Metabolomic NMR fingerprinting to identify and predict survival of patients with metastatic colorectal cancer

Ivano Bertini, Stefano Cacciatore, Benny V. Jensen, Jakob V. Schou, Julia S. Johansen, Mogens Kruhøffer, Claudio Luchinat, Dorte L. Nielsen, Paola Turano

CERM and Department of Chemistry, University of Florence, Italy
†FiorGen Foundation, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy
†Department of Oncology, Herlev Hospital, Copenhagen University Hospital, Denmark
‡AROS Applied Biotechnology A/S, Aarhus N, Denmark

*To whom correspondence should be addressed:
Prof. Ivano Bertini
CERM, University of Florence
Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy
Phone: +39 055 4574270
Fax: +39 055 4574271
e-mail: ivanobertini@cerm.unifi.it

Precis: The metabolomic signature derived from patients with metastatic colorectal cancer predicts overall survival and provides insight into potential new biomarkers that can be used to predict disease progression and personalize treatment.

Running title: Metabolomic NMR fingerprinting of metastatic colorectal cancer

Keywords: metastatic colorectal cancer, overall survival, metabolomics, NMR fingerprinting.

Grant support: Research funded by the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement n° 222916.
Abstract

Earlier detection of patients with metastatic colorectal cancer (mCRC) might improve their treatment and survival outcomes. In this study, we used proton nuclear magnetic resonance (\(^1\)H-NMR) to profile the serum metabolome in mCRC patients and determine whether a disease signature may exist that is strong enough to predict overall survival (OS). In 153 mCRC patients and 139 healthy subjects (HS) from three Danish hospitals, we profiled two independent sets of serum samples in a prospective Phase II study. In the training set, \(^1\)H-NMR metabolomic profiling could discriminate mCRC patients from HS with a cross-validated accuracy of 100%. In the validation set, 96.7% of subjects were correctly classified. Patients from the training set with maximally divergent OS were chosen to construct an OS predictor. After validation, patients predicted to have short OS had significantly reduced survival (HR = 3.4; 95% CI, 2.06 to 5.50; \(P = 1.33 \times 10^{-6}\)). A number of metabolites concurred with the \(^1\)H-NMR fingerprint of mCRC, offering insights into mCRC metabolic pathways. Our findings establish that \(^1\)H-NMR profiling of patient serum can provide a strong metabolomic signature of mCRC and that analysis of this signature may offer an independent tool to predict overall survival.
Introduction

Pattern recognition technologies in the omics world have been used for the diagnosis and prognosis of several tumor types using a variety of experimental platforms (1-3). Metabolomics is a post-genomic research field concerned with the study of metabolic profiles in biological samples such as isolated cells (4), tissues (5) or body fluids (6-10). In recent years, metabolomic studies have provided a significant contribution to the detailed understanding of the biochemical network and pathways in oncology (11), highlighting the potential of metabolomic analyses of blood samples for the detection of cancers, including epithelial ovarian (12), leukemia (13), oral (14), breast (15, 16), prostate (11), liver (17) as well as colorectal cancer (CRC) (18-25). CRC is the second most common cause of cancer related deaths in Europe (26) and the third most common cause of cancer related deaths in the United States (27). The impact of colorectal carcinogenesis on the metabolic profile of tissues and serum has been tested by mass spectrometry (MS) (18-21), nuclear magnetic resonance (NMR) (22, 23) or a combination of the two analytical tools (24). Most of the symptoms associated with CRC do not manifest themselves until late in the progression of the disease and many patients have metastases at the time of initial diagnosis (27). The overall survival (OS) of a patient is reduced considerably when the disease is diagnosed at a stage where it has spread to distant sites. In the past decade the OS of patients with metastatic CRC (mCRC) has improved due to new combinations of chemotherapy, including 5-Fluorouracil (FU), irinotecan and oxaliplatin (28-31). The introduction of the new targeted therapies directed against the Epidermal Growth Factor Receptor (EGFR) has further increased response rates (32) in some patients. Nevertheless, early selection of patients with mCRC for optimal treatment with these biologics is very difficult. For example, KRAS mutations occur in 35-45% of patients with mCRC and predict poor response to EGFR targeted therapy with cetuximab and
panitumumab (33, 34). The remaining mCRC patients are KRAS wild type, and approximately 25-40% of them are non-responders and do not benefit from this treatment (33, 34). OS hazard ratios (HR) for mCRC have been evaluated from KRAS mutations (HR 2.4) (35), Eastern Cooperative Oncology Group (ECOG) performance status (PS) (HR 1.78-1.88) (29, 36) and serum levels of alkaline phosphatase (HR 1.71) (36), bilirubin (HR 1.89) (29), and lactate dehydrogenase (HR 2.12) (29). Somewhat higher HR values for primary CRC are obtained from monocyte (HR 2.86) and neutrophil (HR 2.90) count (37). Novel non-invasive, sensitive and inexpensive indicators for mCRC OS are therefore highly desirable.

In this study, conducted within the frame of the biomarker discovery activities of the FP7 project SPIDIA (www.spidia.eu), we performed 1H-NMR profiling of serum samples collected from 153 Danish patients with mCRC before 3rd line treatment with cetuximab and irinotecan (Table 1 and Table S1), and of serum samples from 139 healthy subjects (HS) (Table 1). The statistical model generated from 1H-NMR profiles of serum samples of subjects from two of the three hospitals can robustly discriminate HS (n=96) from patients with mCRC (n=45) with 100% cross-validated accuracy. General applicability of the resultant classifier was successfully validated in an independent set of 43 HS and 108 patients with mCRC from the third hospital. The capability of the 1H-NMR profiles to predict OS after start of treatment with cetuximab and irinotecan using pre-treatment serum samples was tested on a set of 20 subjects of the training set with maximally divergent OS. The classifier was validated on the independent set of 108 mCRC patients. Furthermore, through multivariate analysis, a number of serum metabolites were identified whose levels were significantly different in patients with mCRC as compared to HS, as well as between patients with short and long OS. These metabolites can provide hints both to define new biomarkers and to better understand the biochemistry involved in mCRC.
Methods

Patients with metastatic colorectal cancer

Serum samples were collected from a cohort of 181 Danish patients with mCRC resistant to 5-FU, oxaliplatin and irinotecan. Details on patient selection are provided in Supplementary Methods. The patients were included in a prospective Phase II study evaluating cetuximab and irinotecan every second week. The patients were included at three Hospitals in Denmark during the period October 17th, 2006 to October 3rd, 2008. The patients were treated until disease progression with 3rd line irinotecan (180 mg/m² of body-surface area on day 1 of each 14-day period during the study) and cetuximab (500 mg/m² of body-surface area every second week) independent of their KRAS status. Serum samples used for metabolomics were collected before initiating the treatment. In the present study, we did not include any patient with hereditary CRC.

Patients were included in the “Phase II study of treatment with cetuximab and irinotecan administered bi-weekly to irinotecan-resistant patients with metastatic colorectal cancer – efficacy and biological markers”. Inclusion criteria were: 1) Patients with histologically verified adenocarcinoma of the colon or rectum with non-resectable or metastatic disease; 2) Patients with measurable disease as per Response Evaluation Criteria in Solid Tumors (RECIST) (38) criteria; 3) Patients with progressive disease following either oxaliplatin-based or irinotecan-based treatment; 4) Patients with irinotecan-resistant disease – defined as progressive disease after at least 6-weeks treatment with irinotecan-based treatment or within 6 months following discontinuation of such treatment. Irinotecan-based treatment should not necessarily be given immediately before inclusion, but oxaliplatin-based treatment may be; 5) ECOG-PS less than 3; 6) An expected survival time of at least 3 months; 7) Neutrophil count ≥1.5 x 10⁹/l, and thrombocytes ≥ 100 x 10⁹/l; 8) Normal liver function with bilirubin < 1.5 x upper limit of normal and ASAT/ALAT < 5 x upper limit of
normal; 9) All should be exposed to oxaliplatin-based treatment; and 10) The signed informed consent as per requirements of the Scientific Ethical Commitees.

The patients were followed until death or July 7th, 2010 (39). The study (including biomarker analysis) was approved by the Regional Ethics Committee (VEK ref. KA-20060094). For 95% of the patients mutations on the KRAS gene were determined, as reported in Supplementary Methods. Further clinical data about the study will be published elsewhere by Dr. Benny V. Jensen. One hundred sixty-eight pretreatment serum samples were available for $^1$H-NMR profiling.

Controls

Serum samples were collected from 96 Danish healthy blood donors and 43 healthy subjects (hospital staff) (Table 1).

Serum sample preparation

Blood was collected into standard blood collection tubes and allowed to clot at room temperature for 30-120 minutes before centrifugation (1500 g for 10 minutes at 4 °C). Serum was aliquoted and stored at -80 °C in Greiner® cryogenic vials. This procedure is compatible with the Standard Operating Procedures (SOPs) for metabolomic-grade serum samples recently defined (40).

NMR sample preparation

Frozen serum samples were thawed at room temperature and shaken before use. A total of 300 µl of buffer (70 mM Na$_2$HPO$_4$; 20% (v/v) $^2$H$_2$O; 6.15 mM NaN$_3$; 6.64 mM sodium trimethylsilyl [2,2,3,3-$^2$H$_4$] propionate; pH 7.4) was added to 300 µl of each serum sample, and the mixture was homogenized by vortexing for 30 seconds. A total of 450 µl of this mixture was transferred into a 4.25 mm NMR tube for analysis.
NMR spectra

Carr-Purcell-Meiboom-Gill (CPMG) $^1$H-NMR spectra were acquired to observe metabolites signals and suppress resonances arising from high molecular weight molecules, and J-resolved spectra projections (pJRES) to help identify biomarkers. Further details are provided in Supplementary Methods. The quality of the serum samples was assessed by visual inspection of the resulting NMR spectra according to the criteria defined in Bernini et al. (40). Thirteen (8%) serum samples were not included in the statistical analysis due to the insufficient quality of the resulting NMR spectra that contained very broad features attributable to high molecular weight impurities. Two patients with mCRC younger than 45 years were excluded since Hereditary Non Polyposis Colorectal Cancer is typically associated with increased inflammatory response and different prognosis than sporadic CRC. Demographic and clinical characteristics of the patient cohort studied by NMR are provided in Table 1 and Table S1.

Statistical Analysis

Multivariate data analyses were performed on processed data by combining established methods (6, 41). Projection to Latent Structure (PLS) was applied on processed data for dimension reduction using the SIMPLS algorithm as implemented in the R library “plsgenomics”; the Support Vector Machines (SVM) method was used for data classification using the “libsvm” module of the R library “e1071”. Canonical Correlation Analysis (CA) was performed using the standard R function “cancor”.

To assess the prediction ability of the model, a 10-fold Cross-Validation (CV) was performed by generating splits with a ratio 1:9 for the data set, that is by removing 10% of samples prior to any step of the statistical analysis, including PLS component selection.

Parameter selection (best number of components for PLS, kernel type, cost of constraints
violation for SVM) was carried out by means of 5-fold CV on the remaining 90%. The whole procedure was repeated 100 times. Data classification using SVM was performed by applying the classifier on PLS scores. Accuracy, sensitivity and specificity were calculated using standard definitions. To estimate the classification/feature selection performance of our predictors, the statistical significance of all observations was further assessed through permutation tests.

Score contribution weights were examined to assess which metabolites were most responsible for the observed discrimination between groups (42). The relative concentrations of each metabolite were calculated by integrating the signals in the spectra. Statistical significance was assessed using the nonparametric Kruskal-Wallis test. A $P < 0.05$ was considered statistically significant. Group-wise comparisons of distributions of clinical and demographic data were performed with the Fisher’s exact test for the categorical variables and Kruskal-Wallis test for the continuous variables.

Kaplan-Meier survival curves were created using the R library “survival”. The Wald test was used to calculate the statistical significance ($P$) of the differences between survival curves. Prognostic factors for OS were analyzed using the Cox proportional hazard regression.

All calculations were made using R (41) scripts developed in-house.

Results

Fingerprinting of mCRC

A schematic diagram of the present strategy for constructing and validating a metabolic signature in patients with mCRC before 3rd line treatment with cetuximab and irinotecan is shown in Figure 1A. Patients with mCRC and HS (Table 1) were discriminated by using SVM on PLS scores (PLS-SVM) calculated on $^1$H-NMR serum spectra of samples from Aalborg and Odense hospitals (n=141). Accuracy, sensitivity and specificity results for the
training set were averaged on a total of 1000 random subsamplings (100-times-repeated 10-fold CV). This procedure consists of two nested loops of CV (Fig. 1A): an outer loop (10-fold CV) for the estimation of the prediction accuracy and an inner loop (5-fold CV) for the optimization of the parameters of the classifier training procedure. Classification performances on the validation set were calculated from the model built on the training set with a 10-fold parameter optimization procedure. Accuracy, sensitivity and specificity of the PLS-SVM classifier are given in Table 2; excellent classification accuracies were found both for the training (100.0%) and the validation sets (96.7%). On a descriptive level, CA on PLS score (PLS-CA) was used to ascertain whether the spectral patterns of serum of patients with mCRC carry unique features with respect to those of HS. The resulting plot is shown in Figure 1B: a very good discrimination is clearly apparent.

The different age distribution between the patients with mCRC and HS could be a possible source of inhomogeneity in the training set ($P < 0.01$). Therefore, the same statistical analysis described above was repeated on a subgroup aged between 45 and 65 years ($P = 0.23$), providing a classification accuracy of 100.0% in the training set and 92.1% in the validation set (Table 2). These results demonstrate that the difference in age distribution in the two groups does not affect significantly the classification accuracy values.

PLS-SVM CV analysis was also repeated on a dataset built with 96 HS and the 17 patients with mCRC and ECOG-PS 0, obtaining a classification accuracy of 100.0% on the training set and 86.4% on the validation set (Table 2), which indicates that a clear metabolic signature of the disease exists even in the serum of patients with mCRC and ECOG-PS 0.

**Prediction of survival**

The strong metabolic signature of mCRC observed in the present study suggests that even within patients with mCRC there may be statistically significant differences in the metabolic profile. This could be related to OS. The prognostic classification capabilities of
\(^1\)H-NMR profiling were therefore tested. A model was built by selecting from the training set a sub-set of 20 patients with mCRC with maximally divergent OS: i.e. the 10 patients with the shortest OS and the 10 patients with longest OS were studied. The PLS-SVM classifier, similar to that employed in the previous analysis, was used to assess whether patients with short or long OS could be identified by \(^1\)H-NMR profiling. The resulting accuracy was 78.5\% (Table 2), suggesting that a signature for the progression state of the tumor is present in the \(^1\)H-NMR profiles. The statistical confidence of the prediction performance of the PLS-SVM classifier on the multi-mode dataset with 10-fold CV evaluation was investigated using a permutation test. Permutation test results (number of permutations = 1000) showed that the accuracy difference between PLS-SVM and a random classifier was statistically significant (\(P = 0.026\)). PLS-CA scores are shown in Figure 1C.

The patients with mCRC from the validation set were examined by using the model built on the basis of the \(^1\)H-NMR profiles of the serum samples of the training set of patients with extreme values of OS, as described before. The model identified 85 (78.7\%) patients as belonging to the long OS group and 23 (21.3\%) patients to the short OS group, with features as in Table 3.

Figure 1D shows the Kaplan-Meier plots of the survival data for the 108 patients with mCRC of the independent validation set, segregated according to the \(^1\)H-NMR-based clustering described above. Statistically significant differences (\(P = 1.33 \cdot 10^{-6}\)) in the curve of OS were calculated based on the Wald test. Cox hazard analysis shows that a significantly high hazard ratio exists between short and long OS groups (HR = 3.37; 95\% CI, 2.06 to 5.50, \(P = 1.33 \cdot 10^{-6}\); Table 4).

ECOG-PS also provides a meaningful prediction (HR = 1.57; 95\% CI, 1.20 to 2.06, \(P = 1.13 \cdot 10^{-5}\); Table 4), whereas the presence of \(KRAS\) mutations does not result a meaningful predictor of OS (\(P = 2.11 \cdot 10^{-1}\); Table 4). Consistently, mutated and wild-type
KRAS patients were found to be almost equally distributed between the long OS and short OS groups (Table 3). When a multivariate analysis on NMR and ECOG-PS is performed, only NMR profiling remains meaningful (Table 4). The predictive power of NMR profiling with respect to that of classical serological markers (YKL-40, CRP, CEA) status was also tested (full analysis to be published elsewhere by Dr. Benny V. Jensen); NMR profiling remains an independent and meaningful predictor, together with CRP levels.

**Metabolites contributing to the mCRC fingerprint**

An analysis of the PLS loadings was performed to identify the metabolites contributing to the mCRC fingerprint using the entire sets of patients with mCRC and HS. The values of the relative serum concentrations of metabolites were estimated through the integration of the signals in the NMR spectra. By comparing the spectra of the serum samples of patients with mCRC with those of HS, it appears that the patients with mCRC are characterized by lower serum levels \((P < 0.05)\) of alanine, citrate, creatine, glutamine, peptide NHs, lactate, leucine, pyruvate, tyrosine and valine, and higher serum levels \((P < 0.05)\) of 3-hydroxybutyrate, acetate, formate, glycerol, lipid \((-\text{CH}_2-\text{OCOR})\), N-acetyl signal of glycoproteins, phenylalanine and proline (Table S2 and Figure 2A).

To identify the metabolites that contribute to discriminate patients with short and long OS, we selected from the entire group of patients with mCRC those with extreme OS values. By setting OS thresholds of > 24 months and < 3 months, 18 and 17 patients were selected, respectively. Analysis of the PLS loadings shows that patients with short OS are characterized by lower serum levels \((P < 0.05)\) of creatine, lipid \((-\text{C} = \text{C}-\text{CH}_2-\text{C} = \text{C}-)\), lipid \((-\text{CH}=\text{CH}-)\) and valine, and higher serum levels \((P < 0.05)\) of lipid \((-\text{CH}_2-\text{OCOR})\) and N-acetyl signal of glycoproteins (Table S3 and Figure 2B). The worsening of the clinical conditions thus corresponds to a more marked difference in the concentration of several of
the metabolites that are responsible of the disease fingerprint.

By comparing the spectra of the serum samples of the underweight (N=7) and normal weight (N=71) versus overweight (N=52) and obese patients with mCRC (N=21), it appears that the overweight and obese patients are characterized by lower levels of formate ($P=7.78 \cdot 10^{-3}$) and LDL/HDL ($P=1.42 \cdot 10^{-4}$), and higher levels of valine ($P=1.98 \cdot 10^{-3}$), N-acetyl signal of glycoproteins ($P=1.28 \cdot 10^{-2}$), CH$_2$CH$_2$CO and CH$_3$CO signals due to lipids ($P=1.24 \cdot 10^{-2}; P=7.37 \cdot 10^{-3}$) and VLDL ($P=3.07 \cdot 10^{-2}$).

**Discussion**

$^1$H-NMR profiling was utilized to characterize the metabolomic signature of patients with mCRC before 3rd line treatment with cetuximab and irinotecan. The classification based on $^1$H-NMR profiling was highly accurate (100.0%, as estimated by cross-validation) and confirmed on the validation set (96.7%) showing a clear separation between the patients with mCRC and HS. The validity of the approach was strengthened by the demonstrated ability to find a clear signature in serum also of patients with mCRC in good performance (i.e. ECOG-PS 0).

The capability of $^1$H-NMR profiling to predict OS is interesting. It probably reflects the fact that metabolites are the end products of the ensemble of processes occurring in living organisms, and can be regarded as the ultimate response of the organisms to disease-induced metabolic alterations, inflammatory processes and changes in life-style (diet, sedentary vs. active life, etc.) as a consequence of the pathological status. The information content derived from classical approaches based on single-molecule markers is somehow embedded in the $^1$H-NMR patterns. On the other hand, both NMR profiling and quantification of patients' general well-being like the definition of ECOG-PS are both complex multi-parametric indicators. However, the former is based on an objective, albeit complex, evaluation of spectral data that contain molecular level information. On the
contrary, the ECOG-PS is based on criteria that reflect how a patient's disease is progressing and how the disease affects the daily living abilities of the patient; it is therefore intrinsically associated with a certain degree of subjectivity (43).

It has been reported that indicators of an inflammatory response in serum NMR are represented by increased intensity for the signals of CH$_2$-COOR of lipids and for the N-acetyl resonance of glycoproteins (44). We show here that the metabolomic fingerprint in serum of patients with mCRC is characterized by a higher intensity of these resonances; the effect is larger for patients with short OS after starting 3rd line treatment with cetuximab and irinotecan. This finding may reflect a nonspecific inflammatory response and is in agreement with the proposal by Galon et al. (45, 46) that the OS time is “governed in large part by the state of the local adaptive immune response”.

Patients with short survival showed decreased serum concentrations of polyunsaturated lipids. Reduced levels of hydroxylated, polyunsaturated ultra long-chain fatty acids have been reported for patients with CRC on the basis of NMR/MS studies and proposed as a signature of this disease (21).

Moreover, in the serum of the patients with mCRC we found a depletion of several potential precursors of glucose in gluconeogenesis (47, 48), such as lactate, pyruvate, alanine, glutamine and other gluconeogenic amino acids, with no significant changes in glucose concentration. The observed trend may suggest an increased uptake of glucose precursors by the liver (49, 50) that could be consistent with an increased hepatic gluconeogenesis (51). Such an effect is found in patients with cachexia, that occurs frequently in patients with metastatic cancer and is associated with more than 20% of cancer deaths (52). By comparing the spectra of the serum samples of the underweight and normal weight versus overweight and obese patients with mCRC, it appears that the overweight and obese patients are characterized by lower levels of formate and LDL/HDL, and higher levels of valine, N-acetyl signal of glycoproteins, CH-CH$_2$-CO and CH$_2$-CO
signals due to lipids and VLDL. Although information on nutritional status is not available, the N-acetyl signal of glycoproteins suggests that underweight and normal weight patients with mCRC may be characterized by a significantly higher immune response, highlighting a link between cachexia and inflammatory response.

The relative low serum levels of lactate and the non-statistically significant changes in glucose levels are different from the observations in previous studies of patients with CRC (20, 22) and in other types of cancer where, commonly, serum lactate levels are high (12, 53). However, relatively high serum levels of glucose and low levels of lactate, alanine and gluconeogenic amino acids were observed in patients with oral cancer (14). This pattern was attributed to an altered energy metabolism (14). Gluconegogenesis in the present study of patients with mCRC may mask anaerobic dissimilation of glucose (48). From the comparison between patients with mCRC and HS it also emerges that high serum levels of 3-hydroxybutyrate are present in the patients with mCRC. This molecule is formed during fatty acid oxidation and its high levels may suggest a liver response to an increased energy demand needed for an increased hepatic gluconeogenesis (14). Depletion of polyunsaturated lipids observed in patients with mCRC also fits in this scheme.

In summary, we found that the serum of patients with mCRC before 3rd line therapy with cetuximab and irinotecan shows a signature of altered energy metabolism, which may reflect an increased gluconeogenesis and an accumulation of 3-hydroxybutyrate. Our findings also suggest that the overall 1H-NMR profile of serum from patients with mCRC may reflect an inflammatory status, more serious in patients with short OS.
References


17. Yang YX, Li CL, Nie X, Feng XS, Chen WX, Yue Y et al. Metabonomic studies of...


38. Therasse P, Arbuck SQ, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L et


Captions to the figures

**Fig 1.** A, schematic overview of the procedure for the generation and validation of the $^1$H-NMR metabolomic profile classifier. Different approaches were used for the validation of the signature of mCRC (1) and for the prediction of OS (2). B, PLS-CA clustering for the HS (triangles) and mCRC patients (circles) of the validation set based on the model built on the NMR spectra of the training set. Triangles are predicted as HS and circles as mCRC. C, PLS-CA clustering for the long OS group (triangles) and short OS group (circles) of the validation set based on the model built on the NMR spectra of the training set. D, Kaplan-Meier curves showing survival probability based on $^1$H-NMR profiling predictive model ($P = 1.33 \times 10^{-6}$).

**Fig 2.** Values of log$_2$(FC) resulting from the comparison between (A) HS and patients with mCRC; and (B) patients with mCRC and long OS (> 24 months) and short OS (< 3 months). A, log$_2$(FC) with a positive value indicates a relatively higher serum concentration present in the patients with mCRC, while a negative value means a relatively lower serum concentration in the patients with mCRC. B, log$_2$(FC) with a positive value indicates a relatively higher serum concentration present in the patients with mCRC and short survival (< 3 months), while a negative value means a relatively lower serum concentration in the patients with mCRC and short survival (< 3 months).
**Table 1.** Patient demographic and clinical characteristics in the training and the validation sets

<table>
<thead>
<tr>
<th>Demographic or Clinical Characteristics</th>
<th>Training set (n=141)</th>
<th>Validation set (n=151)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HS</td>
<td>mCRC patients</td>
</tr>
<tr>
<td>Hospital</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herlev</td>
<td>43 100.0</td>
<td>108 100.0</td>
</tr>
<tr>
<td>Aalborg</td>
<td>96 100.0</td>
<td>25 55.6</td>
</tr>
<tr>
<td>Odense</td>
<td>20 44.4</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>50 52.1</td>
<td>31 68.9</td>
</tr>
<tr>
<td>Female</td>
<td>46 47.9</td>
<td>22 51.2</td>
</tr>
<tr>
<td>Age, years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>42 64</td>
<td>40 63</td>
</tr>
<tr>
<td>Range</td>
<td>30-66</td>
<td>26-69</td>
</tr>
<tr>
<td>BMI&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Underweight</td>
<td>0 0.0</td>
<td>7 6.6</td>
</tr>
<tr>
<td>Normal weight</td>
<td>23 51.1</td>
<td>48 45.3</td>
</tr>
<tr>
<td>Overweight</td>
<td>17 37.8</td>
<td>35 33.0</td>
</tr>
<tr>
<td>Obese</td>
<td>5 11.1</td>
<td>16 15.1</td>
</tr>
<tr>
<td>PS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECOG 0</td>
<td>17 37.8</td>
<td>60 55.6</td>
</tr>
<tr>
<td>ECOG 1</td>
<td>18 40.0</td>
<td>34 31.5</td>
</tr>
<tr>
<td>ECOG 2</td>
<td>10 22.2</td>
<td>14 13.0</td>
</tr>
<tr>
<td>OS, months</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>12.1</td>
<td>11.2</td>
</tr>
<tr>
<td>Range</td>
<td>1.5-31.9</td>
<td>0.5-45.4</td>
</tr>
<tr>
<td>KRAS&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wild-type</td>
<td>29 67.4</td>
<td>65 63.7</td>
</tr>
<tr>
<td>Mutated</td>
<td>14 32.6</td>
<td>37 36.3</td>
</tr>
</tbody>
</table>

Abbreviations: mCRC, metastatic colorectal cancer; HS, healthy subjects; OS, overall survival; ECOG, Eastern Cooperative Oncology Group; PS, performance status; BMI, body mass index.

<sup>a</sup>Data not available for 2 patients of the validation set.

<sup>b</sup>Data not available for 2 patients of the training set and 6 patients of the validation set.
Table 2. Classification parameters for the training set and validation set

<table>
<thead>
<tr>
<th>Study and Model</th>
<th>No. of Subjects</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS vs. CRC</td>
<td>141 (96+45)</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>HS vs. CRC (45 – 65 years)</td>
<td>61 (37+24)</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>HS vs. CRC (ECOG-PS 0)</td>
<td>113 (96+17)</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Long OS vs. Short OS</td>
<td>20 (10+10)</td>
<td>79.9</td>
<td>76.4</td>
<td>78.5</td>
</tr>
<tr>
<td>Validation set</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS vs. CRC</td>
<td>151 (43+108)</td>
<td>100.0</td>
<td>95.5</td>
<td>96.7</td>
</tr>
<tr>
<td>HS vs. CRC (45 – 65 years)</td>
<td>76 (15+61)</td>
<td>80.0</td>
<td>95.1</td>
<td>92.1</td>
</tr>
<tr>
<td>HS vs. CRC (ECOG-PS 0)</td>
<td>103 (43+60)</td>
<td>95.3</td>
<td>80.0</td>
<td>86.4</td>
</tr>
</tbody>
</table>

Abbreviations: mCRC, metastatic colorectal cancer; HS, healthy subjects; OS, overall survival.
Table 3. Patient demographic and clinical characteristics in the long and short OS classes predicted by the model built on the validation set

<table>
<thead>
<tr>
<th></th>
<th>Long OS(a)</th>
<th>Short OS(a)</th>
<th>(P^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Patients</td>
<td>85</td>
<td>23</td>
<td>1.36 (\times) 10^{-5}</td>
</tr>
<tr>
<td>OS, months</td>
<td>1.36 (\times) 10^{-5}</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>12.9</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>1.1-45.4</td>
<td>0.5-16.9</td>
<td></td>
</tr>
<tr>
<td>Age, years</td>
<td>5.14 (\times) 10^{-1}</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>64</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>46-87</td>
<td>55-76</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>54</td>
<td>11</td>
<td>2.40 (\times) 10^{-1}</td>
</tr>
<tr>
<td>Female</td>
<td>31</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>BMI(c)</td>
<td></td>
<td></td>
<td>1.39 (\times) 10^{-1}</td>
</tr>
<tr>
<td>Underweight</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Normal weight</td>
<td>35</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Overweight</td>
<td>29</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Obese</td>
<td>15</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PS</td>
<td></td>
<td></td>
<td>3.36 (\times) 10^{-5}</td>
</tr>
<tr>
<td>ECOG 0</td>
<td>55</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>ECOG 1</td>
<td>25</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>ECOG 2</td>
<td>5</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>KRAS(d)</td>
<td></td>
<td></td>
<td>7.97 (\times) 10^{-1}</td>
</tr>
<tr>
<td>Wild-type</td>
<td>53</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Mutated</td>
<td>29</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Bold indicates significant \(P\) values.

Abbreviations: ECOG, Eastern Cooperative Oncology Group; PS, performance status; BMI, body mass index; OS, overall survival.

\(a\)Predicted classes by the model built on the training set.

\(b\)Group-wise comparisons of distributions of clinical and demographic data were performed with the nonparametric Kruskal-Wallis rank sum test for the continuous variables and the Fisher’s exact test for the categorical variables. Significant at \(P < 0.05\).

\(c\)Data not available for 2 patients of the validation set.

\(d\)Data not available for 6 patients of the validation set.
Table 4. Univariate and multivariate Cox regression analysis of clinical factors associated with OS of the validation set (n=108)

<table>
<thead>
<tr>
<th>Variables</th>
<th>HR</th>
<th>95% CI</th>
<th>(P^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Univariate analysis</strong>&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMR profiling (long OS vs. short OS)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3.37</td>
<td>2.06 to 5.50</td>
<td>(1.33 \cdot 10^-6)</td>
</tr>
<tr>
<td>ECOG-PS (grade 0 vs. grade 1 vs. grade 2)</td>
<td>1.57</td>
<td>1.20 to 2.06</td>
<td>(1.13 \cdot 10^-3)</td>
</tr>
<tr>
<td>BMI (underweight vs. normal weight vs. overweight vs. obese)</td>
<td>0.83</td>
<td>0.65 to 1.06</td>
<td>(1.32 \cdot 10^-1)</td>
</tr>
<tr>
<td>Sex (female vs. male)</td>
<td>0.92</td>
<td>0.61 to 1.38</td>
<td>(6.80 \cdot 10^-1)</td>
</tr>
<tr>
<td>Age (&lt;60 years vs. ≥60 years)</td>
<td>0.93</td>
<td>0.62 to 1.40</td>
<td>(7.26 \cdot 10^-1)</td>
</tr>
<tr>
<td>KRAS (wild type vs. mutant)</td>
<td>1.31</td>
<td>0.86 to 2.02</td>
<td>(2.11 \cdot 10^-1)</td>
</tr>
<tr>
<td><strong>Multivariate analysis</strong>&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMR profiling (long OS vs. short OS)</td>
<td>2.79</td>
<td>1.61 to 4.84</td>
<td>(2.75 \cdot 10^-4)</td>
</tr>
<tr>
<td>ECOG-PS (grade 0 vs. grade 1 vs. grade 2)</td>
<td>1.27</td>
<td>0.94 to 1.71</td>
<td>(1.23 \cdot 10^-1)</td>
</tr>
</tbody>
</table>

**NOTE:** Bold indicates significant \(P\) values.

Abbreviations: NMR, nuclear magnetic resonance; ECOG, Eastern Cooperative Oncology Group; PS, performance status; BMI, body mass index; HR, hazard ratio; OS, overall survival.

<sup>a</sup>Univariate analysis, Cox proportional hazards regression.

<sup>b</sup>Predicted classes by the model built on the training set.

<sup>c</sup>Multivariate analysis, Cox proportional hazards regression adjusting for ECOG-PS.

<sup>d</sup>\(P\) values were calculated according to the Wald method. Significant at \(P < 0.05\).
Aalborg cohort and Odense cohort

Random subsampling

Outer loop (10-fold CV)

9/10 of samples
1/10 of samples

Inner loop (5-fold CV)

Random subsampling

8/10 of samples
2/10 of samples

Classifier training
Classifier validation

First component
Second component

Parameter of classification algorithm are optimized
e averaged over 100 times and its data are averaged

Internal Test set is predicted by classifier

Final classifier construction

Validation Set Herlev cohort

Evaluation of predictive value by comparison with HS/mCRC groups
Evaluation of predictive value of short/long OS by Kaplan-Meier plot

B

C

D

Evaluation of predictive value by comparison with HS/mCRC groups
Evaluation of predictive value of short/long OS by Kaplan-Meier plot

long OS
short OS

Survival probability

Time (months)
Metabolomic NMR fingerprinting to identify and predict survival of patients with metastatic colorectal cancer

Ivano Bertini, Stefano Cacciatore, Benny V. Jensen, et al.

_Cancer Res_ Published OnlineFirst November 11, 2011.

Updated version  Access the most recent version of this article at: doi:10.1158/0008-5472.CAN-11-1543

Supplementary Material  Access the most recent supplemental material at: http://cancerres.aacrjournals.org/content/suppl/2011/11/11/0008-5472.CAN-11-1543.DC1

Author Manuscript  Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.