A Statistical framework to model the meeting-in-the-middle principle using metabolomic data: application to hepatocellular carcinoma in the EPIC study.

Nada Assi[1], Anne Fages[2,a], Paolo Vineis[3], Marc Chadeau-Hyam[3],Magdalena Stepien[1], Talita Duarte-Salles[1], Graham Byrnes[1] ,Houda Boumaza[2], Sven Knüppel[4], Tilman Kühn[5], Domenico Palli[6], Christina Bamia[7], Hendriek Boshuizen[8], Catalina Bonet[9], Kim Overvad[10], Mattias Johansson[1,11], Ruth Travis[12], Marc J Gunter[3], Eiliv Lund[13], Laure Dossus[14-15], Bénédicte Elena-Herrmann2, Mazda Jenab[1], Vivian Viallon[16-18†], Pietro Ferrari[1†]*

[1] International Agency for Research in Cancer (IARC-WHO), 150 Cours Albert Thomas, 69372 Lyon CEDEX 08, France

[2] Centre de RMN à Très Hauts Champs, Institut des Sciences Analytiques (CNRS/ENS Lyon/UCB Lyon 1), Université de Lyon, 69100 Villeurbanne, France

[3] Department of Epidemiology and Biostatistics, MRC-HPA Centre for Environment and Health, School of Public Health, Imperial College London, Norfolk Place, London, W2 1PG, United Kingdom

[4] Department of Epidemiology, German Institute of Human Nutrition, Potsdam-Rehbruecke, 14558 Nuthetal, Germany

[5] Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

[6] Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Institute – ISPO, Florence- Italy

[7] WHO Collaborating Center for Food and Nutrition Policies, Department of Hygiene, Epidemiology and Medical Statistics, University of Athens Medical School, Athens, Greece

[8] National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands

[9] Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Institut Català d'Oncologia, L'Hospitalet de Llobregat, Spain

[10] The Department of Epidemiology, School of Public Health, Aarhus University, Aarhus, Denmark

[11] The Department for Biobank Research, Umeå University, Sweden

[12] Cancer Epidemiology Unit, Nuffield Department of Population Health University of Oxford, Oxford, UK

[13] The Institute of Community Medicine, University of Tromsø, Tromsø, Norway

[14] Inserm, Centre for research in Epidemiology and Population Health (CESP), U1018, Lifestyle, genes and health: integrative trans-generational epidemiology team, Villejuif, France
[15] Université Paris Sud, Villejuif, France
[16] Université de Lyon, F-69622, Lyon, France
[17] Université Lyon 1, UMRESTTE, F-69373 Lyon, France
[18] IFSTTAR, UMRESTTE, F-69675 Bron, France
[a] Present address: Chemical Physics Department, Weizmann Institute of Science, Rehovot, Israel
[†] Joint last authors.

[*]Corresponding author: Pietro Ferrari, International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon cedex 08, France. Tel: +33 472 73 8031; Fax: +33 472 73 8361. E-mail: ferrarip@iarc.fr

Abstract

Metabolomics is a potentially powerful tool for identification of biomarkers associated with lifestyle exposures and risk of various diseases. This is the rationale of the "meeting-in-the-middle" concept, for which an analytical framework was developed in this study. In a nested case-control study on hepatocellular carcinoma (HCC) within the European Prospective Investigation into Cancer and nutrition (EPIC), serum ^{1}H NMR spectra (800 MHz) were acquired for 114 cases and 222 matched controls. Through Partial Least Square (PLS) analysis, 21 lifestyle variables (the "predictors", including information on diet, anthropometry and clinical characteristics) were linked to a set of 285 metabolic variables (the "responses"). The three resulting scores were related to HCC risk by means of conditional logistic regressions. The first PLS factor was not associated with HCC risk. The second PLS metabolomic factor was positively associated with tyrosine and glucose, and was related to a significantly increased HCC risk with OR= 1.11 (95%CI: 1.02, 1.22, p=0.02) for a 1-SD change in the responses score, and a similar association was found for the corresponding lifestyle component of the factor. The third PLS lifestyle factor was associated with lifetime alcohol consumption, hepatitis and smoking, and had negative loadings on vegetables intake. Its metabolomic counterpart displayed positive loadings on ethanol, glutamate and phenylalanine. These factors were positively and statistically significantly associated with HCC risk, with 1.37 (1.05, 1.79, p=0.02) and 1.22 (1.04, 1.44, p=0.01), respectively. Evidence of mediation was found in both the second and third PLS factors, where the metabolomic signals mediated the relation between the lifestyle component and HCC outcome. This study devised a way to bridge lifestyle variables to HCC risk through NMR metabolomics data. This implementation of the "meeting-in-the-middle" approach finds natural applications in settings characterized by high-dimensional data, increasingly frequent in the –omics generation.

Keyword: partial least square, lifestyle factors, metabolomics, hepatocellular carcinoma, meeting-in-the-middle, molecular epidemiology.

Introduction

Metabolomic profiles from blood and other biological samples collected from large-scale epidemiologic studies are increasingly being investigated [1], following recent developments in nuclear magnetic resonance (NMR) and mass spectrometry (MS) enabling the assessment of metabolic profiles for large numbers of individuals. As a result, metabolomic data is gradually playing a key part in clinical and observational studies; and new statistical methodologies [2] are increasingly being sought to explore insights into pathological processes that metabolomics may provide in order to better understand determinants of disease development. These approaches explore a variety of etiological hypotheses; however they usually focus on one aspect at a time, combining metabolomics with either epidemiologic/phenotypic data on lifestyle exposures [3] or with disease outcomes [4,5]. The main aim of this work is to jointly use all aspects that are potentially informative to apprehend the contrivances of disease development.

Metabolomic data offers the opportunity to identify signatures and biomarkers associated with environmental exposures and the risk of a disease. Prospective studies are conceptually suitable for this purpose, since they rely on biological samples collected before disease onset, and are thus marginally influenced by metabolic changes due to processes of disease development. In this scenario, the "meeting-in-the-middle" (MITM) approach [6] has been conceived as a research strategy to identify biomarkers that are related to specific exposures and that are, at the same time, predictive of disease outcome. Finding this overlap between exposure and disease of "intermediate" biomarkers can potentially disclose useful information on the exposure-to-disease pathway, and may serve as an objective risk exposure measure, ultimately allowing the identification of a targeted prevention scheme. The MITM was previously implemented as a proof of concept in a case-control study nested within a cohort of healthy individuals [7], where a list of putative intermediate ^{1}H NMR biomarkers

linking exposure to dietary compounds, mainly micro- and macronutrients, and disease outcomes (colon and breast cancer) were investigated.

In this study we extend previous attempts to model the MITM by fully integrating metabolomics, lifestyle and disease risk in a single analytical framework. A strategy was developed to simultaneously investigate a broad range of metabolites and lifestyle variables with a partial least square (PLS) regression model [8]. The resulting scores were related to the risk of hepatocellular carcinoma (HCC), in a case-control study nested within the European Prospective Investigation into Cancer and nutrition (EPIC). HCC is the most frequent primary form of cancer affecting the liver, an organ that plays a critical role in many metabolic pathways [9]. HCC is a disease with multifactorial origins embracing lifestyle and dietary exposures whose intersection may reveal metabolomic signals [10] relevant to cancer onset. The system of relationships between metabolomic profiles and lifestyle factors in relation to HCC was evaluated by means of mediation analysis. The methodological challenges characterizing the analysis of large and complex metabolomic datasets are described and discussed.

Methods

EPIC design. The European Prospective Investigation into Cancer and nutrition (EPIC) is a large cohort established to investigate the association of diet, lifestyle and environmental factors with cancer incidence and other chronic disease outcomes. Between 1992-2000, over 520,000 participants aged 20-85 years, were recruited from 23 centers in 10 Western European countries including Denmark, France, Germany, Greece, Italy, Norway, Spain, Sweden, The Netherlands and the United Kingdom [11]. The design, rationale and methods of the EPIC study including information on dietary assessment methodology, blood collection protocols and follow-up procedures were previously detailed [11].

Between 1992 and 1998, standardized lifestyle data, anthropometric measures and biological samples were collected at recruitment, prior to onset of any disease [11]. Validated country-specific questionnaires ensuring high compliance were used to measure diet over the previous 12 months [12]. Blood samples are stored at the International Agency for Research on Cancer (IARC, Lyon, France) in -196˚C liquid nitrogen for all countries, exceptions being Denmark (nitrogen vapour,-150˚C) and Sweden (freezers, -80˚C).

The nested case-control study. The present study focused on data with available sera samples from a nested case-control study in EPIC on hepatocellular carcinoma (HCC) [13]. Cases of HCC were identified from all participating EPIC centres except for Norway and France (n=117) from recruitment (1993-1998) up to 2007. Two controls (n=232) were selected for each case from all cohort members alive and free of cancer (except non-melanoma skin cancer) by incidence-density sampling and were matched on age at blood collection (±1 year), sex, study centre, date (±2 months), time of the day at blood collection (±3 hours) and fasting status at blood collection (<3, 3-6, >6 hours); among women, additional matching criteria included menopausal status (pre-, peri-, post-menopausal) and hormone replacement therapy (HRT) use at time of blood collection (yes/no). In the present study, cases and controls were both included in the analyses as the subjects were all cancer-free at blood collection. Out of the total 349 subjects, 7 subjects (3 cases and 4 controls) had too little serum volume for NMR spectral acquisition with sufficient sensitivity; 6 additional control subjects were excluded following the exclusion of their corresponding case subject. The final analysis included 114 HCC cases and 222 matched controls of which 108 case–control sets with two matched control subjects and 6 sets with one matched control subject.

NMR spectra acquisition. Sera were processed using standard procedure for ^{1}H NMR metabolic measurement and profiling protocols [14]. Details on the sera sample preparation as well as NMR data acquisition and processing have been described elsewhere [15]. In brief,

149 each spectrum was reduced to 8,500 bins of 0.001 ppm width over the chemical shift range of
150 0.5 to 9 ppm. Spectra were normalized to total intensity, centred and Pareto scaled, and
151 additionally normalized for batch-effects using the batch profiling calibration method [16].
152 After removal of the structured noise (characterized by a specific mean and standard
153 deviation) located in a well-known noise region (8.5-9ppm) and variables with identical
154 characteristics, the statistical recoupling of variables (SRV) [17], a bucketing procedure, was
155 applied to the metabolomic spectra. The SRV procedure identifies clusters of variables with
156 respect to the ratio of covariance and correlation between consecutive variables along the
157 chemical shift axis, allowing the restauration of the spectral dependency and the recovery of
158 complex NMR signals corresponding to potential physical, chemical or biological entities.
159 More details on the SRV procedure are available in the **Mathematical Appendix**. This
160 permitted a reduction of the number of NMR variables from 8,500 bins to 285 clusters of
161 variables corresponding to reconstructed peak entities which constituted the Y-set of
162 metabolic variables. All steps to obtain the data were done without knowledge of the case-
163 control status of the subjects. Quality control (QC) samples were included to ensure
164 reproducibility of the NMR data acquisition.

165 *Metabolite identification.* The assignment of NMR signals observed in the ^{1}H one-
166 dimensional fingerprints to metabolites has been achieved by the analysis of additional 2D
167 NMR experiments ^{1}H-^{13}C HSQC and ^{1}H-^{1}H TOCSY obtained on a subset of representative
168 samples (one control and one case). The measured chemical shifts were compared to
169 reference shifts of pure compounds using HMDB [18], MMCD [19] and ChenomX,
170 (ChenomX NMR suite, ChenomxInc, Edmonton, Canada) databases.

171 *Lifestyle variables.* The predictors (what will be referred to later on as the X-set) included 13
172 dietary variables from main EPIC food groups compiled from validated country-specific food
173 frequency questionnaires (FFQ) [11,20] (potatoes and other tubers; vegetables; legumes;

fruits, nuts and seeds; dairy products; cereal and cereal products; meat and meat products; fish and shellfish; egg and egg products; fat; sugar and confectionary; cakes and biscuits; non-alcoholic beverages), alcohol average lifetime intake (continuous, g/day), anthropometric measures including body mass index (continuous, kg/m^2) and height (continuous, cm) that were measured by trained interviewers in the majority of participants [11], highest level of education achieved (categorical: none or primary school completed, technical/professional school, secondary school, longer education (incl. university degree), unspecified), smoking status (categorical: never, former, current smoker, unknown), a measure of physical activity (continuous, metabolic equivalents of task (MET)/h), hepatitis status (yes/no, from biomarker measures of HBV and HCV seropositivity [ARCHITECT HBsAg and anti-HCV chemiluminescent microparticle immunoassays; Abbott Diagnostics, France]) and baseline self-reported diabetes status (yes/no). Descriptive information on these variables can be found in **Supplementary table 1**.

Statistical analysis

PC-PR2 analysis. Principal component partial R-square (PC-PR2) was primarily used to identify and quantify sources of systematic variability within metabolomic data [15]. PC-PR2 combines aspects of principal component analysis (PCA) and the $R^2_{partial}$ statistic in multiple linear regression, and allows for (some) inter-correlation between the explanatory variables under scrutiny [15]. In short, PCA is performed on the 285 clusters of ^{1}H NMR variables and a number of components is retained explaining an amount of total variability above a designated threshold (here, 80%). Then, multiple linear regression models are fitted where each component's variability is explained in terms of relevant covariates, e.g. specific characteristics of samples like country of origin, smoking status, laboratory treatment, etc. For each given component, the $R^2_{partial}$ statistic is computed for all covariates, quantifying the amount of variability each independent variable explains, conditional on all other covariates

included in the model. Finally, an overall $R^2_{partial}$ is calculated as a weighted average for every covariate, using the eigenvalues as components' weights. Mathematical details pertaining to the PC-PR2 method are described elsewhere [15].

In this study, PC-PR2 was applied to the 285 clusters of NMR variables, whereas the explanatory variables examined for systematic variability were NMR batch, country of origin, sex, age at blood collection, serum clot contact time (centrifugation at the day of blood collection d, or the following day, d+1), length of freezing time (<= 15 vs. >15 years), and fasting status at blood collection (< 3, 3-6, > 6 hours). With the similar motivation of identifying sources of variability within lifestyle data, a similar PC-PR2 analysis was applied to the 21 lifestyle factors; the examined covariates for systematic variability were country of origin, sex and age at recruitment. For both metabolomics and lifestyle data, residuals on the variable accounting for most variability, identified through PC-PR2 analyses, were computed in a series of univariate linear regression models [21] and were used in the subsequent PLS.

PLS analysis. A PLS model was used to relate lifestyle variables to metabolomic profiles. PLS is a multivariate technique that generalizes features of PCA and multiple linear regression. PLS iteratively extracts linear combinations of, in turn, predictors (the X-set) and responses (the Y-set), which in this study, were lifestyle variables and metabolomic profiles, respectively. First, components or latent factors are extracted allowing a simultaneous decomposition of the X- and Y-sets, in order to maximize their covariance [22]. The factors extracted from the predictors' set are orthogonal. Computational details of PLS are described in the **Mathematical Appendix**. As a standard step for the PLS algorithm, the X- and Y-sets were centered and standardized for the analysis and a simple expectation-maximization (EM) algorithm, adapted from the PLS kernel algorithm [23,24], was used to compute covariance matrices when missing values were present in the lifestyle data. This was done as follows: a first pass of PLS was computed filling in the missing values by the average of the non-

missing values for each corresponding variable. A second pass was then performed whereby the missing data were assigned their predicted values based on the first model, and the PLS regression is recomputed.

Then, a seven-fold cross validation analysis was carried out to select the number *h* of significant PLS factors to retain [8] (see **Mathematical appendix**).This was achieved by splitting the data into seven groups of observations. In turn, each group of observations was considered as the test set, whilst the other six were the training sets, used to perform PLS analysis. A measure of PLS performance was determined for each step through the predicted residual sum of squares (PRESS) statistic, whereby the predicted values in the test set, the $\tilde{Y}_h$ matrix, based on the X-components estimated through the model in the training set, were compared to the observed responses, the Y matrix. This comparison is quantified by the squared Euclidean distance between these two matrices. In turn for an increasing number *h* of components, the process is iterated seven times, until each group of observations serves as a test set. Eventually, the number *h* of selected PLS factors is the one minimizing the PRESS statistic.

For each PLS factor, loadings were computed for the lifestyle (X-set) and the NMR (Y-set) variables. The loadings, i.e. coefficients quantifying the contribution of each original variable to the PLS factor, were used to characterize the various factors. As the analysis involved many variables in the X-set and, particularly, in the Y-set, the interpretation focused primarily on variables with loading values lower than the 10^{th} percentile and larger than the 90^{th} percentile for the X variables, and lower than the 5^{th} and larger than the 95^{th} percentiles for the Y variables, that were deemed the most significant contributors to the PLS factor.

Logistic regression analysis. Last, scores of each PLS factor were related to HCC risk in conditional logistic regression models to compute HCC odds ratios (ORs) and associated 95% confidence intervals (95% CI) where ORs express the change in HCC risk associated to one

standard deviation (1-SD) increase in the score. Models were adjusted for C-reactive protein concentration, alpha-fetoprotein concentration and for a composite score indicative of liver damage. The score summarizes the number of abnormal values of circulating enzymes measured in the hepatic tissue in six liver function tests (alanine aminotransferase >55 U/L, aspartate aminotransferase >34 U/L, gamma-glutamyltransferase: men>64 U/L and women>36 U/L, alkaline phosphatase >150 U/L, albumin<35 g/L, total bilirubin>20.5 µmol/L; cut-points were provided by the clinical biochemistry laboratory that conducted the analyses and were based on assay specifications) [25]. These biomarkers were measured on the ARCHITECT c Systems™ and the AEROSET System (Abbott Diagnostics) using standard protocols. Laboratory analyses were performed at the Centre de Biologie République laboratory, Lyon, France. These adjustments were deemed necessary to address potential confounding stemming from metabolic disorders, inflammation or underlying liver dysfunction [25–28]. Adjustments for total dietary fibre, vitamin D, calcium and iron intakes (continuous) were evaluated but not retained in the final models for lack of confounding exerted by these variables. The receiver operating characteristic (ROC) curve and the associated area under the curve (AUC) were determined from conditional logistic regressions to evaluate the predictive performance of PLS models. AUC values were computed for conditional logistic models including progressively the PLS scores, separately for lifestyle and metabolomic factors (as shown in **Table 4**, column 1). The sensitivity, specificity and accuracy were calculated for a cut-off point, selected as the minimal distance between the ROC curve and the upper left corner of the diagram [29,30]. The corrected positive predictive value (PPV), taking into account the nested case-control design [31,32] was computed by including the prevalence of HCC in the EPIC population(π= 0.0004), computed over a 7-year period (1992-2010) where 191 HCC cases were ascertained from a total of 477,206 participants included for case identification after relevant exclusions [33]. The AUC

unavoidably increases with the number of covariates added to the conditional logistic model. To address this issue, a resampling scheme was devised to compute an objective/ unbiased estimate of the AUC, inspired by the work of Uno et al [34]. For each one of the 1000 drawn bootstrap samples, a 10-fold cross-validation was performed, repeated ten times to remove variation due to random partitioning of data and to yield more stable estimates. The predicted values from each of the conditional logistic models in the training set were used to derive AUC values in the test set. The 2.5^{th} and 97.5^{th} percentile values made up the 95% confidence intervals.

Sensitivity analyses. A sensitivity analysis was performed by running PLS on data excluding sets where cases were diagnosed within the first two years of follow-up. The model was conducted on 271 observations (92 cases, 179 controls), to investigate the performance of the PLS model, ruling out potential reverse causation. The metabolomic profiles of HCC cases diagnosed within two years from enrolment could reflect the presence of the tumour rather than informing about tumour aetiology. The variable importance in the projection (VIP) statistic was used to facilitate the comparison of the sensitivity analysis with the main analysis. The VIP expresses the explanatory power of a predictor variable X across all response variables Y (see **Mathematical Appendix**).

Mediation analysis. The mediating role of the Y-scores in the association between lifestyle profiles and HCC risk was assessed. Separately for each extracted combination of lifestyle and metabolomic PLS factors, mediation analyses were performed with the 'paramed' Stata function that allows for exposure-mediator interaction based on Valeri and VanderWeele's work [35]. Briefly, mediation was computed using a Baron and Kenny approach adapted to dichotomous outcomes [36], where two models were specified. In the mediator model, the mediator (the Y-score) was linearly regressed on the exposure (the X-score), while in the outcome model the exposure (X-score) and the mediator (Y-score) were related to the HCC

indicator in unconditional logistic regressions. Both models accounted for the concentration of C-reactive protein, alpha-fetoprotein and the composite score of liver damage, and additionally accommodated the other extracted metabolic profiles (Y-scores) to control for mediator-outcome confounders that may occur when estimating the Natural Indirect Effect (NIE) [35]. As the outcome (HCC) is rare, direct and indirect effects can be estimated taking into account the case-control design. This is done by using the same formulas for the effects, while running the mediator regression only for the controls [36]. As mediation packages do not yet accommodate conditional logistic models, the outcome and the mediator models, which were accommodated in unconditional logistic regressions, were adjusted for center and age at blood collection for sake of consistency with previous steps of the analysis.

Statistical analyses were performed using R [37] and SAS [38] in general, with the following packages for specific purposes: PROC PLS in SAS 9.4 for PLS analyses, 'paramed' in Stata 12 [39] for mediation analyses, 'OptimalCutpoints' in R for ROC-related assessments.

The different steps of the analytical framework developed in this study to model the MITM are presented in **Figure 1**.

Results

In the PC-PR2 analyses, a total of 17 and 14 principal components were retained to explain an amount of total variability exceeding 80% in metabolomics and lifestyle data respectively. **Figure 2** shows that the ensemble of explanatory variables accounted for 19.4% and 26.7% of total variance, respectively in metabolomics and lifestyle data, of which the highest contributor was 'country of origin' with consistently 8% and 22%. ~~. Major sources of variation in the Y-set displaying large $R^2_{partial}$ value were country of origin (8.3%), NMR batch (4.0%) and fasting status at blood collection (1.6%). In the X-set, country of origin (22.6%) and sex (5.1%) showed the highest contributions. As PC-PR2 analysis showed that~~

~~'country of origin' accounted for about 8% of the variability within the metabolomic data, and 22% in the lifestyle variables, the~~ PLS analysis was carried controlling for this variable.

After a seven-fold cross-validation, three PLS factors were retained accounting for 21.7% and 8.5% of the overall variability observed in predictor and response variables, respectively (**Table 1**). Lifestyle variables and clusters of NMR variables contributing highly to PLS factors were identified using factor loading values (**Table 2**). The first PLS factor was predominantly positively associated with dairy products and cakes and biscuits intake, while lifetime alcohol intake, smoking status and diabetes displayed negative loadings for this lifestyle component (**Table 2**). On the same PLS factor, signals mainly associated with glucose and bonds of lipids with negative loading values, and with aspartate, glutamine and lysine with positive loadings emerged on the metabolomic profile (**Table 2**). Lifestyle variables characterizing the second PLS factor included cereal products, height and education level with negative loadings, and hepatitis with positive loadings. The metabolic signature included NMR variables with positive loadings associated with aromatic amino acids (phenylalanine, tyrosine) and glucose; and those with negative loadings associated mainly with bonds of lipids, threonine and mannose (**Table 2**). The third PLS factor had a lifestyle pattern outlining intake of vegetables (high negative loadings values), lifetime alcohol consumption, smoking, and hepatitis infection (positive loadings). Its counterpart NMR pattern highlighted signals of glucose and aspartate, with high negative loadings, along with signals of ethanol, myo-inositol, proline and glutamate as prominent metabolites with positive loadings (**Table 2).**

Conditional logistic regression models relating HCC risk with the X- and Y-scores are shown in **Table 3**. The first PLS factor was associated to a non-significant decreased HCC risk (23% and 4% in the X- and Y-scores respectively), while the second and third factors were associated to a statistically significant increased HCC risk (54% and 11%; and 37% and

22% respectively). Results for the ROC curves parameters are reported in **Table 4**, including AUC, sensitivity, specificity, accuracy and PPV for different combinations of the X- and Y-scores. The AUC of the X-scores and Y-scores for all 3 PLS factors, adjusted for C-reactive protein concentration, alpha-fetoprotein concentration and the score of liver damage, was respectively 0.859 and 0.853. An increase in the resampled cross-validated AUC values was also observed for all three X- and Y-scores, albeit smaller, with respectively 0.836 and 0.827. Results from the sensitivity analysis conducted on data excluding sets where cases were diagnosed within the first two years of follow-up, showed similarities in terms of lifestyle variables' and metabolites' loadings on the PLS factors (**Supplementary Table 2**). Notable differences pertained to the identification of new signals for the first PLS factor including ethanol, histidine and an unknown compound. On the second lifestyle factor, BMI (positive loadings) replaced education level (negative loadings) while the reflected metabolomic profile was comparable to its counterpart from the main analysis **(Supplementary Table 2)**. On the third factor, smoking status and hepatitis (positive loadings) were replaced by sugar and confectionary intake (negative loadings); signals contributing to the associated metabolic profile remained the same but the direction of the association was inversed as loadings had opposite signs as compared to the counterpart PLS factor of the main model (**Supplementary Table 2**). Corresponding ORs from conditional logistic regression models relating the X- and Y-scores to HCC risk are available in **Table 5**. The scores showed a statistically significant association in the second factor for both sets and in the third factor for the Y-set. ROC-associated statistics for different models are presented in **Supplementary Table 3**. The VIP plot (**Figure 3**) displayed the results for the importance of the lifestyle variables in the prediction of the Y-set computed for the main PLS model performed including all subjects (panel **A**) and for the sensitivity model (panel **B**). The results suggested a potential gain in stability as prominent lifestyle variables for prediction were maintained

(hepatitis/diabetes/cakes and biscuits), the magnitude of the VIP was improved for some (fat/lifetime alcohol intake) and less emphasis was put on others (BMI/physical activity).

Finally, the natural indirect effect was assessed in the mediation analyses and the results are presented in **Table 6**. Overall, there was limited evidence that metabolomic signals mediated the association between lifestyle components and HCC risk in the first PLS factor. Evidence of a significant mediated effect by the Y-scores was found in the second and third PLS factors when models were adjusted for exposure-mediator interaction (**Table 6**).

Discussion

In this work, an analytical strategy based on PLS analysis was conceived to extract relevant information from sets of lifestyle and NMR metabolomic variables, and to relate the resulting components to the risk of disease. This offered a way to implement the MITM approach [6] in a nested case-control study on HCC within the EPIC study. MITM has been suggested as a way to link specific putative metabolites to lifestyle exposures and disease outcomes, thus leading to the identification of potential intermediate biomarkers [6].

An implementation of MITM was previously carried out in a nested case-control study in the Turin sub-cohort of EPIC [7] based on prospectively collected plasma samples from a pilot study on colon and breast cancers. In their work, a list of intermediate markers was identified by an in-parallel evaluation of the relationships between untargeted ^{1}H NMR profiles with dietary exposures and risk of colon and breast cancers using correlation analysis and logistic regression. In our study, a different analytical framework was developed, largely exploiting features of PLS analysis, a multivariate technique that iteratively extracts components capturing co-variability in sets of predictors and response variables [8,40]. A set of lifestyle predictor variables were related to NMR responses. In a second step, PLS predictors' and responses' scores were linked to the risk of HCC.

Another sensitive issue in this analysis was the choice of lifestyle variables. Two disease-indicator variables reflecting environmental exposures, diabetes and hepatitis, were included in the set of predictors, as they turned out to have an important role in the characterization of metabolomic signatures. In addition, diabetes is the main metabolic risk factor for HCC alongside with fatty liver disease [41,42], and chronic infection with hepatitis B (HBV) and particularly hepatitis C (HCV) viruses were classified as class I carcinogens for HCC by IARC [43].

Other relevant biomarkers were not part of the list of predictors in PLS analysis, but were controlled for in logistic regression models. This included C-reactive protein, alpha-fetoprotein, and a score for liver damage, an index of different circulating enzymes measured in the hepatic tissue indicating potential underlying liver function impairment [25]. The alpha-fetoprotein was ~~not~~ included as an adjustment factor in the analyses not because of its established part as a serum marker for HCC diagnosis [26,44], but rather to account for it as a potential confounder that may cloud the relation between scores and HCC, both in conditional logistic regressions and in mediation analyses.

Similarly to other multivariate techniques, a key aspect of PLS analysis is the choice of the number of factors to retain, in an effort of exhaustively summarizing data variability through a limited number of factors. Based on a seven-fold cross-validation, three linear combinations of variables were extracted in this work. A challenging aspect of this analysis is the interpretation of these factors, with respect to lifestyle and metabolomic variables. A subjective criterion based on the distribution of loading values was used throughout. The variables displaying the most extreme loading values (in absolute terms) were the ones characterizing each factor.

The first lifestyle factor highlighted a healthy pattern with negative loadings for diabetes status, smoking status and lifetime alcohol intake, and was not associated to HCC

risk, similarly to its metabolomics counterpart. The lifestyle component of the second PLS factor, was reflective of a lifestyle pattern reflective of "higher-risk exposures", and was related to a significant 54% increase in HCC risk. Likewise, its associated metabolic component displayed a significant HCC risk augmentation by 11%. The lifestyle component of the third PLS factor described participants with lower vegetables intake, elevated lifetime alcohol consumption, more likely to be ever smokers and hepatitis positive; one standard deviation increase of this component was associated to a statistically significant 37% increase in HCC risk. Similarly, a 22% significant increase in HCC risk was observed for its metabolic counterpart, characterized by positive signals of ethanol and myo-inositol, and displayed negative loadings for glucose.

The MITM is captured by the rationale of PLS analysis, in the sense that each set of lifestyle profiles and metabolic signatures of the extracted PLS factors mirrored one another. In addition, mediation was observed for the second and third PLS factors, whereby the metabolomic component mediated the relation between the lifestyle component and HCC, for which statistically significant associations with HCC risk were estimated, emphasising the presence of a MITM. Mediation analysis relies on the assumption that there is no mediator-outcome confounder that is affected by the exposure [35]. In our study C-reactive protein, alpha-fetoprotein and liver damage score were weakly correlated to lifestyle factor score, thus introducing potential bias in the estimation of direct and indirect effects in our mediation analysis. Additionally, a number of background confounders (mediator-outcome and exposure-outcome confounders) were present that we have tried to control for, either by adjustments or by accounting for potential interactions, however some degree of bias can remain and caution should be employed when interpreting the results.

The predictive performance of PLS factors in relation to HCC occurrence was evaluated through an analysis of AUC values. The performance of the model ~~was~~ improved

progressively, with all 3 X- and Y-scores added; ~~and reached an AUC of 0.859 with all 3 X-scores and 0.853 with all 3 Y-scores, with adjustment for concentrations of C-reactive and alpha-fetoprotein, and for the liver damage score. A~~after a bootstrapped cross-validation, the AUC estimates were lower ~~with respectively 0.836 and 0.827,~~ but the increase in the performance was nevertheless present. The ROC methodology allows estimation of PPV, which expresses the risk of disease after a positive test [45]. In a setting with low HCC prevalence (π=0.0004), in line with Western populations [46], extremely low PPV estimates were observed. In the absence of a very specific test, many positive tests arise from disease-free individuals [45], thus leading to a dilution of PPV.

A sensitivity analysis was carried out excluding the first two years of follow-up, but results were virtually unchanged, both in terms of relative risk estimates in logistic regression models, and of percentage of variability explained in PLS analysis. These findings suggest that reverse causation bias, if present, was minimal.

This study had the ambition of integrating in the same analytical framework study participants' lifestyle characteristics with a large number of NMR metabolic profiles. These data pose a number of methodological challenges due to their size and the complexity of exhaustively capturing and interpreting the biological processes they reflect. To address these issues, techniques involving multivariate statistics have been progressively revived in the recent years [2]. Epidemiologic evaluations of metabolomic data frequently combined PLS with discriminant analysis, such as PLS-DA or O-PLS-DA. The main objective of these methods is to identify a series of metabolomic features distinguishing between two very distinct groups of study participants [47,48]. In such strategies, only one set of variables is multi-dimensional and the response is one variable only. Similar multivariate techniques for pattern extraction, belonging to the family of regression methods, include reduced rank regression. This multivariate method relates an ensemble of response variables to a set of

predictor variables where the estimated matrix of the regression coefficients is of reduced rank [49–51]. In addition, canonical correlation analysis (CCA) [52] is a method applied to identify the optimum structure or dimensionality of each variable set that maximizes the relationship between two sets of multi-dimensional variables. The main difference between CCA and PLS regression is that CCA maximizes the correlation between the two new dimensions, i.e. extracted factors, whereas PLS maximizes their covariance. PLS can be considered as a trade-off between CCA and PCA, since maximizing the covariance corresponds to maximizing the product of the correlation and standard deviation, given that cov(X,Y)=cor(X,Y)*SD(X)*SD(Y).

Untargeted NMR was used in this work to acquire metabolomic signals. Prior to PLS analysis, a bucketing procedure, the statistical recoupling of variables (SRV) [17,53], was applied to reduce the number of NMR variables to 285 clusters. This was done by aggregating consecutive NMR bins based on their covariance to correlation ratio. ~~, thus reconstructing peak entities. Neighbouring variables were then merged into clusters, to recover NMR multiplets, corresponding to NMR variables of interest.~~ This allow~~eds~~ the identification of informative components of the spectra, thus acting as an efficient noise-removing filter. Subsequently the annotation effort remains challenging, for a number of reasons. The majority of published metabolomics studies often identified a limited number of metabolites at a time [54], and the Human Metabolome Database (HMDB) and other related resources [18,55], that offer richly annotated information continuously increasing the metabolite coverage for users, are mostly exploited through time consuming interactive procedures. In addition, individual metabolites often overlap in NMR signals, which can hinder ~~interpretation of the annotated metabolic profiles~~annotations. ~~Untargeted NMR approaches are useful for the identification of metabolites of moderate abundance; however they may miss low-abundance metabolites due to the intrinsically low NMR sensitivity.~~ These

challenges, as well as large variability in metabolite concentrations, and disentangling informative signals from noise, are not specific to NMR and pertain to any type of untargeted technique. Such investigations may profit from complementary targeted metabolomic analytical strategies [55].

Throughout the different steps of this work, the scaling problem was first tackled by normalizing spectra to total intensity. NMR data were also centered and Pareto-scaled, together with correction for potential batch effects [16]. The PC-PR2 method offered a way to investigate major sources of systematic variability in NMR and lifestyle data [15]. The variable "country of origin" emerged as the variable accounting for the largest proportion of total variability, and the residual method was used to control for this variable in the following steps of the analysis. While this may lead to removing regional gradients of dietary variability, this step is instrumental to avoid unwanted systematic regional-specific bias in the data in country-specific questionnaire assessments. In addition, technical aspects like storage and handling of biological samples, fasting status at blood collection are specific to each country [15]. In any case, variability due to "country of origin" is not exploited in conditional logistic models, as cases and controls were also matched on center.

One of the limitations of this study is the restricted sample size which raises concerns with regards to power to detect associations. While a larger sample size would possibly result in more statistically significant findings, we used the data that was available with NMR profiles measured. In this work we have developed a framework to analyse complex data integrating lifestyle and metabolomics in relation to risk of disease. The approach described in this study has merits but also pitfalls among which it is worth mentioning that statistical methods are used repeatedly on the same set of data, notably the PLS model, the conditional logistic regression, the AUC estimation and mediation analysis. To partially address this, a cross-validation approach was devised for AUC estimation which involved conditional

logistic regression, whereby PLS was done without knowledge of the case/control status. However, conditional logistic regression models and mediation analyses were implemented on the same data, and our analysis did not account for this limitation. This may have led to spuriously increase the nominal level of statistical significance of statistical tests.

Conclusion

The MITM emerged as a method for the identification of relevant biomarkers, with great potential to unravel utmost important steps in the aetiology of disease. The analytical strategy for MITM was developed to use all potentially informative aspects of high-throughput data by integrating metabolomic, dietary and lifestyle exposures together with disease indicators. While the framework was applied towards the investigation of HCC determinants, it can be easily extended to similar aetiological contexts and applied to other –omics settings.

Funding

This work was supported by the French National Cancer Institute (L'Institut National du Cancer; INCA) [grant number 2009-139; PI: M. Jenab]. The coordination of the European Prospective Investigation into Cancer and nutrition is financially supported by the European Commission (Directorate General for Health and Consumer Affairs) and the International Agency for Research on Cancer. The national cohorts are supported by Health Research Fund (FIS) of the Spanish Ministry of Health RTICC 'Red Temática de Investigación Cooperativa en Cáncer [Grant numbers: Rd06/0020/0091, Rd12/0036/0018], Regional Governments of Andalucía, Asturias, Basque Country, Murcia [project 6236] and Navarra, Instituto de Salud Carlos III, Redes de Investigacion Cooperativa (RD06/0020) (Spain); the Danish Cancer Society (Denmark); the Ligue Contre le Cancer, the Institut Gustave Roussy, the Mutuelle Générale de l'Education Nationale, and the Institut National de la Santé et de la Recherche Médicale (France); the Deutsche Krebshilfe, the Deutsches Krebsforschungszentrum, and the Federal Ministry of Education and Research (Germany); the Hellenic Health Foundation, the Stavros Niarchos Foundation, and the Hellenic Ministry of Health and Social Solidarity (Greece); the Italian Association for Research on Cancer AIRC and the National Research Council (Italy); the Dutch Ministry of Public Health, Welfare and Sports, the Netherlands

Cancer Registry, LK Research Funds, Dutch Prevention Funds, the Dutch Zorg Onderzoek Nederland, the World Cancer Research Fund, and Statistics Netherlands (Netherlands); European Research Council-2009-AdG 232997 and the Nordforsk, Nordic Centre of Excellence program on Food, Nutrition and Health (Norway); the Swedish Cancer Society, the Swedish Research Council, and the Regional Governments of Skåne and Västerbotten (Sweden); Cancer Research UK, the Medical Research Council, the Stroke Association, the British Heart Foundation, the Department of Health, the Food Standards Agency, and the Wellcome Trust (United Kingdom). The work undertaken by N Assi was supported by by the Université de Lyon I through a doctoral fellowship awarded by the EDISS doctoral school.

Acknowledgments

We would like to acknowledge the assistance of Dr Elodie Jobard from the ISA-CRMN in obtaining the annotation of the NMR data.

Reference List

1. Nicholson, J.K., Holmes, E., and Elliott, P. (2008) The metabolome-wide association study: a new look at human disease risk factors. *J. Proteome Res.*, **7**,3637–3638.

2. Chadeau-Hyam, M., Campanella, G., Jombart, T., Bottolo, L., Portengen, L., Vineis, P., Liquet, B., and Vermeulen, R.C.H. (2013) Deciphering the Complex: Methodological Overview of Statistical Models to Derive OMICS-Based Biomarkers. *Environ. Mol. Mutagen.*, **54**,542–557.

3. Floegel, A., Wientzek, A., Bachlechner, U., Jacobs, S., Drogan, D., Prehn, C., Adamski, J., Krumsiek, J., Schulze, M.B., Pischon, T., and Boeing, H. (2014) Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study. *Int. J. Obes. (Lond).*,1–9.

4. Trushina, E. and Mielke, M.M. (2013) Recent advances in the application of metabolomics to Alzheimer's Disease. *Biochim. Biophys. Acta*, **1842**,1232–1239.

5. Jin, X., Yun, S.J., Jeong, P., Kim, I.Y., Kim, W.-J., and Park, S. (2014) Diagnosis of bladder cancer and prediction of survival by urinary metabolomics. *Oncotarget*, **5**,1635–1645.

6. Vineis, P. and Perera, F. (2007) Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. *Cancer Epidemiol. Biomarkers Prev.*, **16**,1954–1965.

7. Chadeau-Hyam, M., Athersuch, T.J., Keun, H.C., De Iorio, M., Ebbels, T.M.D., Jenab, M., Sacerdote, C., Bruce, S.J., Holmes, E., and Vineis, P. (2011) Meeting-in-the-middle using metabolic profiling - a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers*, **16**,83–88.

8. Tenenhaus, M. (1998) La régression PLS. Technip. Paris.

9. Mitra, V. and Metcalf, J. (2009) Metabolic functions of the liver. *Anaesth. Intensive Care Med.*, **10**,334–335.

10. Fages, A., Duarte-Salles, T., Stepien, M., Ferrari, P., Fedirko, V., Pontoizeau, C., Trichopoulou, A., Aleksandrova, K., Tjønneland, A., Olsen, A., Clavel-Chapelon, F., Boutron-Ruault, M.-C., Severi, G., Kaaks, R., Kuhn, T., Floegel, A., Boeing, H., Lagiou, P., Bamia, C., Trichopoulos, D., Palli, D., Pala, V., Panico, S., Tumino, R., Vineis, P., Bueno-de-Mesquita, H.B., Peeters, P.H.M., Weiderpass, E., Agudo, A., Molina-Montes, E., Huerta, J.M., Ardanaz, E., Dorronsoro, M., Sjoberg, K., Ohlsson, B., Khaw, K.-T., Wareham, N.J., Travis, R.C., Schmidt, J.A., Cross, A.J., Gunter, M.J., Riboli, E., Scalbert, A., Romieu, I., Elena-Herrmann, B., and Jenab, M. (2015) Metabolomic Profiles of Hepatocellular Carcinoma in a European Prospective Cohort. *Submitt. to Am. J. Gastroenterol.*,.

11. Riboli, E., Hunt, K.J., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondière, U.R., Hémon, B., Casagrande, C., Vignat, J., Overvad, K., Tjønneland, A., Clavel-Chapelon, F., Thiébaut, A., Wahrendorf, J., Boeing, H., Trichopoulos, D.,

Commented [NA1]: The references, format has been modified according to the Editor's request to include all authors.

Trichopoulou, A., Vineis, P., Palli, D., Bueno-De-Mesquita, H.B., Peeters, P.H.M., Lund, E., Engeset, D., González, C. a, Barricarte, A., Berglund, G., Hallmans, G., Day, N.E., Key, T.J., Kaaks, R., and Saracci, R. (2002) European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.*, **5**,1113–1124.

12. Kaaks, R., Slimani, N., and Riboli, E. (1997) Pilot Phase Studies on the Accuracy of Dietary Intake Measurements in the EPIC Project : Overall Evaluation of Results. **26**,26–36.

13. Trichopoulos, D., Bamia, C., Lagiou, P., Fedirko, V., Trepo, E., Jenab, M., Pischon, T., Nöthlings, U., Overved, K., Tjønneland, A., Outzen, M., Clavel-Chapelon, F., Kaaks, R., Lukanova, A., Boeing, H., Aleksandrova, K., Benetou, V., Zylis, D., Palli, D., Pala, V., Panico, S., Tumino, R., Sacerdote, C., Bueno-De-Mesquita, H.B., Van Kranen, H.J., Peeters, P.H.M., Lund, E., Quirós, J.R., González, C. a, Sanchez Perez, M.-J., Navarro, C., Dorronsoro, M., Barricarte, A., Lindkvist, B., Regnér, S., Werner, M., Hallmans, G., Khaw, K.-T., Wareham, N., Key, T., Romieu, I., Chuang, S.-C., Murphy, N., Boffetta, P., Trichopoulou, A., and Riboli, E. (2011) Hepatocellular carcinoma risk factors and disease burden in a European cohort: a nested case-control study. *J. Natl. Cancer Inst.*, **103**,1686–1695.

14. Beckonert, O., Keun, H.C., Ebbels, T.M.D., Bundy, J., Holmes, E., Lindon, J.C., and Nicholson, J.K. (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat. Protoc.*, **2**,2692–2703.

15. Fages, A., Ferrari, P., Monni, S., Dossus, L., Floegel, A., Mode, N., and Al., E. (2014) Investigating sources of variability in metabolomic data in the EPIC study: the Principal Component Partial R-square (PC-PR2) method. *Metabolomics*, **10**,1074–1083.

16. Fages, A., Pontoizeau, C., Jobard, E., Lévy, P., Bartosch, B., and Elena-Herrmann, B. (2013) Batch profiling calibration for robust NMR metabonomic data analysis. *Anal. Bioanal. Chem.*, **405**,8819–8827.

17. Blaise, B.J., Shintu, L., Elena, B., Emsley, L., Dumas, M.-E., and Toulhoat, P. (2009) Statistical Recoupling Prior to Significance Testing in Nuclear Resonance Based Metabonomics. *Anal. Chem.*, **81**,6242–6251.

18. Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. a, Lim, E., Sobsey, C. a, Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H.J., and Forsythe, I. (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**,D603–D610.

19. Cui, Q., Lewis, I.A., Hegeman, A.D., Anderson, M.E., Li, J., Schulte, C.F., Westler, W.M., Eghbalnia, H.R., Sussman, M.R., and Markley, J.L. (2008) Metabolite

identification via the Madison Metabolomics Consortium Database. *Nat. Biotechnol.*, **26**,162–164.

20. Slimani, N., Deharveng, G., Unwin, I., Southgate, D.A., Viignat, J., Skeie, G., Salvini, S., Parpinel, M., Moller, A., and et al. (2007) The EPIC nutrient database project (ENDB): a first attempt to standardize nutrient databases across the 10 European countries participating in the EPIC study - DTU Orbit. *Eur. J. Clin. Nutr.*, **61**,1037–1056.

21. Kleinbaum, D.G., Kupper, L.K., and Muller, K.E. (1987) Applied regression analysis and other multivariable methods. Belmont, CA: Duxbury Press.

22. Wold, S., Sjostrom, M., and Ericksson, L. (2001) PLS-regression : a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, **58**,109–130.

23. Rannar, S., Geladi, P., Lindgren, F., and Wold, S. (1995) A PLS kernel algorithm for data sets with many variables and few objects. Part II: cross-validation, missing data and examples. *J. Chemom.*, **9**,459–470.

24. Bastien, P. (2008) Régression PLS et Données Censurées. Conservatoire National des Arts et Métiers - CNAM.

25. Fedirko, V., Trichopolou, A., Bamia, C., Duarte-Salles, T., Trepo, E., Aleksandrova, K., Nöthlings, U., Lukanova, A., Lagiou, P., Boffetta, P., Trichopoulos, D., Katzke, V. a, Overvad, K., Tjønneland, A., Hansen, L., Boutron-Ruault, M.C., Fagherazzi, G., Bastide, N., Panico, S., Grioni, S., Vineis, P., Palli, D., Tumino, R., Bueno-de-Mesquita, H.B., Peeters, P.H., Skeie, G., Engeset, D., Parr, C.L., Jakszyn, P., Sánchez, M.J., Barricarte, A., Amiano, P., Chirlaque, M., Quirós, J.R., Sund, M., Werner, M., Sonestedt, E., Ericson, U., Key, T.J., Khaw, K.T., Ferrari, P., Romieu, I., Riboli, E., and Jenab, M. (2013) Consumption of fish and meats and risk of hepatocellular carcinoma: the European Prospective Investigation into Cancer and Nutrition (EPIC). *Ann. Oncol.*, **24**,2166–2173.

26. Akuta, N., Suzuki, F., Kobayashi, M., Hara, T., Sezaki, H., Suzuki, Y., Hosaka, T., Kobayashi, M., Saitoh, S., Ikeda, K., and Kumada, H. (2014) Correlation Between Hepatitis B Virus Surface antigen Level and Alpha-Fetoprotein in Patients Free of Hepatocellular Carcinoma or Severe Hepatitis. *J. Med. Virol.*, **86**,131–138.

27. Kanazir, M., Boricic, I., Delic, D., Tepavcevic, D.K., Knezevic, A., Jovanovic, T., and Pekmezovic, T. (2010) Risk factors for hepatocellular carcinoma: A case-control study in Belgrade (Serbia). *Tumori*, **96**,911–917.

28. Zheng, Z., Zhou, L., Gao, S., Yang, Z., Yao, J., and Zheng, S. (2013) Prognostic role of C-reactive protein in hepatocellular carcinoma: A systematic review and meta-analysis. *Int. J. Med. Sci.*, **10**,653–664.

29. Metz, C.D. (1978) Basic Principles of ROC Analysis. *Semin. Nucl. Med.*, **8**,283–298.

30. Vermont, J., Bosson, J.L., François, P., Robert, C., Rueff, A., and Demongeot, J. (1991) Strategies for graphical threshold determination. *Comput. Methods Programs Biomed.*, **35**,141–150.

31. Biesheuvel, C.J., Vergouwe, Y., Oudega, R., Hoes, A.W., Grobbee, D.E., and Moons, K.G.M. (2008) Advantages of the nested case-control design in diagnostic research. *BMC Med. Res. Methodol.*, **8**,48.

32. Van Zaane, B., Vergouwe, Y., Donders, a R.T., and Moons, K.G.M. (2012) Comparison of approaches to estimate confidence intervals of post-test probabilities of diagnostic test results in a nested case-control study. *BMC Med. Res. Methodol.*, **12**,166.

33. Stepien, M., Duarte-Salles, T., Fedirko, V., Floegel, A., Kumar-Barupal, D., Rinaldi, S., Achaintre, D., Assi, N., Tjønneland, A., Overvad, K., Bastide, N., Boutron-Ruault, M.-C., Severi, G., Kuhn, T., Kaaks, R., Aleksandrova, K., Boeing, H., Trichopoulou, A., Bamia, C., Lagiou, P., Saieva, C., Agnoli, C., Panico, S., Tumino, R., Naccarati, A., Bueno-de-Mesquita, H.B., Peeters, P.H., Weiderpass, E., Quirós, J.R., Agudo, A., Sanchez, M.-J., Dorronsoro, M., Gavrila, D., Barricarte, A., Ohlsson, B., Sjoberg, K., Werner, M., Sund, M., Wareham, N., Khaw, K.-T., Travis, R.C., Schmidt, J.A., Gunter, M., Cross, A.J., Vineis, P., Romieu, I., Scalbert, A., and Jenab, M. (2015) Alteration of Amino Acid and Biogenic Amine Metabolism in Hepatobiliary Cancers: Findings from a Prospective Cohort Study. *Submitt. to Int. J. Cancer*,.

34. Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B., and Wei, L.J. (2011) On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Stat. Med.*, **30**,1105–1117.

35. Valeri, L. and Vanderweele, T.J. (2013) Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol. Methods*, **18**,137–150.

36. Vanderweele, T.J. and Vansteelandt, S. (2010) Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol.*, **172**,1339–1348.

37. R Foundation for Statistical Computing and R Core Team. (2013) R: A language and environment for statistical computing.

38. SAS Institute Inc. and Cary, N. (2012) Base SAS® 9.4 Procedures Guide.

39. College Station TX : StataCorp. (2011) Stata Statistical Software: Release 12.

40. Abdi, H. (2010) Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdiscip. Rev. Comput. Stat.*, **2**,97–106.

41. Yang, W.-S., Va, P., Bray, F., Gao, S., Gao, J., Li, H.-L., and Xiang, Y.-B. (2011) The role of pre-existing diabetes mellitus on hepatocellular carcinoma occurrence and prognosis: a meta-analysis of prospective cohort studies. *PLoS One*, **6**,e27326.

42. Gomaa, A.-I. (2008) Hepatocellular carcinoma: Epidemiology, risk factors and pathogenesis. *World J. Gastroenterol.*, **14**,4300–4308.

43. Cogliano, V.J., Baan, R., Straif, K., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Freeman, C., Galichet, L., and Wild, C.P. (2011) Preventable exposures associated with human cancers. *J. Natl. Cancer Inst.*, **103**,1827–1839.

44. Bialecki, E.S. and Di Bisceglie, A.M. (2005) Diagnosis of hepatocellular carcinoma. *HPB (Oxford).*, **7**,26–34.

45. Wentzensen, N. and Wacholder, S. (2013) From differences in means between cases and controls to risk stratification: a business plan for biomarker development. *Cancer Discov.*, **3**,148–157.

46. Leong, T.Y.-M. and Leong, A.S.-Y. (2005) Epidemiology and carcinogenesis of hepatocellular carcinoma. *HPB (Oxford).*, **7**,5–15.

47. Rothwell, J., Fillâtre, Y., Martin, J.-F., Lyan, B., Pujos-Guillot, E., Fezeu, L., Hercberg, S., Comte, B., Galan, P., Touvier, M., and Manach, C. (2014) New biomarkers of coffee consumption identified by the non-targeted metabolomic profiling of cohort study subjects. *PLoS One*, **9**,e93474.

48. Guo, M., Zhao, B., Liu, H., Zhang, L., Peng, L., Qin, L., Zhang, Z., Li, J., Cai, C., and Gao, X. (2014) A Metabolomic Strategy to Screen the Prototype Components and Metabolites of Shuang-Huang-Lian Injection in Human Serum by Ultra Performance Liquid Chromatography Coupled with Quadrupole Time-of-Flight Mass Spectrometry. *J. Anal. Methods Chem.*, **2014**,241505.

49. Anderson, T.W. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Stat.*, **22**,327–351.

50. Izenman, A.J. (1975) Reduced-Rank Regression for the Multivariate Linear Model. *J. Multivar. Anal.*, **5**,248–264.

51. Aldrin, M. (2002) Reduced-rank regression. In: El-Shaarawi AH, Piegorsch WW, editors. Encyclopedia of Environmetrics. Chichester: John Wiley & Sons, Ltd. pp. 1724–1728.

52. Hotelling, H. (1936) Relations between two sets of variates. *Biometrika*, **28**,321–377.

53. Navratil, V., Pontoizeau, C., Billoir, E., and Blaise, B.J. (2013) SRV : an opensource toolbox to accelerate the recovery of metabolic biomarkers and correlations from metabolic phenotyping data sets. *Bioinformatics*, **29**,1348–1349.

54. Wishart, D.S. (2008) Quantitative metabolomics using NMR. *TrAC Trends Anal. Chem.*, **27**,228–237.

55. Psychogios, N., Hau, D.D., Peng, J., Guo, A.C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., Gautam, B., Young, N., Xia, J., Knox, C., Dong, E.,

754 Huang, P., Hollander, Z., Pedersen, T.L., Smith, S.R., Bamforth, F., Greiner, R.,
755 McManus, B., Newman, J.W., Goodfriend, T., and Wishart, D.S. (2011) The human
756 serum metabolome. *PLoS One*, **6**,e16957.

757

758 **Legends to figures**

759 **Figure 1**: General scheme of the analytical framework developed in the study. A PC-PR2
760 analysis is carried out beforehand to identify relevant sources of variation. In the PLS model
761 the X- and Y- sets are related to each other, and scores are computed (1). X- and Y-scores are,
762 in turn, associated to a case-control indicator of HCC status in conditional logistic regression
763 models (2). A mediation analysis is carried out to explore the role of metabolomics in the
764 association between lifestyle factors and risk of HCC (3).

765 **Figure 2**: PC-PR2 analysis results* identifying the sources of variability in the NMR data
766 (panel A) and in the lifestyle data (panel B).

767 * 17 and 14 components were retained to account for 80 % (threshold used) of total NMR (A)
768 and lifestyle variability (B), respectively. The R2 value represents the amount of variability in
769 NMR / lifestyle variable explained by the ensemble of investigated predictors.

770 **Figure 3**: Variable importance plot (VIP) displaying the variable importance for projection
771 statistic of the predictor variables for the PLS analyses.

772 Panel **A**: Results from the main PLS model run on all observations (N=336, X-set=21, Y-
773 set=285).

774 Panel **B**: Results from the PLS sensitivity analysis run on a subsample (N=271, 92 cases, 179
775 controls) excluding sets where cases were diagnosed within the first two years of follow-up
776 (X-set=21, Y-set=285).

777 The horizontal line corresponds to Wold's criterion (0.8), the threshold used to rule if a
778 variable has an important contribution to the construction of the Y variables (see
779 **Mathematical Appendix** for further details).

Table 1: Individual and cumulative variation (%) explained by the first 3 PLS factors in 21 lifestyle (X-set) and 285 NMR (Y-set) variables.

# of PLS	Lifestyle Variables		NMR Variables	
Factors	**Individual**	**Cumulative**	**Individual**	**Cumulative**
1	6.17	-	5.51	-
2	6.23	12.40	2.38	7.89
3	9.27	21.67	0.59	8.48

Table 2: Lifestyle and NMR cluster variables contributing to each of the 3 PLS factors (N=336, X-set=21, Y-set=285).

PLS Factor	Lifestyle Variable*	Loading value	CS*‡ (ppm)	Metabolite**	Loading value
1	Dairy Products	0.28	5.22		-0.06
	Cakes and Biscuits	0.32	3.88		-0.05
	Lifetime Alcohol Consumption	-0.25	3.82		-0.06
	Smoking Status	-0.39	3.76	Glucose	-0.06
	Diabetes	-0.63	3.71		-0.05
			3.54		-0.05
			3.50		-0.07
			3.48		-0.07
			3.44	Acetoacetate	-0.07
			3.23	Choline + Glycerphosphocholine	-0.04
			3.01	Lysine	0.10
			2.94	Albumin	0.10
			2.65	Aspartate	0.10
			2.42	Glutamine	0.10
			2.28	Acetoacetate	0.10
			2.22	CH_2-C**H**$_2$-COOC bond of lipids + Acetone	-0.04
			1.86	Lysine	0.09
			1.87		0.10
			1.53	C**H**$_2$-CH_2-COOC bond of lipids	-0.03
2	Cereal and Cereal Products	-0.16	7.17	Tyrosine	0.13
	Height	-0.34	6.87		0.13
	Education Level	-0.26	5.27	C**H**=C**H** bond of lipids	-0.13
	Hepatitis	0.49	5.22	Glucose	0.16
			5.18	Mannose + Lipid O-CH2	-0.12
			4.27	Lipid O-CH2	-0.12
			4.25	Threonine	-0.14
			4.07	Choline + Lipid O-CH2 + Myo-inositol	-0.12
			4.05	Creatinine	-0.14
			3.88		0.15
			3.82		0.16
			3.76		0.15
			3.71	Glucose	0.15
			3.54		0.15
			3.50		0.16
			3.48		0.16
			3.44	Acetoacetate	0.16
			3.23	Choline + Glycerphosphocholine	0.15
			2.80	Aspartate	-0.12
			2.22	CH_2-C**H**$_2$-COOC bond of lipids + Acetone	-0.11
			2.19	CH_2-C**H**$_2$-COOC bond of lipids	-0.15
			2.02	Proline + Glutamate + CH2=C bonds of lipids	-0.13
			1.53	C**H**$_2$-CH_2-COOC bond of lipids	-0.13
			1.25	C**H**$_2$ bond of lipids	-0.12
			0.86	Cholesterol + CH3 bond of lipids	-0.12
3	Vegetables	-0.42	7.32	Phenylalanine	0.11
	Lifetime Alcohol Consumption	0.29	5.22	Glucose	-0.13
	Smoking Status	0.25	4.28	Lipid O-C**H**$_2$	0.11
	Hepatitis	0.26	3.88		-0.11
			3.82		-0.11
			3.76	Glucose	-0.12
			3.71		-0.11
			3.69		-0.11
			3.63	Myo-inositol	0.16
			3.50	Glucose	-0.13
			3.48		-0.12
			3.44	Acetoacetate	-0.12
			3.35	Proline	0.11
			3.33		0.13
			3.28	Myo-inositol	0.12
			3.23	Choline + Glycerphosphocholine	-0.12
			2.80	Aspartate	-0.13
			2.76	part of =CH-CH2-CH= bond of lipids	-0.13
			2.35	Proline + Glutamate	0.12
			2.33		0.13

		1.20	3-hydroxybutyrate + CH2 bond of lipids	0.11
		1.16	Ethanol	0.15
		0.66	Cholesterol	0.11

*****Relevant lifestyle and NMR variables contributing to each PLS factor selected based on their associated loading values <10th percentile (pctl) and >90th pctl or <5th pctl and >95th pctl respectively.
‡ CS: ^{1}H chemical shift (in ppm) of the cluster (center value).
******Some of the identified clusters were found to be background noise during the annotation phase and were removed from this table.

Table 3: HCC odds ratios* and 95% confidence interval (OR, 95% CI) associated with the lifestyle (X-set) and the NMR clusters (Y-set) PLS scores in the main analysis (N=336, X-set=21, Y-set=285).

PLS Lifestyle Variables X-scores			PLS NMR Variables Y-scores		
Factor	**OR** ** **(95% CI)**	**P-Wald†**	**Factor**	**OR** ** **(95% CI)**	**P-Wald†**
1	0.77 (0.58, 1.02)	0.07	1	0.96 (0.91, 1.01)	0.09
2	1.54 (1.06, 2.25)	0.02	2	1.11 (1.02, 1.22)	0.02
3	1.37 (1.05, 1.79)	0.02	3	1.22 (1.04, 1.44)	0.01

*Models were adjusted for C-reactive protein concentration, alpha-fetoprotein concentration and a composite score for liver damage. Cases and controls were matched on age at blood collection (± 1 year), sex, study centre, date (± 2 months) and time of the day at blood collection (± 3 hours), fasting status at blood collection (<3/3-6/>6 hours); among women, additional matching criteria included menopausal status (pre-/peri-/postmenopausal) and hormone replacement therapy use at time of blood collection (yes/no). ** ORs expressing the change in HCC risk associated to 1-SD increase in the score. † Wald's test was for continuous exposure compared with a Chi-square distribution with 1 degree of freedom (dof).

Table 4: Area under the curve (AUC), sensitivity, specificity, accuracy and positive predictive value (PPV) of ROC models (with 95% CI), from the main PLS analysis (N=336, X-set=21, Y-set=285).

	AUC	AUC_b**	Sensitivity	Specificity	Accuracy	PPV
Adjustment Covariates (ADJ)*	0.842 (0.794, 0.891)	0.821 (0.766, 0.868)	0.752 (0.662, 0.829)	0.802 (0.743, 0.852)	0.785	0.0015
X1 scores + ADJ	0.846 (0.797, 0.894)	0.825 (0.766, 0.875)	0.743 (0.653, 0.821)	0.838 (0.783, 0.884)	0.806	0.0018
X1+X2 scores + ADJ	0.854 (0.808, 0.900)	0.831 (0.772, 0.881)	0.743 (0.653, 0.821)	0.824 (0.768, 0.872)	0.797	0.0017
X1+X2+X3 scores + ADJ	0.859 (0.811, 0.907)	0.836 (0.778, 0.887)	0.796 (0.710, 0.866)	0.788 (0.729, 0.840)	0.791	0.0015
Y1 scores + ADJ	0.841 (0.793, 0.890)	0.817 (0.760, 0.865)	0.735 (0.643, 0.813)	0.820 (0.763, 0.868)	0.791	0.0016
Y1+Y2 scores + ADJ	0.845 (0.795, 0.894)	0.820 (0.762, 0.872)	0.735 (0.643, 0.813)	0.851 (0.798, 0.895)	0.812	0.0020
Y1+Y2+Y3 scores + ADJ	0.853 (0.804, 0.902)	0.827 (0.771, 0.877)	0.726 (0.634, 0.805)	0.883 (0.833, 0.922)	0.890	0.0025

*The model is run on the adjustment covariates (ADJ) including the C-reactive protein concentration, alpha-fetoprotein concentration and a composite score for liver damage. ** AUC_b is the bootstrapped-cross validated estimate of the AUC. X1, X2 and X3 are the lifestyle component scores of the first, second and third PLS factors, respectively. Y1, Y2, and Y3 are the metabolomics component of the first, second and third PLS factors, respectively.

Table 5: HCC odds ratios* and 95% confidence intervals (OR, 95%CI) associated with the lifestyle (X-set) and the NMR clusters (Y-set) PLS scores. Results from the sensitivity analysis (N=271, 92 cases, 179 controls) conducted excluding sets where cases were diagnosed within the first two years of follow-up (X-set=21, Y-set=285).

PLS Lifestyle Variables X-scores			PLS NMR Variables Y-scores		
Factor	**OR** ** **(95% CI)**	**P-Wald†**	**Factor**	**OR** ** **(95% CI)**	**P-Wald†**
1	0.80 (0.60, 1.08)	0.15	1	0.96 (0.94, 1.04)	0.56
2	1.56 (1.02, 2.40)	0.04	2	1.18 (1.03, 1.36)	0.02
3	0.86 (0.67, 1.11)	0.26	3	0.86 (0.73, 0.99)	<0.05

*Models were adjusted for C-reactive protein concentration, alpha-fetoprotein concentration and a composite score for liver damage. Cases and controls were matched on age at blood collection (± 1 year), sex, study centre, date (± 2 months) and time of the day at blood collection (± 3 hours), fasting status at blood collection (<3/3-6/>6 hours); among women, additional matching criteria included menopausal status (pre-/peri-/postmenopausal) and hormone replacement therapy use at time of blood collection (yes/no). ** ORs expressing the change in HCC risk associated to 1-SD increase in the score. † Wald's test was for continuous exposure compared with a Chi-square distribution with 1 degree of freedom (dof).

Table 6: Results from the mediation analysis (N= 336, X-set=21, Y-set=285): Natural Indirect Effect (NIE) and 95%CI*.

Model**				Natural Indirect Effect (NIE)	
Exposure (A)	Mediator (M)	Outcome	A*M interaction term	Estimate (95%CI)	p-value
X1 score	Y1 score	HCC	No	0.91 (0.77, 1.06)	0.23
X2 score	Y2 score	HCC	No	1.11 (0.97, 1.25)	0.12
X3 score	Y3 score	HCC	No	1.08 (0.94, 1.23)	0.28
X1 score	Y1 score	HCC	Yes	0.96 (0.79, 1.17)	0.70
X2 score	Y2 score	HCC	Yes	1.15 (1.01, 1.31)	0.04
X3 score	Y3 score	HCC	Yes	1.13 (1.01, 1.28)	0.04

* The standard errors used to compute the 95%CI were obtained using the delta method.
**Models were adjusted for the C-reactive protein concentration, alpha-fetoprotein concentration and a composite score for liver damage, as well as for the other Y-scores, as potential mediator-outcome confounders. Additionally, the outcome and the mediator models were adjusted for centre and age at blood collection.

Mathematical Appendix

A statistical framework to model the meeting-in-the-middle principle using metabolomic data: application to hepatocellular carcinoma in the EPIC study.

---

1 PLS regression

1.1 Introduction

PLS (partial least squares) regression is a widely used method in multivariate statistics to relate two sets of variables while reducing their dimensionality. It was first developed as a method to predict a set of variables Y from another set X; and also to depict their common structure. The main aim of PLS is to regress a set Y of $\mathbf{q}$ variables $(y_1, y_2, \ldots, y_q)$ of interest, which are called responses, on a set X of $\mathbf{p}$ predictor variables $(x_1, x_2, \ldots, x_p)$ that may display high levels of correlation. PLS combines and generalizes features of principal component analysis (PCA) and multiple linear regression (MLR); and results in a set of PLS latent factors as linear combinations of variables, in turn, in the X- and Y-sets. By simultaneously decomposing X and Y, PLS finds components that explain as much as possible of the inter-relations of X and Y. The latent factors obtained from the decomposition can be used to predict Y. The following details of the algorithm are adapted from Michel Tenenhaus' book *La régression PLS, Théorie et Pratique* [1].

1.2 The PLS algorithm

Two different, but closely related, techniques exist under the name of PLS regression. The canonical or symmetric PLS regression assumes that the X- and Y- sets play a symmetrical role. The version presented here is the regression mode where latent variables are computed from a succession of singular value decompositions (SVD) followed by deflation of both the X- and Y- matrices. These sets are assumed to play the asymmetric roles of predictors and responses, respectively. Next, we briefly describe the landmark algorithm NIPALS Nonlinear estimation by Iterative Partial Least Squares. As a first step, two substitute matrices X_0 and Y_0 are initialized with $X_0 = X_{(n\times p)}$ and $Y_0 = Y_{(n\times q)}$, where variables were standardized to have means and standard deviations equal to zero and one, respectively. For $h = 1, \ldots, H$, where $H = min(p, q)$, the PLS factors are obtained iteratively. PLS regression focuses on finding two sets of weights, $w_{h(p\times 1)}$ and $c_{h(q\times 1)}$, in order to create respectively a linear combination of the columns of X and Y, known as the PLS factors, such that these two linear combinations have maximum covariance and are unique. These weights define a first pair of vectors, called the X- and Y-scores, $t_h = Xw_h$ and $u_h = Yc_h$ where we have $t_h^\intercal u_h$ maximal. PLS can be written as the following optimisation problem where maximum covariance is sought between $t_{h(1\times n)}$ and $u_{h(1\times n)}$ for each $h = 1 \cdots H$:

$$\text{Max } cov(Xw_h, Yc_h) \tag{1}$$

under the following normality constraints

$$\|w_h\| = 1 \tag{2}$$

$$\|c_h\| = 1 \tag{3}$$

and the following orthogonality constraint

$$t_h^\intercal(t_1, \ldots, t_{h-1}) = 0 \tag{4}$$

By construction we also have the following property:

$$u_h^\intercal(t_1, \ldots, t_{h-1}) = 0 \tag{5}$$

The first pair of X- and Y- scores can equivalently be obtained via a singular value decomposition. Indeed, the SVD of the cross-product matrix $X_{h-1}^\intercal Y_{h-1}$ leads to the identification of the first left and right singular vectors and of the weights w_h and c_h. The scores t_h and u_h are obtained as follows:

$$t_h = X_{h-1}w_h \tag{6}$$

$$u_h = Y_{h-1}c_h \tag{7}$$

The vector t_h is then normalized (a scaling of u_h is optional). Regressing the predictor and response matrices on the t_h vector yields the corresponding loadings.

$$p_h = X_{h-1}^{\intercal} t_h \tag{8}$$

$$c_h = Y_{h-1}^{\intercal} t_h \tag{9}$$

Next is the deflation step, where information based on the extracted latent factor h is subtracted from the current data matrices.

$$X_h = X_{h-1} - t_h p_h^{\intercal} \tag{10}$$

$$Y_h = Y_{h-1} - t_h c_h^{\intercal} \tag{11}$$

The described steps of the algorithm are iterated until one of the following criteria is met:

- If H is specified, and the algorithm stops when the H-th PLS factor is extracted and its associated statistics computed.
- If H is not specified, the algorithm stops when X_H becomes a null matrix. In this case however, H cannot exceed $min(p, q)$.

Algorithm 1 PLS1 classic algorithm steps - When Y is univariate.

$X_0 \leftarrow X$; $y_0 \leftarrow y$

for $(h = 1; h \leq H; h++)$ **do**

  $w_h = {X_{h-1}^{\intercal} y_{h-1}}/{y_{h-1}^{\intercal} y_{h-1}}$

  $w_h = {w_h}/{\sqrt{w_h^{\intercal} w_h}}$

  $t_h = {X_{h-1} w_h}/{w_h^{\intercal} w_h}$

  $p_h = {X_{h-1}^{\intercal} t_h}/{t_h^{\intercal} t_h}$

  $X_h = X_{h-1} - t_h p_h^{\intercal}$

  $c_h = {y_{h-1}^{\intercal} t_h}/{t_h^{\intercal} t_h}$

  $u_h = {y_{h-1}}/{c_h}$

  $y_h = y_{h-1} - c_h t_h$

When Y is univariate, the PLS algorithm carried out is PLS1 (See Algorithm 1, following the notation of M. Tenenhaus [1]). PLS2 (Algorithm 2) is used when Y is multivariate. When there are missing data in either the X- or Y- sets, the coordinates of the vectors w_h, t_h, c_h, u_h, and p_h are computed as slopes of the least squares straight line that passes through the origin, using the available data as follows:

Algorithm 2 PLS2 classic algorithm steps - When Y is multivariate.

1: $X_0 \leftarrow X$; $Y_0 \leftarrow Y$
2: **for** $(h = 1; h \leq H; h++)$ **do**
3: $u_h = Y_{h-1}[, 1]$ i.e. the first column of the matrix
4: **while** w_h has not converged **do**
5: $w_h = {}^{X_{h-1}^\intercal u_h}/_{u_h^\intercal u_h}$
6: $w_h = {}^{w_h}/_{\sqrt{w_h^\intercal w_h}}$
7: $t_h = {}^{X_{h-1} w_h}/_{w_h^\intercal w_h}$
8: $c_h = {}^{Y_{h-1}^\intercal t_h}/_{t_h^\intercal t_h}$
9: $u_h = {}^{Y_{h-1} c_h}/_{c_h^\intercal c_h}$
10: $p_h = {}^{X_{h-1}^\intercal t_h}/_{t_h^\intercal t_h}$
11: $X_h = X_{h-1} - t_h p_h^\intercal$
12: $Y_h = Y_{h-1} - t_h c_h^\intercal$

- $w_h = (w_{h1}, \ldots, w_{hp})^\intercal$, is a normalized vector, where w_{hj} is the slope of the least squares line passing through the origin of the plane defined by $(u_h, X_{h-1,j})$. $X_{h-1,j}$ is the j-th X variable of the $h-1$ PLS factor.

- $t_h = (t_{h1}, \ldots, t_{hn})^\intercal$, where t_{hi} is the slope of the least squares line passing through the origin of the plane defined by $(w_h, x_{h-1,i})$. $x_{h-1,i}$ is the i-th x observation of the $h-1$ PLS factor.

- $c_h = (c_{h1}, \ldots, c_{hq})^\intercal$, where c_{hk} is the slope of the least squares line passing through the origin of the plane defined by $(t_h, Y_{h-1,k})$. $Y_{h-1,k}$ is the k-th Y variable of the $h-1$ PLS factor.

- $u_h = (u_{h1}, \ldots, u_{hn})^\intercal$, where u_{hi} is the slope of the least squares line passing through the origin of the plane defined by $(c_h, y_{h-1,i})$. $y_{h-1,i}$ is the i-th y observation of the $h-1$ PLS factor.

- $p_h = (p_{h1}, \ldots, p_{hp})^\intercal$, where p_{hj} is the slope of the least squares line passing through the origin of the plane defined by $(t_h, X_{h-1,j})$. $X_{h-1,j}$ is the j-th X variable of the $h-1$ PLS factor.

1.3 Tools for interpretation

1.3.1 Choice of number of components

The number of PLS latent factors or components to be retained can be decided based on a cross-validation.

For each model with a number h of extracted factors, this is done by running the PLS analysis on only a part of the data called the training set, and then evaluating how well the model fits observations in the test set. This includes the part of the data not involved in the PLS modelling of the training set.

The dataset comprised of n observations is split into z approximately equal sets of observations. The training set consists of the data in the first $z-1$ folds and the remaining fold is used as test set. Predicted values for the Y-set are computed on this test set along with the sum of the squared error of prediction. This process is repeated z times so that each fold can in turn serve as a test set. In practice, for each number of possible latent factors $h = 1, \ldots, H$, we compute the prediction of y_i by the PLS model with results obtained on the training set with a number h of components applied to observations in the test set in order to yield $\hat{y}_{h(-i)}$. The Prediction Error Sum of Squares (PRESS) is the resulting sum of all squared errors of prediction statistic computed across all test sets as defined in the following equation:

$$PRESS_h = \sum (y_i - \hat{y}_{h(-i)})^2 \tag{12}$$

The Residual Sum of Squares (RSS) is computed in a standard way:

$$RSS_h = \sum (y_i - \hat{y}_{hi})^2 \tag{13}$$

Different criteria can be used to determine the number of components h to retain. One such criterion, Q_h^2 was first introduced by H. Wold [2] and is mainly used in the software SIMCA-P. It is based on the following statistic:

$$Q_h^2 = 1 - \frac{PRESS_h}{RSS_{h-1}} \tag{14}$$

As pointed out by M. Tenenhaus, the initial value for RSS when y is univariate centred-scaled and $h = 0$ is:

$$RSS_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = n - 1 \tag{15}$$

In the software SIMCA-P the PLS component is kept when the following condition is met:

$$\sqrt{PRESS_h} \leq 0.95\sqrt{RSS_{h-1}} \tag{16}$$

$$\iff Q_h^2 \geqslant 0.0975 \tag{17}$$

The default threshold 0.0975 is equal to $1 - 0.95^2$. In SAS, the criteria to select the number h of components to be retained is by minimizing the $PRESS_h$ statistic.
The above described formulae can be generalized for multivariate Y, thus we have for any given variable y_k, $k = 1, \ldots, q$:

$$Q^2_{kh} = 1 - \frac{PRESS_{kh}}{RSS_{k(h-1)}} \tag{18}$$

$$Q^2_h = 1 - \frac{\sum_{k=1}^{q} PRESS_{kh}}{\sum_{k=1}^{q} RSS_{k(h-1)}} \tag{19}$$

The criteria for keeping a PLS factor are identical to what was established for the univariate case. One can alternately use one of the following rules, where the equivalence defined in formula (17) still holds true:

- $Q^2_h \geqslant 0.0975$
- At least one value of $Q^2_{hk} \geqslant 0.0975$

If the criteria are met by several values of h, the one retained is the smallest h, to achieve a better dimensionality reduction.
The Q^2 and $PRESS$ criteria are relatively robust to the choice of number of folds (blocks) used for cross-validation. A number of folds between 5 and 10 is recommended (Tenenhaus 1998, p.238) [1]. The default choice in the SIMCA-P and SAS softwares is 7, and is the parameter used in this study.

1.3.2 Variable Importance in the Projection (VIP)

The Variable Importance in the Projection (VIP) is a measure of the explanatory power of a given variable x_j over Y. The VIP_{hj} of a given component h of the j-th variable x_j is defined as:

$$VIP_{hj} = \sqrt{\frac{p}{Rd(Y; t_1, \ldots, t_h)} \sum_{l=1}^{h} Rd(Y, t_l) w^2_{lj}} \tag{20}$$

and one has:

$$\sum_{j=1}^{p} VIP^2_{hj} = p \tag{21}$$

where $Rd(Y; t_1, \ldots, t_h)$ is the redundancy of Y with respect to the t scores $(t_1, \ldots, t_h)$. It describes the amount of variance of Y explained by the component t_h of the X-set. It is defined

as follows:

$$Rd(Y, t_h) = \frac{1}{q} \sum_{k=1}^{q} cor^2(y_k, t_h) \tag{22}$$

It can be equivalently computed as:

$$Rd(Y, t_h) = r_h^2 \frac{1}{q} \sum_{k=1}^{q} cor^2(y_k, u_h) \tag{23}$$

where $r_h = cor(Xw_h, Yc_h)$ is called a canonical correlation and r_h^2 is the h^{th} largest eigenvalue of the crossproduct matrix decomposition.

The contribution of a variable x_j to the construction of a component t_l is measured by the weight w_{lj}^2. For each l, with $l = 1, \ldots, h$, the sum of these weights across the p variables x_j equals 1. To measure the contribution of the variable x_j to the construction of Y through the components t_l, one should consider the explanatory power of the component t_l, measured by the redundancy $Rd(Y; t_l)$. An equal weight w_{lj}^2 indicates an explanatory power of the x_j variable over the Y-set whose importance increases with the level of redundancy $Rd(Y; t_l)$.

The VIP enables the ranking of the predictors x_j according to their explanatory power on Y, and summarizes their contribution to the model. A VIP is considered small if its value is less than 0.8 and high when its value is greater than 1. Variables with a high VIP ($VIP > 1$) are the most important for the reconstruction and prediction of Y.

2 Statistical Recoupling of Variables (SRV)

The SRV procedure was introduced by *Blaise et al.(2009)* [3] and for which a matlab toolbox was later implemented [4]. The SRV is an "intelligent bucketing" algorithm that aims at regrouping variables (typically the smallest unit of the NMR spectrum) in clusters corresponding to a wider biological and chemical entity.

SRV exploits the spectral structure of data, without forming any metabolic hypothesis to reduce the dimensionality of spectra. A typical NMR ^1H 9 ppm spectrum is often partitioned into $9,000$ buckets of 0.001 ppm width. The main idea of the algorithm is to exploit the spectral dependency landscape L which is the covariance to correlation ratio between two neighbouring variables along the chemical shift axis to assemble them within a cluster. If one considers a matrix Z of serum spectra acquired by NMR with n observations and r columns $(z_1, \ldots, z_r)$ corresponding to neighbouring bins of NMR signals. The first bin-variable starts the first

cluster, then L is computed for each z_i as follows with $i = 1, \ldots, r$:

$$L(z_i) = \frac{cov(z_i, z_{i+1})}{cor(z_i, z_{i+1})} = sd(z_i) * sd(z_{i+1}) \tag{24}$$

where sd is the standard deviation.

The variable then joins a cluster according to the following rules:

- $L(z_i)$ values are used to locate local minima i.e. borders between clusters.
- If $L(z_{i-1}) > L(z_i)$ then z_{i-1} and z_i are associated in the same cluster, otherwise z_i and z_{i+1} start a new cluster.
- The minimum number of variables belonging to a cluster is set a priori as it is based on the resolution of the NMR spectra. When acquired at 700 MHz, the typical peak base width of a well-resolved singlet is equal to 7 Hz. Therefore, the threshold was set to 10 in our analysis, meaning that if a cluster has less than 10 variables, it is discarded.
- The super-cluster intensity is computed as the mean of the intensities of the signal in the bins assigned to the super-cluster.
- If two neighbouring clusters have a correlation > 0.9, they are aggregated to form a super-cluster. In these analyses, the association is limited to 3 clusters per super-cluster (this value is empirical and was discussed in the original paper [3]).

References

[1] Michel Tenenhaus. *La régression PLS: Théorie et Pratique*. Paris, 1998.

[2] Herman Wold. *Multivariate Analysis*, chapter Estimation of principal components and related models by iterative least squares, pages 391–420. Academic Press, New York, 1966.

[3] Benjamin J Blaise, Laetitia Shintu, Bénédicte Elena, Lyndon Emsley, Marc-Emmanuel Dumas, and Pierre Toulhoat. Statistical recoupling prior to significance testing in nuclear magnetic resonance based metabonomics. *Analytical Chemistry*, 81(15):6242–6251, August 2009.

[4] Vincent Navratil, Clément Pontoizeau, Elise Billoir, and Benjamin J Blaise. Srv: an open-source toolbox to accelerate the recovery of metabolic biomarkers and correlations from metabolic phenotyping data sets. *Bioinformatics*, 29(10):1348–1349, May 2013.

A Statistical framework to model the meeting-in-the-middle principle using metabolomic data: application to hepatocellular carcinoma in the EPIC study.

Nada Assi[1], Anne Fages[2,a], Paolo Vineis[3], Marc Chadeau-Hyam[3],Magdalena Stepien[1], Talita Duarte-Salles[1], Graham Byrnes[1] ,Houda Boumaza[2], Sven Knüppel[4], Tilman Kühn[5], Domenico Palli[6], Christina Bamia[7], Hendriek Boshuizen[8], Catalina Bonet[9], Kim Overvad[10], Mattias Johansson[1,11], Ruth Travis[12], Marc J Gunter[3], Eiliv Lund[13], Laure Dossus[14-15], Bénédicte Elena-Herrmann[2], Elio Riboli[3], Mazda Jenab[1], Vivian Viallon[16-18†], Pietro Ferrari[1†]*

[1] International Agency for Research in Cancer (IARC-WHO), 150 Cours Albert Thomas, 69372 Lyon CEDEX 08, France

[2] Centre de RMN à Très Hauts Champs, Institut des Sciences Analytiques (CNRS/ENS Lyon/UCB Lyon 1), Université de Lyon, 69100 Villeurbanne, France

[3] Department of Epidemiology and Biostatistics, MRC-HPA Centre for Environment and Health, School of Public Health, Imperial College London, Norfolk Place, London, W2 1PG, United Kingdom

[4] Department of Epidemiology, German Institute of Human Nutrition, Potsdam-Rehbruecke, 14558 Nuthetal, Germany

[5] Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

[6] Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Institute – ISPO, Florence- Italy

[7] WHO Collaborating Center for Food and Nutrition Policies, Department of Hygiene, Epidemiology and Medical Statistics, University of Athens Medical School, Athens, Greece

[8] National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands

[9] Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Institut Català d'Oncologia, L'Hospitalet de Llobregat, Spain

[10] The Department of Epidemiology, School of Public Health, Aarhus University, Aarhus, Denmark

[11] The Department for Biobank Research, Umeå University, Sweden

[12] Cancer Epidemiology Unit, Nuffield Department of Population Health University of Oxford, Oxford, UK

[13] The Institute of Community Medicine, University of Tromsø, Tromsø, Norway

[14] Inserm, Centre for research in Epidemiology and Population Health (CESP), U1018, Lifestyle, genes and health: integrative trans-generational epidemiology team, Villejuif, France
[15] Université Paris Sud, Villejuif, France
[16] Université de Lyon, F-69622, Lyon, France
[17] Université Lyon 1, UMRESTTE, F-69373 Lyon, France
[18] IFSTTAR, UMRESTTE, F-69675 Bron, France
[a] Present address: Chemical Physics Department, Weizmann Institute of Science, Rehovot, Israel
[†] Joint last authors.

[*]Corresponding author: Pietro Ferrari, International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon cedex 08, France. Tel: +33 472 73 8031; Fax: +33 472 73 8361. E-mail: ferrarip@iarc.fr

Abstract

Metabolomics is a potentially powerful tool for identification of biomarkers associated with lifestyle exposures and risk of various diseases. This is the rationale of the "meeting-in-the-middle" concept, for which an analytical framework was developed in this study. In a nested case-control study on hepatocellular carcinoma (HCC) within the European Prospective Investigation into Cancer and nutrition (EPIC), serum ^{1}H NMR spectra (800 MHz) were acquired for 114 cases and 222 matched controls. Through Partial Least Square (PLS) analysis, 21 lifestyle variables (the "predictors", including information on diet, anthropometry and clinical characteristics) were linked to a set of 285 metabolic variables (the "responses"). The three resulting scores were related to HCC risk by means of conditional logistic regressions. The first PLS factor was not associated with HCC risk. The second PLS metabolomic factor was positively associated with tyrosine and glucose, and was related to a significantly increased HCC risk with OR= 1.11 (95%CI: 1.02, 1.22, p=0.02) for a 1-SD change in the responses score, and a similar association was found for the corresponding lifestyle component of the factor. The third PLS lifestyle factor was associated with lifetime alcohol consumption, hepatitis and smoking, and had negative loadings on vegetables intake. Its metabolomic counterpart displayed positive loadings on ethanol, glutamate and phenylalanine. These factors were positively and statistically significantly associated with HCC risk, with 1.37 (1.05, 1.79, p=0.02) and 1.22 (1.04, 1.44, p=0.01), respectively. Evidence of mediation was found in both the second and third PLS factors, where the metabolomic signals mediated the relation between the lifestyle component and HCC outcome. This study devised a way to bridge lifestyle variables to HCC risk through NMR metabolomics data. This implementation of the "meeting-in-the-middle" approach finds natural applications in settings characterized by high-dimensional data, increasingly frequent in the –omics generation.

Keyword: partial least square, lifestyle factors, metabolomics, hepatocellular carcinoma, meeting-in-the-middle, molecular epidemiology.

Introduction

Metabolomic profiles from blood and other biological samples collected from large-scale epidemiologic studies are increasingly being investigated [1], following recent developments in nuclear magnetic resonance (NMR) and mass spectrometry (MS) enabling the assessment of metabolic profiles for large numbers of individuals. As a result, metabolomic data is gradually playing a key part in clinical and observational studies; and new statistical methodologies [2] are increasingly being sought to explore insights into pathological processes that metabolomics may provide in order to better understand determinants of disease development. These approaches explore a variety of etiological hypotheses; however they usually focus on one aspect at a time, combining metabolomics with either epidemiologic/phenotypic data on lifestyle exposures [3] or with disease outcomes [4,5]. The main aim of this work is to jointly use all aspects that are potentially informative to apprehend the contrivances of disease development.

Metabolomic data offers the opportunity to identify signatures and biomarkers associated with environmental exposures and the risk of a disease. Prospective studies are conceptually suitable for this purpose, since they rely on biological samples collected before disease onset, and are thus marginally influenced by metabolic changes due to processes of disease development. In this scenario, the "meeting-in-the-middle" (MITM) approach [6] has been conceived as a research strategy to identify biomarkers that are related to specific exposures and that are, at the same time, predictive of disease outcome. Finding this overlap between exposure and disease of "intermediate" biomarkers can potentially disclose useful information on the exposure-to-disease pathway, and may serve as an objective risk exposure measure, ultimately allowing the identification of a targeted prevention scheme. The MITM was previously implemented as a proof of concept in a case-control study nested within a cohort of healthy individuals [7], where a list of putative intermediate ^{1}H NMR biomarkers

100 linking exposure to dietary compounds, mainly micro- and macronutrients, and disease
101 outcomes (colon and breast cancer) were investigated.

102 In this study we extend previous attempts to model the MITM by fully integrating
103 metabolomics, lifestyle and disease risk in a single analytical framework. A strategy was
104 developed to simultaneously investigate a broad range of metabolites and lifestyle variables
105 with a partial least square (PLS) regression model [8]. The resulting scores were related to the
106 risk of hepatocellular carcinoma (HCC), in a case-control study nested within the European
107 Prospective Investigation into Cancer and nutrition (EPIC). HCC is the most frequent primary
108 form of cancer affecting the liver, an organ that plays a critical role in many metabolic
109 pathways [9]. HCC is a disease with multifactorial origins embracing lifestyle and dietary
110 exposures whose intersection may reveal metabolomic signals [10] relevant to cancer onset.
111 The system of relationships between metabolomic profiles and lifestyle factors in relation to
112 HCC was evaluated by means of mediation analysis. The methodological challenges
113 characterizing the analysis of large and complex metabolomic datasets are described and
114 discussed.

115 **Methods**

116 *EPIC design.* The European Prospective Investigation into Cancer and nutrition (EPIC) is a
117 large cohort established to investigate the association of diet, lifestyle and environmental
118 factors with cancer incidence and other chronic disease outcomes. Between 1992-2000, over
119 520,000 participants aged 20-85 years, were recruited from 23 centers in 10 Western
120 European countries including Denmark, France, Germany, Greece, Italy, Norway, Spain,
121 Sweden, The Netherlands and the United Kingdom [11]. The design, rationale and methods
122 of the EPIC study including information on dietary assessment methodology, blood collection
123 protocols and follow-up procedures were previously detailed [11].

Between 1992 and 1998, standardized lifestyle data, anthropometric measures and biological samples were collected at recruitment, prior to onset of any disease [11]. Validated country-specific questionnaires ensuring high compliance were used to measure diet over the previous 12 months [12]. Blood samples are stored at the International Agency for Research on Cancer (IARC, Lyon, France) in -196˚C liquid nitrogen for all countries, exceptions being Denmark (nitrogen vapour,-150˚C) and Sweden (freezers, -80˚C).

The nested case-control study. The present study focused on data with available sera samples from a nested case-control study in EPIC on hepatocellular carcinoma (HCC) [13]. Cases of HCC were identified from all participating EPIC centres except for Norway and France (n=117) from recruitment (1993-1998) up to 2007. Two controls (n=232) were selected for each case from all cohort members alive and free of cancer (except non-melanoma skin cancer) by incidence-density sampling and were matched on age at blood collection (±1 year), sex, study centre, date (±2 months), time of the day at blood collection (±3 hours) and fasting status at blood collection (<3, 3-6, >6 hours); among women, additional matching criteria included menopausal status (pre-, peri-, post-menopausal) and hormone replacement therapy (HRT) use at time of blood collection (yes/no). In the present study, cases and controls were both included in the analyses as the subjects were all cancer-free at blood collection. Out of the total 349 subjects, 7 subjects (3 cases and 4 controls) had too little serum volume for NMR spectral acquisition with sufficient sensitivity; 6 additional control subjects were excluded following the exclusion of their corresponding case subject. The final analysis included 114 HCC cases and 222 matched controls of which 108 case–control sets with two matched control subjects and 6 sets with one matched control subject.

NMR spectra acquisition. Sera were processed using standard procedure for ^{1}H NMR metabolic measurement and profiling protocols [14]. Details on the sera sample preparation as well as NMR data acquisition and processing have been described elsewhere [15]. In brief,

each spectrum was reduced to 8,500 bins of 0.001 ppm width over the chemical shift range of 0.5 to 9 ppm. Spectra were normalized to total intensity, centred and Pareto scaled, and additionally normalized for batch-effects using the batch profiling calibration method [16]. After removal of the structured noise (characterized by a specific mean and standard deviation) located in a well-known noise region (8.5-9ppm) and variables with identical characteristics, the statistical recoupling of variables (SRV) [17], a bucketing procedure, was applied to the metabolomic spectra. The SRV procedure identifies clusters of variables with respect to the ratio of covariance and correlation between consecutive variables along the chemical shift axis, allowing the restauration of the spectral dependency and the recovery of complex NMR signals corresponding to potential physical, chemical or biological entities. More details on the SRV procedure are available in the **Mathematical Appendix**. This permitted a reduction of the number of NMR variables from 8,500 bins to 285 clusters of variables corresponding to reconstructed peak entities which constituted the Y-set of metabolic variables. All steps to obtain the data were done without knowledge of the case-control status of the subjects. Quality control (QC) samples were included to ensure reproducibility of the NMR data acquisition.

Metabolite identification. The assignment of NMR signals observed in the ^{1}H one-dimensional fingerprints to metabolites has been achieved by the analysis of additional 2D NMR experiments ^{1}H-^{13}C HSQC and ^{1}H-^{1}H TOCSY obtained on a subset of representative samples (one control and one case). The measured chemical shifts were compared to reference shifts of pure compounds using HMDB [18], MMCD [19] and ChenomX, (ChenomX NMR suite, ChenomxInc, Edmonton, Canada) databases.

Lifestyle variables. The predictors (what will be referred to later on as the X-set) included 13 dietary variables from main EPIC food groups compiled from validated country-specific food frequency questionnaires (FFQ) [11,20] (potatoes and other tubers; vegetables; legumes;

fruits, nuts and seeds; dairy products; cereal and cereal products; meat and meat products; fish and shellfish; egg and egg products; fat; sugar and confectionary; cakes and biscuits; non-alcoholic beverages), alcohol average lifetime intake (continuous, g/day), anthropometric measures including body mass index (continuous, kg/m^2) and height (continuous, cm) that were measured by trained interviewers in the majority of participants [11], highest level of education achieved (categorical: none or primary school completed, technical/professional school, secondary school, longer education (incl. university degree), unspecified), smoking status (categorical: never, former, current smoker, unknown), a measure of physical activity (continuous, metabolic equivalents of task (MET)/h), hepatitis status (yes/no, from biomarker measures of HBV and HCV seropositivity [ARCHITECT HBsAg and anti-HCV chemiluminescent microparticle immunoassays; Abbott Diagnostics, France]) and baseline self-reported diabetes status (yes/no). Descriptive information on these variables can be found in **Supplementary table 1**.

Statistical analysis

PC-PR2 analysis. Principal component partial R-square (PC-PR2) was primarily used to identify and quantify sources of systematic variability within metabolomic data [15]. PC-PR2 combines aspects of principal component analysis (PCA) and the $R^2_{partial}$ statistic in multiple linear regression, and allows for (some) inter-correlation between the explanatory variables under scrutiny [15]. In short, PCA is performed on the 285 clusters of ^1H NMR variables and a number of components is retained explaining an amount of total variability above a designated threshold (here, 80%). Then, multiple linear regression models are fitted where each component's variability is explained in terms of relevant covariates, e.g. specific characteristics of samples like country of origin, smoking status, laboratory treatment, etc. For each given component, the $R^2_{partial}$ statistic is computed for all covariates, quantifying the amount of variability each independent variable explains, conditional on all other covariates

included in the model. Finally, an overall $R^2_{partial}$ is calculated as a weighted average for every covariate, using the eigenvalues as components' weights. Mathematical details pertaining to the PC-PR2 method are described elsewhere [15].

In this study, PC-PR2 was applied to the 285 clusters of NMR variables, whereas the explanatory variables examined for systematic variability were NMR batch, country of origin, sex, age at blood collection, serum clot contact time (centrifugation at the day of blood collection d, or the following day, d+1), length of freezing time (<= 15 vs. >15 years), and fasting status at blood collection (< 3, 3-6, > 6 hours). With the similar motivation of identifying sources of variability within lifestyle data, a similar PC-PR2 analysis was applied to the 21 lifestyle factors; the examined covariates for systematic variability were country of origin, sex and age at recruitment. For both metabolomics and lifestyle data, residuals on the variable accounting for most variability, identified through PC-PR2 analyses, were computed in a series of univariate linear regression models [21] and were used in the subsequent PLS.

PLS analysis. A PLS model was used to relate lifestyle variables to metabolomic profiles. PLS is a multivariate technique that generalizes features of PCA and multiple linear regression. PLS iteratively extracts linear combinations of, in turn, predictors (the X-set) and responses (the Y-set), which in this study, were lifestyle variables and metabolomic profiles, respectively. First, components or latent factors are extracted allowing a simultaneous decomposition of the X- and Y-sets, in order to maximize their covariance [22]. The factors extracted from the predictors' set are orthogonal. Computational details of PLS are described in the **Mathematical Appendix**. As a standard step for the PLS algorithm, the X- and Y-sets were centered and standardized for the analysis and a simple expectation-maximization (EM) algorithm, adapted from the PLS kernel algorithm [23,24], was used to compute covariance matrices when missing values were present in the lifestyle data. This was done as follows: a first pass of PLS was computed filling in the missing values by the average of the non-

missing values for each corresponding variable. A second pass was then performed whereby the missing data were assigned their predicted values based on the first model, and the PLS regression is recomputed.

Then, a seven-fold cross validation analysis was carried out to select the number *h* of significant PLS factors to retain [8] (see **Mathematical appendix**).This was achieved by splitting the data into seven groups of observations. In turn, each group of observations was considered as the test set, whilst the other six were the training sets, used to perform PLS analysis. A measure of PLS performance was determined for each step through the predicted residual sum of squares (PRESS) statistic, whereby the predicted values in the test set, the $\tilde{Y}_h$ matrix, based on the X-components estimated through the model in the training set, were compared to the observed responses, the Y matrix. This comparison is quantified by the squared Euclidean distance between these two matrices. In turn for an increasing number *h* of components, the process is iterated seven times, until each group of observations serves as a test set. Eventually, the number *h* of selected PLS factors is the one minimizing the PRESS statistic.

For each PLS factor, loadings were computed for the lifestyle (X-set) and the NMR (Y-set) variables. The loadings, i.e. coefficients quantifying the contribution of each original variable to the PLS factor, were used to characterize the various factors. As the analysis involved many variables in the X-set and, particularly, in the Y-set, the interpretation focused primarily on variables with loading values lower than the 10^{th} percentile and larger than the 90^{th} percentile for the X variables, and lower than the 5^{th} and larger than the 95^{th} percentiles for the Y variables, that were deemed the most significant contributors to the PLS factor.

Logistic regression analysis. Last, scores of each PLS factor were related to HCC risk in conditional logistic regression models to compute HCC odds ratios (ORs) and associated 95% confidence intervals (95% CI) where ORs express the change in HCC risk associated to one

standard deviation (1-SD) increase in the score. Models were adjusted for C-reactive protein concentration, alpha-fetoprotein concentration and for a composite score indicative of liver damage. The score summarizes the number of abnormal values of circulating enzymes measured in the hepatic tissue in six liver function tests (alanine aminotransferase >55 U/L, aspartate aminotransferase >34 U/L, gamma-glutamyltransferase: men>64 U/L and women>36 U/L, alkaline phosphatase >150 U/L, albumin<35 g/L, total bilirubin>20.5 µmol/L; cut-points were provided by the clinical biochemistry laboratory that conducted the analyses and were based on assay specifications) [25]. These biomarkers were measured on the ARCHITECT c Systems™ and the AEROSET System (Abbott Diagnostics) using standard protocols. Laboratory analyses were performed at the Centre de Biologie République laboratory, Lyon, France. These adjustments were deemed necessary to address potential confounding stemming from metabolic disorders, inflammation or underlying liver dysfunction [25–28]. Adjustments for total dietary fibre, vitamin D, calcium and iron intakes (continuous) were evaluated but not retained in the final models for lack of confounding exerted by these variables. The receiver operating characteristic (ROC) curve and the associated area under the curve (AUC) were determined from conditional logistic regressions to evaluate the predictive performance of PLS models. AUC values were computed for conditional logistic models including progressively the PLS scores, separately for lifestyle and metabolomic factors (as shown in **Table 4**, column 1). The sensitivity, specificity and accuracy were calculated for a cut-off point, selected as the minimal distance between the ROC curve and the upper left corner of the diagram [29,30]. The corrected positive predictive value (PPV), taking into account the nested case-control design [31,32] was computed by including the prevalence of HCC in the EPIC population(π= 0.0004), computed over a 7-year period (1992-2010) where 191 HCC cases were ascertained from a total of 477,206 participants included for case identification after relevant exclusions [33]. The AUC

unavoidably increases with the number of covariates added to the conditional logistic model. To address this issue, a resampling scheme was devised to compute an objective/ unbiased estimate of the AUC, inspired by the work of Uno et al [34]. For each one of the 1000 drawn bootstrap samples, a 10-fold cross-validation was performed, repeated ten times to remove variation due to random partitioning of data and to yield more stable estimates. The predicted values from each of the conditional logistic models in the training set were used to derive AUC values in the test set. The 2.5^{th} and 97.5^{th} percentile values made up the 95% confidence intervals.

Sensitivity analyses. A sensitivity analysis was performed by running PLS on data excluding sets where cases were diagnosed within the first two years of follow-up. The model was conducted on 271 observations (92 cases, 179 controls), to investigate the performance of the PLS model, ruling out potential reverse causation. The metabolomic profiles of HCC cases diagnosed within two years from enrolment could reflect the presence of the tumour rather than informing about tumour aetiology. The variable importance in the projection (VIP) statistic was used to facilitate the comparison of the sensitivity analysis with the main analysis. The VIP expresses the explanatory power of a predictor variable X across all response variables Y (see **Mathematical Appendix**).

Mediation analysis. The mediating role of the Y-scores in the association between lifestyle profiles and HCC risk was assessed. Separately for each extracted combination of lifestyle and metabolomic PLS factors, mediation analyses were performed with the 'paramed' Stata function that allows for exposure-mediator interaction based on Valeri and VanderWeele's work [35]. Briefly, mediation was computed using a Baron and Kenny approach adapted to dichotomous outcomes [36], where two models were specified. In the mediator model, the mediator (the Y-score) was linearly regressed on the exposure (the X-score), while in the outcome model the exposure (X-score) and the mediator (Y-score) were related to the HCC

299 indicator in unconditional logistic regressions. Both models accounted for the concentration
300 of C-reactive protein, alpha-fetoprotein and the composite score of liver damage, and
301 additionally accommodated the other extracted metabolic profiles (Y-scores) to control for
302 mediator-outcome confounders that may occur when estimating the Natural Indirect Effect
303 (NIE) [35]. As the outcome (HCC) is rare, direct and indirect effects can be estimated taking
304 into account the case-control design. This is done by using the same formulas for the effects,
305 while running the mediator regression only for the controls [36]. As mediation packages do
306 not yet accommodate conditional logistic models, the outcome and the mediator models,
307 which were accommodated in unconditional logistic regressions, were adjusted for center and
308 age at blood collection for sake of consistency with previous steps of the analysis.

309 Statistical analyses were performed using R [37] and SAS [38] in general, with the following
310 packages for specific purposes: PROC PLS in SAS 9.4 for PLS analyses, 'paramed' in Stata
311 12 [39] for mediation analyses, 'OptimalCutpoints' in R for ROC-related assessments.

312 The different steps of the analytical framework developed in this study to model the MITM
313 are presented in **Figure 1**.

314 **Results**

315 In the PC-PR2 analyses, a total of 17 and 14 principal components were retained to
316 explain an amount of total variability exceeding 80% in metabolomics and lifestyle data
317 respectively. **Figure 2** shows that the ensemble of explanatory variables accounted for 19.4%
318 and 26.7% of total variance, respectively in metabolomics and lifestyle data, of which the
319 highest contributor was 'country of origin' with consistently 8% and 22%. PLS analysis was
320 carried controlling for this variable.

321 After a seven-fold cross-validation, three PLS factors were retained accounting for
322 21.7% and 8.5% of the overall variability observed in predictor and response variables,

323 respectively (**Table 1**). Lifestyle variables and clusters of NMR variables contributing highly
324 to PLS factors were identified using factor loading values (**Table 2**). The first PLS factor was
325 predominantly positively associated with dairy products and cakes and biscuits intake, while
326 lifetime alcohol intake, smoking status and diabetes displayed negative loadings for this
327 lifestyle component (**Table 2**). On the same PLS factor, signals mainly associated with
328 glucose and bonds of lipids with negative loading values, and with aspartate, glutamine and
329 lysine with positive loadings emerged on the metabolomic profile (**Table 2**). Lifestyle
330 variables characterizing the second PLS factor included cereal products, height and education
331 level with negative loadings, and hepatitis with positive loadings. The metabolic signature
332 included NMR variables with positive loadings associated with aromatic amino acids
333 (phenylalanine, tyrosine) and glucose; and those with negative loadings associated mainly
334 with bonds of lipids, threonine and mannose (**Table 2**). The third PLS factor had a lifestyle
335 pattern outlining intake of vegetables (high negative loadings values), lifetime alcohol
336 consumption, smoking, and hepatitis infection (positive loadings). Its counterpart NMR
337 pattern highlighted signals of glucose and aspartate, with high negative loadings, along with
338 signals of ethanol, myo-inositol, proline and glutamate as prominent metabolites with positive
339 loadings (**Table 2).**

340 Conditional logistic regression models relating HCC risk with the X- and Y-scores are
341 shown in **Table 3**. The first PLS factor was associated to a non-significant decreased HCC
342 risk (23% and 4% in the X- and Y-scores respectively), while the second and third factors
343 were associated to a statistically significant increased HCC risk (54% and 11%; and 37% and
344 22% respectively). Results for the ROC curves parameters are reported in **Table 4**, including
345 AUC, sensitivity, specificity, accuracy and PPV for different combinations of the X- and Y-
346 scores. The AUC of the X-scores and Y-scores for all 3 PLS factors, adjusted for C-reactive
347 protein concentration, alpha-fetoprotein concentration and the score of liver damage, was

respectively 0.859 and 0.853. An increase in the resampled cross-validated AUC values was also observed for all three X- and Y-scores, albeit smaller, with respectively 0.836 and 0.827. Results from the sensitivity analysis conducted on data excluding sets where cases were diagnosed within the first two years of follow-up, showed similarities in terms of lifestyle variables' and metabolites' loadings on the PLS factors (**Supplementary Table 2**). Notable differences pertained to the identification of new signals for the first PLS factor including ethanol, histidine and an unknown compound. On the second lifestyle factor, BMI (positive loadings) replaced education level (negative loadings) while the reflected metabolomic profile was comparable to its counterpart from the main analysis **(Supplementary Table 2)**. On the third factor, smoking status and hepatitis (positive loadings) were replaced by sugar and confectionary intake (negative loadings); signals contributing to the associated metabolic profile remained the same but the direction of the association was inversed as loadings had opposite signs as compared to the counterpart PLS factor of the main model (**Supplementary Table 2**). Corresponding ORs from conditional logistic regression models relating the X- and Y-scores to HCC risk are available in **Table 5**. The scores showed a statistically significant association in the second factor for both sets and in the third factor for the Y-set. ROC-associated statistics for different models are presented in **Supplementary Table 3**. The VIP plot (**Figure 3**) displayed the results for the importance of the lifestyle variables in the prediction of the Y-set computed for the main PLS model performed including all subjects (panel **A**) and for the sensitivity model (panel **B**). The results suggested a potential gain in stability as prominent lifestyle variables for prediction were maintained (hepatitis/diabetes/cakes and biscuits), the magnitude of the VIP was improved for some (fat/lifetime alcohol intake) and less emphasis was put on others (BMI/physical activity).

Finally, the natural indirect effect was assessed in the mediation analyses and the results are presented in **Table 6**. Overall, there was limited evidence that metabolomic signals mediated

373 the association between lifestyle components and HCC risk in the first PLS factor. Evidence
374 of a significant mediated effect by the Y-scores was found in the second and third PLS factors
375 when models were adjusted for exposure-mediator interaction (**Table 6**).

376 **Discussion**

377 In this work, an analytical strategy based on PLS analysis was conceived to extract
378 relevant information from sets of lifestyle and NMR metabolomic variables, and to relate the
379 resulting components to the risk of disease. This offered a way to implement the MITM
380 approach [6] in a nested case-control study on HCC within the EPIC study. MITM has been
381 suggested as a way to link specific putative metabolites to lifestyle exposures and disease
382 outcomes, thus leading to the identification of potential intermediate biomarkers [6].

383 An implementation of MITM was previously carried out in a nested case-control study
384 in the Turin sub-cohort of EPIC [7] based on prospectively collected plasma samples from a
385 pilot study on colon and breast cancers. In their work, a list of intermediate markers was
386 identified by an in-parallel evaluation of the relationships between untargeted ^{1}H NMR
387 profiles with dietary exposures and risk of colon and breast cancers using correlation analysis
388 and logistic regression. In our study, a different analytical framework was developed, largely
389 exploiting features of PLS analysis, a multivariate technique that iteratively extracts
390 components capturing co-variability in sets of predictors and response variables [8,40]. A set
391 of lifestyle predictor variables were related to NMR responses. In a second step, PLS
392 predictors' and responses' scores were linked to the risk of HCC.

393 Another sensitive issue in this analysis was the choice of lifestyle variables. Two
394 disease-indicator variables reflecting environmental exposures, diabetes and hepatitis, were
395 included in the set of predictors, as they turned out to have an important role in the
396 characterization of metabolomic signatures. In addition, diabetes is the main metabolic risk

factor for HCC alongside with fatty liver disease [41,42], and chronic infection with hepatitis B (HBV) and particularly hepatitis C (HCV) viruses were classified as class I carcinogens for HCC by IARC [43].

Other relevant biomarkers were not part of the list of predictors in PLS analysis, but were controlled for in logistic regression models. This included C-reactive protein, alpha-fetoprotein, and a score for liver damage, an index of different circulating enzymes measured in the hepatic tissue indicating potential underlying liver function impairment [25]. The alpha-fetoprotein was included as an adjustment factor in the analyses not because of its established part as a serum marker for HCC diagnosis [26,44], but rather to account for it as a potential confounder that may cloud the relation between scores and HCC, both in conditional logistic regressions and in mediation analyses.

Similarly to other multivariate techniques, a key aspect of PLS analysis is the choice of the number of factors to retain, in an effort of exhaustively summarizing data variability through a limited number of factors. Based on a seven-fold cross-validation, three linear combinations of variables were extracted in this work. A challenging aspect of this analysis is the interpretation of these factors, with respect to lifestyle and metabolomic variables. A subjective criterion based on the distribution of loading values was used throughout. The variables displaying the most extreme loading values (in absolute terms) were the ones characterizing each factor.

The first lifestyle factor highlighted a healthy pattern with negative loadings for diabetes status, smoking status and lifetime alcohol intake, and was not associated to HCC risk, similarly to its metabolomics counterpart. The lifestyle component of the second PLS factor, was reflective of a lifestyle pattern reflective of “higher-risk exposures”, and was related to a significant 54% increase in HCC risk. Likewise, its associated metabolic component displayed a significant HCC risk augmentation by 11%. The lifestyle component

of the third PLS factor described participants with lower vegetables intake, elevated lifetime alcohol consumption, more likely to be ever smokers and hepatitis positive; one standard deviation increase of this component was associated to a statistically significant 37% increase in HCC risk. Similarly, a 22% significant increase in HCC risk was observed for its metabolic counterpart, characterized by positive signals of ethanol and myo-inositol, and displayed negative loadings for glucose.

The MITM is captured by the rationale of PLS analysis, in the sense that each set of lifestyle profiles and metabolic signatures of the extracted PLS factors mirrored one another. In addition, mediation was observed for the second and third PLS factors, whereby the metabolomic component mediated the relation between the lifestyle component and HCC, for which statistically significant associations with HCC risk were estimated, emphasising the presence of a MITM. Mediation analysis relies on the assumption that there is no mediator-outcome confounder that is affected by the exposure [35]. In our study C-reactive protein, alpha-fetoprotein and liver damage score were weakly correlated to lifestyle factor score, thus introducing potential bias in the estimation of direct and indirect effects in our mediation analysis. Additionally, a number of background confounders (mediator-outcome and exposure-outcome confounders) were present that we have tried to control for, either by adjustments or by accounting for potential interactions, however some degree of bias can remain and caution should be employed when interpreting the results.

The predictive performance of PLS factors in relation to HCC occurrence was evaluated through an analysis of AUC values. The performance of the model improved progressively, with all 3 X- and Y-scores added; after a bootstrapped cross-validation, the AUC estimates were lower but the increase in the performance was nevertheless present. The ROC methodology allows estimation of PPV, which expresses the risk of disease after a positive test [45]. In a setting with low HCC prevalence (π=0.0004), in line with Western populations

447 [46], extremely low PPV estimates were observed. In the absence of a very specific test, many
448 positive tests arise from disease-free individuals [45], thus leading to a dilution of PPV.
449 A sensitivity analysis was carried out excluding the first two years of follow-up, but results
450 were virtually unchanged, both in terms of relative risk estimates in logistic regression
451 models, and of percentage of variability explained in PLS analysis. These findings suggest
452 that reverse causation bias, if present, was minimal.

453 This study had the ambition of integrating in the same analytical framework study
454 participants' lifestyle characteristics with a large number of NMR metabolic profiles. These
455 data pose a number of methodological challenges due to their size and the complexity of
456 exhaustively capturing and interpreting the biological processes they reflect. To address these
457 issues, techniques involving multivariate statistics have been progressively revived in the
458 recent years [2]. Epidemiologic evaluations of metabolomic data frequently combined PLS
459 with discriminant analysis, such as PLS-DA or O-PLS-DA. The main objective of these
460 methods is to identify a series of metabolomic features distinguishing between two very
461 distinct groups of study participants [47,48]. In such strategies, only one set of variables is
462 multi-dimensional and the response is one variable only. Similar multivariate techniques for
463 pattern extraction, belonging to the family of regression methods, include reduced rank
464 regression. This multivariate method relates an ensemble of response variables to a set of
465 predictor variables where the estimated matrix of the regression coefficients is of reduced
466 rank [49–51]. In addition, canonical correlation analysis (CCA) [52] is a method applied to
467 identify the optimum structure or dimensionality of each variable set that maximizes the
468 relationship between two sets of multi-dimensional variables. The main difference between
469 CCA and PLS regression is that CCA maximizes the correlation between the two new
470 dimensions, i.e. extracted factors, whereas PLS maximizes their covariance. PLS can be
471 considered as a trade-off between CCA and PCA, since maximizing the covariance

corresponds to maximizing the product of the correlation and standard deviation, given that cov(X,Y)=cor(X,Y)*SD(X)*SD(Y).

Untargeted NMR was used in this work to acquire metabolomic signals. Prior to PLS analysis, a bucketing procedure, the statistical recoupling of variables (SRV) [17,53], was applied to reduce the number of NMR variables to 285 clusters. This was done by aggregating consecutive NMR bins based on their covariance to correlation ratio. This allowed the identification of informative components of the spectra, thus acting as an efficient noise-removing filter. Subsequently the annotation effort remains challenging, for a number of reasons. The majority of published metabolomics studies often identified a limited number of metabolites at a time [54], and the Human Metabolome Database (HMDB) and other related resources [18,55], that offer richly annotated information continuously increasing the metabolite coverage for users, are mostly exploited through time consuming interactive procedures. In addition, individual metabolites often overlap in NMR signals, which can hinder annotations. These challenges, as well as large variability in metabolite concentrations, and disentangling informative signals from noise, are not specific to NMR and pertain to any type of untargeted technique. Such investigations may profit from complementary targeted metabolomic analytical strategies [55].

Throughout the different steps of this work, the scaling problem was first tackled by normalizing spectra to total intensity. NMR data were also centered and Pareto-scaled, together with correction for potential batch effects [16]. The PC-PR2 method offered a way to investigate major sources of systematic variability in NMR and lifestyle data [15]. The variable “country of origin” emerged as the variable accounting for the largest proportion of total variability, and the residual method was used to control for this variable in the following steps of the analysis. While this may lead to removing regional gradients of dietary variability, this step is instrumental to avoid unwanted systematic regional-specific bias in the

data in country-specific questionnaire assessments. In addition, technical aspects like storage and handling of biological samples, fasting status at blood collection are specific to each country [15]. In any case, variability due to "country of origin" is not exploited in conditional logistic models, as cases and controls were also matched on center.

One of the limitations of this study is the restricted sample size which raises concerns with regards to power to detect associations. While a larger sample size would possibly result in more statistically significant findings, we used the data that was available with NMR profiles measured. In this work we have developed a framework to analyse complex data integrating lifestyle and metabolomics in relation to risk of disease. The approach described in this study has merits but also pitfalls among which it is worth mentioning that statistical methods are used repeatedly on the same set of data, notably the PLS model, the conditional logistic regression, the AUC estimation and mediation analysis. To partially address this, a cross-validation approach was devised for AUC estimation which involved conditional logistic regression, whereby PLS was done without knowledge of the case/control status. However, conditional logistic regression models and mediation analyses were implemented on the same data, and our analysis did not account for this limitation. This may have led to spuriously increase the nominal level of statistical significance of statistical tests.

Conclusion

The MITM emerged as a method for the identification of relevant biomarkers, with great potential to unravel utmost important steps in the aetiology of disease. The analytical strategy for MITM was developed to use all potentially informative aspects of high-throughput data by integrating metabolomic, dietary and lifestyle exposures together with disease indicators. While the framework was applied towards the investigation of HCC determinants, it can be easily extended to similar aetiological contexts and applied to other –omics settings.

Funding

This work was supported by the French National Cancer Institute (L'Institut National du Cancer; INCA) [grant number 2009-139; PI: M. Jenab]. The coordination of the European Prospective Investigation into Cancer and nutrition is financially supported by the European Commission (Directorate General for Health and Consumer Affairs) and the International Agency for Research on Cancer. The national cohorts are supported by Health Research Fund (FIS) of the Spanish Ministry of Health RTICC 'Red Temática de Investigación Cooperativa en Cáncer [Grant numbers: Rd06/0020/0091, Rd12/0036/0018], Regional Governments of Andalucía, Asturias, Basque Country, Murcia [project 6236] and Navarra, Instituto de Salud Carlos III, Redes de Investigacion Cooperativa (RD06/0020) (Spain); the Danish Cancer Society (Denmark); the Ligue Contre le Cancer, the Institut Gustave Roussy, the Mutuelle Générale de l'Education Nationale, and the Institut National de la Santé et de la Recherche Médicale (France); the Deutsche Krebshilfe, the Deutsches Krebsforschungszentrum, and the Federal Ministry of Education and Research (Germany); the Hellenic Health Foundation, the Stavros Niarchos Foundation, and the Hellenic Ministry of Health and Social Solidarity (Greece); the Italian Association for Research on Cancer AIRC and the National Research Council (Italy); the Dutch Ministry of Public Health, Welfare and Sports, the Netherlands Cancer Registry, LK Research Funds, Dutch Prevention Funds, the Dutch Zorg Onderzoek Nederland, the World Cancer Research Fund, and Statistics Netherlands (Netherlands); European Research Council-2009-AdG 232997 and the Nordforsk, Nordic Centre of Excellence program on Food, Nutrition and Health (Norway); the Swedish Cancer Society, the Swedish Research Council, and the Regional Governments of Skåne and Västerbotten (Sweden); Cancer Research UK, the Medical Research Council, the Stroke Association, the British Heart Foundation, the Department of Health, the Food Standards Agency, and the Wellcome Trust (United Kingdom). The work undertaken by N Assi was supported by by the Université de Lyon I through a doctoral fellowship awarded by the EDISS doctoral school.

Acknowledgments

We would like to acknowledge the assistance of Dr Elodie Jobard from the ISA-CRMN in obtaining the annotation of the NMR data.

Reference List

1. Nicholson, J.K., Holmes, E., and Elliott, P. (2008) The metabolome-wide association study: a new look at human disease risk factors. *J. Proteome Res.*, **7**,3637–3638.

2. Chadeau-Hyam, M., Campanella, G., Jombart, T., Bottolo, L., Portengen, L., Vineis, P., Liquet, B., and Vermeulen, R.C.H. (2013) Deciphering the Complex: Methodological Overview of Statistical Models to Derive OMICS-Based Biomarkers. *Environ. Mol. Mutagen.*, **54**,542–557.

3. Floegel, A., Wientzek, A., Bachlechner, U., Jacobs, S., Drogan, D., Prehn, C., Adamski, J., Krumsiek, J., Schulze, M.B., Pischon, T., and Boeing, H. (2014) Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study. *Int. J. Obes. (Lond).*,1–9.

4. Trushina, E. and Mielke, M.M. (2013) Recent advances in the application of metabolomics to Alzheimer's Disease. *Biochim. Biophys. Acta*, **1842**,1232–1239.

5. Jin, X., Yun, S.J., Jeong, P., Kim, I.Y., Kim, W.-J., and Park, S. (2014) Diagnosis of bladder cancer and prediction of survival by urinary metabolomics. *Oncotarget*, **5**,1635–1645.

6. Vineis, P. and Perera, F. (2007) Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. *Cancer Epidemiol. Biomarkers Prev.*, **16**,1954–1965.

7. Chadeau-Hyam, M., Athersuch, T.J., Keun, H.C., De Iorio, M., Ebbels, T.M.D., Jenab, M., Sacerdote, C., Bruce, S.J., Holmes, E., and Vineis, P. (2011) Meeting-in-the-middle using metabolic profiling - a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers*, **16**,83–88.

8. Tenenhaus, M. (1998) La régression PLS. Technip. Paris.

9. Mitra, V. and Metcalf, J. (2009) Metabolic functions of the liver. *Anaesth. Intensive Care Med.*, **10**,334–335.

10. Fages, A., Duarte-Salles, T., Stepien, M., Ferrari, P., Fedirko, V., Pontoizeau, C., Trichopoulou, A., Aleksandrova, K., Tjønneland, A., Olsen, A., Clavel-Chapelon, F., Boutron-Ruault, M.-C., Severi, G., Kaaks, R., Kuhn, T., Floegel, A., Boeing, H., Lagiou, P., Bamia, C., Trichopoulos, D., Palli, D., Pala, V., Panico, S., Tumino, R., Vineis, P., Bueno-de-Mesquita, H.B., Peeters, P.H.M., Weiderpass, E., Agudo, A., Molina-Montes, E., Huerta, J.M., Ardanaz, E., Dorronsoro, M., Sjoberg, K., Ohlsson, B., Khaw, K.-T., Wareham, N.J., Travis, R.C., Schmidt, J.A., Cross, A.J., Gunter, M.J., Riboli, E., Scalbert, A., Romieu, I., Elena-Herrmann, B., and Jenab, M. (2015) Metabolomic Profiles of Hepatocellular Carcinoma in a European Prospective Cohort. *Submitt. to Am. J. Gastroenterol.*,.

11. Riboli, E., Hunt, K.J., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondière, U.R., Hémon, B., Casagrande, C., Vignat, J., Overvad, K., Tjønneland, A., Clavel-Chapelon, F., Thiébaut, A., Wahrendorf, J., Boeing, H., Trichopoulos, D.,

Trichopoulou, A., Vineis, P., Palli, D., Bueno-De-Mesquita, H.B., Peeters, P.H.M., Lund, E., Engeset, D., González, C. a, Barricarte, A., Berglund, G., Hallmans, G., Day, N.E., Key, T.J., Kaaks, R., and Saracci, R. (2002) European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.*, **5**,1113–1124.

12. Kaaks, R., Slimani, N., and Riboli, E. (1997) Pilot Phase Studies on the Accuracy of Dietary Intake Measurements in the EPIC Project : Overall Evaluation of Results. **26**,26–36.

13. Trichopoulos, D., Bamia, C., Lagiou, P., Fedirko, V., Trepo, E., Jenab, M., Pischon, T., Nöthlings, U., Overved, K., Tjønneland, A., Outzen, M., Clavel-Chapelon, F., Kaaks, R., Lukanova, A., Boeing, H., Aleksandrova, K., Benetou, V., Zylis, D., Palli, D., Pala, V., Panico, S., Tumino, R., Sacerdote, C., Bueno-De-Mesquita, H.B., Van Kranen, H.J., Peeters, P.H.M., Lund, E., Quirós, J.R., González, C. a, Sanchez Perez, M.-J., Navarro, C., Dorronsoro, M., Barricarte, A., Lindkvist, B., Regnér, S., Werner, M., Hallmans, G., Khaw, K.-T., Wareham, N., Key, T., Romieu, I., Chuang, S.-C., Murphy, N., Boffetta, P., Trichopoulou, A., and Riboli, E. (2011) Hepatocellular carcinoma risk factors and disease burden in a European cohort: a nested case-control study. *J. Natl. Cancer Inst.*, **103**,1686–1695.

14. Beckonert, O., Keun, H.C., Ebbels, T.M.D., Bundy, J., Holmes, E., Lindon, J.C., and Nicholson, J.K. (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat. Protoc.*, **2**,2692–2703.

15. Fages, A., Ferrari, P., Monni, S., Dossus, L., Floegel, A., Mode, N., and Al., E. (2014) Investigating sources of variability in metabolomic data in the EPIC study: the Principal Component Partial R-square (PC-PR2) method. *Metabolomics*, **10**,1074–1083.

16. Fages, A., Pontoizeau, C., Jobard, E., Lévy, P., Bartosch, B., and Elena-Herrmann, B. (2013) Batch profiling calibration for robust NMR metabonomic data analysis. *Anal. Bioanal. Chem.*, **405**,8819–8827.

17. Blaise, B.J., Shintu, L., Elena, B., Emsley, L., Dumas, M.-E., and Toulhoat, P. (2009) Statistical Recoupling Prior to Significance Testing in Nuclear Resonance Based Metabonomics. *Anal. Chem.*, **81**,6242–6251.

18. Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. a, Lim, E., Sobsey, C. a, Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H.J., and Forsythe, I. (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**,D603–D610.

19. Cui, Q., Lewis, I.A., Hegeman, A.D., Anderson, M.E., Li, J., Schulte, C.F., Westler, W.M., Eghbalnia, H.R., Sussman, M.R., and Markley, J.L. (2008) Metabolite

identification via the Madison Metabolomics Consortium Database. *Nat. Biotechnol.*, **26**,162–164.

20. Slimani, N., Deharveng, G., Unwin, I., Southgate, D.A., Viignat, J., Skeie, G., Salvini, S., Parpinel, M., Moller, A., and et al. (2007) The EPIC nutrient database project (ENDB): a first attempt to standardize nutrient databases across the 10 European countries participating in the EPIC study - DTU Orbit. *Eur. J. Clin. Nutr.*, **61**,1037–1056.

21. Kleinbaum, D.G., Kupper, L.K., and Muller, K.E. (1987) Applied regression analysis and other multivariable methods. Belmont, CA: Duxbury Press.

22. Wold, S., Sjostrom, M., and Ericksson, L. (2001) PLS-regression : a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, **58**,109–130.

23. Rannar, S., Geladi, P., Lindgren, F., and Wold, S. (1995) A PLS kernel algorithm for data sets with many variables and few objects. Part II: cross-validation, missing data and examples. *J. Chemom.*, **9**,459–470.

24. Bastien, P. (2008) Régression PLS et Données Censurées. Conservatoire National des Arts et Métiers - CNAM.

25. Fedirko, V., Trichopolou, A., Bamia, C., Duarte-Salles, T., Trepo, E., Aleksandrova, K., Nöthlings, U., Lukanova, A., Lagiou, P., Boffetta, P., Trichopoulos, D., Katzke, V. a, Overvad, K., Tjønneland, A., Hansen, L., Boutron-Ruault, M.C., Fagherazzi, G., Bastide, N., Panico, S., Grioni, S., Vineis, P., Palli, D., Tumino, R., Bueno-de-Mesquita, H.B., Peeters, P.H., Skeie, G., Engeset, D., Parr, C.L., Jakszyn, P., Sánchez, M.J., Barricarte, A., Amiano, P., Chirlaque, M., Quirós, J.R., Sund, M., Werner, M., Sonestedt, E., Ericson, U., Key, T.J., Khaw, K.T., Ferrari, P., Romieu, I., Riboli, E., and Jenab, M. (2013) Consumption of fish and meats and risk of hepatocellular carcinoma: the European Prospective Investigation into Cancer and Nutrition (EPIC). *Ann. Oncol.*, **24**,2166–2173.

26. Akuta, N., Suzuki, F., Kobayashi, M., Hara, T., Sezaki, H., Suzuki, Y., Hosaka, T., Kobayashi, M., Saitoh, S., Ikeda, K., and Kumada, H. (2014) Correlation Between Hepatitis B Virus Surface antigen Level and Alpha-Fetoprotein in Patients Free of Hepatocellular Carcinoma or Severe Hepatitis. *J. Med. Virol.*, **86**,131–138.

27. Kanazir, M., Boricic, I., Delic, D., Tepavcevic, D.K., Knezevic, A., Jovanovic, T., and Pekmezovic, T. (2010) Risk factors for hepatocellular carcinoma: A case-control study in Belgrade (Serbia). *Tumori*, **96**,911–917.

28. Zheng, Z., Zhou, L., Gao, S., Yang, Z., Yao, J., and Zheng, S. (2013) Prognostic role of C-reactive protein in hepatocellular carcinoma: A systematic review and meta-analysis. *Int. J. Med. Sci.*, **10**,653–664.

29. Metz, C.D. (1978) Basic Principles of ROC Analysis. *Semin. Nucl. Med.*, **8**,283–298.

30. Vermont, J., Bosson, J.L., François, P., Robert, C., Rueff, A., and Demongeot, J. (1991) Strategies for graphical threshold determination. *Comput. Methods Programs Biomed.*, **35**,141–150.

31. Biesheuvel, C.J., Vergouwe, Y., Oudega, R., Hoes, A.W., Grobbee, D.E., and Moons, K.G.M. (2008) Advantages of the nested case-control design in diagnostic research. *BMC Med. Res. Methodol.*, **8**,48.

32. Van Zaane, B., Vergouwe, Y., Donders, a R.T., and Moons, K.G.M. (2012) Comparison of approaches to estimate confidence intervals of post-test probabilities of diagnostic test results in a nested case-control study. *BMC Med. Res. Methodol.*, **12**,166.

33. Stepien, M., Duarte-Salles, T., Fedirko, V., Floegel, A., Kumar-Barupal, D., Rinaldi, S., Achaintre, D., Assi, N., Tjønneland, A., Overvad, K., Bastide, N., Boutron-Ruault, M.-C., Severi, G., Kuhn, T., Kaaks, R., Aleksandrova, K., Boeing, H., Trichopoulou, A., Bamia, C., Lagiou, P., Saieva, C., Agnoli, C., Panico, S., Tumino, R., Naccarati, A., Bueno-de-Mesquita, H.B., Peeters, P.H., Weiderpass, E., Quirós, J.R., Agudo, A., Sanchez, M.-J., Dorronsoro, M., Gavrila, D., Barricarte, A., Ohlsson, B., Sjoberg, K., Werner, M., Sund, M., Wareham, N., Khaw, K.-T., Travis, R.C., Schmidt, J.A., Gunter, M., Cross, A.J., Vineis, P., Romieu, I., Scalbert, A., and Jenab, M. (2015) Alteration of Amino Acid and Biogenic Amine Metabolism in Hepatobiliary Cancers: Findings from a Prospective Cohort Study. *Submitt. to Int. J. Cancer*,.

34. Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B., and Wei, L.J. (2011) On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Stat. Med.*, **30**,1105–1117.

35. Valeri, L. and Vanderweele, T.J. (2013) Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol. Methods*, **18**,137–150.

36. Vanderweele, T.J. and Vansteelandt, S. (2010) Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol.*, **172**,1339–1348.

37. R Foundation for Statistical Computing and R Core Team. (2013) R: A language and environment for statistical computing.

38. SAS Institute Inc. and Cary, N. (2012) Base SAS® 9.4 Procedures Guide.

39. College Station TX : StataCorp. (2011) Stata Statistical Software: Release 12.

40. Abdi, H. (2010) Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdiscip. Rev. Comput. Stat.*, **2**,97–106.

41. Yang, W.-S., Va, P., Bray, F., Gao, S., Gao, J., Li, H.-L., and Xiang, Y.-B. (2011) The role of pre-existing diabetes mellitus on hepatocellular carcinoma occurrence and prognosis: a meta-analysis of prospective cohort studies. *PLoS One*, **6**,e27326.

42. Gomaa, A.-I. (2008) Hepatocellular carcinoma: Epidemiology, risk factors and pathogenesis. *World J. Gastroenterol.*, **14**,4300–4308.

43. Cogliano, V.J., Baan, R., Straif, K., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Freeman, C., Galichet, L., and Wild, C.P. (2011) Preventable exposures associated with human cancers. *J. Natl. Cancer Inst.*, **103**,1827–1839.

44. Bialecki, E.S. and Di Bisceglie, A.M. (2005) Diagnosis of hepatocellular carcinoma. *HPB (Oxford).*, **7**,26–34.

45. Wentzensen, N. and Wacholder, S. (2013) From differences in means between cases and controls to risk stratification: a business plan for biomarker development. *Cancer Discov.*, **3**,148–157.

46. Leong, T.Y.-M. and Leong, A.S.-Y. (2005) Epidemiology and carcinogenesis of hepatocellular carcinoma. *HPB (Oxford).*, **7**,5–15.

47. Rothwell, J., Fillâtre, Y., Martin, J.-F., Lyan, B., Pujos-Guillot, E., Fezeu, L., Hercberg, S., Comte, B., Galan, P., Touvier, M., and Manach, C. (2014) New biomarkers of coffee consumption identified by the non-targeted metabolomic profiling of cohort study subjects. *PLoS One*, **9**,e93474.

48. Guo, M., Zhao, B., Liu, H., Zhang, L., Peng, L., Qin, L., Zhang, Z., Li, J., Cai, C., and Gao, X. (2014) A Metabolomic Strategy to Screen the Prototype Components and Metabolites of Shuang-Huang-Lian Injection in Human Serum by Ultra Performance Liquid Chromatography Coupled with Quadrupole Time-of-Flight Mass Spectrometry. *J. Anal. Methods Chem.*, **2014**,241505.

49. Anderson, T.W. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Stat.*, **22**,327–351.

50. Izenman, A.J. (1975) Reduced-Rank Regression for the Multivariate Linear Model. *J. Multivar. Anal.*, **5**,248–264.

51. Aldrin, M. (2002) Reduced-rank regression. In: El-Shaarawi AH, Piegorsch WW, editors. Encyclopedia of Environmetrics. Chichester: John Wiley & Sons, Ltd. pp. 1724–1728.

52. Hotelling, H. (1936) Relations between two sets of variates. *Biometrika*, **28**,321–377.

53. Navratil, V., Pontoizeau, C., Billoir, E., and Blaise, B.J. (2013) SRV : an opensource toolbox to accelerate the recovery of metabolic biomarkers and correlations from metabolic phenotyping data sets. *Bioinformatics*, **29**,1348–1349.

54. Wishart, D.S. (2008) Quantitative metabolomics using NMR. *TrAC Trends Anal. Chem.*, **27**,228–237.

55. Psychogios, N., Hau, D.D., Peng, J., Guo, A.C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., Gautam, B., Young, N., Xia, J., Knox, C., Dong, E.,

Huang, P., Hollander, Z., Pedersen, T.L., Smith, S.R., Bamforth, F., Greiner, R., McManus, B., Newman, J.W., Goodfriend, T., and Wishart, D.S. (2011) The human serum metabolome. *PLoS One*, **6**,e16957.

Legends to figures

Figure 1: General scheme of the analytical framework developed in the study. A PC-PR2 analysis is carried out beforehand to identify relevant sources of variation. In the PLS model the X- and Y- sets are related to each other, and scores are computed (1). X- and Y-scores are, in turn, associated to a case-control indicator of HCC status in conditional logistic regression models (2). A mediation analysis is carried out to explore the role of metabolomics in the association between lifestyle factors and risk of HCC (3).

Figure 2: PC-PR2 analysis results* identifying the sources of variability in the NMR data (panel A) and in the lifestyle data (panel B).

* 17 and 14 components were retained to account for 80 % (threshold used) of total NMR (A) and lifestyle variability (B), respectively. The R2 value represents the amount of variability in NMR / lifestyle variable explained by the ensemble of investigated predictors.

Figure 3: Variable importance plot (VIP) displaying the variable importance for projection statistic of the predictor variables for the PLS analyses.

Panel **A**: Results from the main PLS model run on all observations (N=336, X-set=21, Y-set=285).

Panel **B**: Results from the PLS sensitivity analysis run on a subsample (N=271, 92 cases, 179 controls) excluding sets where cases were diagnosed within the first two years of follow-up (X-set=21, Y-set=285).

765 The horizontal line corresponds to Wold's criterion (0.8), the threshold used to rule if a
766 variable has an important contribution to the construction of the Y variables (see
767 **Mathematical Appendix** for further details).

Table 1: Individual and cumulative variation (%) explained by the first 3 PLS factors in 21 lifestyle (X-set) and 285 NMR (Y-set) variables.

# of PLS	Lifestyle Variables		NMR Variables	
Factors	Individual	Cumulative	Individual	Cumulative
1	6.17	-	5.51	-
2	6.23	12.40	2.38	7.89
3	9.27	21.67	0.59	8.48

Table 2: Lifestyle and NMR cluster variables contributing to each of the 3 PLS factors (N=336, X-set=21, Y-set=285).

PLS Factor	Lifestyle Variable*	Loading value	CS*‡ (ppm)	Metabolite**	Loading value
1	Dairy Products	0.28	5.22		-0.06
	Cakes and Biscuits	0.32	3.88		-0.05
	Lifetime Alcohol Consumption	-0.25	3.82		-0.06
	Smoking Status	-0.39	3.76	Glucose	-0.06
	Diabetes	-0.63	3.71		-0.05
			3.54		-0.05
			3.50		-0.07
			3.48		-0.07
			3.44	Acetoacetate	-0.07
			3.23	Choline + Glycerphosphocholine	-0.04
			3.01	Lysine	0.10
			2.94	Albumin	0.10
			2.65	Aspartate	0.10
			2.42	Glutamine	0.10
			2.28	Acetoacetate	0.10
			2.22	CH_2-C**H**$_2$-COOC bond of lipids + Acetone	-0.04
			1.86	Lysine	0.09
			1.87		0.10
			1.53	C**H**$_2$-CH_2-COOC bond of lipids	-0.03
2	Cereal and Cereal Products	-0.16	7.17	Tyrosine	0.13
	Height	-0.34	6.87		0.13
	Education Level	-0.26	5.27	C**H**=C**H** bond of lipids	-0.13
	Hepatitis	0.49	5.22	Glucose	0.16
			5.18	Mannose + Lipid O-CH2	-0.12
			4.27	Lipid O-CH2	-0.12
			4.25	Threonine	-0.14
			4.07	Choline + Lipid O-CH2 + Myo-inositol	-0.12
			4.05	Creatinine	-0.14
			3.88		0.15
			3.82		0.16
			3.76		0.15
			3.71	Glucose	0.15
			3.54		0.15
			3.50		0.16
			3.48		0.16
			3.44	Acetoacetate	0.16
			3.23	Choline + Glycerphosphocholine	0.15
			2.80	Aspartate	-0.12
			2.22	CH_2-C**H**$_2$-COOC bond of lipids + Acetone	-0.11
			2.19	CH_2-C**H**$_2$-COOC bond of lipids	-0.15
			2.02	Proline + Glutamate + CH2=C bonds of lipids	-0.13
			1.53	C**H**$_2$-CH_2-COOC bond of lipids	-0.13
			1.25	C**H**$_2$ bond of lipids	-0.12
			0.86	Cholesterol + CH3 bond of lipids	-0.12
3	Vegetables	-0.42	7.32	Phenylalanine	0.11
	Lifetime Alcohol Consumption	0.29	5.22	Glucose	-0.13
	Smoking Status	0.25	4.28	Lipid O-C**H**$_2$	0.11
	Hepatitis	0.26	3.88		-0.11
			3.82		-0.11
			3.76	Glucose	-0.12
			3.71		-0.11
			3.69		-0.11
			3.63	Myo-inositol	0.16
			3.50	Glucose	-0.13
			3.48		-0.12
			3.44	Acetoacetate	-0.12
			3.35	Proline	0.11
			3.33		0.13
			3.28	Myo-inositol	0.12
			3.23	Choline + Glycerphosphocholine	-0.12
			2.80	Aspartate	-0.13
			2.76	part of =CH-CH2-CH= bond of lipids	-0.13
			2.35	Proline + Glutamate	0.12
			2.33		0.13

		1.20	3-hydroxybutyrate + CH2 bond of lipids	0.11
		1.16	Ethanol	0.15
		0.66	Cholesterol	0.11

*Relevant lifestyle and NMR variables contributing to each PLS factor selected based on their associated loading values <10th percentile (pctl) and >90th pctl or <5th pctl and >95th pctl respectively.
‡ CS: ^{1}H chemical shift (in ppm) of the cluster (center value).
**Some of the identified clusters were found to be background noise during the annotation phase and were removed from this table.

Table 3: HCC odds ratios* and 95% confidence interval (OR, 95% CI) associated with the lifestyle (X-set) and the NMR clusters (Y-set) PLS scores in the main analysis (N=336, X-set=21, Y-set=285).

PLS Lifestyle Variables X-scores			PLS NMR Variables Y-scores		
Factor	**OR** (95% CI)**	**P-Wald†**	**Factor**	**OR** (95% CI)**	**P-Wald†**
1	0.77 (0.58, 1.02)	0.07	1	0.96 (0.91, 1.01)	0.09
2	1.54 (1.06, 2.25)	0.02	2	1.11 (1.02, 1.22)	0.02
3	1.37 (1.05, 1.79)	0.02	3	1.22 (1.04, 1.44)	0.01

*Models were adjusted for C-reactive protein concentration, alpha-fetoprotein concentration and a composite score for liver damage. Cases and controls were matched on age at blood collection (± 1 year), sex, study centre, date (± 2 months) and time of the day at blood collection (± 3 hours), fasting status at blood collection (<3/3-6/>6 hours); among women, additional matching criteria included menopausal status (pre-/peri-/postmenopausal) and hormone replacement therapy use at time of blood collection (yes/no). ** ORs expressing the change in HCC risk associated to 1-SD increase in the score. † Wald's test was for continuous exposure compared with a Chi-square distribution with 1 degree of freedom (dof).

Table 4: Area under the curve (AUC), sensitivity, specificity, accuracy and positive predictive value (PPV) of ROC models (with 95% CI), from the main PLS analysis (N=336, X-set=21, Y-set=285).

	AUC	AUC_b**	Sensitivity	Specificity	Accuracy	PPV
Adjustment Covariates (ADJ)*	0.842 (0.794, 0.891)	0.821 (0.766, 0.868)	0.752 (0.662, 0.829)	0.802 (0.743, 0.852)	0.785	0.0015
X1 scores + ADJ	0.846 (0.797, 0.894)	0.825 (0.766, 0.875)	0.743 (0.653, 0.821)	0.838 (0.783, 0.884)	0.806	0.0018
X1+X2 scores + ADJ	0.854 (0.808, 0.900)	0.831 (0.772, 0.881)	0.743 (0.653, 0.821)	0.824 (0.768, 0.872)	0.797	0.0017
X1+X2+X3 scores + ADJ	0.859 (0.811, 0.907)	0.836 (0.778, 0.887)	0.796 (0.710, 0.866)	0.788 (0.729, 0.840)	0.791	0.0015
Y1 scores + ADJ	0.841 (0.793, 0.890)	0.817 (0.760, 0.865)	0.735 (0.643, 0.813)	0.820 (0.763, 0.868)	0.791	0.0016
Y1+Y2 scores + ADJ	0.845 (0.795, 0.894)	0.820 (0.762, 0.872)	0.735 (0.643, 0.813)	0.851 (0.798, 0.895)	0.812	0.0020
Y1+Y2+Y3 scores + ADJ	0.853 (0.804, 0.902)	0.827 (0.771, 0.877)	0.726 (0.634, 0.805)	0.883 (0.833, 0.922)	0.890	0.0025

*The model is run on the adjustment covariates (ADJ) including the C-reactive protein concentration, alpha-fetoprotein concentration and a composite score for liver damage. ** AUC_b is the bootstrapped-cross validated estimate of the AUC. X1, X2 and X3 are the lifestyle component scores of the first, second and third PLS factors, respectively. Y1, Y2, and Y3 are the metabolomics component of the first, second and third PLS factors, respectively.

Table 5: HCC odds ratios* and 95% confidence intervals (OR, 95%CI) associated with the lifestyle (X-set) and the NMR clusters (Y-set) PLS scores. Results from the sensitivity analysis (N=271, 92 cases, 179 controls) conducted excluding sets where cases were diagnosed within the first two years of follow-up (X-set=21, Y-set=285).

PLS Lifestyle Variables X-scores			PLS NMR Variables Y-scores		
Factor	**OR** (95% CI)**	**P-Wald†**	**Factor**	**OR** (95% CI)**	**P-Wald†**
1	0.80 (0.60, 1.08)	0.15	1	0.96 (0.94, 1.04)	0.56
2	1.56 (1.02, 2.40)	0.04	2	1.18 (1.03, 1.36)	0.02
3	0.86 (0.67, 1.11)	0.26	3	0.86 (0.73, 0.99)	<0.05

*Models were adjusted for C-reactive protein concentration, alpha-fetoprotein concentration and a composite score for liver damage. Cases and controls were matched on age at blood collection (± 1 year), sex, study centre, date (± 2 months) and time of the day at blood collection (± 3 hours), fasting status at blood collection (<3/3-6/>6 hours); among women, additional matching criteria included menopausal status (pre-/peri-/postmenopausal) and hormone replacement therapy use at time of blood collection (yes/no). ** ORs expressing the change in HCC risk associated to 1-SD increase in the score. † Wald's test was for continuous exposure compared with a Chi-square distribution with 1 degree of freedom (dof).

Table 6: Results from the mediation analysis (N= 336, X-set=21, Y-set=285): Natural Indirect Effect (NIE) and 95%CI*.

Model**				Natural Indirect Effect (NIE)	
Exposure (A)	Mediator (M)	Outcome	A*M interaction term	Estimate (95%CI)	p-value
X1 score	Y1 score	HCC	No	0.91 (0.77, 1.06)	0.23
X2 score	Y2 score	HCC	No	1.11 (0.97, 1.25)	0.12
X3 score	Y3 score	HCC	No	1.08 (0.94, 1.23)	0.28
X1 score	Y1 score	HCC	Yes	0.96 (0.79, 1.17)	0.70
X2 score	Y2 score	HCC	Yes	1.15 (1.01, 1.31)	0.04
X3 score	Y3 score	HCC	Yes	1.13 (1.01, 1.28)	0.04

* The standard errors used to compute the 95%CI were obtained using the delta method.
**Models were adjusted for the C-reactive protein concentration, alpha-fetoprotein concentration and a composite score for liver damage, as well as for the other Y-scores, as potential mediator-outcome confounders. Additionally, the outcome and the mediator models were adjusted for centre and age at blood collection.

Supplementary Tables

A Statistical framework to model the meeting-in-the-middle principle using metabolomic data: application to hepatocellular carcinoma in the EPIC study.

Supplementary Table 1: Summary statistics of the predictors variables (X-set) of the study subjects in the EPIC liver nested case–control study (N=336, 114 Cases, 222 Controls).

	Mean / N*	sd / %*	p5	p95	N missing
Dietary Variables (g/day)					
Potatoes and other tubers	100.57	78.15	9.34	266.97	0
Vegetables	194.20	143.22	45.03	473.45	0
Legumes	9.85	18.03	0.00	41.18	0
Fruits, nuts and seeds	232.80	197.94	23.55	585.22	0
Dairy products	334.40	261.46	49.92	777.48	0
Cereal and cereral products	227.04	117.67	76.39	458.94	0
Meat and meat products	115.97	62.29	37.83	236.32	0
Fish and shellfish	32.88	32.26	3.78	81.43	0
Egg and egg products	18.67	18.72	1.88	55.57	0
Fat	34.61	18.48	11.01	70.76	0
Sugar and confectionary	47.26	51.51	1.93	138.73	0
Cakes and biscuits	41.33	49.68	0.00	147.26	0
Non-alcoholic beverages	1053.91	793.31	85.00	2391.90	0
Anthropometric variables					
BMI (kg/m2)	27.41	4.41	21.22	36.16	0
Height (cm)	169.70	9.99	152.00	184.80	0
Lifestyle Variables					
Lifetime alcohol intake (g/day)	23.27	41.38	0	91.998	61
Physical activity (Mets/h)	77.13	49.45	11.5	173.63	0
Highest Education Level					
None or primary school completed	167	49.7	-	-	-
Technical/professional school	75	22.32	-	-	-
Secondary school	27	8.04	-	-	-
Longer education (incl. university degree)	62	18.45	-	-	-
Unspecified or Unknown	5	1.49	-	-	-
Smoking status					
Never	124	36.9	-	-	-
Former	125	37.2	-	-	-
Current smoker	85	25.3	-	-	-
Unspecified or Unknown	2	0.6	-	-	-
Pathology variables indicative of lifestyle					
Hepatitis status					1
No	291	86.87	-	-	-
Yes	44	13.13	-	-	-
Diabetes					0
No	307	91.37	-	-	-
Yes	29	8.63	-	-	-

*Mean and standard deviation (sd), were reported for continuous variables and frequencies and percentages (%) were reported for categorical variables.
p5: 5th percentile, p95:95th percentile.

Supplementary Table 2: Results from the sensitivity analysis run on a subsample (N=271, 92 cases, 179 controls) excluding sets where cases were diagnosed within the first two years of follow-up (X-set=21, Y-set=285). Lifestyle and NMR cluster variables contributing to each PLS factor.

PLS Factor	Lifestyle Variable*	Loading value	CS*‡ (ppm)	Metabolite**	Loading value
1	Dairy Products	0.33	7.03	Histidine	0.09
	Cakes and Biscuits	0.34	5.22	Glucose	-0.07
	Lifetime Alcohol Consumption	-0.34	3.88		-0.06
	Smoking Status	-0.26	3.82		-0.07
	Diabetes	-0.59	3.76		-0.06
			3.71		-0.06
			3.54		-0.05
			3.50		-0.07
			3.48		-0.08
			3.44	Acetoacetate	-0.08
			3.23	Choline + Glycerphosphocholine	-0.05
			3.03	Creatine	0.10
			3.01	Albumin	0.10
			2.28	Acetoacetate	0.10
			2.22	CH_2-C**H**$_2$-COOC bond of lipids + Acetone	-0.03
			2.06	Proline + Glutamate	0.09
			1.91	Lysine + Arginine	-0.03
			1.87	Lysine	0.09
			1.16	Ethanol	-0.04
			1.08	Unknown 1	0.09
			0.91	C**H**$_3$ bond of lipids	0.09
2	Cereal and Cereal Products	-0.24	7.17	Tyrosine	0.14
	BMI	0.34	6.87		0.14
	Height	-0.39	5.27	C**H**=C**H** bond of lipids	-0.14
	Hepatitis	0.55	5.22	Glucose	0.13
			5.18	Mannose + Lipid O-CH2	-0.13
			4.27	Lipid O-C**H**$_2$	-0.12
			4.25	Threonine	-0.14
			4.05	Creatinine	-0.14
			3.88	Glucose	0.13
			3.82		0.13
			3.76		0.13
			3.75		0.12
			3.71		0.12
			3.54		0.15
			3.50		0.13
			3.48		0.13
			3.44	Acetoacetate	0.13
			3.23	Choline + Glycerphosphocholine	0.12
			2.80	Aspartate	-0.13
			2.76	=CH-CH2-CH= bond of lipids	-0.12
			2.19	CH_2-C**H**$_2$-COOC bond of lipids	-0.16
			2.02	Proline + Glutamate	-0.14
			1.53	C**H**$_2$-CH_2-COOC bond of lipids	-0.13
			1.25	C**H**$_2$ bond of lipids	-0.12
			0.86	Cholesterol + CH3 bond of lipids	-0.12
3	Vegetables	0.39	5.25	Glucose	0.17
	Sugar and Confectionnary	-0.21	4.28	Lipid O-CH2	-0.07
	Lifetime Alcohol Consumption	-0.29	4.14	Proline	-0.08
			4.07	Choline + Lipid O-CH2 + Myo-inositol	-0.07
			3.88	Glucose	0.16
			3.82		0.16
			3.76		0.16
			3.75		0.14
			3.71		0.15
			3.69		0.16
			3.63	Myo-inositol	-0.16

	3.54		0.12
	3.50	Glucose	0.17
	3.48		0.17
	3.44	Acetoacetate	0.16
	3.41		-0.10
	3.35	Proline	-0.15
	3.34		-0.12
	3.28	Myo-inositol	-0.09
	3.23	Choline + Glycerphosphocholine	0.15
	1.91	Lysine + Arginine	-0.07
	1.16	Ethanol	-0.16
	0.68	Cholesterol	-0.06
	0.66		-0.08

*Relevant lifestyle and NMR variables contributing to each PLS factor selected based on their associated loading values <10th percentile (pctl) and >90th pctl or <5th pctl and >95th pctl respectively.
‡ CS: ^{1}H chemical shift (in ppm) of the cluster (center value).
**Some of the identified clusters were found to be background noise during the annotation phase and were removed from this table.

Supplementary Table 3: Results from the sensitivity analysis (N=271, 92 cases, 179 controls) conducted excluding sets where cases were diagnosed within the first two years of follow-up (X-set=21, Y-set=285). Area under the curve (AUC), sensitivity, specificity, accuracy and positive predictive value (PPV) of ROC models (with 95% CI).

	AUC	**AUC$_b$****	**Sensitivity**	**Specificity**	**Accuracy**	**PPV**
Adjustment Covariate (ADJ)*	0.846 (0.793, 0.899)	0.827 (0.765,0.879)	0.750 (0.649, 0.834)	0.838 (0.776, 0.889)	0.808	0.0018
X1 scores + ADJ	0.853 (0.800, 0.905)	0.834 (0.774, 0.890)	0.728 (0.626, 0.816)	0.872 (0.813, 0.917)	0.823	0.0023
X1+X2 scores + ADJ	0.860 (0.811, 0.910)	0.837 (0.772, 0.893)	0.750 (0.649, 0.834)	0.832 (0.769, 0.884)	0.804	0.0018
X1+X2+X3 scores + ADJ	0.861 (0.810, 0.912)	0.837 (0.773, 0.893)	0.761 (0.661, 0.844)	0.838 (0.776, 0.889)	0.812	0.0019
Y1 scores + ADJ	0.847 (0.794, 0.900)	0.827 (0.768, 0.884)	0.739 (0.637, 0.825)	0.838 (0.776, 0.889)	0.804	0.0018
Y1+Y2 scores + ADJ	0.848 (0.794, 0.901)	0.827 (0.764, 0.883)	0.717 (0.614, 0.806)	0.899 (0.846, 0.939)	0.838	0.0028
Y1+Y2+Y3 scores + ADJ	0.853 (0.800, 0.907)	0.826 (0.763, 0.882)	0.717 (0.614, 0.806)	0.911 (0.859, 0.948)	0.845	0.0032

*The model is run on the adjustment covariates (ADJ) including the C-reactive protein concentration, alpha-fetoprotein concentration and a composite score for liver damage. ** AUC$_b$ is the bootstrapped-cross validated estimate of the AUC. X1, X2 and X3 are the lifestyle component scores of the first, second and third PLS factors, respectively. Y1, Y2, and Y3 are the metabolomics component of the first, second and third PLS factors, respectively.