and Zhan,M. (2003) A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC.Bioinformatics.*, 4, 24.
15. 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.
16. Baumann,S., Ceglarek,U., Fiedler,G.M., Lembcke,J., Leichtle,A., and Thiery,J. (2005) Standardized Approach to Proteome Profiling of Human Serum Based on Magnetic Bead Separation and Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. *Clin. Chem.*, 51, 973-980.

17. Yamanishi,H., Kimura,S., Iyama,S., Yamaguchi,Y., and Yanagihara,T. (1997) Fully automated measurement of total iron-binding capacity in serum. *Clin.Chem.*, 43, 2413-2417.
18. Eilers,P.H. (2003) A perfect smoother. *Anal.Chem.*, 75, 3631-3636.
19. Eilers,P.H. (2004) Parametric time warping. *Anal.Chem.*, 76, 404-411.
20. Richter,R., Schulz-Knappe,P., Schrader,M., Standker,L., Jurgens,M., Tammen,H., and Forssmann,W.G. (1999) Composition of the peptide fraction in human blood plasma: database of circulating human peptides. *J.Chromatogr.B Biomed.Sci.Appl.*, 726, 25-35.
21. Semmes,O.J., Feng,Z., Adam,B.L., Banez,L.L., Bigbee,W.L., Campos,D., Cazares,L.H., Chan,D.W., Grizzle,W.E., Izbicka,E., Kagan,J., Malik,G., McLerran,D., Moul,J.W., Partin,A., Prasanna,P., Rosenzweig,J., Sokoll,L.J., Srivastava,S., Srivastava,S., Thompson,I., Welsh,M.J., White,N., Winget,M., Yasui,Y., Zhang,Z., and Zhu,L. (2005) Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. *Clin.Chem.*, 51, 102-112.
22. Morris,J.S., Coombes,K.R., Koomen,J., Baggerly,K.A., and Kobayashi,R. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics.*.
23. Knochenmuss R., Dubois F., Dale M.J., and Zenobi R. The matrix Suppression Effect and Ionization Mechanisms in Matrix-assisted Laser Desoprtion/ Ionization. Rapid Commun.Mass Spectrom. 10, 871-877. 9-5-1996. John Wiley & Sons. Ltd. Ref Type: Generic
24. Cohen,L.H. and Gusev,A.I. (2002) Small molecule analysis by MALDI mass spectrometry. *Anal.Bioanal.Chem.*, 373, 571-586.
25. Annesley,T.M. (2003) Ion suppression in mass spectrometry. *Clin.Chem.*, 49, 1041-1044.
26. Jaenicke,R. and Zavodszky,P. (1990) Proteins under extreme physical conditions. *FEBS Lett.*, 268, 344-349.
27. Wang,M.Z., Howard,B., Campa,M.J., Patz,E.F., Jr., and Fitzgerald,M.C. (2003) Analysis of human serum proteins by liquid phase isoelectric focusing and matrix-assisted laser desorption/ionization-mass spectrometry. *Proteomics.*, 3, 1661-1666.

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 °C 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 α-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

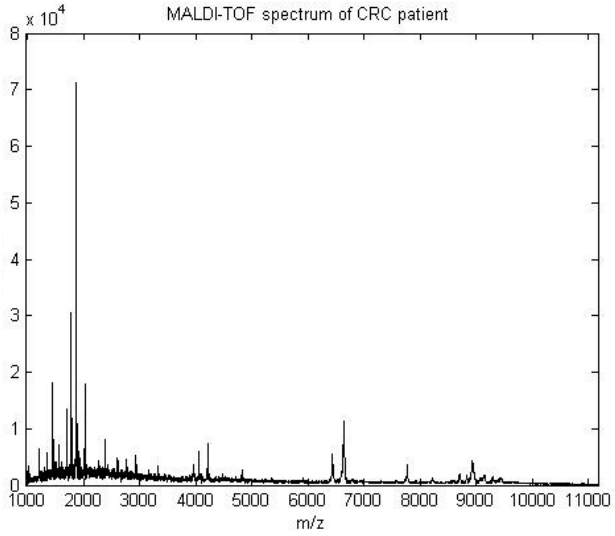


Figure 1a

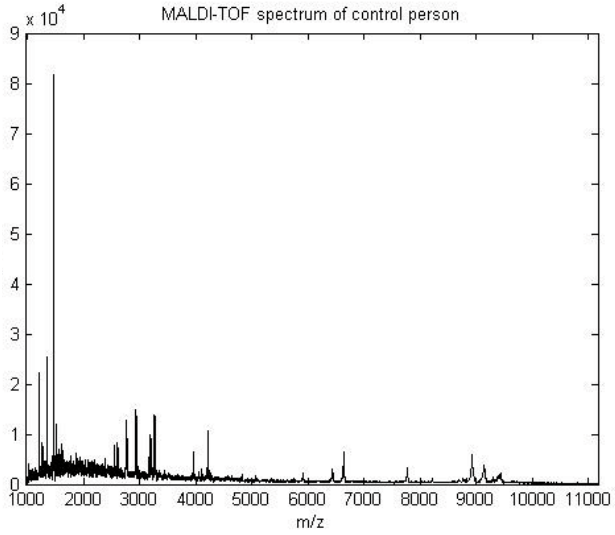


Figure 1b

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

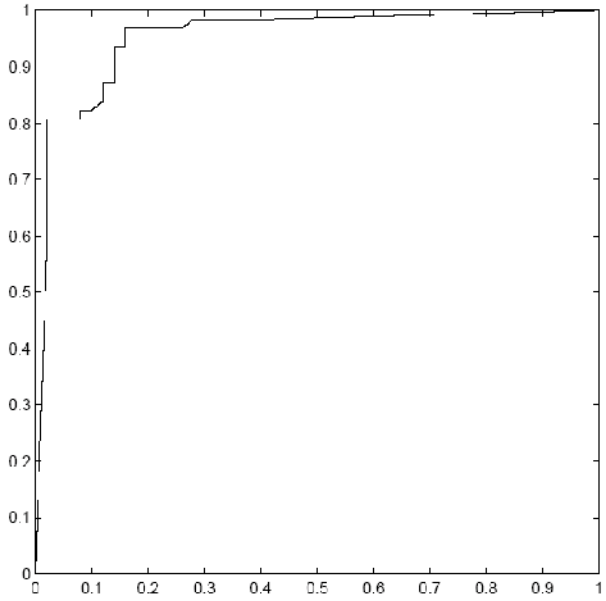


Figure 2. ROC-curve for the double cross-validated two-group classifier. The true positive recognition rate (sensitivity) is demonstrated on the y-axis against the false negative recognition rate (1-specificity) on the x-axis of the classifier.

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 ($p<0.0001$).

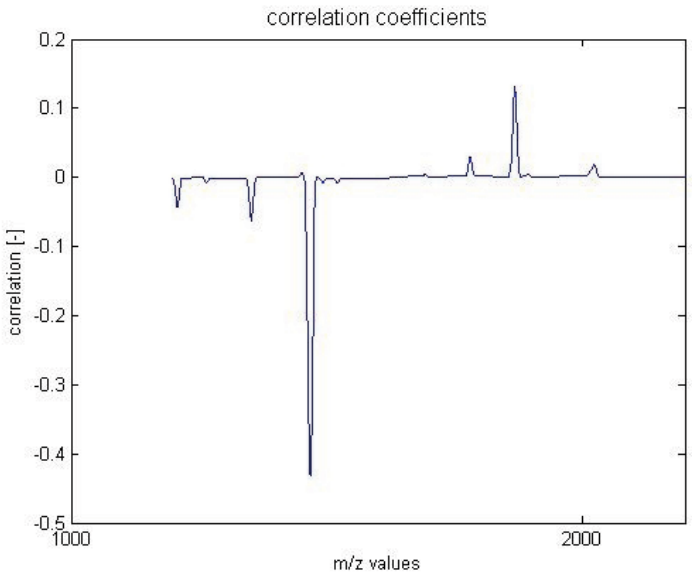


Figure 3. Correlation coefficients of two first principal components with the class indicator. The correlation coefficients were calculated from the linear discriminant weightings. The negative correlation of the first peak is an indicator for the control group and the positive correlation of the second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

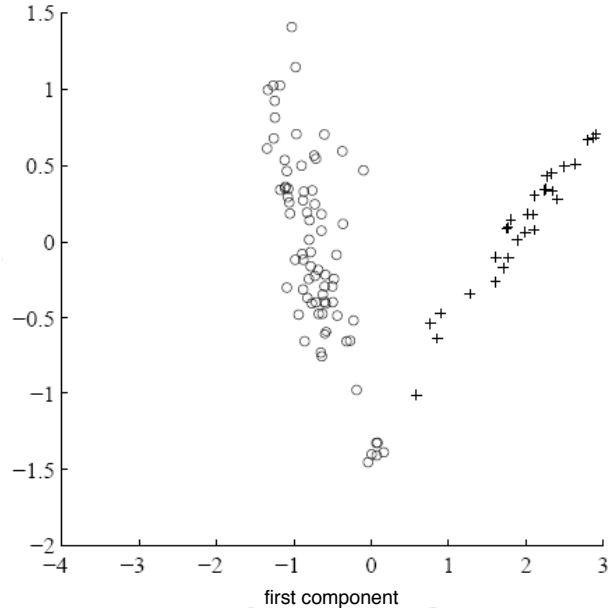


Figure 4. Scatter plot of the first two principle components on basis of which the classification patient-control group was made.

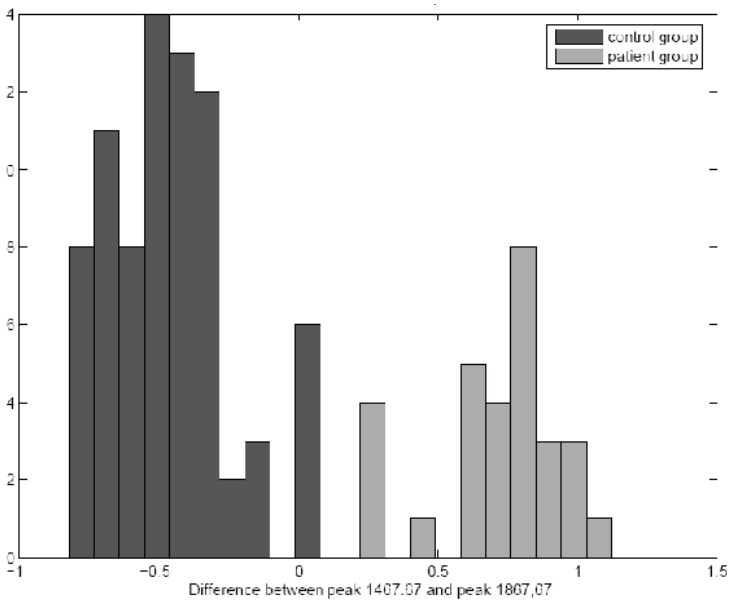


Figure 5. Histogram showing the difference between the normalized intensities of the two most discriminating "peaks" (bins). The X-axis shows the difference between the normalized intensities of the peaks. On the Y-axis the number of subjects is displayed.

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 $\hat{D}_g(\boldsymbol{x}) = (\boldsymbol{x} - \bar{\boldsymbol{x}}_g)\boldsymbol{S}_{(k)}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_g)^T$.

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]

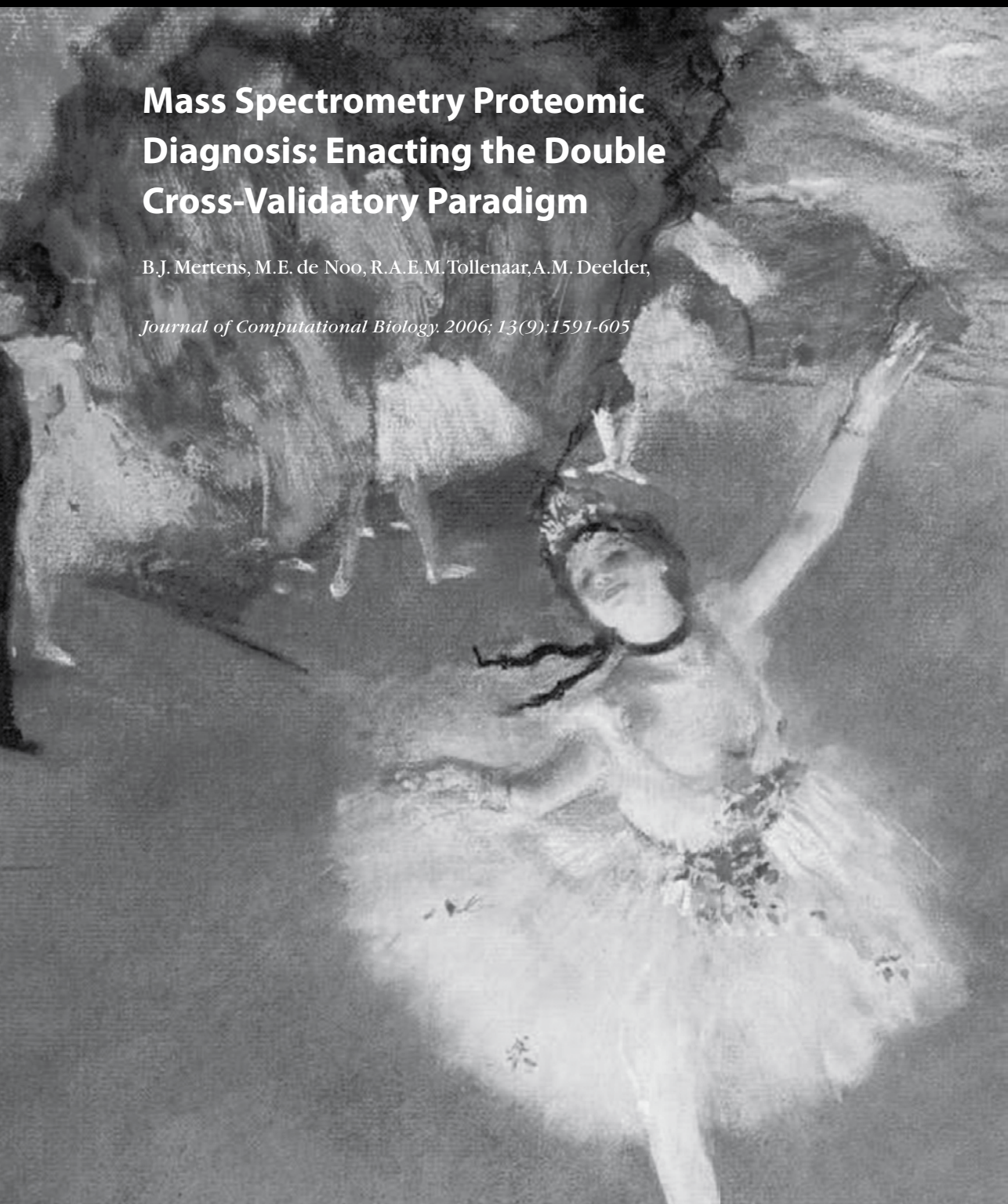
Appendix 1

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.

Chapter 5



Mass Spectrometry Proteomic Diagnosis: Enacting the Double Cross-Validatory Paradigm

B.J. Mertens, M.E. de Noo, R.A.E.M. Tollenaar, A.M. Deelder,

Journal of Computational Biology. 2006; 13(9):1591-605

ABSTRACT

This paper presents an approach to the evaluation and validation of the diagnostic potential of mass spectrometry data in an application on the construction of an 'early warning' diagnostic procedure. Our approach is based on a full implementation and application of double cross-validatory calibration and evaluation. It is a key feature of this methodology that we can jointly optimize the classifiers for prediction while simultaneously calculating validated error rates. The methodology leaves the size of the training data nearly intact. We present application to data from a designed experiment in a colon-cancer study. Subsequent to presentation of results from the double cross-validatory analysis, we explore a post-hoc analysis of the calibrated classifiers to identify the markers that drive the classification.

INTRODUCTION

There is currently much interest in application of mass spectrometry for the construction of new diagnostic proteomic approaches for the early detection of disease. This is particularly the case in oncology, where there is need for new and reliable diagnostic tests. In this paper, we discuss the problem of ascertaining the presence of discriminatory information in mass spectra of serum samples in a case-control study for the detection of colorectal cancer. In other words, we describe -in essence -an early-stage feasibility study for subsequent construction of a diagnostic test based on proteomic mass spectra. A crucial objective of such research is to provide information which allows researchers to make informed decisions as to the continuation of the research effort (which may involve experiments of much greater cost and complexity in comparison to the first-stage evaluation). Hence, it is essential to get a fully validated and unbiased assessment of predictive error rates that may be achieved, based on the proteomic data. At the same time, in a high-dimensional setting such as mass spectrometry, it is desirable that construction of the diagnostic classifier would involve calibration of the predictive potential of the allocation rule itself.

Mass spectrometry proteomics, sample size and clinical science

In problems such as these and related settings (*e.g.*: microarray diagnostics, chemometric discriminant studies), a key difficulty is often the collection of a sufficient number of samples. In oncology applications this may tend to happen, due to logistical and ethical reasons. Our example is a typical one, as our study is a first-stage evaluation within the context of an academic center, which has a typical patient population with more advanced disease. This limits the number of patients available for research. On the other hand, clinicians and biomedical researchers who wish to explore application of proteomic mass spectrometry for the construction of new diagnostic procedures, will be interested first to get an indication of whether there is information in the spectra to allow groups to be separated and what the likely error rates of misclassification will be. This is particularly the case since ethical review boards (or funding authorities) are not inclined to give permission for large-scale collaborative trials between hospitals, which would ease the patient recruitment problem, without preliminary evidence from smaller within-center trials. Both these reasons may conspire to cause proteomic studies to be of small sample size initially.

Classical statistics will often use a separate validation set to optimize a chosen diagnostic classifier for prediction first. Assessment of the rule is then carried out on yet another test set, which must often be set-aside from the available data [1]. Unfortunately, when the amount of experimental data is small to begin with, the

training set left over may be too small to allow researchers to apply this paradigm fully. In this paper, we present a double cross-validatory approach which allows for simultaneous predictive calibration and assessment of the allocation rule, without (substantial) reduction of the size of the calibration data.

Mass spectrometry data

The experiment and data discussed and analysed in this paper are derived from a MALDI-TOF (Matrix Assisted Laser Desorption Ionisation Time-Of-Flight) mass spectrometer (Ultraflex TOF/TOF, Bruker Daltonics, equipped with a SCOUT ion source which was operated in linear mode). The spectrometer produces a sequence of intensity readings for each sample on an ordered set of contiguous bins in the m/z range from 960 to 11,160 Dalton. Bin sizes (length) of the unprocessed spectra gradually increase with increasing m/z values, ranging from 0.07 Dalton at the lower end of the mass/charge scale up to 0.24 Dalton at the upper end of the scale. This gives intensity readings on a fixed grid of 4483 bins within the mass-charge range across all samples. We refer to an earlier paper by our group for detailed information on experimental setup and measurement protocols.[2]

We will discuss the essential aspects of the study design first, followed by a description of the discriminant method and the double cross-validatory approach to joint predictive estimation (calibration) and validation of the allocation rule, which allows for validated error rate evaluation. Subsequent to description of the methodological approach, we consider application to the colon cancer data and present a *post hoc* exploratory data analysis to interpretation of the results. While we will focus on our example to structure the discussion, the issues apply quite generally to similar problems in proteomics and many other related problems in bioinformatics, chemometrics, statistical prediction and beyond. We will assume that the reader has some knowledge of standard leave-one-out cross-validation.

DESIGN AND SAMPLE REPLICATION

Design

A characteristic problem of proteomic mass spectrometry design is the need to cope with the presence of what we may loosely refer to as so-called 'batch effects'. Examples are plate-to-plate variability, day-to-day variation and so on, whose presence is in reality unavoidable. To accommodate these effects, we identify each plate by day combination as a block and employ standard *randomised block design* by randomly distributing the available samples from each group (colon cancer and controls) across the blocks such that proportions are (as near as) equal within and across

blocks for each group. For colon cancer, we randomised samples to plates in such a manner that the distribution of disease stages is in approximately equal proportions within and across plates. The position on the plates of samples allocated to each plate was also randomised. Each plate was then assigned to a distinct day, which completes the design. Table 1 summarizes the design as executed on the first week, which provides mass spectra on 63 colon cancer patients and 50 healthy controls.

Table 1. Design as executed on the first week. A replicate of the entire experiment was run on the subsequent week using plate duplicates. 'Stage' refers to the distribution of cases across the four respective disease stages.

	TNM stage	Plate 1	Plate 2	Plate 3	Total
Cases	I	4	4	3	
	II	10	10	8	63
	III	4	4	4	
	IV	4	4	4	
		22	22	19	
Controls		17	17	16	50
Total		39	39	35	113

In our case, it was decided to carry out the experiment in a single week using three plates only, each of which was assigned to a consecutive day in the middle of the week - Tuesday to Thursday. We refer the reader to the statistical literature on design of experiments for further discussion and details of the issues involved, as well as many other examples of these basic design principles.[3-6]

Sample replication

We can exploit design to augment cross-validatory analysis. This is because while sample sizes may be small (*i.e.* it is difficult to get new independent samples), the amount of sample material available for each sample may be more abundant. This allows the introduction of so-called replicate samples into the design. Since the samples are pre-arranged on rectangular plates, a second 'copy' of any plate can be made provided sufficient sample material is available from each sample. (In our case, sufficient sample material was available for a second copy only). Thus, we can duplicate the entire design from the first week and remeasure the replicate plates through the same design on the second subsequent week, using new sample material from each sample (but of course not new samples themselves). With this

approach, we thus generally have available from each i^{th} sample an observation $x^1_i = (x^1_{i1}, \ldots, x^1_{ip})$ of the associated recorded mass spectrum in the first week, where the vector elements refer to the measured mass/charge intensities on a predefined and ordered grid of mass/charges of dimensionality p. In addition, we have for each sample a duplicate measurement $x^2_i = (x^2_{i1}, \ldots, x^2_{ip})$ obtained from the corresponding replicate on the corresponding plate measured on the same day one week later. We may denote the associated class label from each i^{th} observation as $c(i)$ which takes value in the set of group indicators $\{1,\ldots,G\}$, where G is the number of groups. [Note we will drop use of the suffixes 1,2 when the context makes clear to which week the data relates.] Unfortunately, due to a technical malfunction which occurred on the last day of the second week the replicate measurements from the third plate are unavailable. As a consequence we only have available the 78 replicates from the first 2 plates in week 2 for further analysis.

INTEGRATED CALIBRATION AND VALIDATION FOR CLASSIFICATION BY DOUBLE CROSS-VALIDATION

We restrict attention to double cross-validated linear discrimination for joint calibration and validation.[7] First we discuss shrinkage-based estimation and the need for it in linear discrimination. Then we explain the double cross-validatory implementation.

Linear classification and shrinkage estimation methodology

We base classification on Fisher linear discrimination. There is voluminous literature on the method, which is well established in the applied sciences, such as biology and medicine.[1;8-10] An article by Hastie et al. contains an up-to-date account of many new applications which demonstrate the continuing success of the approach.[1]

Fisher linear discriminant allocation may be defined as assigning a new observation with feature vector $\mathbf{x}$ to the group for which the distance measure

$$Dg(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_g)\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_g)^T$$

is minimal, where g denotes the group indicator with $g \in \{1,\ldots,G\}$, $\boldsymbol{\mu}_g$ the population means and Σ the population within-group dispersion matrix which is assumed equal across groups. In practice, the population means and dispersion matrix will be unknown and hence must be estimated from the data. In a high-dimensional problem such as in mass spectrometry proteomics, this leaves us with a difficulty in estimating the dispersion matrix as we will typically not be able to achieve a full rank estimate.

At the risk of some oversimplification of the discussion, there are basically two ways in which we may remedy the problem so that the above methodology may again be applied. The first is through either selection or construction (or a combination of both) of a set of features which is reduced in dimensionality, while capturing most of the variability in the data. In essence, this is the approach which is currently applied in most of the mass spectrometry proteomics literature. Typical examples are found in papers by Baggerly, Yasui, Sauve and Morris, among others.[11-14] We do not consider this approach to be fundamentally flawed for mass spectrometry proteomic data. On the contrary, it is self evident that mass spectra consist of mixtures of possibly overlaid intensity peaks corresponding to substances present in the analyte. Thus, to elucidate this structure (first) is in principle of interest.

The alternative is not to select in the first instance, but instead explicitly utilize the correlations which are induced between intensities on the mass-charge bins through the associated discretisation of the continuous signal (peaks). The simplest approach is through principal components decomposition [15], which has a long history of successful application in classical spectroscopy such as in near infrared spectroscopy for example Krzanowski et al. [16]

Within this approach, we leave the dimensionality of the data intact and instead introduce a regularised estimation of the dispersion matrix to cope with the singularity of the sample dispersion matrix, based on the component decomposition. We explore two distinct forms of regularization, both of which may be expressed in terms of the spectral decomposition of the 'observed' (or sample) pooled dispersion matrix $\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathrm{T}}$ where $\mathbf{Q}$ and $\mathbf{\Lambda} = \mathrm{diag}\,(\lambda 1, \ldots, \lambda r)$ are the matrices of principal component weights (or loadings) and variances respectively, with $\lambda 1 > \ldots > \lambda r > 0$ respectively (r is the rank of the pooled covariance matrix). The within-group covariance matrix is re-estimated by only retaining the first $1 \leq k \leq r$ components only, which gives an estimate

$$\mathbf{S}(k) = \mathbf{Q}_{(k)} \Lambda_{(k)} \mathbf{Q}^{\mathrm{T}}{}_{(k)},$$

where $\Lambda_{(k)} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_k)$ and $\mathbf{Q}_{(k)}$ denotes the corresponding reduced matrix of component loadings. The associated linear discriminant allocation rule hence assigns observations to the group for which the smallest sample-based distance estimates

$$\hat{D}_g(\mathrm{x}) = (\mathrm{x} - \bar{\mathrm{x}}_g)\, \mathbf{S}^{-1}{}_{(k)}\, (\mathrm{x} - \bar{\mathrm{x}}_g)^{\mathrm{T}}$$

are observed, with $\bar{\mathrm{x}}_g$ the sample group means for $g \in \{1, \ldots G\}$. 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 ($k<r$) is equivalent to so-called shrunken least-squares regression.[8;17] Alternatively, we may employ ridge regularization

$$\mathbf{S}(\gamma) = \mathbf{Q}[(1 - \gamma)\Lambda + \gamma\mathbf{I}]\mathbf{Q}^{T},$$

where $0<\gamma\leq 1$ is the ridge regularization or 'tuning' parameter, in which case the sample distance measures are $(x - \bar{x}_g)\ \mathbf{S}^{-1}_{(\lambda)}\ (x - \bar{x}_g)^T$.

Double cross-validatory estimation and validation

Application of the above described classification approaches still require choice of the tuning parameters k or γ involved. As we are specifically interested in an evaluation of predictive performance of any diagnostic allocation rule, it becomes crucial that any optimization -such as the choice of the tuning parameters - does not take place on the same data used for validation. On the other hand, predictive tuning is clearly highly desirable if diagnosis is of interest, so we would not wish to base the choice of tuning parameters on the full calibration data itself (and thus effectively drop predictive tuning from the analysis), but use a truly validatory choice instead. This implies we either set aside a so-called separate 'tuning set' from the available calibration data prior to validation of predictive performance itself or appeal to some form of cross-validation. Good predictive optimization or tuning becomes particularly important in a high-dimensional setting, such as proteomics, as it provides an opportunity to safeguard model choice against over-fitting (in other words: over-interpreting the data). Meanwhile, even if we were able to effectively choose good tuning parameters, the predictive performance (in our case essentially the error rates) of any implied allocation rule should again be validated, which again introduces a need for yet another set-aside validation set or cross-validation.

We may solve both problems by carrying out a so-called double cross-validatory approach, which avoids the need to introduce separate test (tuning) and validation sets. The method has been first proposed and investigated by Stone[7] and integrates predictive optimization and unbiased validated error rate estimation in a single validatory procedure. While the principle of the methodology is sound and well described, this procedure has until recently not been applied in practice due to the considerable computational cost and (algebraic) complexity of the method. [18] This paper describes a first full implementation in the related setting of discriminant allocation on microarray data. Other papers by our group give further details on computational background, application for leave-one-out in spectroscopy and further references.[19;20]

Similar to with ordinary leave-one-out cross-validation, double cross-validation removes each individual (sample) in turn from the data, after which the discriminant

rule is fully recalibrated (and optimised for prediction) on the leftover data and using the same procedure in each case. 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 misclassification rates are calculated on the basis of the thus validated classifications. The double-validatory aspect results from the fact that the discriminant rule constructed to classify each left-out datum is 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 combine predictive optimization and predictive unbiased validation in the same procedure, without loss of data -which is an important requirement to get realistic estimates of error rate with high-dimensional data.

APPLICATION AND EVALUATION

Preprocessing of mass spectra

Some pre-processing can be beneficial when it removes variation from the data which does not relate to the group separation and might obscure an existing group separation. We describe the pre-processing steps carried out prior to the double cross-validatory classification analysis.

First, we calculated for each sample the average intensity within each bin across the four mass spectra from the associated spots on the target plate. Then, we aggregated contiguous bins on the m/z scale, such that the new aggregated bin size spans approximately one Dalton at the left side of the spectrum and gradually increases to a width of approximately 3 Dalton at the right hand side. For each of these new aggregated bins, we calculated for each spectrum the associated aggregate intensity by summing the intensities across the bins being aggregated. Subsequently, spectral baseline was removed from each of the thus aggregated spectra separately using an asymmetric least squares algorithm.[21]

Suppose $x_{bi} = (x_{bi1}, \ldots, x_{bip})$ denotes the ordered sequence of baseline corrected m/z intensity values for the i^{th} sample at this stage of preprocessing. We then correct the spectrum for the typical intensity and variability across the spectrum by calculating the standardised values

$$x_{sbij} = \frac{x_{bij} - \text{medain}(x_{bi})}{(q_{0.75}(x_{bi}) - q_{0.25}(x_{bi}))} ,$$

where $q_{0.25}(x_{bi})$ and $q_{0.75}(x_{bi}))$ denote the 25th and 75th percentiles of the baseline corrected intensity values for the i^{th} sample. These steps bear close resemblance to the

preprocessing procedure proposed by Satten et al , although our cruder version does not employ local estimates.[22] The final preprocessing step is a log-transformation

$$x_{ij} = \log(x_{sbij} + \alpha)$$

of each spectrum, where α is a real constant. We chose $\alpha = 100$. The main purpose of the log-transform is to ensure numerical stability of calculations. The above preprocessing steps were applied for each sample and within each week separately, which thus gives us the observations x^1_i and x^2_i from the first and second weeks. It is important to stress that the preprocessing of the data of any i^{th} sample does not involve use of any information based on the remaining samples $\{k \mid k \neq i\}$, nor of the duplicate replicate measured spectrum of the same sample on another week. This is an important requirement to ensure the validity of the cross-validatory evaluation described subsequently.

Double cross-validatory error rates

First, we restrict ourselves to the data from the first week. Table 2 displays the estimated recognition rates and performance measures from an analysis of the first week data (leftmost 3 columns). All of the estimates are based on double cross-validation. We used the average of sensitivity (Se) and specificity (Sp) as our estimate of the total recognition rate (T), which implies we assume prior class probabilities to equal 0.5. A threshold of 0.5 was also used to assign observations on the basis of the a-posteriori class probabilities within the cross-validatory calculations. B denotes the Brier distance defined

$$B = \frac{1}{n}\sum_i [1 - p(c(i) \mid \mathrm{x}_i)]^2$$

where $p(c(i) \mid \mathbf{x}_i)$ is the double cross-validated predicted a-posteriori class probability for the correct class $c(i)$ for each i^{th} sample and n is the total sample size. Likewise, AUC is a double cross-validation estimate of the area under the empirical ROC curve defined as

$$AUC = \frac{1}{n_1 n_2} \sum_{i \in G1} \sum_{i \in G2} [I(p(1 \mid \mathrm{x}_i) > p(1 \mid \mathrm{x}_i)) + 0.5 * I(p(1 \mid \mathrm{x}_i) = p(1 \mid \mathrm{x}_j))],$$

where $G1$ and $G2$ refer to the sample index labels for samples from the first and second group respectively. Use of the threshold at 0.5 is appropriate and sufficient for an evaluation of diagnostic potential only. Application in *e.g.* a screening type application would require a more careful choice of prior probability, which is how-

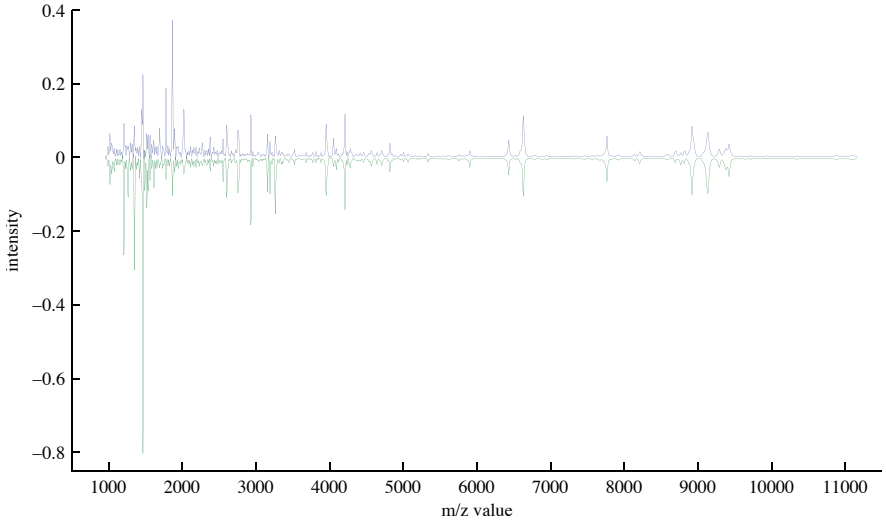


Figure 1. Mean spectra for each group separately, after preprocessing. We plot negative intensity value for the control group (bottom mean spectrum).

ever a subtly different and also subsequent research question and not the focus of this paper.

The rightmost three columns of the table refers to a repetition of this entire double cross-validatory exercise, which replaces each sample feature vector x^1_i with the corresponding replicate measurement x^2_i immediately prior to classification of that i^{th} sample (*i.e.* replacing the feature vectors with the data from week 2 in the outermost layer (only!) of the double cross-validatory calculation). Crucially and importantly, construction of the corresponding discriminant rule for the classification of each such i^{th} sample in the internal 'calibration' layer of the double cross-validatory procedure does of course remain based on the data from week 1. Note that as the replicate data from the third plate are not available, these results are based on the double cross-validated predictions for the remaining 78 replicate samples from week 2 only.

Table 2. Double cross-validated classification results for the colon cancer data. T is the total recognition rate. Se and Sp are sensitivity and specificity, respectively. B is the Brier distance and AUC is the estimated area under the ROC curve.

Method	First week			Second week		
	T (Se, Sp)	B	AUC	T (Se, Sp)	B	AUC
Moore-Penrose $S_{(r)}$	92.6 (95.2, 90.0)	0.0618	97.6	94.4 (91.7,97.1)	0.0600	97.4
PCA Selection $S_{(k)}$	92.6 (95.2, 90.0)	0.0606	97.3	88.8 (80.6,97.1)	0.0914	96.8
Moore-Penrose Euclidian $S_{(r)}$ $\lambda_{(r)} = I_{(r)}$	89.4 (88.9,90.0)	0.0829	96.0	87.2 (86.1,88.2)	0.0770	97.0
PCA Selection Euclidian $S_{(k)}$ $\lambda_{(k)} = I_{(k)}$	88.7 (87.3, 90.0)	0.0865	96.0	90.0 (88.9,91.2)	0.0795	97.0
Ridge $S_{(r)}$	92.0 (95.2,88.0)	0.0602	98.4	95.8 (91.7,100.0)	0.0469	97.9

At first sight, the Moore Penrose implementation (top line of the table, both weeks one and two) would seem to be the best performing and most consistent method. In week 1, Moore-Penrose, PCA-selection (both using the Mahalanobis distance) and ridge estimation perform equally well, but there seems to be an increase in error rate for week 2 for both the PCA-selection and ridge implementation. The Euclidean distance based implementations are worse in the evaluation on the first week, but recognition rates are consistent across both weeks when compared to the other methods. These results should be interpreted with some caution and require some explanation. First of all, the 'plain' Moore-Penrose is leave-one-out only as it does not involve choice of shrinkage or data reduction parameter (k or λ). The deterioration of the PCA-selection implementations is partly due to the uncertainty in estimating the shrinkage terms or choice which is introduced by the double-cross-validatory estimation. For the ridge implementation, performance is comparable to that from Moore-Penrose in week 1, which is not surprising since the chosen ridge shrinkage parameter $\lambda < 0.0001$ for most observations. The effects of uncertainty in the determination of the shrinkage term become particularly apparent for PCA-selection using Mahalanobis distance (second line in the table) in week 2. The two Euclidean distance based implementations on the other hand seem more consistent across both weeks. The reason is that component selection is much more stringent for these two implementations, which selects only the first 2 components for nearly all observations (with exception of two observations out of 113 for which only the first principal component is retained). This explains the reduced performance but also the greater consistency of the classification results. It is precisely because of this reason that these results (from the Euclidean based implementations) are more credible and may well turn out to be more repeatable if the classifier were applied in the future to data from a new repeat experiment. For comparison, component selection in the Mahalanobis distance based PCA implementation is much less stringent and selects (k = 23 for 53 observations, k = 28 for 28 observations and the remainder of

the samples uses even more components). There is thus some evidence of insufficient shrinkage for this method, and similarly for the ridge implementation.

Investigating bias: a permutation exercise

We have proposed double cross-validatory integrated estimation and assessment of statistical diagnostic rules on the basis of the argument that it should protect against optimistically biased evaluations. We may check this property by 'removing' the class labels $c(i)$ from the samples $i \in \{1, \ldots, n\}$, randomly permute and then reassign them to the samples. We then carry out the double cross-validatory procedure again for any of our classification methods. Repeating this procedure several times will give an indication of the biases involved, as the typical recognition rate -for example -should equal 50% across a large number of permutations for an unbiased method.

Table 3. Permutation-based evaluation of double cross-validatory calculations for linear discrimination using principal component selection. DBCV refers to the actual double cross-validatory results (see table 2). $q2{:}5$ and $q97{:}5$ are the 2.5 and 97.5 percentiles. B is the Brier distance and AUC is the estimated area under the ROC curve.

		Permutation results		
Measure	DBCV	median	$q_{2.5}$	$q_{72.5}$
Misclassification rate	7.4	50.0	36.3	72.7
AUC	97.3	49.4	24.8	64.2
B	0.0606	0.324	0.200	0.446

Table 3 shows results from such an exercise for the pca-selection based algorithm across more than 600 such permutations. The results, both for misclassification rate as we find median rates and areas of 50% exactly. Table 3 also includes 95% confidence intervals for the permutation-based performance measures. These give an indication of the variability which can be expected with purely random data and can be compared with the actually observed double-cross-validation results in our study (second column of the table). Clearly, the distance between the validated measures actually observed and even the extreme bounds of the random permutation confidence intervals is considerable, demonstrating the presence of discriminating information in the mass spectra.

Data reduction and post-hoc exploratory analysis

We wish to get an indication of which markers drive the classification. To explore these aspects, we can complement the double cross-validatory analysis with post-hoc exploratory analyses. We consider two analyses, the first of which is based on a very ad hoc algorithmic approach through pre-selection of a small set of adjacent

bins which together account for most of the variation in the spectra. The second explores the linear discriminant weights from a post-hoc fit on the full data.

Data reduction

Initialize $I = \{1,\ldots,p\}$ as the ordered set of bin indices and $V = \{v_1,\ldots,v_p\}$ the associated set of variances for all p bins in the preprocessed spectra and across all n samples, such that $v_j = \sum_i [(x_{ij} - \bar{x}_j)^2]/(n - 1)$, where $\bar{x}_j = \sum x_{ij}/n$ is the sample mean and j is the bin index number. Calculate the constant $vref = q_{0.95}(V)$ as the 95% percentile of all p bin variances. Now initialize the bin selection set B as the set containing the bin indicator j for which the maximum variance vj is observed in the set V. Initialize the set of intensity readings $X_s = \{x_{|j|} \mid j \in B\}$ corresponding to the set B, where $x_{|j|} = (x_{1i}, \ldots, x_{nj})^T$. We write $\mathbf{m} = (m_1,\ldots,m_n)^T$ as the set of means $m_1 = \text{mean}(\{x_{ij} \mid j \in B\})$, i :1, ... ,n. Define cor(**a**, **b**) to be the coefficient of correlation between two vectors **a** and **b**.

Now run the following algorithm.

{*Start of outer loop*}

{Start of inner loop}

Set k=1, $I = I - \{j\}$ and $V = V - \{v_j\}$

Now iterate the following procedure until termination.

Calculate $\rho_{lower} = \text{cor}(\mathbf{m}_i, \mathbf{x}_{[j-k]})$ and $\rho_{upper} = \text{cor}(\mathbf{m}_i, \mathbf{x}_{[j+k]})$

If $\rho_{lower} > 0.9$ and $\rho_{upper} > 0.9$ then

1. Add $j - k$ and $j + k$ to the bin selection set: $B = \{j - k\} \cup B \cup \{j + k\}$.
2. Update the means m_i,$i : 1, \ldots, n$.
3. Remove indices $j - k$ and $j + k$ from the index set I, such that $I = I - \{j - k, j + k\}$. Similarly update $V = V - \{v_{j-k}, v_{j+k}\}$
4. set k=k+1

Else

k=k-1

End iteration.

Now select the bin index j for which $v_j = \max(V)$.

If $v_j > v_{ref}$ then

Update the index set $B = B + \{j\}$ and likewise X_s and m.

Go to {Start of inner loop}

Else End algorithm.

The algorithm identifies a set of 'clusters' of bins. There is no assumption on either shape of the signal or of monotonicity involved (a single cluster may span mixture of underlying peaks). Running this algorithm on the data from the first week finds the set of indices B that corresponds to the bins which account for most of the variation

in the data. Applying this to our data results in a subset of 330 bins (in 32 bin clusters -but note it is possible that we visit the same contiguous region of bins several times). Repeating the entire double cross-validatory procedure using the principal component selection shrinkage procedure on this reduced set yields recognition rates as described in table 4, which are not inconsistent with those from the full double cross-validatory evaluation shown in table 2. (Note however, "double-cross" error rates from this algorithmic approach will be biased as they are based on feature selection from the full data.)

Table 4. Results from re-running double cross-validatory calculations after bin-selection for the colon cancer data (week 1 data only). T is the total recognition rate. Se and Sp are sensitivity and speci¯city respectively. B is the Brier distance and AUC is the estimated area under the ROC curve.

Method	T (Se,Sp)	B	AUC
PCA-selection $S_{(k)}$	90.0 (92.1, 88.0)	0.0807	96.5
PCA-selection Euclidisch $S_{(k)}$ $\lambda_{(k)} = I_{(k)}$	89.0 (92.1, 86.0)	0.0824	95.4

Post-hoc data exploration

The second aspect which is of interest is a post-hoc exploration of the (linear) discriminant coefficients β $(\beta_1, \beta_2, \ldots, \beta_p)^T = \mathbf{S}^{-1}_{(k)}(\bar{x}_1 - \bar{x}_2)^T$ [see (Seber 1984) or (Hand 1997)], where $\bar{x}_1$ and $\bar{x}_2$ are the two sample group means (for cases and controls). [9;17] An appropriate and convenient way to summarize the information contained in these coefficients is via the associated correlations of the measured intensities for each j^{th} bin with the class indicator, which are easily calculated as $\rho_j = s_{xj}\beta_j / s_g$, for $j = 1, \ldots, p$ where $s_{xj} = \sqrt{v_j}$ is the standard deviation at the j^{th} bin and s_g the standard deviation of class indicators. We will base this investigation on the linear discriminant fit using the Euclidean distance on the first two principal components (use $\mathbf{S}_{(k)}$, with $k = 2$ and $\Lambda_{(k)} = \mathbf{I}_{(k)}$), as the double validatory assessment of this classifier clearly identifies the first 2 components as containing the discriminatory information.

At this point, we can carry out the analysis starting from a linear discriminant fit based on the full data. Alternatively, we may equally well base the evaluation on a recomputation of the linear discriminant fit on the reduced data described in previous subsection (in both cases we use the data from the first week). Figure 2 (middle section) shows a plot of the correlation coefficients, subsequent to data reduction (previously described selection of 330 bins, but of course now using all 113 samples from the first week). We only show results within the m/z region between 1200 and 2200 Dalton, as the correlations are effectively zero in the remainder of the m/z range. Evidently, this immediately implies that the separating information is to be found within the 1200 to 2200 m/z range.

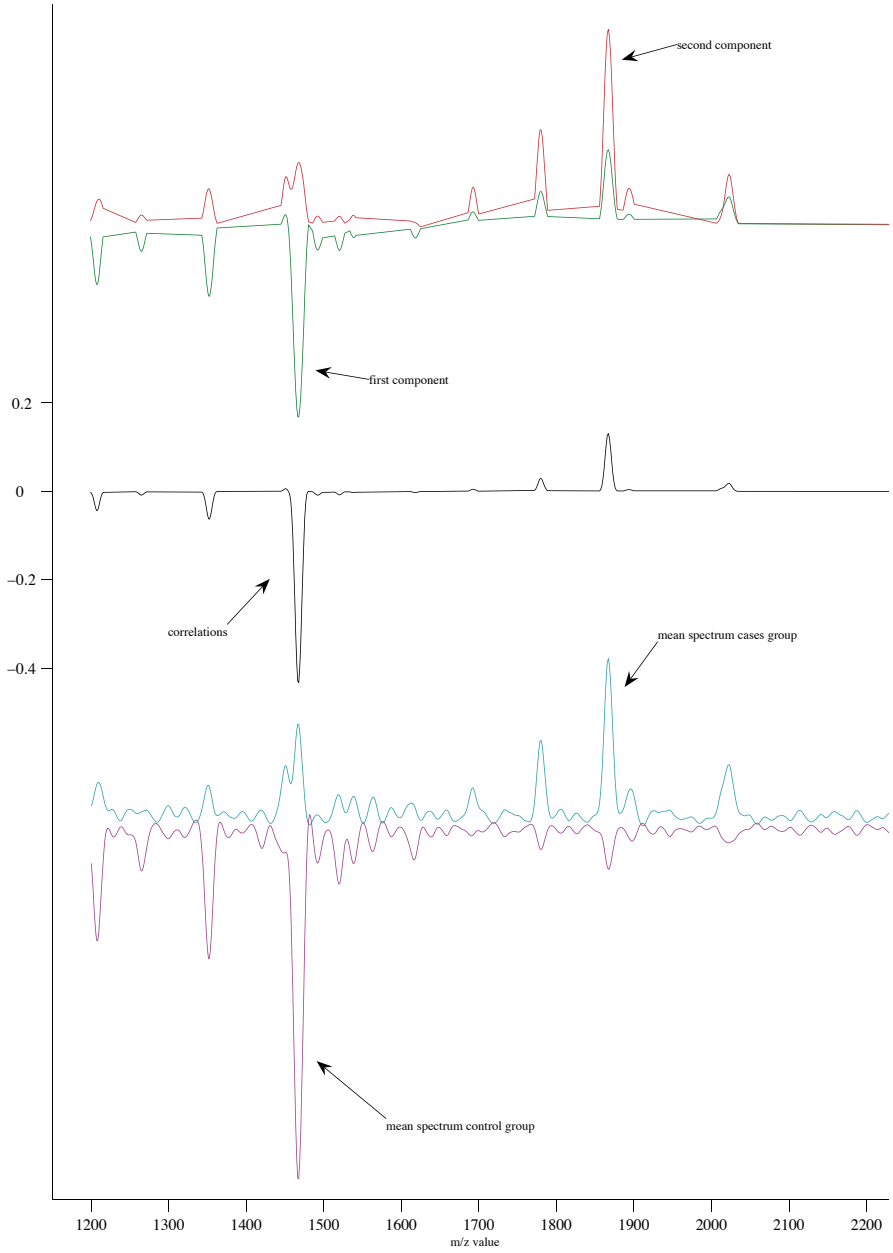


Figure 2. Discriminant correlation coefficients $\rho_j = s_{xj}\beta_j / s_g$ of observed intensity values with the class indicators in the m/z range from 1200 up to 2200 Dalton. We have plotted the first two principal components above these correlations for visual comparison and interpretation. Below the correlations, we plot mean spectra per group (*i.e.*, the vectors x1 and x2, as in figure 1). The y-axis is only relevant to the correlation coefficient, while we have vertically offset and rescaled both components and mean spectra to aid visual comparison across the m/z range.

We note that the picture shown is virtually indistinguishable by eye from that which results from an analysis of the full data (not shown to save space). The reason for this is that the data reduction restricts attention to the dominant sources of variation, which is not very different from what is achieved through principal component reduction. Immediately above the correlation coefficients graph, figure 2 displays the first two principal components (vertically offset and rescaled to aid visual interpretation) and again based on the reduced data. In this case, the distinct bin subsets selected by the previous data reduction step are clearly visible in the two components, and display the characteristic 'peaks' we would expect to identify. Disjoint neighbouring bin sets are connected with straight lines. The thus calculated components are a close approximation to those which would result from an analysis of the full data, as we should expect (results not shown). As for the correlation coefficients, any conclusions are therefore identical whether we use the reduced data or not, although the data reduction step perhaps makes the component plot easier to 'read'. At the bottom of the graph we give the mean spectrum again for each group separately and from the original data within the m/z range of interest, as shown in figure 1 also, along the complete m/z range.

From this graphical analysis, it is clear how the linear discriminant correlation coefficients identify two major discriminating contributions, the first of which is centered at 1467.7 Dalton and the second at 1867.7 Dalton. Furthermore, the correlations have opposite signs at these locations, which would indicate that the discriminating information can be summarised through a contrast effect between corresponding measured intensities in the spectra. An investigation of the principal components plots above learns that the contribution at 1467.7 Dalton is primarily accounted for by the first component, which also already contains the contrast with intensities recorded at 1867.7 Dalton. This contrast is then further amplified by the second component which identifies a second orthogonal source of variation relative to the first component, centered predominately at the already identified peak at 1867.7 Dalton. Note how each component identifies several other smaller contributions, which could also be of interest for further investigation. Comparing these graphs with the within-group mean spectra, the resemblance with the principal components plots at the top of the figure are striking and would suggest that the first component may be primarily explained through variation within the control group at 1467.7 Dalton. Likewise, the second component accounts for a substantial intensity peak at 1867.7 Dalton within the colon cancer group.

To investigate this further, figure 3 provides scatter plots of cases and controls versus the first 2 components (left plot) and between intensities at 1467.7 and 1867.7 Dalton respectively (right plot). The resemblance between both graphs is striking as the right plot can be obtained (virtually) after clockwise rotation of the left plot. As

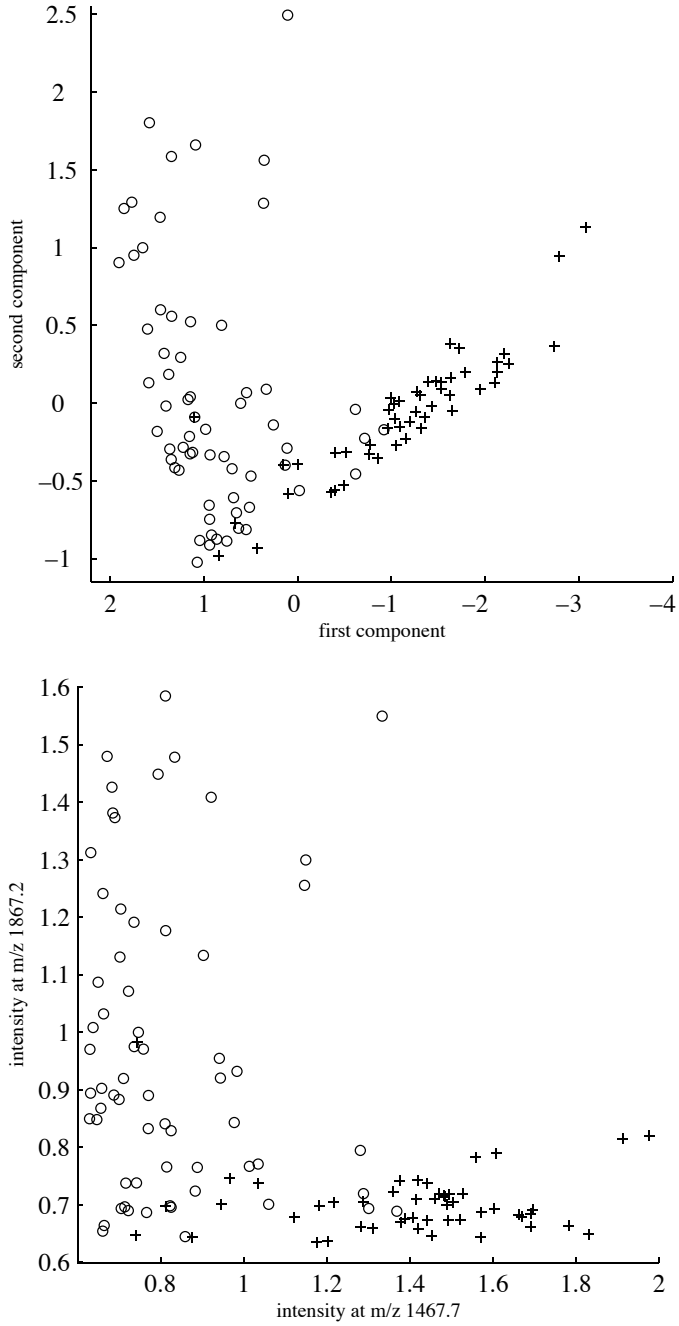


Figure 3. Scatter plots distinguishing cases (o) from controls (+). On the left we plot the second versus the first principal component. The right plot shows intensity values at 1867.2 m/z versus those at 1467.7 m/z.

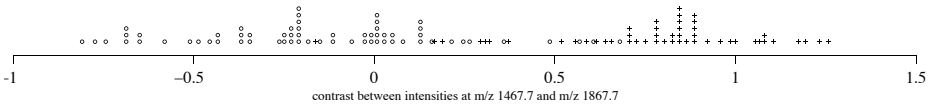


Figure 4. Plot of the diferences between intensities at 1467.7 m/z and 1867.7 m/z across all observations, using distinct plotting symbols for each group: cases (o) and controls (+).

we can see, an increase in intensity at 1467.7 Dalton separates controls from cases. Similarly, an increase in intensity at 1867.7 Dalton separates cases from controls. The same interpretation applies to the principal components scatter plot, which confirms our interpretation of the data in figure 2. Figure 4 provides a concise summary graphical illustration of the results. We calculate the contrast (difference) for all 113 individuals participating in the study between the measured intensities at 1467.7 and 1867.7 Dalton and display the differences in a dot plot using distinct plotting symbols for cases and controls respectively, which demonstrates the separation between both groups.

For further discussion of the clinical background, study rationale, setup, execution and interpretation of results from a substantive clinical perspective, we refer to (Noo 2006) and subsequent papers from these authors.

DISCUSSION

Double validatory analysis

Use of a separate validation or test set is often precluded in high dimensional problems, due to sample size restrictions. In our case, this arises because the experiment was carried out in an academic medical center, which implies (colon cancer) cases are restricted to a maximum of about 50 patients yearly and with more advanced disease. Selection of appropriate control samples may be more difficult still, even if we use surrogate serum samples -as in this experiment. Larger numbers of cases may be recruited by setting up multi-center trials and using longer recruitment periods. However, researchers may need some justification in the form of a small feasibility study before setting up such complex trials. It is in such situations that double cross-validatory analysis can be most useful to help researchers make the maximum use of the scarce data available. The other option of reducing the available calibration data prior to optimization of any discriminant rule by setting aside data (perhaps for both a 'predictive tuning' as well as 'validation' set) is not as innocent as appears at first sight. This is because it will often reduce the calibration set beyond what is

needed for reasonable calibration. Moreover, reducing the size of the calibration data changes the condition of the estimation itself. To put this simply: we are not only reducing the data by setting-aside data from the calibration set, but also changing the discriminant problem itself. This is again particularly the case in high-dimensional cases such as in proteomics where the problem will typically be ill-conditioned.

The approach we have described in this paper avoids these difficulties through application of double cross-validation to combine the two aspects of *predictive optimization* and *validation*. Subsequent to this basic evaluation of the discriminatory potential of the spectral data, a more exploratory analysis can be carried out, provided we are carefully to interpret results cautiously without contradicting the primary validated evaluation. We discuss a number of issues related to application of (double) cross-validation.

Full validation

One potential cause for concern is whether double cross-validation precludes the need for a completely separate validation set entirely. Is 'double-cross' also 'full' validation? The simple answer to this question is that it can not be, as any form of cross-validation must typically always remain 'within-study' validation and there can be factors beyond our knowledge which have influenced the study results. Good scientific practice requires that we replicate results in a separate repeat study. This caution applies particularly to the definition of the case and control group, as the impact of systematic effects due to measurement can be minimised through use of randomised block design. Repeat studies may help to detect such problems. Note however, that these criticisms would also have applied to the standard practice of using within-study set-aside test and validation sets. Meanwhile, double cross-validation should give reasonable protection against overfitting and unbiased estimates of error rate *at the time of study*. Double-cross represents the maximum usage we can make of the data for joint predictive optimization and validation *within a single experiment*. Even when separate test and validation sets are available however, researchers may still be interested to compare the thus validated re-search findings with those from a fully double cross-validated analysis on the combined data in order to evaluate whether the greater sample size would have allowed for better calibrations -possibly because of improved detection of the smaller signal sources in the spectra.[23] More generally, we could speculate where the validation process should stop. Typically, the performance of any decision rule or classifier has a tendency to 'decay' over time. To assess this, subsequent experiments are needed to verify the estimated error rates.

What classifier are we evaluating?

Two related questions to the previous discussion are 'What classifier does double cross-validation evaluate?' and 'How to assign a new observation?'. Indeed, each observation has its own classifier in the double cross-validatory evaluation. This seems to run counter to the intuition that we calibrate a discriminant rule first and only then evaluate. In that case, the estimated error rate is taken as a reflection of the diagnostic abilities of that particular classifier and the allocation of a new sample is immediate. There is however no logical inconsistency here. Double cross-validation estimates the error rate we would get 'if we were to apply leave-one-out' on the whole data. Once we know what the error rate is, we may choose the specific classifier (choice of k or λ in our case) for allocation of future samples (if required) through application of ordinary leave-one-out on the whole data (this is in line with the discussion presented by Mervin Stone.[7] With double cross-validation, there are however other options to allow allocation of new samples which have not yet been discussed in the literature. In our case for example, we may use the mode of the number of components selected (k) across all samples and then re-estimate the discriminant model with this choice from the full data. More adventurous still, we could retain each of the n classification rules which are calibrated within the double-cross procedure and use this ensemble (of classifiers) for allocation of any future new observation x. This could be done by calculating the associated a-posteriori class probabilities $p_i(g \mid \mathrm{x})$, for each $i \in \{1, \ldots, n\}$ and $g \in \{1, \ldots, G\}$, where $\mathrm{p_i}$ is obtained from the discriminant model calibrated in the double-cross procedure when the $\mathrm{i^{th}}$ datum has been removed from the data (in the outer shell of the double-cross procedure). Classification may then be based on the mean across these n a-posteriori class probabilities for any $\mathrm{g^{th}}$ class. We will not pursue these options further in this paper.

Validation and the future of (statistical) proteomics

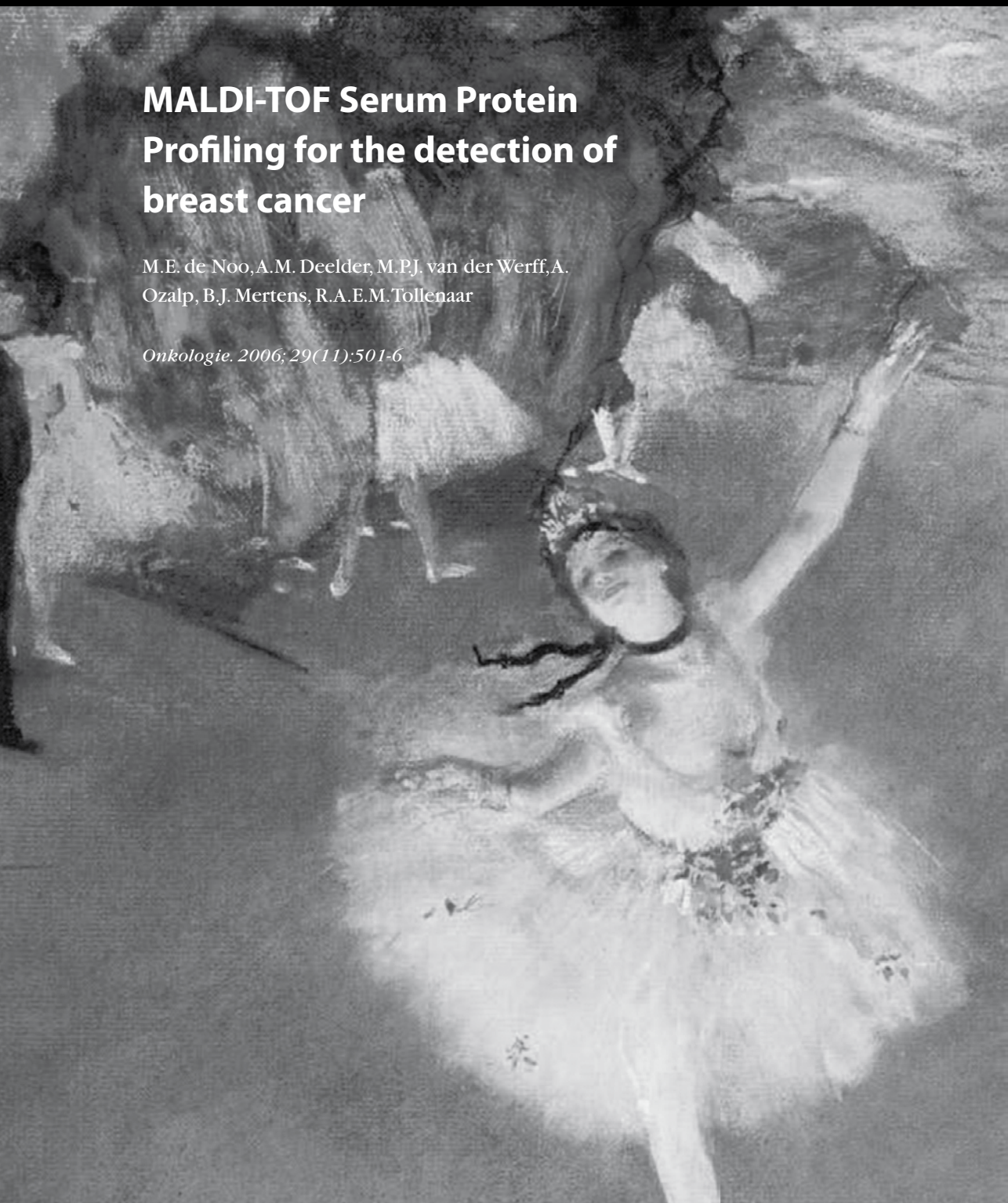
Rigorous emphasis on validation and proper design can help to establish long-term credibility for proteomic research and more general bioinformatics applications. The double-cross approach with randomised block design described in this paper represents one contribution towards this goal. Many other steps may however be taken to enhance the quality of such research studies. One example is to promote use of 'truly' separate validation sets, as obtained from subsequent separate and additional sampling from the population of interest and measurement through identical protocols as applied in the first study. In practice, this will be particularly relevant for those studies which indicate potential from the first within-study verification of diagnostic ability. Editors of scientific journals can also contribute much to inspire a conservative attitude by careful scrutiny of the papers presented for publication. Perhaps simple check lists could be developed to help reviewers establish the degree

to which validatory evaluation did (or should) contribute to the research findings presented. This may also prevent mistakes from slipping through the net. Although this may cause considerable annoyance in some cases when we face the difficulties of establishing results in the short term, but may enhance scientific credibility of (proteomic) research as a whole in the long run. Results from the present study show that, with good designed experimentation, these precautions need not form insurmountable obstacles.

REFERENCES

1. Hastie,T., Tibshirani,R., and Friedman,J. (2001) *The elements of statistical learning*. Springer-verlag.
2. de Noo,M.E., Mertens,B.J., Ozalp,A., Bladergroen,M.R., van der Werff,M.P., van de Velde,C.J., Deelder,A.M., and Tollenaar,R.A. (2006) Detection of colorectal cancer using MALDI-TOF serum protein profiling. *Eur.J Cancer*, 42, 1068-1076.
3. Cox D.R. and Reid N. (2000) *The theory of the design of experiments*. Chapmann/Hall CRC.
4. Box,G.E.P., Hunter W.G., and Hunter J.S. (1978) *Statistics for experimenters*. John Wiley & Sons, Inc..
5. Neter,J.et.al. (1996) *Applied Linear Statistical Models*. McGraw-Hill/Irwin.
6. Fisher,R.A. (1935) *The Design of Experiments*. Oliver and Boyd: Edinburgh..
7. Stone,M. (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36, 111-147.
8. Ripley,B.D. (1996) *Pattern recognition and neural networks*. Cambridge University Press.
9. Seber,G.A.F. (1984) *Multivariate Observations*. Whiley Chichester.
10. McLachlan,G.J. (1992) *Discriminant analysis and statistical pattern recognition*.
11. 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.
12. Morris,J.S., Coombes,K.R., Koomen,J., Baggerly,K.A., and Kobayashi,R. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics.*, 21, 1764-1775.
13. Yasui,Y., Pepe,M., Thompson,M.L., Adam,B.L., Wright,G.L., Jr., Qu,Y., Potter,J.D., Winget,M., Thornquist,M., and Feng,Z. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics.*, 4, 449-463.
14. Sauve, A. C. and Speed T. P. Normalization, baseline correction and alignment of high-throughput mass spectrometry data. *Proceedings Gensips*. 2004.
15. Jolliffe,I.T. (2002) *Principal Component Analysis*. Springer-Verlag, New York.
16. Krzanowski,W.J et al. (1995) Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied Statistics*, 44, 101-115.
17. Hand,D.J. (1997) *Construction and assessment of classification rules*. John Wiley and Sons; Inc.
18. Mertens,B.J.A. (2003) Microarrays, pattern recognition and exploratory data analysis. *Statistics in Medicine*, 22, 1879-1899.
19. Mertens,B.J.A. (1998) Exact principal component in?uence measures applied to the analysis of spectroscopic data on rice. *Applied Statistics*, 47, 527-542.
20. Mertens,B.J.A. (2001) Downdating: interdisciplinary research between statistics and computing. *Statistica Neerlandica*, 55, 358-366.
21. Eilers,P.H. (2004) Parametric time warping. *Anal.Chem.*, 76, 404-411.
22. Satten,G.A.et al. (2004) Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. *Bioinformatics*, 20, 3136.
23. Ransohoff,D.F. (2004) Evaluating discovery-based research: when biologic reasoning cannot work. *Gastroenterology*, 127, 1028.

Chapter 6



MALDI-TOF Serum Protein Profiling for the detection of breast cancer

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

Onkologie. 2006; 29(11):501-6

ABSTRACT

Purpose

With a lifetime risk currently estimated one in nine, breast cancer is among the most common diagnosed malignancies and remains a leading cause of cancer-related morbidity and mortality. Proteomic expression profiling generated by mass spectrometry has been suggested as a potential tool for the early diagnosis of cancer and other diseases. The objective of our study was to assess the feasibility of this approach for the discrimination of breast cancer patients from healthy individuals.

Experimental design

In a randomised block design pre-operative serum samples obtained from 77 breast cancer patients and 29 controls were used to generate high-resolution MALDI-TOF protein profiles. The median age of the patient group and control group was respectively, 57.2 years and 50.0 years. All available 106 samples from both groups were randomly distributed across 3 plates in roughly equal proportions. The MALDI-TOF spectra generated using C8 magnetic beads assisted mass spectrometry (Ultraflex, Bruker Daltonics, Germany) were smoothed, binned and normalised after baseline correction. After pre-processing of the spectra, linear discriminant analysis with double cross-validation, based on principal component analysis, was used to classify the protein profiles.

Results

A total recognition rate of 99%, a sensitivity of 100% and a specificity of 97.0% for the detection of breast cancer were shown. The area under the curve of the classifier was 98.3%, which demonstrates the high, significant separation power of the classifier. The first 2 principal components account for most of the between-group separation.

Conclusions

Double cross-validation showed that classification could be attributed to actual information in the protein profiles rather than to chance. Although preliminary, the high sensitivity and specificity indicate the potential usefulness of serum protein profiles for the detection of breast cancer.

INTRODUCTION

With a lifetime risk currently estimated one in nine, breast cancer is among the most common diagnosed malignancies and remains a leading cause of cancer-related morbidity and mortality. Although the precise pathways of tumour genesis remain poorly defined, it appears that most invasive breast cancers arise from gene alterations that result in an initial transformation of normal breast tissue to in situ carcinoma.[1] Currently, mammography remains the most important diagnostic tool, although MRI and ultrasonography are used in case of impairment of the latter diagnostic results.[2] However, up to 20% of new breast cancers are not detected or visible on a mammogram.[3] Prognosis and selection of therapy may be influenced by the age and menopausal status of the patient, Bloom-Richardson stage, histological and nuclear grade of the primary tumour, oestrogen-receptor (ER) and progesterone-receptor (PR) status, measures of proliferative capacity, and HER2/neu gene amplification.[4] Currently, serum tumour markers play no role of importance in the diagnosis of breast cancer due to a lack of sensitivity and specificity.

Proteomic expression profiles generated with mass spectrometry have been suggested as potential tools for the early diagnosis of cancer and other diseases. After the initial 'hype' of biomarker detection on the basis of multiple low-molecular-weight serum proteins stringent demands have been proposed on both study design and experimental procedures for proteomic profiling.[5-11] Subsequently, several studies appeared showing the importance of standardised protocols and homogeneity of subject groups and especially validation of the classification method.[12-16] This study aims to live up to all these demands.

Since no serum biomarker is currently known to reliably detect breast cancer, the present study was designed to test and validate whether serum protein profiles generated with mass spectrometry could be indicative of the presence breast cancer.

MATERIAL AND METHODS

Subjects

Serum samples were obtained from a total of 77 patients one day prior to surgery for a breast disease. All surgical specimens were histologically examined and if malignant, the extent of tumour spread was assessed by TNM classification. All stages of breast cancer were present in the patient group. The median age of the patient group was 57.2 years (range 32.6-90.3). Patients were included from October 2002 till July 2005 in our center. The control group consisted of 29 healthy female volunteers. The

median age of the healthy symptom-free control group was 50.0 years (25.9-76.7). The 29 controls were included in November and December 2004 (Table 1).

Table 1. Patient characteristics.

	Patients	Controls
n =	78	29
Age (mean)	56.6	49.9
(range)	36.2-90.3	25.9-76.7

Serum samples

Informed consent was obtained from all patients and the study was approved by the Medical Ethical Committee of the LUMC. All samples were collected and processed following a standardised protocol: 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 (BD Diagnostics, Plymouth, UK) 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. After thawing on ice the serum samples were randomised over different 96-well microtitration racks (Matrix) and then stored at -70ºC until the experiment.

Study design

We used a randomised blocked design to avoid any potential batch effects.[17;18] All the available 106 samples from both groups were randomly distributed across 3 plates in roughly equal proportions (Table 2). For breast 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. Analysis was carried out on 3 consecutive days, Tuesday to Thursday, processing a single plate each day.

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 random fashion and approximately equal proportions.

	Plate 1	**Plate 2**	**Plate 3**	**Total**
Breast cancer	26	26	26	78
Controls	11	9	9	29
Total	37	35	35	107

Table 3. Distribution and randomisation of serum samples of breast cancer group over the three MS target plates.

Stage	Plate 1	Plate 2	Plate 3	Total
DCIS	5	4	3	12
I	6	6	8	22
IIA	7	8	3	18
IIB	4	6	4	14
IIIA	1	2	4	5
IIIB	1	0	2	3
IIIC	1	0	2	3
Total	25	26	26	77

Isolation of peptides and protein profiling

The isolation of peptides from serum was performed using the C8 magnetic beads based hydrophobic interaction chromatography (MB-HIC) kit from Bruker Daltonics (Bremen, Germany) mainly according to manufacturers instructions, adapted for automation on a 8-channel Hamilton STAR® pipetting robot (Hamilton, Martinsried, Germany) as previously described by our group. Each sample was spotted in quadruplicate on a MALDI AnchorChip™. Matrix Assisted Laser Desorption Ionisation Time-Of-Flight (MALDI-TOF) mass spectrometry measurements were performed using an Ultraflex I 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.

Data processing and statistical analysis

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 1.160 - 11,600 Dalton. To increase robustness, the average of four spots was used to represent one serum sample. Subsequently, we lightly smoothed, binned and normalised the spectra after baseline correction. 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 previously described.[19]

RESULTS

Three different randomised target plates were successfully measured on three consecutive days in the middle of the week. 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. 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. One sample from the breast cancer group was excluded from analysis due to its poor quality spectra.

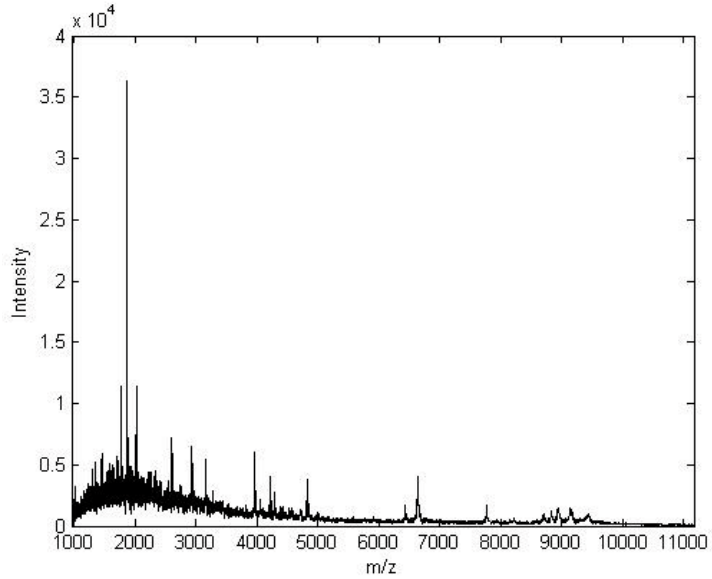


Figure 1. MALDI-TOF spectrum of a breast cancer patient after peptide isolation with C8 magnetic beads. On the Y-axis the relative intensity is shown. The mass to charge ration (m/z) is demonstrated on the X-axis in Dalton.

Double cross-validatory analysis and evaluation carried out on the protein spectra correctly classified 28 of 29 controls as non-cancer. All breast cancer patients were correctly classified as malignant (Table 4). These validated results yield a total recognition rate of 98.2%, a sensitivity of 100% and a specificity of 97.6% for the detection of breast cancer. 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 98.3%. The median AUC was 49.4% with confidence interval of

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

	Test results for detection of BC		
	Neg	Pos	Total
Controls	77	0	77
CRC patients	1	28	29
	78	28	106

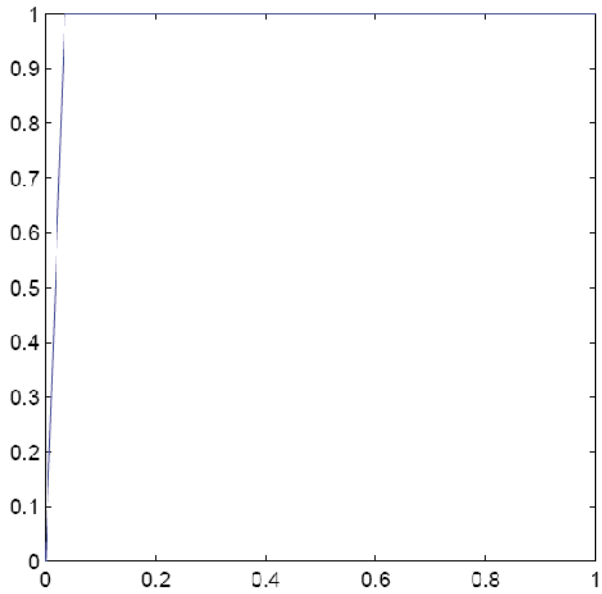


Figure 2. ROC-curve for the double cross-validated two-group classifier. The true positive recognition rate (sensitivity) is demonstrated on the y-axis against the false negative recognition rate (1-specificity) on the x-axis of the classifier.

[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.

We then proceeded to a post hoc exploration of the classification model. In the present study the first two principal components provided 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 1.160 and 11,600 Dalton.[20;21] As illustrated, the classification is achieved primarily through a contrast in peak intensities between the first and sec-

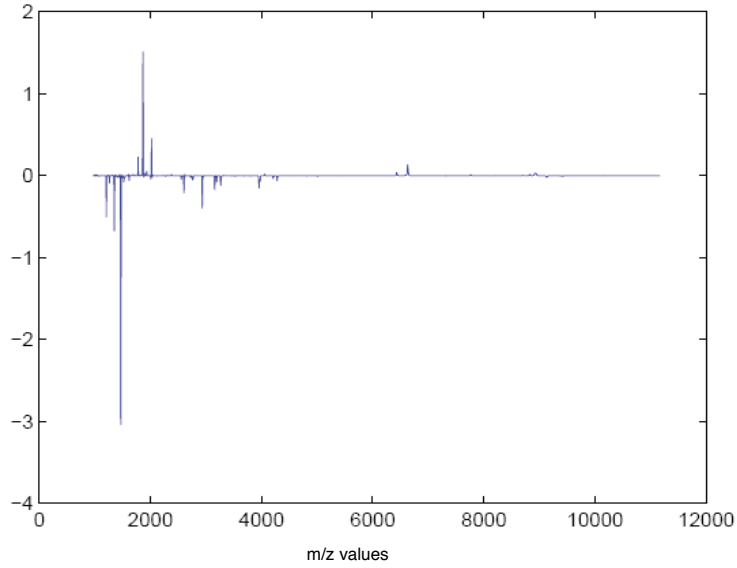


Figure 3. Correlation coefficients of two first principal components with the class indicator. The correlation coefficients were calculated from the linear discriminant weightings. The negative correlation of the first peak is an indicator for the control group and the positive correlation of the second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

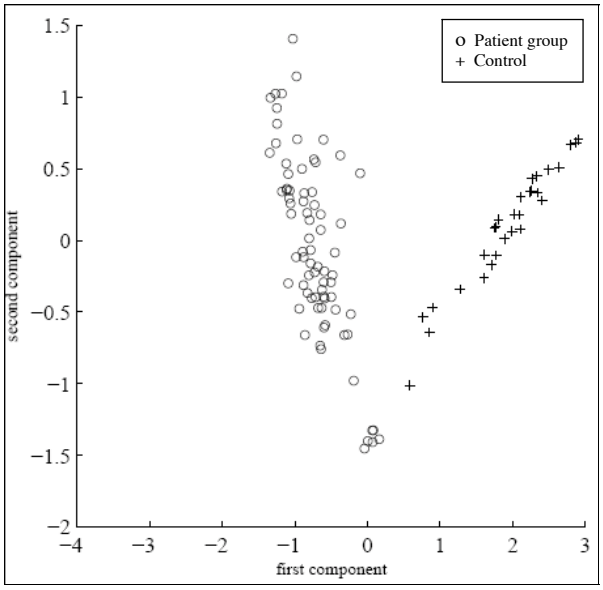


Figure 4. Scatter plot of the first two principle components on basis of which the classification patient-control group was made.

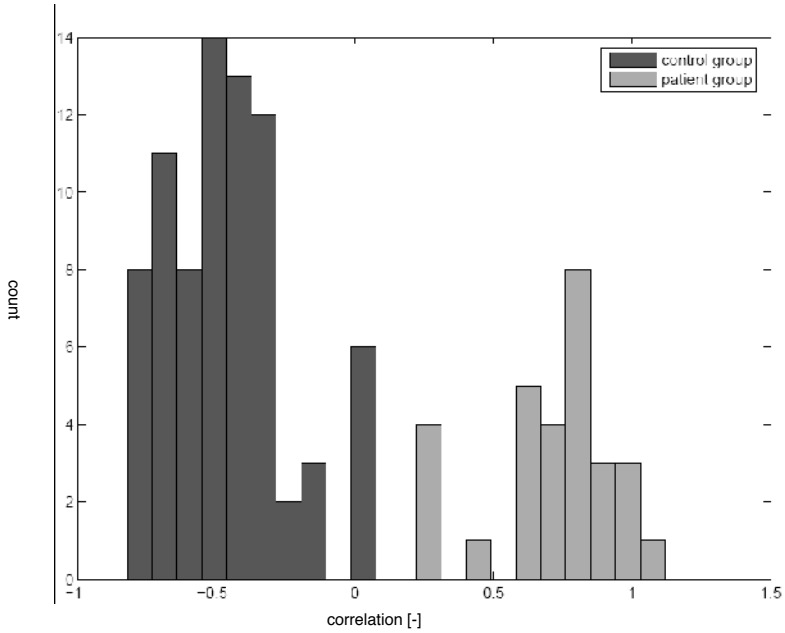


Figure 5. Histogram showing the difference between the normalized intensities of the two most discriminating "peaks" (bins). The X-axis shows the difference between the normalized intensities of the peaks. On the Y-axis the number of subjects is displayed.

ond principal component. This can also be seen in 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 also quantified the significance of this difference by performing a two-sample Student t-test on this contrast, ($p<0.0001$).

DISCUSSION

This study underlines the potential of serum protein profiling for the detection of breast cancer. We were able to classify breast cancer patients and healthy individuals very accurately based upon information in MALDI-TOF serum spectra. The classifier, calibrated and validated on spectra of the entire dataset demonstrated a sensitivity and specificity of 100% and 97.6% respectively. Only one subject from the control

group was misclassified in the malignant group (Table 4). Moreover, with a lifetime risk of one in nine, it might even be the case that one of 29 control subjects currently is developing or carrying the disease. Unfortunately, since the control group consisted of anonymous symptom-free subjects it was impossible to retrieve the current physical state. All patients with various stages of breast cancer were correctly classified, including DCIS patients. The fact that all DCIS patients are recognised in the cancer group, adds to its possible future applicability as a tool for early detection. If these results are validated, future studies could be performed to screen women at high risk for breast cancer by both mammography and serum protein profiling.[22] In that way the positive predictive value of the proteomic pattern approach could be assessed. Further, efforts have to be made to correlate different stages of breast cancer with serum protein profiles because this may contribute better prognostication and may eventually lead to more individualised treatment. Obviously, validation and sufficient sample size are once again of paramount importance for the reliability and its potential in a clinical setting.

We favour a thorough and stringent study design and double cross-validation of our classification model.[19] We feel that the use of standardised serum collection and mass spectrometry protocols, as advocated in various studies, has lifted serum protein profiling to a more reliable level.[12;13;16;23] To avoid the most common pitfalls in clinical proteomics sample collection, pre-analytical conditions and biological variables were in the present study matched for both groups and were rigorously standardised. The location of blood collection, i.e. the outdoor clinic for controls and the surgical ward for the patient group showed no influence on serum protein profiles (data not shown). Furthermore, patient samples from all stages of breast cancer were randomly distributed over three different target plates, excluding these factors as a discriminator in the current classifier. Ideally, the control group should consist of precisely age-matched individuals undergoing a mammography showing no aberrations. However, in practice this is difficult to realize, due to ethical and logistical issues. Notwithstanding, we performed an analysis to examine the differences in intensity of most discriminating peaks based on 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).

Regarding the bioinformatic and statistical approach of these high dimensional data there were two main points to consider: avoiding batch effects and validation of the classification model. To avoid observational bias, a randomised block design was used as an additional precaution. The randomised block design ensured that no batch effects were introduced and excluded artificial between-group separations.[17;18] Another recurrent topic of debate in serum protein profiling is validation of the classification model.[11;24] Consensus is achieved that, ideally, discriminating

protein profiles for the detection of a certain malignancy should be validated using an independent dataset. In the current phase of this study, the use of an independent validation set was excluded since the relatively small sample size did not allow this. Until a larger sample set is obtained, we advocate the use of a double cross validation of classification. This procedure 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.[19;25]

The classification between cancer and non-cancer was mostly performed using the first two principal components, corresponding to two most discriminating peaks. 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. Until nowadays, little success has been booked in assigning reproducible discriminating biomarkers.[14;24] 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. Some have argued that low molecular weight proteins in serum, the serum peptidome, is nothing but aspecific biological trash and therefore does not yield any reliable biomarkers in the currently technically available mass range.[26-28] However, very recently Villanueva et al. published a study in which they proposed that although discriminating peptides do indeed belong to well known coagulation and complement pathways, their patterns or signatures do most certainly indicate the presence of cancer.[29] This study showed that most of the cancer-type specific biomarker fragments were generated in patient serum by enzymatic cleavage at previously known endoproteage cleavage sites after the blood sample was collected.[30] They postulated that these cancer-specific low molecular weight proteins in the serum peptidome are an indirect snapshot of the enzyme activity in tumour cells. We support to their hypothesis that discriminating serum protein profiles are a compilation of surrogate markers for the detection and classification of certain types of tumours.

In conclusion, the present study demonstrated that patterns of proteomic signatures from high dimensional mass spectrometry data can be used as highly reliable diagnostic classifiers for the detection of breast cancer. With the double crossvalidatory study in a randomised block design we obtained maximal reliability in classification while maintaining protection against overfitting. Surely, independent validation and follow up studies are necessary and currently in progress. Nevertheless, the

extremely high sensitivity and specificity of the present study are highly promising for a new diagnostic approach in breast cancer.

REFERENCES

1. Burstein,H.J., Polyak,K., Wong,J.S., Lester,S.C., and Kaelin,C.M. (2004) Ductal carcinoma in situ of the breast. *N.Engl.J Med.*, 350, 1430-1441.
2. Veronesi,U., Boyle,P., Goldhirsch,A., Orecchia,R., and Viale,G. (2005) Breast cancer. *Lancet*, 365, 1727-1741.
3. Astley,S.M. (2004) Computer-based detection and prompting of mammographic abnormalities. *Br J Radiol*, 77, S194-S200.
4. Simpson,J.F., Gray,R., Dressler,L.G., Cobau,C.D., Falkson,C.I., Gilchrist,K.W., Pandya,K.J., Page,D.L., and Robert,N.J. (2000) Prognostic Value of Histologic Grade and Proliferative Activity in Axillary Node-Positive Breast Cancer: Results From the Eastern Cooperative Oncology Group Companion Study, EST 4189. *J Clin Oncol*, 18, 2059-2069.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. Ransohoff,D.F. (2004) Rules of evidence for cancer molecular-marker discovery and validation. *Nat.Rev.Cancer*, 4, 309-314.
12. Boguski,M.S. and McIntosh,M.W. (2003) Biomedical informatics for proteomics. *Nature*, 422, 233-237.
13. 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.
14. 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.
15. 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.
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. Box,G.E.P., Hunter W.G., and Hunter J.S. (1978) *Statistics for experimenters.* John Wiley & Sons, Inc..
18. Cox D.R. and Reid N. (2000) *The theory of the design of experiments.* Chapmann/Hall CRC.

19. de Noo,M.E., Mertens,B.J., Ozalp,A., Bladergroen,M.R., van der Werff,M.P., van de Velde,C.J., Deelder,A.M., and Tollenaar,R.A. (2006) Detection of colorectal cancer using MALDI-TOF serum protein profiling. *Eur.J Cancer*, 42, 1068-1076.
20. Seber,G.A.F. (2005) *Multivariate Observations*. John Wiley & Sons Inc.
21. Ripley,B.D. (2005) *Pattern recognition and neural networks*. Cambridge University Press.
22. Espinosa,E., Redondo,A., Vara,J.A., Zamora,P., Casado,E., Cejas,P., and Baron,M.G. (2006) High-throughput techniques in breast cancer: A clinical perspective. *Eur.J Cancer*.
23. Baumann,S., Ceglarek,U., Fiedler,G.M., Lembcke,J., Leichtle,A., and Thiery,J. (2005) Standardized Approach to Proteome Profiling of Human Serum Based on Magnetic Bead Separation and Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. *Clin. Chem.*, 51, 973-980.
24. 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.
25. Mertens,B.J.A. (2003) Microarrays, pattern recognition and exploratory data analysis. *Statistics in Medicine*, 22, 1879-1899.
26. Diamandis,E.P. and van der Merwe,D.E. (2005) Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations. *Clin Cancer Res.*, 11, 963-965.
27. Sorace,J.M. and Zhan,M. (2003) A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC.Bioinformatics.*, 4, 24.
28. Koomen,J.M., Li,D., Xiao,L.C., Liu,T.C., Coombes,K.R., Abbruzzese,J., and Kobayashi,R. (2005) Direct tandem mass spectrometry reveals limitations in protein profiling experiments for plasma biomarker discovery. *J Proteome.Res.*, 4, 972-981.
29. Villanueva,J., Shaffer,D.R., Philip,J., Chaparro,C.A., Erdjument-Bromage,H., Olshen,A.B., Fleisher,M., Lilja,H., Brogi,E., Boyd,J., Sanchez-Carbayo,M., Holland,E.C., Cordon-Cardo,C., Scher,H.I., and Tempst,P. (2006) Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J Clin Invest*, 116, 271-284.
30. Liotta,L.A. and Petricoin,E.F. (2006) Serum peptidome for cancer detection: spinning biologic trash into diagnostic gold. *J Clin Invest*, 116, 26-30.

Chapter 7

Validation of serum protein profiling for the detection of breast cancer

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

Submitted

ABSTRACT

Background

With over 1 million new cases in the world each year, breast cancer is the commonest malignancy in women and comprises 18% of all female cancers. Proteomic expression profiling generated by mass spectrometry has been suggested as a potential tool for the early diagnosis of cancer. The objective of our study was to assess and validate the feasibility of this approach for the detection of breast cancer.

Methods

In a randomised block design pre-operative serum samples obtained from 63 breast cancer patients and 73 controls were used to generate high-resolution MALDI-TOF protein profiles as a calibration set. The median age of the patient and control group was respectively, 52 (20-81) and 57 years (39-87). The MALDI-TOF spectra generated using WCX magnetic beads assisted mass spectrometry (Ultraflex) were smoothed, binned and normalised after baseline correction. After pre-processing of the spectra, linear discriminant analysis with double cross-validation, based on principal component analysis, was used to classify the protein profiles. Consequently, the classifier constructed on the first 2 plates was applied on the spectra of an independent validation set. This validation set consisted of serum samples from 29 breast cancer patients and 38 controls. The median age was 59 years (26-87) and 57 years (24-71) for the patient and control group respectively.

Results

Double cross-validatory analysis carried out on the protein spectra of the calibration set yielded a total recognition rate of 86%, a sensitivity of 88% and a specificity of 84% for the detection of breast cancer within the calibration set. The AUC of this classifier was 90.3%. When this classifier was applied on the spectra of the independent validation set a total recognition rate of 80.9%, a sensitivity of 72% and a specificity of 89% were found.

Conclusions

The use of a randomised block design, but mainly an independent validation set proves that discriminating protein profiles can be detected between breast cancer patients and healthy controls. Although further validation in larger series and identification of the discriminating proteins must be achieved, the high sensitivity and specificity indicate that serum protein profiles could be an option for the detection of breast cancer.

INTRODUCTION

With over 1 million new cases in the world each year, breast cancer is the commonest malignancy in women and comprises 18% of all female cancers. In 2005, breast cancer caused 502,000 deaths (7% of cancer deaths; almost 1% of all deaths) worldwide.[1] Despite increasing incidence rates, annual mortality rates from breast cancer have decreased over the last decade (2.3% per year from 1990 to 2002).[2] The effect of reduction due to early diagnosis of breast cancer has been outlined with patients' data by the Surveillance, Epidemiology, and End Results program in a competing-risk analysis calculating probabilities of death from breast cancer and other causes according tot stage, race and age at diagnosis.[3]

Currently, mammography remains the most important diagnostic tool in women with breast tissue that is not dense, although MRI and ultrasonography are used in case of impairment of the latter diagnostic results.[4] In many countries mammography is used as a population based screening in woman older than 50 years. The effect of breast screening in terms of breast cancer mortality reduction persists after long-term follow-up. A recent meta-analysis of seven randomised trials concluded that there was a 15-20% reduction in risk of death from breast cancer in women attending mammography.[5] However, up to 20% of new breast cancers are not detected or visible on a mammogram.[6]

Prognosis and selection of therapy may be influenced by the age and menopausal status of the patient, Bloom-Richardson stage, histological and nuclear grade of the primary tumour, oestrogen-receptor (ER) and progesterone-receptor (PR) status and HER2/neu gene amplification.[7] Currently, serum tumour markers play no role of importance in the diagnosis of breast cancer due to a lack of sensitivity and specificity.

Proteomic expression profiles generated with mass spectrometry have been suggested as a potential tool for the early diagnosis of cancer and other diseases. Although promising results have been shown in classifying cancer studies based on biomarker detection with multiple low-molecular-weight serum proteins[8-11] stringent demands have been proposed on both study design and experimental procedures for proteomic profiling.[12-14] Subsequently, several groups have stressed the importance of standardised protocols and homogeneity of subject groups and especially validation of the classification method.[15-19] Since no serum biomarker is currently available to detect breast cancer, the present study was designed to test and validate whether serum protein profiles generated with mass spectrometry could be indicative of the presence of breast cancer in an independent set.

MATERIAL AND METHODS

Subjects

Serum samples were obtained from a total of 111 patients one day prior to surgery for breast disease. All surgical specimens were histological examined and if malignant, the extent of tumour spread was assessed by TNM classification. Next to invasive stages of breast cancer, ductal carcinoma in situ (DCIS) samples were present in the patient group. The control group consisted of 92 healthy female volunteers. Patients and controls were included from October 2002 till July 2006 in our center.

In the calibration set the mean age of the patient group and control group was 52 (20-81) and 57 years (39-87) respectively. In the validation set the mean age was 59 years (26-87) and 57 years (24-71) for the patient and control group respectively (Table 1).

Table 1. Patient characteristics.

	Calibration set		Validation set	
	Patients	Controls	Patients	Controls
N=	73	63	38	29
Age (median)	52	57	57	59
(range)	20-81	39-87	26-87	24-71

Serum samples

Informed consent was obtained from all subjects and the study was approved by the Medical Ethical Committee of the LUMC. All samples were collected and processed following a standardised protocol: all blood samples were drawn from non-fasting patients or healthy controls while they were seated. The samples were collected in a 8.5 cc Serum Separator Vacutainer Tube (BD Diagnostics, Plymouth, UK) and centrifuged 30 min later at 3000 rpm for 10 minutes. The serum samples were distributed into 0.5 ml aliquots and stored at -70ºC. After thawing on ice the serum samples were randomised over different 96-well microtitration racks (Matrix, Hudson, USA) and then stored at -70ºC until the experiment.

Study design

We used a randomised blocked design to avoid any potential batch effects.[20;21] All the available samples from both groups were randomly distributed across 3 plates in roughly equal proportions (Table 2). For breast cancer, the distribution of disease stages across plates was again in random fashion and in approximately equal pro-

Table 2. Distribution and randomisation of serum samples of breast cancer patients and controls over the 3 MS target plates. Plate 1 and 2 were used as a calibration set, while plate 3 was used as a validation set.

	Plate 1	Plate 2	Plate 3	Total
Breast cancer	36	37	38	111
Controls	30	33	29	92
Total	66	70	67	203

Table 3. Distribution and randomisation of all different stages of breast cancer over the 3 MS target plates. Plate 1 and 2 were used as a calibration set, while plate 3 was used as a validation set.

Stage	Plate 1	Plate 2	Plate 3	Total
DCIS	3	7	7	17
I	12	14	14	40
IIA	11	11	6	28
IIB	5	3	7	15
IIIA	3	1	2	7
IIIB	1	1	2	4
IIIC	1	0	0	1
Total	36	37	38	111

portions (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. Analysis was carried out on 3 consecutive days, Tuesday to Thursday, processing a single plate each day. During the first two days, a calibration set with serum samples from 73 breast cancer patients and 63 controls was used to generate high-resolution MALDI-TOF protein profiles. The last day of the experiment, an independent validation set with serum samples from 38 breast cancer patients and 29 controls was measured.

Isolation of peptides and protein profiling

The isolation of peptides from serum was performed using magnetic beads based weak cation exchange chromatography (MB-WCX) kit from Bruker, mainly according to the manufacturers instructions, and adapted for automation on a 8-channel Hamilton STAR® pipetting robot (Hamilton, Martinsried, Germany). Magnetic beads with WCX-functionality (MB-WCX) were divided in 10 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. MB-WCX binding solution

(10 µl) 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 5 minutes and the magnet was lifted, followed by a 30s 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-WCX washing solution (also provided with the kit) lifting and lowering the magnet as needed. The peptides were eluted from the beads using 10 µl elution solution (from the kit). Stabilization buffer was added (10 µl) and 2 µl of the stabilised eluate was transferred to a fresh 384-well microtiter plate (Greiner). Fifteen µl of α-cyano-4-hydroxycinnamic acid (0.3 g/l in ethanol: acetone 2:1) was added to the 2 µl eluate in the 384-well microtiter plate and mixed carefully. One microliter of this mixture was spotted in quadruplicate on a MALDI AnchorChip™ (Bruker Daltonics, Bremen, Germany).

Data processing and statistical analysis

To increase robustness, the average of four spots was used to represent one serum sample. All unprocessed spectra were exported from the Ultraflex in standard 8-bit binary ASCII format. They consisted of approximately 32,670 mass-to-charge ratio (m/z) values, covering a domain of 960 - 11,168 Dalton. The high-resolution spectra were first lightly smoothed and then, due to the quadratic nature of the TOF-equation, binned using a linear function of the time scale, resulting in bin widths of approximately 0.4 Dalton at the beginning of the spectrum and 1.4 Dalton at the end at the mass/charge scale. Subsequently, we normalised the spectra after baseline correction. In the calibration set, classification error rates were estimated and validated based on a classical Fisher linear discriminant analysis through complete double cross-validation as previously described.[24] This double cross-validated classifier was then applied on the validation set. This set was pre-processed using the exact same procedure as the calibration set. Using the estimated parameters from the calibration set, each sample in the validation set was assigned to the group for which the probability was highest. The error rates are based on sensitivity and specificity values, assuming a prior class probability of 0.5 for each group.

RESULTS

Three different randomised target plates were successfully measured on three consecutive days in the middle of the week. Figure 1 shows a raw data spectrum, directly obtained from the MALDI-TOF mass spectrometer. For further analysis, we first calculated the mean spectrum of each sample across all four spots that were measured for each sample, after pre-processing. The above-described pre-processing

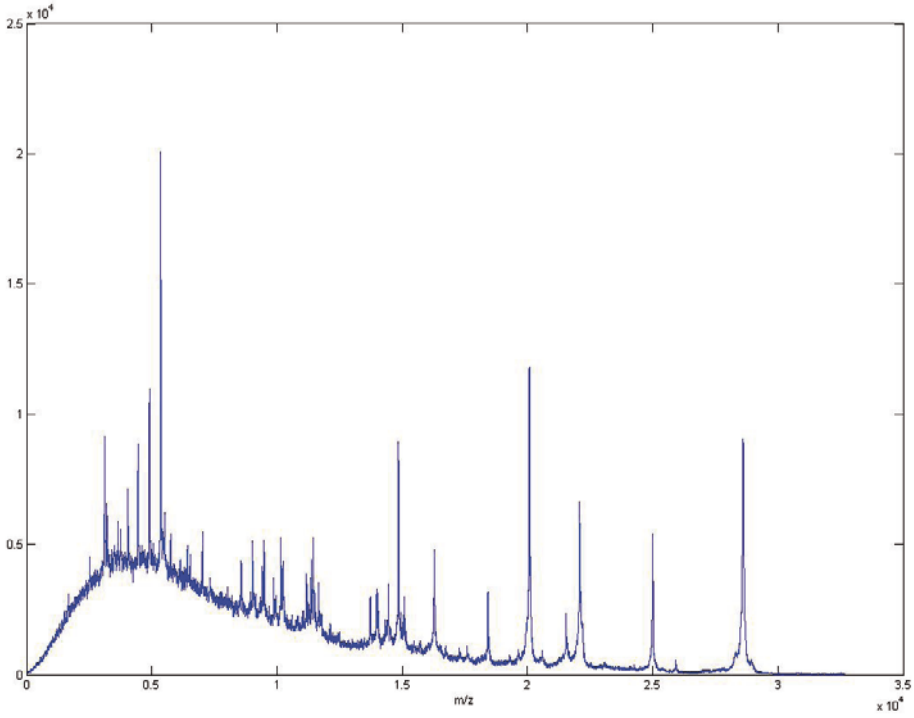


Figure 1. MALDI-TOF spectrum of a breast cancer patient after peptide isolation with WCX magnetic beads. On the Y-axis the relative intensity is shown. The mass to charge ration (m/z) is demonstrated on the X-axis in Dalton.

steps resulted in a sequence of 11,205 normalised m/z values ranging from 960 to 11,168 Dalton, for each individual.

Double cross-validatory analysis and evaluation carried out on the protein spectra of the calibration set (2 target plates) correctly classified 56 of 63 breast cancer patients as malignant. Sixty-one of 73 controls were correctly classified as non-cancer (Table 4a). The misclassified patients in the calibration set included 1 patient with DCIS, 4 stage I patients, 3 with stage IIA and 4 patients with stage IIB. There was no correlation with hormonal status.

These double cross validated results yielded a total recognition rate of 86%, a sensitivity of 88% and a specificity of 84% for the detection of breast cancer within the calibration set. 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 90.3%. To further evaluate possible bias of the double cross-validatory calculations, we performed a permutation exercise, which randomly permutes and reassigns the class labels across subjects and then repeats the entire

Table 4a. Double cross-validatory classification of serum samples **in calibration set**. A positive test results assigns subjects to the breast cancer (BC) group and a negative to the controls.

	Test results for detection of BC		
	Pos	Neg	Total
BC patients	61	12	73
Controls	7	56	63
	68	68	136

Table 4b. Double cross-validatory classification of serum samples **in validation set**.

	Test results for detection of BC		
	Pos	Neg	Total
BC patients	4	34	38
Controls	21	8	29
	25	42	67

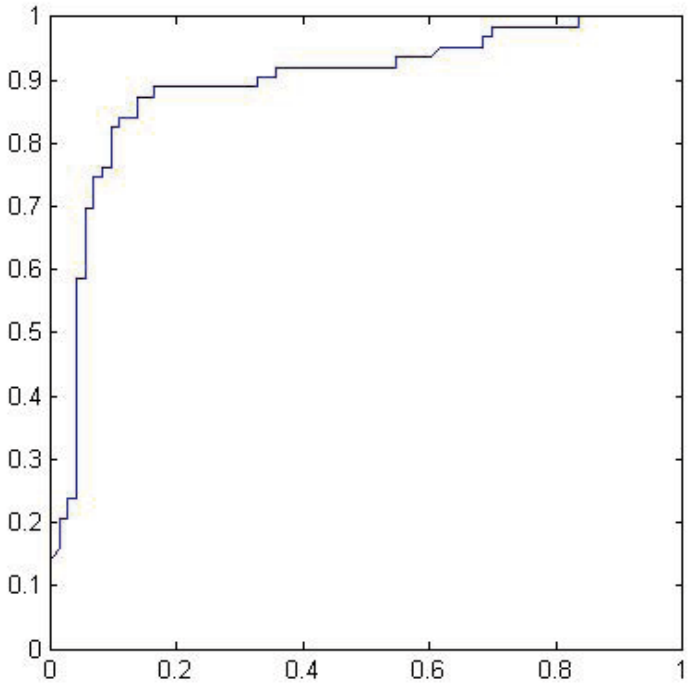


Figure 2. ROC-curve for the double cross-validated two-group classifier. The true positive recognition rate (sensitivity) is demonstrated on the y-axis against the false negative recognition rate (1-specificity) on the x-axis of the classifier.

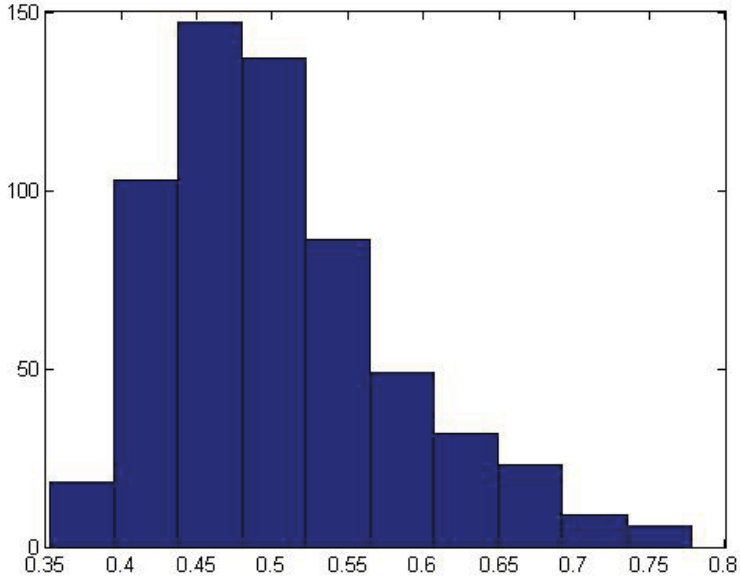


Figure 3. Histogram showing the normal distribution for the misclassification rate in the permutation exercise. The X-axis shows the misclassification rate calculated in the permutation exercise. On the Y-axis the number of permutations is displayed (n=600).

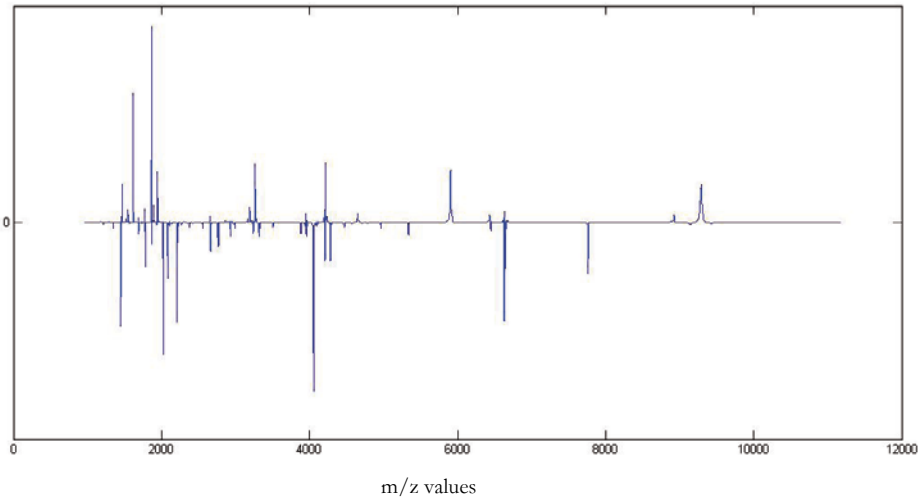


Figure 4. Correlation coefficients of most discriminating principal components with the class indicator. The correlation coefficients were calculated from the linear discriminant weightings.

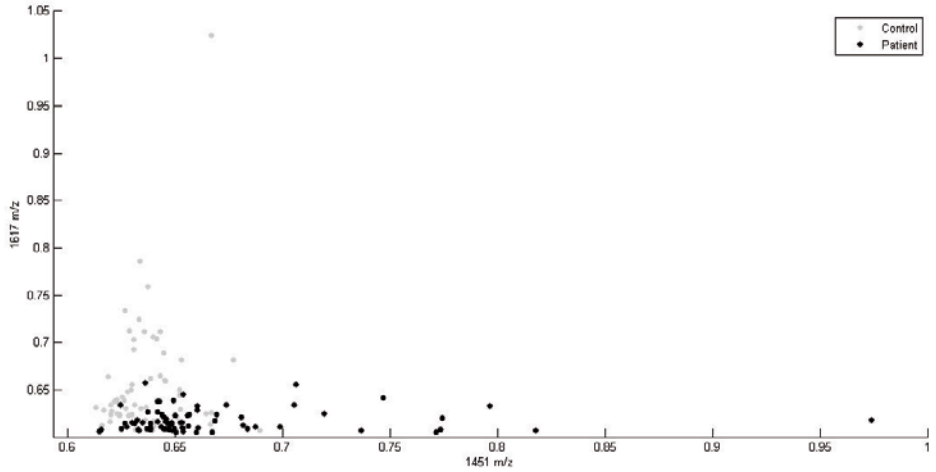


Figure 5a

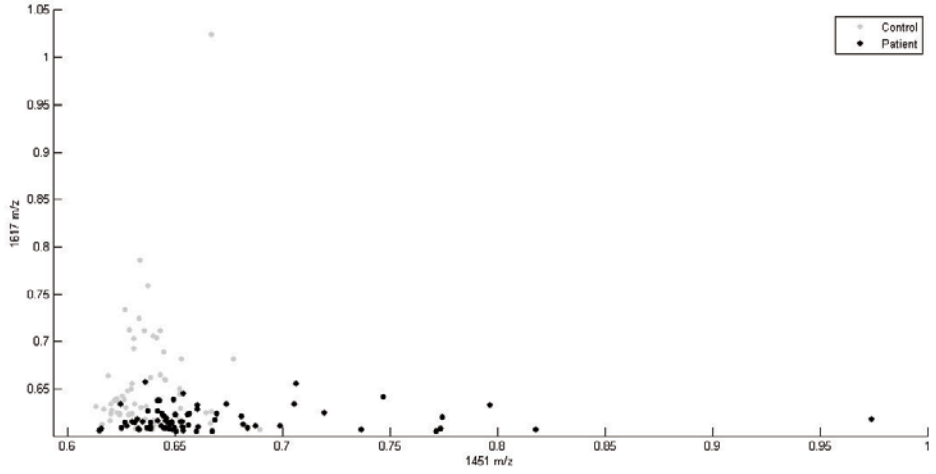


Figure 5b

Figure 5. Scatter plot of the first two principle components on basis of which the classification patient-control group was made.

double cross-validation procedure. Carrying out this procedure more than 600 times resulted in a median recognition rate of 49.0% with a 95% confidence interval of [0.39, 0.69] as shown in figure 3. The median AUC was 49.7% with confidence interval of [0.28, 0.61]. As both median recognition rates and AUC's equal roughly 50.0%, there is thus no substantial evidence of bias remaining within the cross-validatory calculation. Moreover, the actually observed recognition rates as well as AUC are

clearly separated from these permutation-based null-hypothesis confidence intervals, which prove the existence of discriminatory information in the spectra.

A post hoc exploration of the classification model was performed. In the present study 8 peaks that correlated most with the two groups provided most of the between-group separation. Figure 4 shows a plot of the correlation coefficients, with the class indicator, which can be calculated from the linear discriminant weightings in the region between 960 and 11,168 Dalton. Figure 5 shows scatter plots of 4 of these most discriminating peaks between cases from controls.

Consequently, we applied the double cross-validated classifier constructed on the first 2 plates on the spectra of the independent validation set. In this set 21 of 29 controls were correctly classified as non-malignant and only 3 of 38 breast cancer patients were misclassified, as shown in table 4b. From the 4 misclassified patients 1 had DCIS, whereas 2 patients both had stage I disease, according to their post-operative histological report. Nevertheless, these results produce a total recognition rate of 80.9%, a sensitivity of 72% and a specificity of 92% of the classifier in an independent dataset.

DISCUSSION

This validation study shows that breast cancer can be detected by serum protein profiling. We were able to classify breast cancer patients and healthy individuals accurately based upon information in MALDI-TOF serum spectra. In the calibration set the double cross-validated classifier demonstrated a sensitivity and specificity of 88% and 84% respectively. Sixty-one out of 73 controls were correctly classified as non-cancer. Moreover, with a lifetime risk of 1 in 9, it cannot be excluded that some of the control subjects currently are developing or carrying the disease. Unfortunately, since the control group consisted of anonymous symptom-free subjects it was impossible to retrieve the current physical state. The misclassified cancer patients had varying early stages of disease, from stage I to IIB. However, the fact that all but one DCIS patients was recognised in the cancer group, adds to its possible future applicability as a tool for early detection.

More importantly, in the independent validation set the classifier demonstrated a sensitivity and specificity of 72% and 92% respectively. While for the misclassified controls in the validation set the current physical state was also unknown due to their anonymity, the current physical state could easily be retrieved for the misclassified patients. Histological reports showed that one of them had DCIS and two of them were diagnosed with stage I disease. In this case the breast tumour had been resected for 2 years. Interestingly, this protein profile was classified in the

non-cancer group, which was confirmed from the treatment chart at the time of blood collection. When this sample would have been excluded of the analysis, the specificity of the classifier would increase even more in the validation set. Especially this high specificity adds to the potential of serum protein profiles to screen women at high risk for breast cancer, since there appears to be a low chance of false positive test results and unnecessary treatment will be avoided. When combined with mammography the positive predictive value of the proteomic pattern approach could be assessed in a high risk population.[22]

Since a potential drawback of any approach with high dimensional data is the tendency to discover patterns among many variables that may not be a direct result of the pathological state but rather a result of pre-analytical characteristics the need of an independent validation set has been stressed extensively.[14;23;24] In our previous studies, the use of an independent validation set was not possible due the relatively small sample. Therefore, until now we advocated a thorough and stringent study design and double cross-validation of the classification model.[24] This procedure avoided the need for separate test and validation sets to yield unbiased error rate estimates. However, in the current study discriminating protein profiles for the detection of breast cancer could be validated using an independent dataset. Nevertheless, the classifier in the calibration set was constructed following stringent demands. Again, a randomised block design was used to avoid observational bias, ensuring that no batch effects were introduced and artificial between-group separations excluded.[20;21] However, the issue clinically most relevant is the use of an independent validation set for the classification of diseased versus healthy individuals. This is primarily based on a specific problem in the discovery-based research field of clinical proteomics, namely overfitting. Overfitting may occur when multivariate models show apparent discrimination that is actually caused by data over-interpretation, and hence give rise to results that are not reproducible.[14;17;18] Therefore, protection against overfitting of the classifier was maintained by using the double cross validation in the calibration set. In this way, maximal reliability of the classifier was obtained by this procedure. Then, the performance of the classifier was tested in the independent validation set and it proved that breast cancer can indeed reliably be detected by discriminating protein profiles.

Obviously, other most common pitfalls in clinical proteomics such as sample collection, pre-analytical conditions and biological variation were avoided.[15;16;25] Therefore, serum sample collection and pre-analytical factors were rigorously standardised.[19] Furthermore, subjects in both groups were matched for age, although age is recently shown not to bias serum peptodomics.[26] In addition, patient samples from all stages of breast cancer were randomly distributed over three different target plates, excluding these factors as a discriminator in the current classifier.

Interestingly, the classification between cancer and non-cancer was performed using more principal components than in our earlier work when C8 magnetic beads were used. In the present study more than 2 peaks were responsible for most on the between group separation. As shown in figure 2 and 5, these include peaks with 1451, 1617, 5906 and 6644 m/z. Since the primary focus of this study was to assess and validate the feasibility of protein profiles based detection of breast cancer and therefore merely concentrated on pattern diagnostics, we have not identified these potential biomarkers yet. However, the controversy about the use of protein profiles as a pattern diagnostic without analysis of the diagnostic biomarkers still remains to be solved for its clinical application. Some have argued that low molecular weight proteins in serum, the serum peptidome, is nothing but aspecific biological trash and therefore does not yield any reliable biomarkers in the currently technically available mass range.[27-29] Nonetheless, recently it was postulated that although discriminating peptides do indeed belong to well known coagulation and complement pathways, their patterns or signatures do most certainly indicate the presence of cancer.[26] This study showed that most of the cancer-type specific biomarker fragments were generated in patient serum by enzymatic cleavage at previously known endoproteage cleavage sites after the blood sample was collected.[30] Villanueva et al. postulated that these cancer-specific low molecular weight proteins in the serum peptidome are an indirect snapshot of the enzyme activity in tumour cells. We support their hypothesis that discriminating serum protein profiles are a compilation of surrogate markers for the detection and classification of certain types of tumours.

In conclusion, the present study demonstrated that serum protein profiles generated with mass spectrometry could be indicative of the presence breast cancer. In order to obtain most realistic estimates of the discriminating power of serum protein profiles a classifier was constructed in a randomised block design. Maximal reliability in classification was achieved through double cross-validation of the classifier while maintaining protection against overfitting. Principally the potential of proteomic signatures from high dimensional mass spectrometry data as highly reliable diagnostic classifiers for the detection of breast cancer was actually confirmed in the independent validation set. The fact that high sensitivity and specificity could be maintained in the validation set is the first steps towards a new diagnostic approach in breast cancer.

REFERENCES

1. Jemal,A., Siegel,R., Ward,E., Murray,T., Xu,J., Smigal,C., and Thun,M.J. (2006) Cancer statistics, 2006. *CA Cancer J.Clin.*, 56, 106-130.
2. Edwards,B.K., Brown,M.L., Wingo,P.A., Howe,H.L., Ward,E., Ries,L.A., Schrag,D., Jamison,P.M., Jemal,A., Wu,X.C., Friedman,C., Harlan,L., Warren,J., Anderson,R.N., and Pickle,L.W. (2005) Annual report to the nation on the status of cancer, 1975-2002, featuring population-based trends in cancer treatment. *J.Natl.Cancer Inst.*, 97, 1407-1427.
3. Schairer,C., Mink,P.J., Carroll,L., and Devesa,S.S. (2004) Probabilities of death from breast cancer and other causes among female breast cancer patients. *J.Natl.Cancer Inst.*, 96, 1311-1321.
4. Veronesi,U., Boyle,P., Goldhirsch,A., Orecchia,R., and Viale,G. (2005) Breast cancer. *Lancet*, 365, 1727-1741.
5. Gotzsche,P.C. and Nielsen,M. (2006) Screening for breast cancer with mammography. *Cochrane.Database.Syst.Rev.*,CD001877.
6. Astley,S.M. (2004) Computer-based detection and prompting of mammographic abnormalities. *Br J Radiol*, 77, S194-S200.
7. Simpson,J.F., Gray,R., Dressler,L.G., Cobau,C.D., Falkson,C.I., Gilchrist,K.W., Pandya,K.J., Page,D.L., and Robert,N.J. (2000) Prognostic Value of Histologic Grade and Proliferative Activity in Axillary Node-Positive Breast Cancer: Results From the Eastern Cooperative Oncology Group Companion Study, EST 4189. *J Clin Oncol*, 18, 2059-2069.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. Ransohoff,D.F. (2004) Rules of evidence for cancer molecular-marker discovery and validation. *Nat.Rev.Cancer*, 4, 309-314.
15. Boguski,M.S. and McIntosh,M.W. (2003) Biomedical informatics for proteomics. *Nature*, 422, 233-237.
16. 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.
17. 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.

18. 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.
19. 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.
20. Box,G.E.P., Hunter W.G., and Hunter J.S. (1978) *Statistics for experimenters*. John Wiley & Sons, Inc..
21. Cox D.R. and Reid N. (2000) *The theory of the design of experiments*. Chapmann/Hall CRC.
22. Espinosa,E., Redondo,A., Vara,J.A., Zamora,P., Casado,E., Cejas,P., and Baron,M.G. (2006) High-throughput techniques in breast cancer: A clinical perspective. *Eur.J Cancer*.
23. 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.
24. de Noo,M.E., Mertens,B.J., Ozalp,A., Bladergroen,M.R., van der Werff,M.P., van de Velde,C.J., Deelder,A.M., and Tollenaar,R.A. (2006) Detection of colorectal cancer using MALDI-TOF serum protein profiling. *Eur.J Cancer*, 42, 1068-1076.
25. Baumann,S., Ceglarek,U., Fiedler,G.M., Lembcke,J., Leichtle,A., and Thiery,J. (2005) Standardized Approach to Proteome Profiling of Human Serum Based on Magnetic Bead Separation and Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. *Clin. Chem.*, 51, 973-980.
26. Villanueva,J., Shaffer,D.R., Philip,J., Chaparro,C.A., Erdjument-Bromage,H., Olshen,A.B., Fleisher,M., Lilja,H., Brogi,E., Boyd,J., Sanchez-Carbayo,M., Holland,E.C., Cordon-Cardo,C., Scher,H.I., and Tempst,P. (2006) Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J Clin Invest*, 116, 271-284.
27. Diamandis,E.P. and van der Merwe,D.E. (2005) Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations. *Clin Cancer Res.*, 11, 963-965.
28. Sorace,J.M. and Zhan,M. (2003) A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC.Bioinformatics.*, 4, 24.
29. Koomen,J.M., Li,D., Xiao,L.C., Liu,T.C., Coombes,K.R., Abbruzzese,J., and Kobayashi,R. (2005) Direct tandem mass spectrometry reveals limitations in protein profiling experiments for plasma biomarker discovery. *J Proteome.Res.*, 4, 972-981.
30. Liotta,L.A. and Petricoin,E.F. (2006) Serum peptidome for cancer detection: spinning biologic trash into diagnostic gold. *J Clin Invest*, 116, 26-30.

Chapter 8

General discussion

M.E. de Noo, A.M. Deelder, R.A.E.M. Tollenaar,
L.H. Bouwman

*World Journal of Gastroenterology. 2006;
12(41):6594-601*

BACKGROUND

There is an urgent need for new biomarkers in oncology to improve early detection, monitor disease outcome and find targets for more individualised therapy. A field of recent interest is clinical proteomics, which was reported to lead to high sensitivity and specificities for early detection of several solid tumours.[1;2] This emerging field uses mass spectrometry based protein profiles/patterns of easy accessible body fluids to distinguish cancer from none-cancer patients. This would be a solution to the problem that cancer is often diagnosed in late stages, when curative resection of the diseased organ is not possible anymore and the disease has already metastised, dropping survival rates dramatically. However, after the initial hype in early 2002 critical noise has been heard on several aspects of serum proteomics. In this paper we describe the hopes and fears for the introduction of clinical proteomics for (early) detection of CRC.

COLORECTAL CANCER

Colorectal adenocarcinoma is the third most common cancer and the fourth most frequent cause of death due to cancer worldwide. Worldwide almost one million new cases occur yearly, with 492,000 related deaths.[3] In developed countries it is the second most common tumour, with a lifetime risk of 5%, but its incidence and mortality are now decreasing.[4;5] Surgery is the cornerstone of therapy when the disease is confined to the bowel wall. This results in 70 to 80% of patients who have tumours that, at diagnosis, can be resected with curative intent.[6] After curative surgery the five-year survival rate for patients with localised disease is 90%, decreasing to 65% in case of metastised disease in the lymph nodes. Adjuvant radiation therapy, chemotherapy, or both are useful in selected patients. Classification of tumours into pathogenetical subtypes with distinct clinical courses enables clinicians to target therapy. For CRC TNM staging system remains the golden standard and relies entirely on morphological appearance of the tumour. However, tumours with similar histopathological characteristics may have different clinical outcome and responsiveness to therapy.[7] Therefore, more individualised treatment would benefit the individual patient and avoid unnecessary morbidity. Nonetheless, early detection of CRC will increase survival most, in view of the fact that 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.[8-10]

BIOMARKERS

In cancer research biomarkers are molecules that indicate the presence of cancer in the body. Most biomarkers are based on abnormal changes or mutation in genes, RNA, proteins and metabolites. Since the molecular changes that occur during tumour development can take place over a number of years, some biomarkers can potentially be used to detect colorectal cancer early. Furthermore, they might be used to predict prognosis, monitor disease progression and therapeutic response. Gion et al. classified different circulating biomarkers according to their clinical application.[11] These candidate biomarkers however, are frequently found in relatively low concentrations amid a sea of other biomolecules, so biomarker research and possible diagnostic tests depend critically on the ability to make high sensitive and accurate biochemical measurements. Ideally, such biomarkers should be specific to the disease and easy accessible, such as serum, plasma or urine, increasing their clinical applicability.

Carcinoembryonic antigen (CEA) is the best-characterised serologic tumour marker for CRC. However, its use as a population based screening tool for early detection and diagnosis of CRC is hindered by its low sensitivity and specificity. Fletcher showed that for screening purposes in a normal population, a cut-off concentration of 2.5 µg/L CEA would yield a sensitivity of 30-40%. Based on these data he calculated that there would be 250 false positive tests for every true positive test, i.e. a patient with cancer. Furthermore, 60% of the cancers would not be detected. The same poor sensitivity applies for diagnosis of CRC. In addition, as CEA can be elevated in the absence of malignancy, specificity is also impaired.[12-15]

Faecal occult-blood testing (FOBT) is another biomarker for which clinical trials have shown evidence of a decreased risk of death. This approach is a non-invasive option that limits the need for follow-up colonoscopy to patients with evidence of bleeding. Neoplasms bleed intermittently, however, allowing many to escape detection with faecal occult-blood testing. Annual retesting is therefore necessary but is still insufficient, detecting only 25 to 50% of colorectal cancers and 10% of adenomas. The specificity of FOBT is also limited by frequent false positive reactions to dietary compounds, medications, and gastrointestinal bleeding from causes other than colorectal cancer.[16-18]

A NEW DIAGNOSTIC PARADIGM: CLINICAL PROTEOMICS

In 2002 several studies discriminated patients with various cancers from healthy subjects on the basis of presence/absence of multiple low-molecular-weight serum

proteins using SELDI-TOF mass spectrometry technologies.[19-22] The authors hypothesised that proteomic patterns are correlated to biological events occurring in the entire organism and are likely to change in the presence of disease. New types of bioinformatic pattern recognition algorithms were used to identify patterns of protein changes in order to discriminate cancer patients from healthy individuals with promising results.

Petricoin and his co-workers stated that finding a single disease-related biomarker is like searching for a needle in a haystack; each entity has to be separated and identified individually.[23;24] Moreover, they postulated that the blood proteome constantly changes as a consequence of the perfusion of the diseased organ adding, subtracting, or modifying the circulating proteome. These differences might be the result of proteins being abnormally produced or shed and added to the serum proteome, clipped or modified as a consequence of the disease process, or subtracted from the proteome owing to disease-related proteolytic degradation pathways. Therefore, protein pattern diagnostics would provide easier and more reliable tools for detection of cancer. The advantages of the SELDI proteomic pattern approach were stressed in several papers. In addition to the high sensitivity and specificity, cost-effectiveness, easy accessibility of body fluid and especially the high-throughput, ultimately allowing application in future screening studies, were mentioned.[20;25] Next to these hopeful voices, soon critical notes were made on analytical reproducibility and the use of the so-called black box approach, lacking identification of discriminating proteins.

In the next paragraphs this paper will focus on the current status of clinical proteomics research in oncology and will reflect on pitfalls and fears in this relatively new area in clinical medicine: reproducibility issues and pre-analytical factors; statistical issues; and identification and nature of discriminating proteins/peptides.

REPRODUCIBILITY ISSUES AND PRE-ANALYTICAL FACTORS

Boguski and McIntosh were among the first to argue that serum proteomics may be susceptible to observational biases. They stated that any confounding factor could conceivably cause a phenotypic response that might be confused with a specific characteristic of the disease process under study.[26] Confounding factors such as smoking, diet and preoperative stress, but also sample collection and quality, trouble a reliable and clear differentiation of a normal or malignant status. Another cause for concern mentioned in this study, is the sample quality and number. The authors favoured use of homogeneous groups with sufficient sample size and stringent standard procedures for serum collection, an aspect which is also advocated in other

studies.[27;28] Another critical study questioned the reliability of the presence of statistically significant signals at M/Z values less than 500, as used in one of the first studies. Sorace et al. claimed that the presence of statistically significant bands of low M/Z includes degradation products of higher molecular weight macromolecules or a matrix effect. Furthermore, this study cautioned for poor reproducibility of experimental conditions of chip based mass spectrometry.[29] This is also reported by another group, which showed the poor reproducibility of the SELDI-TOF ovarian cancer data. Baggerly and colleagues postulated that this could partly be contributed to baseline correction, poor sample features in noise regions and even a change of protocol mid-experiment.[30] Most importantly, the promising results that were reported earlier could not be reproduced and therefore stressed the importance of standardised approaches, stringent experimental design. Furthermore, their study pointed out that strong pre-processing of the protein spectra is required in order tot obtain reliable classification results in the search for new biomarkers.

Possible confounding factors can be categorised into three sources of variation and bias: biological variation, pre-analytical variation and analytical reproducibility. Biological variation, consist of both environmental and individual factors, such as race, age, diet, smoking, stress, general physical condition, and use of drugs, and may also influence serum protein profiles. However, at the present no data have been published on this source of variation. Nevertheless, in a previous study our group analysed pre-analytical and reproducibility issues of our MALDI-TOF approach.[31] The pre-analytical variations corresponded to the logistical conditions in the routine clinical setting; the effects of sample handling and storage. So far, only few other studies have reported on the effects of different serum sample preparations and the use of a magnetic-beads-based approach to capture and concentrate serum proteins for MALDI-TOF mass spectrometry.[32-34] Where Villanueva et al. mostly focused on influences of different magnetic beads capturing and its automation on the reproducibility of serum protein profiles, Baumann and co-workers mainly studied pre-analytical variation of sample handling. In table 1, different results of sample handling experiments of the above mentioned studies are summarised. For clinical studies the use of two freeze/thaw cycles is recommended by 3 out of 4 manuscripts. This in mainly due to logistical reasons, such as the ´standard´ for centralised sample collection in large hospitals. The point all authors agreed on is the influence of sample handling, i.e. the time venous blood is left to stand before serum centrifugation. This aspect appears to account for the largest effect on serum or plasma protein profiles. Consequently, standardised sample collection and a well documented population are recommended in all performed studies. Standardised protocols should be used from the point of sample collection, sample handling, storage and freezing of the samples. Although the importance of homogeneity and uniformity within sample groups must

Table 1. Recommendations of various pre-analytical variations from three MALDI-TOF based reproducibility studies.

	Blood component	Peptide isolation	Temp before sample handling	Time before centrifugation	Storage of serum	Freeze/thaw cycles	Circadian rhythm effect
Baumann et al.	Serum/Plasma	C3, C8, C18 beads	21° C	< 30 min	-80° C	1	N.A.
de Noo et al.	Serum	C8 beads	21° C	Ideally < 30 min, practically < 2-4 hrs	N.A.	2	No effect
West-Nielsen et al.	Serum/Plasma	C8 beads	21° C	< 8 hrs	-20/ -80 ° C	1	N.A.

once again be stressed, variation of such factors can not totally be excluded in a clinical setting. In all, when these recommendations are strictly followed and both clinical and analytical factors are controlled, we think that the methodology can be standardised to a level which allows application as a tool in biomarker discovery.

STATISTICAL ISSUES

As in all research with high dimensional data, two practical realities constrain the analysis of mass spectra in proteomics. The first is the 'curse of dimensionality': the number of features characterizing these data is in the thousands or tens of thousands. The second is the 'curse of dataset sparsity': the number of samples is limited. Somorjai et al. showed the influences of these two curses on classification outcomes. Both the sample per feature ratio, which should be 5 to 10 ideally, and feature selection are pivotal importance for reliable classification and biological optimal relevance.[35;36]

Previous to any feature selection or classification, raw mass spectra have to be submitted to so-called pre-processing. During this process noise of protein/peptide mass spectra is reduced and spectra are normalised. Furthermore, smoothing, binning and baseline correction are also performed during pre-processing of the data. Currently, there is a lot of discussion between several groups on how to establish the best method, because data pre-processing is extremely important. There are complex interactions between baseline subtraction, normalization, noise estimation, and peak identification, and therefore these steps should not be considered in isolation.[31;37-40]

Another recurrent topic for debate is the bioinformatic approach and statistical analysis of protein spectra. Clinically most relevant is the issue of an independent validation set for the classification of diseased versus healthy individuals. This is pri-

marily based on a specific problem in the discovery-based research field of clinical proteomics, namely 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. [30;41;42] 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.[43] Although we have shown this in a previous study, the general opinion is in favor performing a classification study with independent validation. Furthermore, feature selection is also given a lot of attention by statisticians in the field. Several experimental investigations have been made with different peak feature selection methods. A common approach thus far is analysing the data in two phases. First, the peaks in the spectra are extracted and quantified. Secondly, a resulting matrix of peak quantifications is created. For more detailed information on this statistical matter, we refer to the literature.[37;44-46]

IDENTIFICATION AND NATURE OF DISCRIMINATING PROTEINS

The controversy about the use of protein profiles as a pattern diagnostic without identification of the individual diagnostic biomarkers remains to be solved before its clinical application. Whereas the first clinical proteomics studies published their classification method mainly as a black box study, nowadays identification of the most discriminating proteins or peptides is required for publication in most scientific journals. Identification and functional analysis of these discriminating proteins/peptides might render new insights on tumour development and environmental responsiveness, which could eventually be translated into new diagnostic and prognostic insights for the clinician. Unfortunately, little success has been booked so far in assigning reproducible discriminating biomarkers.[35;42]

Furthermore, several studies have identified their discriminating peaks as components of the coagulation cascade or complement system.[47-51] So, in contrast to the original reflection that discriminating proteomic patterns would identify cancer-specific proteins, it appears that these potential markers belong to the normal serum and plasma proteome. Consequently, some investigators have argued that low molecular weight proteins in serum, the serum peptidome, is nothing but aspecific biological trash and therefore does not yield any reliable biomarkers in the currently technically available mass range.[29;52] Others have proposed that the discriminatory protein peaks represent acute phase reactants that are present in serum in extremely high concentrations.[49;53] Conversely, recently a study reported that although discriminating peptides do indeed belong to the well known coagulation and complement

pathways, their patterns or signatures can nevertheless indicate the presence of cancer. Villanueva et al. showed that most of the cancer-type specific biomarker fragments were generated in patient serum by enzymatic cleavage at previously known endoprotease cleavage sites after the blood sample was collected.[54;55] They postulated that the discriminating peptides originated after ex vivo proteolysis by tumour specific proteases of high abundance protein fragments primarily generated by the coagulation and complement enzymatic cascades. In this view, they consider these cancer-specific low molecular weight proteins in the serum peptidome an indirect snapshot of the enzyme activity in tumour cells. We support their hypothesis that proteolytic process profiles in the serum peptidome hold important information that may have direct clinical utility as a surrogate marker for the detection and classification of certain types of tumours. Unique proteases may be shed by tumour cells or reflect activity of the host immune response, which may contribute to new proteins such as chemokines and lymfokines. These processes result in subtle changes in low molecular proteomic signatures, which may ultimately be used for classification methods in various cancers and disease in the future.[54] Proteases have been extensively implicated in the development and progression of cancer.[56;57] Song et al. recently stated that proteolytic processing of high abundance host-response proteins actually amplifies the signal of potentially low-abundance biologically active disease markers such as proteases. Therefore, it might be expected that more convenient and reliable blood proteins and peptides simply serve as an endogenous substrate pool for proteases as surrogate markers for the detection and classification of cancer.[58]

Another recurrent topic of debate is which blood component is best used for protein profiling and peptidome analysis. Some investigators favour the use of plasma because they presume that, in serum, ongoing enzymatic activity, occurring during clotting, is likely to cleave even proteins that are not involved in biological relevant pathways.[53;59] Others, however, advocate the use of serum. We support the hypothesis that since the kidneys rapidly clear peptides smaller than 4 kDa which are in vivo generated in the circulation, the majority of peptides in blood samples exist from ex vivo proteolysis. This explains that low abundance proteins, including possible tumour markers, may be totally obscured and not retraceable during direct mass spectrometry. However, it has recently been shown that exogenous proteases are functionally measurable in serum, yet in higher concentrations than in plasma.[54]

Functional proteomics studies allow the investigation of environmental factors over time, rendering the monitoring of metabolic responses to various stimuli. Hence, post translational modifications can be studied, whereas they can not be detected by genomic studies. Posttranslational modifications changes like glycosylation of proteins and lipids are a common feature in colorectal cancer and influence cancer

cell behaviour and can be detected using mass spectrometry due to characteristic mass shifts.[60] We expect that both phosphoproteomics and/or glycoproteomics, enabling study of crucial post translational modifications of proteins in the cancer pathway, will revolutionize our understanding of the function of these proteins, and hence render new insights for monitoring and therapy.

CLINICAL PROTEOMICS IN CRC

So far, few protein profiling studies have been published on the detection of CRC, of two were based on SELDI/TOF and one on MALDI-TOF mass spectrometry. The first SELDI/TOF study showed seven potential biomarkers that could differentiate CRC patients from patients with colorectal adenoma with a sensitivity of 89% and specificity of 83%. The seven potential biomarkers have a large range in mass values, differing from 4654 till 21,742 Da.[61] A more recent published study found 5 possible biomarkers to differentiate between healthy control subjects and CRC patients. For three of these potential markers they found a sensitivity and specificity between 65% and 90%. They reported that m/z 3100, 3300, 4500, 6600 and 28,000 were the most important biomarkers.[62] Our group used MALDI-TOF mass spectrometry to differentiate CRC patients from healthy controls. In a randomised block design preoperative serum samples obtained from 66 colorectal cancer patients and 50 controls were used to generate high-resolution MALDI-TOF protein profiles.[43] After preprocessing of the spectra, linear discriminant analysis with double cross-validation, based on principal component analysis was used to classify the protein profiles. 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. In our study two first principal components accounted for most of the between-group separation, both with a m/z between 1000 and 2000 Da.

Although a lot of research has been done using 2D gel electrophoresis to detect possible biomarkers and targets for CRC, this falls outside the scope of this paper since this technique can not be scaled up to a directly applicable diagnostic test. On the other hand, recently a screening assay based on APC protein truncation test has been proposed and other studies mention the potential use of protein microarrays.[2;63-65] However, studies linking large protein expression patterns with clinical outcome in colorectal cancer are still in their infancy. To be able to predict occurrence of disease, and treatment outcome, more studies on genotype-phenotype correlations are needed both in sporadic and in hereditary colorectal cancer.

FUTURE PERSPECTIVES

The best anticancer strategies still rely on early detection followed by close monitoring for early relapse so that therapies can be appropriately adjusted.[66] In addition, new targets for therapy are a constant subject of study in oncology. In fact, increased understanding of the molecular mechanisms of cancer progression may refine treatment and management of patients. Advances in genomics and proteomics may lead to earlier detection of cancer and may enable a more precise classification of (smaller subsets of) patients based on their predicted response to individual therapies. Conceptually, proteomics is more suitable than genomics for novel targeted therapies, since most of protein biomarkers are based on aberrant protein signalling circuits represented by post translational modifications. The dynamic range of the proteome allows more insight in the functional state of a cell, tissue or organ over time. Besides, protein profiling and classification of several components of multiple aberrant cell signalling cascades would be expected to predict disease behaviour better than just single pathways in isolation.[64] Therefore, proteomics could be expected to render better insight in pathogenetic mechanisms, disease progression and treatment response. This is of paramount importance as cancer advances dynamically and affects heterogeneous cell populations, either as a part of cancer or as a part of a tumour-host reaction.[49;67]

Further refinement of serum protein profiles is needed before these mass spectrometry based techniques become part of clinical routine. Nowadays, several studies have carefully evaluated reproducibility, automation, sample throughput and sensitivity of serum proteomic techniques. The first problems related to these factors seem to have been overcome due to stringent standardised approaches as described earlier. However, proteomics studies still have several drawbacks: 1) current tools only allow narrow-range analyses, 2) identification of proteins of interest remains cumbersome, 3) protein studies address mixtures of high complexity. Hence, due to the dynamic ranges of the human proteome and the lack of amplification methods in protein studies, targeted proteomics techniques for (quantitative) identification of low-abundant proteins have to be further investigated.[68] Another approach to study proteins at a functional level might be the use of array-based proteomics platforms. This techniques offers the potential for highly multiplex and sensitive analysis of serum or tumour proteins.[64] Using this direct approach of studying the proteomic circuitry would theoretically allow for the creation of functional signalling maps of cancers, even at the level of the individual patient. Regarding identification of potential biomarkers, limitations of direct MS/MS have been stressed before as well as the fact that antibody-approaches may yield higher sensitivity.[53;54]

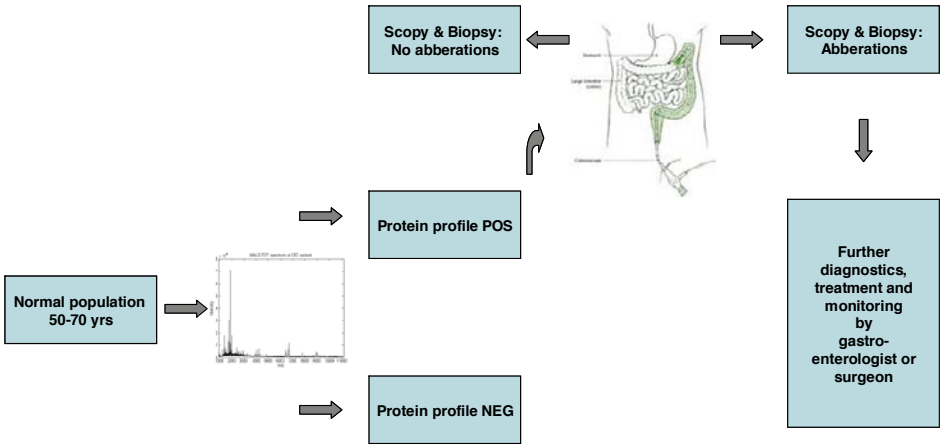


Figure 1. Flow chart of possible clinical application of MALDI-TOF

In the next era research in oncology will drift to more individualised medicine. In this view, molecular profiling forms a welcome addition to the pathology report of cancer. Until now, histopathological staging and demographics have been used to predict disease outcome. However, we believe that protein profiling and other proteomics techniques may lead to more individualised medicine and tailor made therapy.[69;70] At first, both approaches should be used complementary instead of competitively.

It is unlikely that in the next decade, serum protein profiles will certainly replace the current gold standard colonoscopy for the diagnosis of CRC. Nevertheless, we hypothesise that MALDI-TOF based serum protein profiles, once validated in independent studies, could be used as selection criteria for the more invasive and time consuming diagnostic colonoscopy (Figure 1). Eventually, with the present debate on screenings programs for colorectal cancer in several countries, clinical proteomics may replace and surpass the use of faecal occult-blood testing (FOBT). When in independent validation studies sensitivity and specificity remain about 90% protein profiling might even replace FOBT, since this approach has a lower specificity and a number of disadvantages. Non-bleeding tumours and more relevant, polyps and adenomas can not be detected using FOBT, whereas we expect to realise this with serum protein profiling within the next decade.[17;18]

So, although the current reality may not have kept pace with previous expectations and the translation from bench to beside is more laborious than initially thought, there is supporting evidence for the potential great use of clinical proteomics in oncology. Particularly, when efforts for technical innovations to further increase sensitivity and specificity of proteomic techniques will be implemented and more sensi-

tive methods for protein identification on alternations are developed. In combination with the use and set-up of well-defined cases with well documented serum banks, including not only CRC samples, but also inflammatory disease and polyps, serum protein profiling may propel diagnostic research in CRC in the right direction.

REFERENCES

1. Chambers,G., Lawrie,L., Cash,P., and Murray,G.I. (2000) Proteomics: a new approach to the study of disease. *J.Pathol.*, 192, 280-288.
2. Posadas,E.M., Simpkins,F., Liotta,L.A., Macdonald,C., and Kohn,E.C. (2005) Proteomic analysis for the early detection and rational treatment of cancer--realistic hope? *Ann.Oncol.*, 16, 16-22.
3. Weitz,J., Koch,M., Debus,J., Hohler,T., Galle,P.R., and Buchler,M.W. (2005) Colorectal cancer. *Lancet*, 365, 153-165.
4. Russo,M.W., Wei,J.T., Thiny,M.T., Gangarosa,L.M., Brown,A., Ringel,Y., Shaheen,N.J., and Sandler,R.S. (2004) Digestive and liver diseases statistics, 2004. *Gastroenterology*, 126, 1448-1453.
5. Jemal,A., Tiwari,R.C., Murray,T., Ghafoor,A., Samuels,A., Ward,E., Feuer,E.J., and Thun,M.J. (2004) Cancer statistics, 2004. *CA Cancer J Clin*, 54, 8-29.
6. Pfister,D.G., Benson,A.B., III, and Somerfield,M.R. (2004) Clinical practice. Surveillance strategies after curative treatment of colorectal cancer. *N.Engl.J Med.*, 350, 2375-2382.
7. Liefers,G.J. and Tollenaar,R.A. (2002) Cancer genetics and their application to individualised medicine. *Eur.J.Cancer*, 38, 872-879.
8. 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.
9. 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.
10. Hawk,E.T. and Levin,B. (2005) Colorectal cancer prevention. *J Clin Oncol*, 23, 378-391.
11. Gion,M. and Daidone,M.G. (2004) Circulating biomarkers from tumour bulk to tumour machinery: promises and pitfalls. *Eur.J Cancer*, 40, 2613-2622.
12. Duffy,M.J., van Dalen,A., Haglund,C., Hansson,L., Klapdor,R., Lamerz,R., Nilsson,O., Sturgeon,C., and Topolcan,O. (2003) Clinical utility of biochemical markers in colorectal cancer: European Group on Tumour Markers (EGTM) guidelines. *Eur.J.Cancer*, 39, 718-727.
13. Fletcher,R.H. (2002) Rationale for combining different screening strategies. *Gastrointest.Endosc.Clin.N.Am.*, 12, 53-63.
14. Winawer,S., Fletcher,R., Rex,D., Bond,J., Burt,R., Ferrucci,J., Ganiats,T., Levin,T., Woolf,S., Johnson,D., Kirk,L., Litin,S., and Simmang,C. (2003) Colorectal cancer screening and surveillance: clinical guidelines and rationale-Update based on new evidence. *Gastroenterology*, 124, 544-560.
15. Ouyang,D.L., Chen,J.J., Getzenberg,R.H., and Schoen,R.E. (2005) Noninvasive testing for colorectal cancer: a review. *Am.J.Gastroenterol.*, 100, 1393-1403.
16. Pignone,M., Rich,M., Teutsch,S.M., Berg,A.O., and Lohr,K.N. (2002) Screening for colorectal cancer in adults at average risk: a summary of the evidence for the U.S. Preventive Services Task Force. *Ann.Intern.Med.*, 137, 132-141.
17. Ransohoff,D.F. and Lang,C.A. (1997) Screening for colorectal cancer with the fecal occult blood test: a background paper. American College of Physicians. *Ann.Intern.Med.*, 126, 811-822.
18. Ransohoff,D.F. (2005) Colon cancer screening in 2005: status and challenges. *Gastroenterology*, 128, 1685-1695.
19. 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.

20. Petricoin,E.F., Ardekani,A.M., Hitt,B.A., Levine,P.J., Fusaro,V.A., Steinberg,S.M., Mills,G.B., Simone,C., Fishman,D.A., Kohn,E.C., and Liotta,L.A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359, 572-577.
21. 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.
22. 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.
23. Petricoin,E.F. and Liotta,L.A. (2002) Proteomic analysis at the bedside: early detection of cancer. *Trends Biotechnol.*, 20, S30-S34.
24. Wulfkuhle,J.D., Liotta,L.A., and Petricoin,E.F. (2003) Proteomic applications for the early detection of cancer. *Nat.Rev.Cancer*, 3, 267-275.
25. Petricoin,E.E., Paweletz,C.P., and Liotta,L.A. (2002) Clinical applications of proteomics: proteomic pattern diagnostics. *J.Mammary.Gland.Biol.Neoplasia.*, 7, 433-440.
26. Boguski,M.S. and McIntosh,M.W. (2003) Biomedical informatics for proteomics. *Nature*, 422, 233-237.
27. Diamandis,E.P. (2003) Re: Serum proteomic patterns for detection of prostate cancer. *J.Natl. Cancer Inst.*, 95, 489-490.
28. Diamandis,E.P. (2002) Proteomic patterns in serum and identification of ovarian cancer. *Lancet*, 360, 170-171.
29. Sorace,J.M. and Zhan,M. (2003) A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC.Bioinformatics.*, 4, 24.
30. 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.
31. 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.
32. 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.
33. Baumann,S., Ceglarek,U., Fiedler,G.M., Lembcke,J., Leichtle,A., and Thiery,J. (2005) Standardized Approach to Proteome Profiling of Human Serum Based on Magnetic Bead Separation and Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. *Clin. Chem.*, 51, 973-980.
34. West-Nielsen,M., Hogdall,E.V., Marchiori,E., Hogdall,C.K., Schou,C., and Heegaard,N.H. (2005) Sample handling for mass spectrometric proteomic investigations of human sera. *Anal. Chem.*, 77, 5114-5123.
35. 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.
36. Morris,J.S., Coombes,K.R., Koomen,J., Baggerly,K.A., and Kobayashi,R. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics.*.
37. 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.

38. Yasui,Y., Pepe,M., Thompson,M.L., Adam,B.L., Wright,G.L., Jr., Qu,Y., Potter,J.D., Winget,M., Thornquist,M., and Feng,Z. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics.*, 4, 449-463.
39. Eilers,P.H. (2003) A perfect smoother. *Anal.Chem.*, 75, 3631-3636.
40. 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.
41. Ransohoff,D.F. (2004) Rules of evidence for cancer molecular-marker discovery and validation. *Nat.Rev.Cancer*, 4, 309-314.
42. 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.
43. de Noo,M.E., Mertens,B.J., Ozalp,A., Bladergroen,M.R., van der Werff,M.P., van de Velde,C.J., Deelder,A.M., and Tollenaar,R.A. (2006) Detection of colorectal cancer using MALDI-TOF serum protein profiling. *Eur.J Cancer*, 42, 1068-1076.
44. 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.
45. Wang,X., Zhu,W., Pradhan,K., Ji,C., Ma,Y., Semmes,O.J., Glimm,J., and Mitchell,J. (2006) Feature extraction in the analysis of proteomic mass spectra. *Proteomics.*, 6, 2095-2100.
46. Levner,I. (2005) Feature selection and nearest centroid classification for protein mass spectrometry. *BMC.Bioinformatics.*, 6, 68.
47. Koomen,J.M., Zhao,H., Li,D., Nasser,W., Hawke,D.H., Abbruzzese,J.L., Baggerly,K.A., and Kobayashi,R. (2005) Diagnostic protein discovery using liquid chromatography/mass spectrometry for proteolytic peptide targeting. *Rapid Commun.Mass Spectrom.*, 19, 1624-1636.
48. Ye,B., Cramer,D.W., Skates,S.J., Gygi,S.P., Pratomo,V., Fu,L., Horick,N.K., Licklider,L.J., Schorge,J.O., Berkowitz,R.S., and Mok,S.C. (2003) Haptoglobin-alpha subunit as potential serum biomarker in ovarian cancer: identification and characterization using proteomic profiling and mass spectrometry. *Clin Cancer Res.*, 9, 2904-2911.
49. Koomen,J.M., Shih,L.N., Coombes,K.R., Li,D., Xiao,L.C., Fidler,I.J., Abbruzzese,J.L., and Kobayashi,R. (2005) Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. *Clin.Cancer Res.*, 11, 1110-1118.
50. Wang,X., Wang,E., Kavanagh,J.J., and Freedman,R.S. (2005) Ovarian cancer, the coagulation pathway, and inflammation. *J Transl.Med.*, 3, 25.
51. Li,J., Orlandi,R., White,C.N., Rosenzweig,J., Zhao,J., Seregni,E., Morelli,D., Yu,Y., Meng,X.Y., Zhang,Z., Davidson,N.E., Fung,E.T., and Chan,D.W. (2005) Independent validation of candidate breast cancer serum biomarkers identified by mass spectrometry. *Clin Chem.*, 51, 2229-2235.
52. Diamandis,E.P. and van der Merwe,D.E. (2005) Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations. *Clin Cancer Res.*, 11, 963-965.
53. Koomen,J.M., Li,D., Xiao,L.C., Liu,T.C., Coombes,K.R., Abbruzzese,J., and Kobayashi,R. (2005) Direct tandem mass spectrometry reveals limitations in protein profiling experiments for plasma biomarker discovery. *J Proteome.Res.*, 4, 972-981.
54. Villanueva,J., Shaffer,D.R., Philip,J., Chaparro,C.A., Erdjument-Bromage,H., Olshen,A.B., Fleisher,M., Lilja,H., Brogi,E., Boyd,J., Sanchez-Carbayo,M., Holland,E.C., Cordon-Cardo,C., Scher,H.I., and Tempst,P. (2006) Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J Clin Invest*, 116, 271-284.
55. Liotta,L.A. and Petricoin,E.F. (2006) Serum peptidome for cancer detection: spinning biologic trash into diagnostic gold. *J Clin Invest*, 116, 26-30.
56. Matrisian,L.M., Sledge,G.W., Jr., and Mohla,S. (2003) Extracellular proteolysis and cancer: meeting summary and future directions. *Cancer Res.*, 63, 6105-6109.

57. Bank,U., Kruger,S., Langner,J., and Roessner,A. (2000) Review: peptidases and peptidase inhibitors in the pathogenesis of diseases. Disturbances in the ubiquitin-mediated proteolytic system. Protease-antiprotease imbalance in inflammatory reactions. Role of cathepsins in tumour progression. *Adv.Exp.Med.Biol.*, 477, 349-378.
58. Song,J., Patel,M., Rosenzweig,C.N., Chan-Li,Y., Sokoll,L.J., Fung,E.T., Choi-Miura,N.H., Goggins,M., Chan,D.W., and Zhang,Z. (2006) Quantification of Fragments of Human Serum Inter-{alpha}-Trypsin Inhibitor Heavy Chain 4 by a Surface-Enhanced Laser Desorption/Ionization-Based Immunoassay. *Clin Chem.*, 52, 1045-1053.
59. Marshall,J., Kupchak,P., Zhu,W., Yantha,J., Vrees,T., Furesz,S., Jacks,K., Smith,C., Kireeva,I., Zhang,R., Takahashi,M., Stanton,E., and Jackowski,G. (2003) Processing of serum proteins underlies the mass spectral fingerprinting of myocardial infarction. *J Proteome.Res.*, 2, 361-372.
60. Steinert,R., Buschmann,T., van der,L.M., Fels,L.M., Lippert,H., and Reymond,M.A. (2002) The role of proteomics in the diagnosis and outcome prediction in colorectal cancer. *Technol. Cancer Res.Treat.*, 1, 297-304.
61. 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.
62. Engwegen,J.Y., Helgason,H.H., Cats,A., Harris,N., Bonfrer,J.M., Schellens,J.H., and Beijnen,J.H. (2006) Identification of serum proteins discriminating colorectal cancer patients and healthy controls using surface-enhanced laser desorption ionisation-time of flight mass spectrometry. *World J Gastroenterol.*, 12, 1536-1544.
63. Steinert,R., Von Hoegen,P., Fels,L.M., Gunther,K., Lippert,H., and Reymond,M.A. (2003) Proteomic prediction of disease outcome in cancer : clinical framework and current status. *Am.J.Pharmacogenomics.*, 3, 107-115.
64. Gulmann,C., Sheehan,K., Kay,E., Liotta,L., and Petricoin,E., III (2006) Array-based proteomics: mapping of protein circuitries for diagnostics, prognostics, and therapy guidance in cancer. *J Pathol.*, 208, 595-606.
65. Miller,J.C., Zhou,H., Kwekel,J., Cavallo,R., Burke,J., Butler,E.B., Teh,B.S., and Haab,B.B. (2003) Antibody microarray profiling of human prostate cancer sera: antibody screening and identification of potential biomarkers. *Proteomics.*, 3, 56-63.
66. Etzioni,R., Urban,N., Ramsey,S., McIntosh,M., Schwartz,S., Reid,B., Radich,J., Anderson,G., and Hartwell,L. (2003) The case for early detection. *Nat.Rev.Cancer*, 3, 243-252.
67. Kolch,W., Mischak,H., and Pitt,A.R. (2005) The molecular make-up of a tumour: proteomics in cancer research. *Clin Sci.(Lond)*, 108, 369-383.
68. Srinivas,P.R., Verma,M., Zhao,Y., and Srivastava,S. (2002) Proteomics for cancer biomarker discovery. *Clin.Chem.*, 48, 1160-1169.
69. de Noo,M.E., Liefers,G.J., and Tollenaar,R.A. (2005) Translational research in prognostic profiling in colorectal cancer. *Dig.Surg.*, 22, 276-281.
70. Ludwig,J.A. and Weinstein,J.N. (2005) Biomarkers in cancer staging, prognosis and treatment selection. *Nat.Rev.Cancer*, 5, 845-856.

Chapter 9

Nederlandse samenvatting

M.E. de Noo, A.M. Deelder, R.A.E.M. Tollenaar

Tijdschrift kanker. 2006; 30 (4): 20-26

Kanker van de dikke darm is een van de meest voorkomende vormen van kanker in de westerse wereld. In Nederland worden er jaarlijks circa 10.000 nieuwe patiënten gedocumenteerd, terwijl er wereldwijd bijna 900.000 nieuwe gevallen per jaar worden gemeld. Over het algemeen geldt dat hoe eerder de diagnose gesteld wordt, hoe groter de kans dat een curatieve resectie mogelijk is en dus de kans op een betere overleving. Helaas wordt kanker van de dikkedarm vaak pas in een laat stadium gediagnosticeerd en behandeld, met als gevolg een relatief slechte prognose. Daarom bestaat er een dringende behoefte aan een weinig invasief, specifiek onderzoek dat de diagnose reeds in een vroeg ziektestadium stelt, waardoor de behandeling eerder gestart en tevens beter op het individu toegespitst kan worden.

De gouden standaard voor de diagnostiek van primaire colorectale tumoren is de colonoscopie. Afgezien van het invasieve karakter en de voor patiënten onprettige darmvoorbereiding zijn ook de kosten en tijd een beperkende factor om deze methode grootschalig te gebruiken in bijvoorbeeld een bevolkingsonderzoek. Naast de colonoscopie wordt tot op heden de concentratie van het eiwit CarcinoEmbryonic Antigen (CEA) in het bloed gebruikt als diagnostische marker voor de aanwezigheid van colorectale tumoren, vooral tijdens de follow-up. Echter als diagnostische test voor het detecteren van primaire darmtumoren heeft CEA een lage sensitiviteit van 44% en specificiteit van 88%. Door deze matige betrouwbaarheid is de techniek niet geschikt voor vroegdiagnostiek toepassing op grote schaal, laat staan voor screenings doeleinden. Daarentegen is na resectie van de primaire tumor het verloop van de CEA spiegel wel een betrouwbare indicator voor de ziektestatus. Een relatief nieuwe screeningstechniek voor dikke darmkanker is de Fecale Occulte Bloed Test (FOBT), een test die spoortjes bloed aantoont in de ontlasting. Het is een non-invasieve techniek die de noodzaak voor follow-up colonoscopie limiteert. Een nadeel is echter dat tumoren die intermitterend bloeden, aan detectie door FOBT kunnen ontsnappen. Bovendien is de specificiteit van de test beperkt doordat er frequent foutpositieve uitslagen optreden als gevolg van dieet (rauw vlees), medicatie en andere oorzaken van gastro-intestinale bloeding dan darmkanker.

Recapitulerend is er momenteel geen gebruiksvriendelijke diagnostische test met voldoende hoge sensitiviteit en specificiteit om de ziekte in een vroeg stadium op te sporen. Nochtans is er grote behoefte aan gevoelige technieken om tumormarkers te identificeren, waarmee detectie van zowel primaire tumoren als recidiverende ziekte of metastasen in een vroeg stadium mogelijk wordt. Hierdoor kan er eerder tot behandeling van tumor(recidief) of metastasen worden overgegaan, wat tot een betere overleving zal leiden.

In **hoofdstuk 1** wordt een introductie gegeven over een nieuwe en gevoelige techniek, proteomics genaamd; een zich snel ontwikkelend onderzoeksveld dat ten doel heeft om zowel kwalitatief als kwantitatief alle functionele eiwitten van een

organisme in kaart te brengen. Na het afronden van het humane genoomproject is het mogelijk geworden om een groot aantal van deze eiwitten te identificeren. Proteomics is gebaseerd op de separatie en visualisatie van complexe eiwitmengsels. Zo kunnen eiwitten die een veranderd expressiepatroon vertonen, worden opgespoord en in verband worden gebracht met de aanwezigheid van een tumor. Deze discriminerende eiwitprofielen kunnen als diagnostische of prognostische test gebruikt worden, of zelfs targets vormen voor nieuwe behandelingsstrategieën. De meeste clinical proteomics studies binnen de oncologie richten zich op het detecteren van verschillen in eiwitexpressie in serummonsters tussen gezonde proefpersonen en patiënten met een maligniteit. Deze techniek werd in 2002 voor het eerst toegepast bij de detectie van ovariumcarcinoom. Deze studie was, afgezien van de veelbelovende resultaten voor de opsporing van een solide tumor, erg vernieuwend door het gebruik van patroonherkenning van meerdere eiwitten zonder dat deze geïdentificeerd waren, een zogenaamde black box methode. Deze aanpak kreeg al snel navolging en ook de resultaten van detectie van prostaat en longkanker waren zeer veelbelovend in volgende studies.

De eiwitprofielen worden gemeten met behulp van massaspectrometrie, een veelzijdige techniek die gebruikt kan worden voor identificatie, kwantificeren en profilering van isotopen, moleculen en molecuulcomplexen in kleine hoeveelheden van chemische en biologische mengsels. In een massaspectrometer worden individuele eiwitten van het serummonster in de gasfase geïoniseerd. Daarna worden de gevormde ionen versneld in een zeer precies geregeld elektrisch veld en in een vluchtbuis afgeschoten. De ionen worden vervolgens gescheiden op basis van hun massa/ ladingsverhouding (m/z) waarna de detectie volgt. Na de detectie wordt een massaspectrum gegenereerd als grafiek, waarin de intensiteit (Y-as) van iedere m/z (X-as) wordt weergegeven.

Na initiële positieve reacties op de veelbelovende resultaten van het gebruik van eiwitprofielen als patroonherkenning voor de aanwezigheid van een tumor, kwam er in de internationale literatuur kritiek op de reproduceerbaarheid en betrouwbaarheid van de massaspectrometrie. De vraag was of de classificatieresultaten, die op basis van de eiwitprofielen een individu aan de kanker of gezonde groep toewees, vertroebeld zouden worden door andere factoren. Verschil in experimentele omstandigheden en individuele variaties werden aangewezen als potentiële co-factoren op eiwitprofielen. Indien de eiwitprofielen inderdaad zo gevoelig zouden zijn voor variaties in bloedafname, opslag, voorbewerking van het serum en het massaspectrometrie experiment, dan zou de eventuele klinische toepasbaarheid voor de toekomst drastisch afnemen.

Hoofdstuk 2 is een overzicht van de verschillende vormen van translationeel onderzoek die onze onderzoeksgroep verricht naar colorectale tumoren en kan als inleiding van het proteomics onderzoek gelezen worden.

Om te verifiëren of de eiwitprofielen inderdaad onderhevig zijn aan variatie door externe factoren onderzochten wij in **hoofdstuk 3** van dit proefschrift de reproduceerbaarheid van onze methode en de invloed van verschillende logistieke preanalytische variaties de serum eiwitprofielen. Om zoveel mogelijk de reële klinische situatie na te bootsen, werden er experiment gedaan met verschil in tijd tussen afdraaien van de bloedbuis tot serum, verschil in afname tijdstip over de dag en het uitvoeren van meerdere vries -en dooicycli met een serummonster. Uit de resultaten van deze experimenten bleek dat de tijd dat het bloed staat voordat het serum wordt gecentrifugeerd de grootste variatie op serum eiwitprofielen heeft. Er was een te verwaarlozen invloed van het afnametijdstip op de dag en ook het aantal vries -en dooi cycli, mits kleiner dan 4, had weinig invloed op de variatie van de massaspectra. Een ander belangrijk aspect voor de praktische bruikbaarheid van een toekomstige serumtest voor de detectie van colorectale tumoren, is de invloed van maaltijden op de eiwitprofielen. Als klein onderdeel van de experimenten naar de invloed van preanalytische variatie vergeleken wij tevens eiwitprofielen van een individu in nuchtere toestand en na maaltijden. Hierin was echter geen verschil te vinden en daarmee werd geconcludeerd dat dit geen invloed heeft op de serum eiwitprofielen. De gehele studie was één van de eerste studies die de reproduceerbaarheid van de eigen methodiek onderzocht en op basis van deze resultaten een gestandaardiseerde aanpak voor alle verdere studies aanbeval. Het belang van het gebruik van een uniforme en goed gedocumenteerde populatie en een gestandaardiseerd afname protocol voor de serummonsters werd in alle studies onderstreept. Mits aan deze voorwaarden wordt voldaan, kan met behulp van discriminerende eiwitprofielen betrouwbare classificatie resultaten geboekt worden in het kankeronderzoek.

In **hoofdstuk 4 en 5** wordt ingegaan op de resultaten die met behulp van dezelfde aanpak verkregen zijn om op basis van discriminerende serum eiwitprofielen onderscheid te maken tussen patiënten met dikke darmkanker en gezonde controles. Hierbij voldeed elk geselecteerd serummonster aan het in hoofdstuk 3 opgestelde standaardprotocol. Naast de variatie in biologische en preanalytische factoren, is de statistische analyse een ander belangrijk onderwerp van discussie in de internationale literatuur. De eiwitspectra vormen hoogdimensionale data en brengen hierdoor direct enkele praktische bezwaren met zich mee. Gezien het hoge aantal datapunten per spectrum is voor een betrouwbare analyse een groot aantal monsters nodig. Daarnaast is het voor de opzet van de studie van belang dat de verschillende groepen binnen een experiment in dezelfde proporties aanwezig zijn. Echter de belangrijkste valkuil in de statistische analyse van de serum eiwitprofielen is de validatie. Hiervoor

is het gebruik van een aparte training en validatie set in een klinische studie van groot belang. Het classificatiemodel wordt berekend in een zogeheten trainingstest door te zoeken naar discriminerende eiwitprofielen die het beste een verschil maken tussen de gezonde en de zieke groep. Vervolgens wordt in een validatie set met onafhankelijke patiënten en gezonde controles het classificatiemodel getoetst en de sensitiviteit en specificiteit van dit model berekend. Dit classificatiemodel is een eiwitprofiel waarvan de identiteit van de discriminerende eiwitten (pieken) onbekend is. Als aanvulling op deze black box benadering kan met behulp van dezelfde massaspectrometer worden overgegaan op identificatie van de discriminerende eiwitten. Er is nog veel onderzoek noodzakelijk alvorens er consensus bereikt kan worden over de optimale statistische en bioinformatica benadering van de hoogdimensionale data binnen de clinical proteomics studies.

Tot op heden is er slechts een summier aantal klinische studies verschenen die met clinical proteomics een maligniteit betrouwbaar hebben opgespoord. Het grootste manco tot op heden is dat de resultaten niet in een ander laboratorium bevestigd kunnen worden, maar ook dat de validatie van het eigen classificatiemodel vaak te wensen over laat. Om dit en bovengenoemde valkuilen te vermijden heeft onze groep een zeer strikt opgezette classificatie studie uitgevoerd voor de detectie van colorectale tumoren. **Hoofdstuk 4** beschrijft de resultaten van het onderzoek naar de detectie van colorectale tumoren op basis van discriminerende eiwitprofielen. Hiertoe werden 66 patiënten met alle ziekte stadia, inclusief stadium IV patiënten met tot de lever beperkte metastasen, en 50 anonieme gezonde controles geïncludeerd. De monsters werden vervolgens gerandomiseerd, maar in gelijke proporties per groep over 3 platen verdeeld en op 3 achtereenvolgende dagen met de massaspectrometer gemeten. Dit leverde een zogeheten 'randomised block design' op, waarmee wij geprobeerd hebben om de classificatieresultaten zo min mogelijk te beïnvloeden met experimentele of batcheffecten. Voor de statistische analyse hebben wij lineaire discriminantie gebruikt om een optimale scheiding te krijgen tussen beide groepen. In verband met het relatieve kleine aantal monsters dat voldeed aan het standaard inclusie en afname protocol, was het gebruik van een onafhankelijke validatieset niet mogelijk in deze studie. Daarom hebben wij gebruik gemaakt van een methode die maximaal betrouwbare resultaten oplevert binnen de mogelijkheden van een interne validatie. Met deze zogeheten dubbele kruisvalidatie van de eiwitprofielen toonden de resultaten een sensitiviteit van 95% en een specificiteit van 90% voor het de detectie van colorectale tumoren. De oppervlakte onder de ROC-curve voor deze test bedroeg 97% en toont hiermee de significantie van het onderscheidend vermogen van de test aan. Het gebruik van dubbele kruisvalidatie binnen de dataset en de grote oppervlakte onder de ROC-curve tonen aan dat het classificatiemodel

op daadwerkelijke informatie in de spectra en niet op toeval berust. De statistische achtergronden hiervan worden in **hoofdstuk 5** beschreven.

Dezelfde studie opzet, massaspectrometrie en statistische analyse is gebruikt in **hoofdstuk 6**, waar eiwitprofielen van patiënten met mammacarcinomen werden vergeleken met die van gezonde controles met dezelfde gemiddelde leeftijd. Ook deze studie toonde veelbelovende resultaten voor de detectie van mammacarcimonen, met een sensitiviteit van 100% en een specificiteit van 97%. De discriminerende eiwitpieken werden andermaal in de laag moleculaire massa range gevonden, maar in andere correlaties dan bij de studie beschreven in hoofdstuk 3. Belangrijk echter was wederom de optimaal mogelijke betrouwbare classificatie en het uitsluiten van zoveel mogelijk verstorende factoren. Identificatie van deze pieken was ten tijde van het onderzoek nog niet volledig mogelijk, maar dit zal binnen afzienbare tijd wel gerealiseerd worden.

In **hoofdstuk 7** wordt het bewijs geleverd dat discriminerende eiwitprofielen ook in een onafhankelijke validatieset op zeer betrouwbare wijze gebruikt kunnen worden voor de detectie van borstkanker. Het is de eerste proteomics studie die in een gerandomiseerde en onafhankelijke patiëntengroep (in een validatieset) deze veelbelovende resultaten laat zien.

In **hoofdstuk 8** wordt een overzicht gegeven over de stand van zaken van clinical proteomics binnen de oncologie en worden de toekomstige bevindingen en verwachtingen binnen het veld besproken. Hoewel clinical proteomics pas in de kinderschoenen verkeert, lijkt deze methode te kunnen bijdragen aan de urgente zoektocht naar nieuwe biomarkers binnen de oncologie. De eerste kinderziektes, de reproduceerbaarheid en preanalytische variaties, lijken overwonnen en maken betrouwbare classificatiestudies mogelijk mits er volgens strikte standaardprotocollen met een homogene groep patiënten en sera gewerkt wordt. De samenstelling van de populatie blijft veel aandacht vragen om een maximale uniformiteit te waarborgen en zoveel mogelijk externe co-factoren te vermijden. Er is nog veel werk te verrichten aan de statistische benadering van klinische studies op basis van proteomics. Een ander punt van aandacht is de identiteit van de discriminerende eiwitten. De identiteit van de eiwitten zou kunnen leiden tot een beter begrip van pathomechanisme van het ontstaan van colorectale tumoren, maar ook targets voor nieuwe behandelingen kunnen vormen. Er zijn hypothesen dat deze discriminerende eiwitten niet specifiek voor de tumor zelf zijn, maar veelal acute fase eiwitten en afkomstig uit de complement- of stollingscascade. Momenteel wordt dit door meerdere groepen onderzocht. Dit is immers ook essentieel voor het bepalen van de specificiteit van de eiwitprofielen per verschillend type solide tumor. Hoewel er in de eerste jaren al veelbelovende resultaten zijn geboekt met vroegdetectie van kanker met behulp van clinical proteomics, is er nog een lange weg te gaan voordat introductie in de kliniek

plaats kan vinden. Dit proefschrift is echter een kleine stap in de juiste richting en onderstreept de vele hoopvolle mogelijkheden van translationeel onderzoek.

Dankwoord

DANKWOORD

Dit proefschrift is tot stand gekomen met hulp van een groot aantal mensen. Allereerst wil ik mijn coauteurs bedanken voor de uitdagende en wetenschappelijke discussies over de manuscripten. Voor de uitvoering van de experimenten ben ik veel dank verschuldigd aan Aliye Ozalp, Ronald van Vlierberghe en Hans Dalebout. Zonder hun technische ondersteuning was dit werk niet tot stand gekomen. Dit geldt ook voor de vele administratieve hulp rond de logistiek van de groot opgezette serumverzameling. Hiervoor wil ik Ada, Gerda, Annemarie van Heelkunde 2 en op de poli Riek en Tiny van harte danken. Ook het datacenter Heelkunde heeft veel bijgedragen aan de dataverzameling. Initieel was dit vooral Jan Junggeburt, later Marjolijn Duym en Linda Verhoeff. De dames van het stafsecretariaat, vooral Ingrid, Aisha en Margriet wil ik bedanken voor de hand en spandiensten en gezelligheid. De volgende schakel die onontbeerlijk was in de logistieke keten is het Centraal Klinisch Chemisch Laboratorium. Alle medewerkers wil ik danken voor de opslag van spijtserum, waarbij ik met name Marijke Frolich en Wil van der Bent wil danken voor de prettige samenwerking.

De medewerkers van de afdeling Parasitologie ben ik zeer erkentelijk voor hun acceptatie van die vreemde dokter in de biochemische bijt. Het serum protein profiling project is altijd met voorrang behandeld en dat heeft het project zeker versneld. Caroline Remmerswaal was vooral de laatste maanden een enorme hulp. Jouw optimisme en punctualiteit hebben er voor gezorgd dat de promotie daadwerkelijk plaats zou vinden. Veel dank voor je onmisbare ondersteuning. Een andere afdeling zonder wiens ondersteuning dit werk nooit gerealiseerd had kunnen worden, is de afdeling Medische Statistiek, waarvan ik met name Paul Eilers en vooral Bart Mertens wil bedanken voor hun mathematische ondersteuning en hulp.

Voor de dagelijkse dingen en de lol op het werk ben ik menigeen dank verschuldigd. De mensen van het lab SOLIT onder de menselijke begeleiding van Peter Kuppen, maar ook de verpleging en specialisten van Heelkunde 2. Bert Bonsing voor de arcade van Riolan en de motivatie, Gerrit-Jan Liefers voor het aanstekelijke enthousiasme voor de wetenschap, Henk Hartgrink voor het rolmodel van een top clinicus. Mijn kamergenoten, collega's en vrienden van het eerste uur, aan wie ik me kon optrekken: Frederieke van Duijnhoven, Joost Rothbarth, Koen Peeters en Remco Aalbers. Dank voor de scherpte en de vriendschap. Na jullie vertrek was het even wennen, maar gelukkig kwam mijn goede vriend Lee Bouwman de leegte vullen. Hoewel initieel een puur zakelijke set om een student uit Delft te halen voor de complexe statistische berekeningen, bleek Martijn van der Werff niet alleen een zeer waardevolle collega, maar ook een perfecte kamergenoot. De vele uren die wij

gedrieën in ons hok hebben doorgebracht koester ik, net als onze vriendschap en ik ben trots dat jullie mijn paranimfen zijn.

Mijn huidige collega's wil ik ook graag bedanken voor de vele dingen die ik van hen geleerd heb in de kliniek en voor het plezier waarmee ik elke dag naar het werk ga.

Buiten het LUMC heb ik veel steun gehad aan mijn vrienden en vriendinnen van de tennisbaan, die ik wil bedanken voor de perfecte uitlaatklep en de vele (nachtelijke) uren gezelligheid. Uiteraard wil ik mijn andere vrienden, schoonfamilie en mijn broertje danken voor de interesse en steun en het feit dat ze er altijd zijn als het nodig is.

Maar vooral wil ik mijn lieve ouders bedanken voor hun steun, geduld en liefde. Door jullie opvoeding en vorming is dit boekje tot stand gekomen. Naast de kalmerende reflecties heeft vooral jullie vertrouwen in mij me gemotiveerd om dit tot een goed eind te brengen - ondanks dat ik maar een meisje ben, pap…

De persoon die ik de meeste dank verschuldigd ben, is Han. Jouw humor en kalmte, maar vooral liefde maken elke dag weer de moeite waard. Zonder jou waren zowel dit proefschrift als ik nergens.

CURRICULUM VITAE

The author of this thesis was born on February 21st, 1978 in Enschede. She grew up in Enschede and graduated from Kottenpark College in 1996 (Atheneum met Latijn). In the same year she started medical school at the University of Leiden. During her graduation project in 2000 she worked in the laboratory of neuropathology at the University of Oxford, UK. After her medical rotations she performed her requisite scientific traineeship in the laboratory of Surgery of the Leiden University Medical Center on serological protein profiles.

In October 2002 she obtained her medical degree *(cum laude)* and continued working on the project 'Serum protein profiling in oncology' at the Department of Surgery in collaboration with the Department of Parasitology (Prof. dr. R.A.E.M. Tollenaar and Prof. dr. A.M. Deelder respectively). During her thesis she chaired the national Genomics Network for Young Scientists for one year and in 2006 she joined the national Proteomics taskforce from the Dutch Society of Clinical Chemists.

At the end of 2004 she was appointed AGIKO at the Department of Surgery. In July 2006 she started her surgical residency at the Leiden University Medical Center (Prof. dr. J. F. Hamming) and will continue her residency in 2008 at the Westeinde Hospital in Den Haag (Dr. J.C.A. de Mol van Otterloo).

LIST OF PUBLICATIONS

M.E. de Noo, R.A.E.M. Tollenaar, A.M. Deelder, L.H. Bouwman. Current status and prospects of clinical proteomics studies on the detection of colorectal cancer: hopes and fears, W*orld J Gastroenterol 2006*;12(41):6594-601

M.E. de Noo, A.M. Deelder , M.P.J. van der Werff, B.J.A. Mertens, A. Ozalp, R.A.E.M. Tollenaar. MALDI-TOF serum protein profiles for the detection of breast cancer, *Onkologie 2006*;29(11):501-6.

M.E. de Noo, R.A.E.M. Tollenaar, P. Eilers, A.Ozalp, M.R. Bladergroen, M.P.J. van der Werff, P.J.K. Kuppen, A.M. Deelder. The use of serological protein profiles for the detection of colorectal cancer. *Eur J Canc* 2006; 42 (8): 1068-76.

B.J.A. Mertens, M.E.de Noo, R.A.E.M. Tollenaar, A.M. Deelder. Mass spectrometry proteomics diagnosis: Enacting the validation paradigm, *J Comp Biol 2006*; 13(9):1591-605

M.E. de Noo, R.A.E.M. Tollenaar, A.Ozalp, P.J.K. Kuppen, M.R. Bladergroen, P. Eilers, A.M. Deelder. Reliability of human serum protein profiles generated with C8 magnetic beads assisted MALDI-TOF mass spectrometry, *Anal Chem* 2005, 77 (22):7232-41.

M.E.de Noo, G.J. Liefers, R.A.E.M. Tollenaar. Translational Research in Colorectal Cancer, *Dig Surg* 2005**,** *22*, 276-81.

M.E. de Noo, A.M. Deelder, R.A.E.M. Tollenaar. *Tijdschrijft Kanker* 2006, 30 (4): 20-26.

M.E. de Noo, M.P.J. van der Werff, B.J. Mertens, H. Dalebout, M. Bladergroen, R.A.E.M. Tollenaar, A.M. Deelder. Validation of serum protein profiling for the detection of breast cancer. *Submitted.*

M.E. de Noo, R.A.E.M. Tollenaar. Het gebruik van genomics en proteomics voor de diagnostiek en therapie van kanker. Cursusboek "Vorderingen in de Geneeskunde", Boerhaavecursus voor huisartsen, ISBN : 90-6767-536-9

M.E. de Noo, M.P.J. van der Werff, C.H.J. van de Velde, A.M. Deelder, R.A.E.M. Tollenaar. Detection of colorectal cancer using MALDI-TOF. *Annals of Oncology*, 2006, Vol 17 Suppl 1. *Abstract*

M.E. de Noo, A.M. Deelder, A.Ozalp, B.J. Mertens, C.H.J. van de Velde, P.J.K. Kuppen, R.A.E.M. Tollenaar. The use of serological proteomics for early detection of colorectal cancer. *Journal of biological markers*, 2004, 19-S3, S18. *Abstract*

M.E. de Noo, A.M. Deelder, A.Ozalp, B.J. Mertens, C.H.J. van de Velde, P.J.K. Kuppen, R.A.E.M. Tollenaar. Het gebruik van serumproteomics voor vroegdiagnostiek van colorectale tumoren, *NTVG* 2004. *Abstract*