Exploitation de méthodes biostatistiques factorielles pour l'investigation de la relation nutrition-cancer dans la cohorte Européenne sur le Cancer et la Nutrition (EPIC)

Nada Assi

N°d'ordre NNT : 2016LYSE1185

THESE de DOCTORAT DE L'UNIVERSITE DE LYON

Université Claude Bernard Lyon 1
Ecole Doctorale Interdisciplinaire Sciences-Santé - EDISS (ED205)

Spécialité de doctorat :
Epidémiologie, santé publique, recherche sur les services de santé
Discipline :
Biostatistiques

Soutenue publiquement le 19/10/2016 par

Mlle Nada ASSI

"Use of factorial biostatistical methods to investigate the relation between nutrition and cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC) study"

Devant le jury composé de :

MICHIELS, Stefan, phD (Président du jury)
Directeur équipe : « Oncostat, méthodologie et épidémiologie clinique en oncologie moléculaire », CESP, Institut Gustave Roussy, Villejuif (France)

TZOULAKI, Ioanna, phD (Rapporteur)
Senior Lecturer, Department of Epidemiology and Biostatistics, Imperial College, London (UK)

SEVERI, Gianluca, phD (Rapporteur)
Directeur de Recherche, Inserm (U1018), Université Paris-Sud et Université Paris Saclay
CESP, Institut Gustave Roussy, Villejuif (France)

BELLOCO, Rino, phD (Examinateur)
Full Professorship portfolio, Ministry of Education, University and Research (Italy)
Associate professor, Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan (Italy)
Associate professor, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm (Sweden)

FERVERS, Béatrice, MD, phD (Examinatrice)
Professeur associé, Université Claude Bernard Lyon I
Directrice, Département « Cancer et Environnement », Centre Léon Bérard, Lyon (France)

PHILIP, Thierry, MD, phD, PU-PH (Directeur de thèse)
Professeur, Membre du laboratoire « Santé-Individu-Société » (EAM 4128), Université Claude Bernard Lyon I
Centre Léon Bérard, Lyon (France)

CHAJES, Véronique, phD (Co-directrice de thèse)
Chercheur au Centre International de Recherche sur le Cancer, Lyon (France)

FERRARI, Pietro, phD (Co-directeur de thèse)
Directeur équipe : « Nutritional Methodology and Biostatistics »,
Chercheur au Centre International de Recherche sur le Cancer, Lyon (France)

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. X. PERROT
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Résumé en Français

La nutrition est un facteur de risque modifiable pour le cancer. Il est estimé qu'un tiers des cas pourraient être évités en adoptant une meilleure alimentation en adéquation avec les recommandations les plus récentes. La relation entre nutrition et cancer est complexe, et son étude est enrichie par les nouveaux défis apportés par les récentes avancées technologiques dans le domaine des « -omiques » auxquels elle doit répondre. Des approches analytiques combinant des informations provenant de questionnaires alimentaires avec ceux de biomarqueurs et de la métabolomique sont actuellement la cible de nombreuses recherches.

Cette thèse avait pour but de développer de nouvelles approches biostatistiques afin d'étudier la relation entre nutrition et cancer au sein de la cohorte EPIC. Pour ce faire, l'applicabilité de nouvelles méthodologies, principalement factorielles, a été étudiée.

Une nouvelle méthode multivariée pour la réduction de la dimensionnalité, le Treelet Transform (TT), a été examinée afin d'extraire des patterns de nutriments issus de questionnaires. Les patterns ainsi obtenus étaient facilement interprétables puisque le TT est un bon compromis entre analyse en composante principale et clustering hiérarchique.

Ensuite, un cadre analytique pour implémenter le concept du « meeting-in-the-middle » (MITM) a été développé et appliqué dans deux études cas-témoin nichées sur le cancer hépatocellulaire avec des données métabolomiques, ciblé et non-ciblée. Le MITM cherche à identifier des biomarqueurs qui soient à la fois des marqueurs de certaines expositions passées et de conditions pathologiques. L'implémentation s'est focalisée sur l'application de la méthode des moindres carrés partiels (PLS) et de l'analyse de médiation. Des signaux métaboliques qui médiaient la relation des expositions vers le cancer ont été identifiés.

Enfin, nous avons examiné la relation entre les niveaux plasmatiques de 60 acides gras issus de biomarqueurs et le risque de cancer du sein dans une étude cas-témoin nichée dans EPIC. Les résultats issus de cette analyse seront un point de départ pour des développements plus poussés.

Cette thèse servira de base pour des applications épidémiologiques futures examinant la relation nutrition-cancer.

Mots-clefs : Biostatistiques, méthodes multivariées, treelet transform, cancer, nutrition, EPIC, meeting-in-the-middle, PLS, PCA, analyse de médiation

English Abstract

Diet is a modifiable risk factor for many cancers. It has been estimated that about a third of cancer cases can be prevented by complying with a healthy diet and adhering to the recommendations in terms of nutrition. The nutrition-cancer relationship is a complex one, and its study is currently at a turning point with the opportunity and challenges brought by the recent technological advances in the fields of « -omics ». New analytical strategies are being sought to combine and explore information collected through dietary questionnaires, biomarkers along with metabolomic data.

The main objective of this thesis was to develop new biostatistical approaches to investigate the diet-cancer relation within the European Prospective Investigation into Cancer and nutrition (EPIC) study. To this end, the applicability of new methodologies in the field of nutritional epidemiology, mainly multivariate and factorial, has been examined.

First, a new multivariate dimension reduction method, the Treelet Transform (TT) was applied to extract nutrient patterns relying on questionnaire data. The extracted patterns were easily interpretable as TT is a good compromise halfway between principal component analysis and hierarchical clustering.

Then, an analytical framework was conceived for the « meeting-in-the-middle » (MITM) principle and applied to two nested case-control studies on hepatocellular carcinoma, with targeted and untargeted metabolomic data. The MITM aims to identify overlap biomarkers of past exposures that are at the same time predictive of disease outcomes. The implementation focused on the application of partial least squares (PLS) and mediation analyses. Metabolic signatures were identified that mediated the relation from exposures towards cancer risk.

Last, the association between 60 plasma fatty acids levels assessed from biomarkers and breast cancer risk was examined in a nested case-control study in EPIC. Results from this analysis are a stepping stone towards more sophisticated modelling.

This thesis will serve as a basis for future epidemiological applications looking into the nutrition-cancer relation.

Keywords: Biostatistics, multivariate methods, treelet transform, cancer, nutrition, EPIC, meeting-in-the-middle, PLS, PCA, mediation analysis

Acknowledgments

This work was the fruit of three years of the most wonderful learning adventure I have been given thus far. It all started with a fortuitous email of a young master's student looking for an internship that transformed into an incredible opportunity to step inside the world of research and pursue a doctoral thesis. I will be eternally grateful to my supervisor Pietro Ferrari for giving me the chance to prove myself, for his teaching and shared expertise, his valuable day-to-day guidance, and most importantly for his humane qualities and kindness. Thank you for having been a tremendous mentor for me.

I would like to express my sincere gratitude to Thierry Philip, who came on board during the second year, and who was abundantly helpful. Thank you for your open-mindedness and your esteemed advice.

I would like to gratefully acknowledge Véronique Chajès who has been particularly supportive and kind-hearted, gave me insightful counsel and always encouraged me.

Vivian Viallon has my deepest gratitude for his support, his availability and bright advice throughout these years, and for making even the most daunting formulas clear as day.

I would like to thank Gianluca Severi and Ioana Tzoulaki who have kindly accepted to act as external reviewers for this work and I am also grateful for the jury members Stefan Michiels, Béatrice Fervers and Rino Bellocco. I thank them all for their time and for the honour of participating to the defence committee.

Additionally, I would like to thank both Béatrice Fervers and David Cox for their active participation to the thesis follow-up committee overseeing the yearly progress.

I would also like to acknowledge all my financial supports: The EDISS doctoral school that funded this phD through a Université de Lyon doctoral contract at UCBL1, the International Agency for Research on Cancer for completing my doctoral grant, as well as the INCA who funded some of the hepatocellular carcinoma studies I was fortunate to work on.

I would like to warmly thank my colleagues from the Section of Nutrition and Metabolism for their daily support and their patience and for making the workplace a second home. I was very pleased and proud to work alongside all of you, exchanging ideas and learning from you. I looked forward to coming in every morning with the same drive and motivation all because it is the people who make the working environment a great one, not the building or the equipment, not even the reputation. The list of colleagues, past and present, is endless and I am afraid I won't do you justice in forgetting some names but know that I feel truly fortunate and lucky to have met you all. I would like to personally thank each colleague from the NME Section (2013-2016), in the NMB group in particular as well as some other IARC colleagues. Specifically, Amina, Amy, Anne-Sophie, Anouk, Augustin, Behnaz, Bertrand, Carine, Chiara, Diama, Dina, Elom, Elizabeth, Eve, Fiona, Flavie, Graham, Hafed, Heinz, Hwayoung, Isabelle (B and R),

Idlir, Jordi, Julie, Karina, Kayo, Kuanroung, Laura, Latifa, Laure, Magdalena, Marta, Mazda, Michèle, Minkyung, Nadia A, Nicolas T, Olaf, Patricia, Pekka, Robert, Sabina, Sabine, Sahar, Talita, Tracy, Viktoria. And many more.

Special thanks to special people: Alice and Maria-Paula.

This journey would have been completely different had I not been surrounded by my friends and loved ones. Naturally, I am grateful for my friends and their constant support throughout the years, especially my close friends from the Bioinformatics department at INSA, amongst others that I have also met on the University benches. We have been through a lot together and you have been a family away from home for me, and for that, you have my everlasting friendship and consideration. Again in alphabetical order: Amanda, Amandine, Anaïs, Ardi, Aurélie, Azedine, Béryl, Camille, Chiara, Clara, Cindy, Hélène, Julie (x2), Margaux, Marion, Matthieu, Ombeline, Sandra, Vincent and Timothée. And the list goes on.

Last but not least, I want to say a few words of thanks for my family: my parents Marwan and Rita and my brothers Elias and Karim. Without your encouragements, sacrifices and unwavering support, this journey simply wouldn't have seen the light of day. I dedicate this work to you. I love you fiercely.

Before turning the page, I have an enduring thought for all patients who have fought with the big C or are fighting still, to those who survived it and to those who fell in battle. The fight continues on the research front.

A quote for the road, by economist (and later Baruch College professor) Aaron Levenstein: "Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital."

THIS THESIS HAS BEEN PREPARED IN THE FOLLOWING INSTITUTES

- Nutritional Epidemiology Group,
  Section of Nutrition and Metabolism,
  International Agency for Research on Cancer (IARC)
  150 cours Albert Thomas, 69372 Lyon Cedex 08, France

IN COLLABORATION WITH

- EA 4129 Laboratoire "Santé Individu Société",
  Université Claude Bernard Lyon 1, Hospices Civils de Lyon
  Hôtel-Dieu, Place de l'Hôpital,
  Quai des Célestins - 69002 Lyon, France

- Unité cancer et Environnement, UMR INSERM – U 1052,
  Centre Léon Bérard,
  28 Rue Laennec, 69373 Lyon 08 Cedex, France

Résumé substantiel en français

La nutrition est un facteur de risque modifiable pour de nombreux cancers. Environ 35% des cas de cancers pourraient être évités en adoptant une meilleure alimentation en adéquation avec les recommandations les plus récentes. Partant de ce constat, l'épidémiologie nutritionnelle s'est efforcée dans les 30 dernières années d'étudier la relation entre nutrition et cancer, d'appréhender sa complexité et d'en comprendre les mécanismes. Avec les avancées technologiques récentes, notamment dans le domaine de la biologie moléculaire, de nouvelles données dites «-omiques », en particulier les données métabolomiques, ont pu être acquises. Ainsi un nouveau défi s'offre à ce domaine : celui d'allier les nouvelles informations de haute dimensionnalité provenant de la métabolomique aux informations obtenues par des méthodes plus conventionnelles de recueil par questionnaires alimentaires, ainsi qu'avec d'autres biomarqueurs.

Cette thèse avait pour objectif de développer de nouvelles approches biostatistiques dans le but d'étudier la relation entre nutrition et cancer au sein de la cohorte Européenne Prospective sur le Cancer et la nutrition (EPIC). Pour ce faire, l'applicabilité de nouvelles méthodologies, principalement factorielles multivariées, a été étudiée.

Tout d'abord, nous avons appliqué une nouvelle méthode multivariée pour la réduction de la dimensionnalité, le Treelet Transform (TT), afin d'extraire des patterns alimentaires, et nous l'avons comparée à l'Analyse en Composante Principale (PCA) qui est une technique de référence. Cette application a été réalisée dans la sous-cohorte de femmes d'EPIC (n=334 850, dont 11 576 cancers de sein incidents) sur 23 nutriments estimés à partir de questionnaires alimentaires. Ainsi, deux patterns principaux ont été identifiés, pour lesquels l'association avec le risque de développer un cancer du sein (BC) a ensuite été évaluée. Un premier profil apparenté à une consommation élevée en produits d'origine animale a été associé à une augmentation non significative du risque de BC. Un second profil associé à un régime riche en vitamines et minéraux a été relié à une diminution significative du risque de BC. Le TT a produit des résultats comparables à ceux obtenus avec des méthodes plus classiques. Ces patterns étaient plus facilement

interprétables que ceux de la PCA puisque TT permet d'introduire de la sparsité dans les composantes.

Par la suite, nous nous sommes penchés sur des données métabolomiques issues de deux études cas-témoin sur le cancer hépatocellulaire (HCC) nichées dans la cohorte EPIC, avec 114 cas et 222 témoins appariés pour la première et 147 cas et autant de témoins appariés pour la seconde.
Dans la première étude, nous avons développé un cadre analytique pour l'implémentation du concept dit « meeting-in-the-middle » (MITM). L'idée phare du MITM est d'identifier des biomarqueurs qui soient à la fois des marqueurs de certaines expositions passées et qui soient en même temps prédicteurs de conditions pathologiques. Pour ce faire, un ensemble de 21 variables d'expositions « lifestyle » (alimentaires, de mode de vie, anthropométriques) ont été reliées à un set de 285 variables obtenues par résonance magnétique nucléaire (RMN), correspondant à des pics reconstitués, grâce à l'application de la méthode des moindres carrés partiels (PLS). La PLS est une méthode multivariée combinant des aspects de l'ACP avec ceux de la régression linéaire multiple. Elle permet de relier deux sets de données et d'en extraire des composantes dont la covariance est maximale. Les facteurs ainsi obtenus ont été reliés par le biais de leurs scores au risque de HCC par l'intermédiaire de modèles de régression logistique conditionnelle. Enfin, une analyse de médiation a évalué si les profils métaboliques obtenus sont des médiateurs de la relation entre les profils de « lifestyle » et le HCC.

Dans la seconde étude cas-témoins nichée portant cette fois-ci sur la métabolomique ciblée, nous avons pu affiner le cadre statistique mis en place précédemment. Dans un premier temps, nous avons limité le nombre d'expositions à 7 variables provenant d'un indice niveau d'adéquation à un mode de vie sain et nous nous sommes focalisés sur un ensemble de 132 métabolites bien identifiés. Ensuite, après une première analyse PLS générale, nous avons procédé à une analyse de PLS multiple pour obtenir des signatures métaboliques spécifiques à chacune des expositions. Enfin, l'analyse de médiation a été étendue et adaptée à notre design d'étude, et les effets directs et médiés ont été estimés grâce à des modèles de régression logistique conditionnelle.

Le cadre analytique développé lors de ces deux applications pourrait être réutilisé et ajusté aux besoins d'autres études ayant d'autres types de données « -omiques » ou dans des contextes épidémiologiques similaires.

Enfin, nous nous sommes intéressés à une étude cas-témoin sur le BC nichée dans EPIC où 60 mesures d'acides gras (AG) plasmatiques ont été effectuées chez 2 982 cas de BC invasifs et autant de témoins appariés. L'association entre chacun des AG et le risque de BC a été évaluée à travers des régressions logistiques conditionnelles multivariables ajustées. Ces analyses ont été combinées à une correction pour les tests multiples afin de préserver la valeur nominale de significativité des tests statistiques. Ainsi, des niveaux trop élevés en acide palmitoléique et un indice de désaturation DI_{16} fort ont été associés à une augmentation du risque de BC. Cette étude est l'une des plus larges à cette date se basant exclusivement sur des biomarqueurs en ce qui concerne les expositions des AG, avec une bonne séparation pour AG *trans* d'origine animale de ceux d'origine industrielle. Elle constitue une première étape dans des analyses plus poussées à venir, notamment des analyses de patterns afin de caractériser le lipidome ainsi qu'une possible application du MITM.

Les différentes applications et développements statistiques mis en place lors ce travail de thèse viennent répondre à un besoin d'approches dites holistiques qui visent à intégrer des données de natures différentes et de haute dimension. Cette prise en compte des différents facteurs d'expositions et de risques permettra à l'avenir de mieux appréhender les questions de l'épidémiologie nutritionnelle de nouvelle génération. Cette thèse servira également de base pour des applications multidisciplinaires futures examinant la relation nutrition-cancer.

TABLE OF CONTENTS

List of Tables and Figures

Abbreviations

95%CI	95% Confidence Intervals
BC	Breast Cancer
BMI	Body Mass Index
DQ	Dietary Questionnaires
E3N	Etude Epidémiologique auprès de femmes de l'Education Nationale
ENDB	EPIC Nutrient Database
EPIC	European Prospective Investigation into Cancer and nutrition
ER	Estrogen Receptor
FA	Fatty Acids or Factor Analysis
FDR	False Discovery Rates
Fe	Iron
FFQ	Food Frequency Questionnaires
HLI	Healthy Lifestyle Index
HCC	Hepatocellular carcinoma
HR	Hazard Ratio
MET	Metabolic Equivalent of Task
MITM	Meeting-in-the-Middle
MLR	Multiple Linear Regression
MR	Mendelian Randomisation
MS	Mass Spectrometry
MUFA	Monounsaturated Fatty Acids
NCD	Non-Communicable Diseases

NDE	Natural Direct Effect
NIE	Natural Indirect Effect
NMR	Nuclear Magnetic Resonance
OR	Odds Ratio
PC	Principal Component
PCA	Principal Component Analysis
PLS	Partial Least Squares
PR	Progesterone Receptor
PUFA	Polyunsaturated Fatty Acids
Se	Selenium
TE	Total Effect
TFA	Trans Fatty Acids
TC	Treelet Component
TT	Treelet Transform

CHAPTER I:

INTRODUCTION

Ever since Doll and Peto's comprehensive review of 1981 estimating that 30 to 35% of cancers could be avoided by adopting a better diet in western populations [1], the field of nutritional epidemiology strove to investigate nutritional exposures and their link with individual cancer sites. The initial estimate was characterised by a wide range of uncertainty (from 10 to 70%) [2], and the mechanisms through which specific dietary factors contribute to cancer occurrence are still to be understood. Three decades later, the quantitative estimate remained around 30-40% [3]. It has been argued that obesity and physical inactivity accounted for most of the burden of cancer attributable to nutrition, in a broad sense [4]. There is, however, no consensus around these figures since the extent to which diet adds to the burden of cancer remains difficult to assess [3]. Part of this difficulty is imputed to the lack of knowledge with respect to the stage of carcinogenesis on which many nutritional factors may exert their effects and the dose at which they may achieve their protective or harmful impact [5]. Nevertheless, nutritional epidemiology in the past decades has amassed a growing body of evidence establishing diet as an important modifiable risk factor for a substantial proportion of cancers, making it a great public health target for prevention [3,6,7]. Studies in nutrition provided substantial, yet often inconsistent, epidemiologic evidence of the diet-cancer link [7,8] with findings on alcohol consumption [6,9–23], obesity and weight change [24–28], fat intake [29–39], meat consumption [29,30,40–48], plant foods [49–52], glycaemic index/load [53,54], coffee [55–57], *inter alia*. In addition, these studies have canvassed the relationships between a selection of dietary constituents and molecularly [58,59] or anatomically [50,60,61] defined subsets of cancer, and evaluated dietary behaviours in relation to cancer [62] and cancer survival [38].

Nutritional epidemiology is an intricate area due to the fact that diet is not a single simple exposure but rather a complex set of many variables, characterised by profound inter-correlations between dietary constituents. These inter-correlations may arise from food composition, behavioural patterns, e.g. food items are often consumed together, or from differences in the energy balance and total energy intake as people eating a high-energy diet tend to eat a lot of different nutrients [63]. Disentangling the separate effects of each food/nutrient is extremely challenging, largely because of confounding and residual confounding [64]. Adding to the methodologic and conceptual complexity are the potential physiological interactions amongst nutrients, e.g. Selenium (Se) and Vitamin E, Vitamin C and Iron (Fe), including food component synergies or

antagonisms [65–68]. Furthering the nutrient assessment challenge is the common exposure misclassification. In fact, nutritional epidemiology relies on dietary assessment instruments, mainly questionnaires such as food frequency questionnaires or dietary histories, which are subject to random and systematic measurement errors [69]. These errors are frequent in self-reported dietary estimates as a consequence of study subjects' consistent underestimation or overestimation of their dietary intakes.

Traditional approaches initially relied on simple models to evaluate the associations between single dietary constituents, i.e. foods or nutrients, possibly involving statistical adjustment by total energy intake to ensure iso-caloric comparisons [70], and the risk of disease [63]. These models were straightforward to interpret but did not necessarily capture the inherent complexity of individuals' dietary habits, where simultaneous variability of many foods is observed. Approaches became progressively more complex moving towards multivariable models that accounted for more dietary and lifestyle confounders, at times even involving the inclusion of interaction terms. While these models may better capture the inner sophistication of the diet-disease association, parameters expressing these links are more challenging to interpret. In these models the evaluation of the relation between a given dietary exposure and disease is conditional on all other confounders included in the linear predictor, and it is assumed that they remain constant. This turns out to be an unrealistic assumption that does not factor in the dynamism of an intricate system of synergies between foods, nutrients and other lifestyle variables [67,71,72]. The rigorous analysis consistently struggles to find the optimal trade-off between the two extremes: over-simplistic interpretable models on one hand, and increasingly more multifaceted models that progressively lose their ability to provide a realistic overview of individuals' diet on the other, yet involving statistical challenges for their estimation.

In recent years, research focus of nutritional epidemiology has progressively moved towards dietary pattern analysis and the use of multivariate approaches [71]. Pattern analysis allows for a comprehensive mode taking the full complexity of diet into consideration [73]. Two main strategies are often applied: *a priori* hypothesis-driven patterns and *a posteriori* data-driven patterns [74,75]. *A priori* techniques often use predefined criteria based on specific health outcomes to construct dietary scores reflecting the degree to which a person adheres to given dietary patterns [67,71]. These include compliance with guidelines or recommendations such as the WCRF/AICR score

[76] and the healthy eating index (HEI) [77], characteristics of established diets such as the Mediterranean diet [78–84], or even agreement with dietary aspects of a more general healthy lifestyle [85,86]. *A priori* techniques have seen a shift from adherence to a purely dietary predefined pattern towards scores embracing lifestyle factors as healthy eating behaviours are often in conjunction with healthy lifestyle practices [86]. *A posteriori* methods rely on data driven methods that often use dimension reduction techniques such as principal component analysis (PCA) or factor analysis (FA) to yield uncorrelated dietary factors based on data covariance or correlation matrices. These analyses have been successful in identifying distinct food/nutrient intake patterns that were related to different cancer endpoints [87–128]. Statistical research is underway to explore novel multivariate techniques that provide solutions with easier interpretation of the components [129,130] and tools to reduce the number of arbitrary steps involved (number of components to retain, threshold for loadings, etc.) [131]. Investigations are ongoing to assess the validity of these approaches, and evaluate whether they may predict disease risk in studies involving populations characterised by heterogeneous dietary habits and different cancer rates [71].

Most of the early results on the role of diet in cancer aetiology stemmed from retrospective case-control studies. These designs however are subject to selection and recall biases [132], making the retrospective studies not the best suited to effectively capture the diet-disease association leading to somewhat inconsistent findings [63,133]. It was suggested that prospective designs were more rigorous and provided a valid solution to minimise methodological biases [3,63,134,135]. Since information on dietary exposure is collected at baseline in cancer-free individuals illness is less likely to affect the recall of dietary habits. In addition, prospective cohorts provide the opportunity to assess diet over time through repeated measurements and to examine its associations with a wide array of diseases with appropriate statistical power, if a sufficiently large number of study subjects is enrolled [135]. If the latter condition applies and if the follow-up is carried out for several years, prospective designs allow the investigation of rare outcomes. The Nurses Health Studies [136] and the EPIC cohort [137,138] were among the first large-scale retrospective cohorts expressly designed to explore the diet-cancer association. In such large sized investigations diet is assessed through the use of structured, self-administered questionnaires [135], which include food frequency questionnaires (FFQs) for estimation on long-term, or habitual, dietary exposure, i.e.

referring to study subjects' diet during a 12-month period preceding its administration. These instruments are then utilized to provide estimations of frequency of consumption, portion sizes and total energy intake. Long-term assessment can be complemented by short-term instruments, which include food diaries, food records, and 24h dietary recalls. These types of assessments are meant to collect deeper aspects of individuals' diet, like, for example, detailed information on portion sizes, timings of meals, recipes and possibly cooking methods [139]. All self-reported dietary instruments rely on the existence of adequate food and nutrient composition databases, to convert food amounts into nutrient and macronutrient contents [135,139]. All dietary assessment methods in large scale epidemiological investigation rely on study participants' ability to recall their diet, and are therefore prone to systematic and random measurement errors [133,139,140]. Measurement errors can be substantial and can, in turn, bias estimates of associations between diet and cancer risk [139–142], and lead to loss of statistical power to detect associations [142]. It has been argued that a large proportion of inconsistencies and null results observed in population-based studies of diet and cancer could be the consequence of poor dietary assessments [143]. One compelling example is the downgrading by the 2003 IARC Handbooks of Cancer Prevention on Fruits and Vegetables [144] and by the 2007 update of the World Cancer Research Fund (WCRF) comprehensive report [145], of the cancer protective role of intakes of fruits and/ or vegetables from 'convincing' to 'probable', depending on the cancer site, which were established in the 1997 WCRF comprehensive review [146].

Research in the field of nutrition has strived to develop better methods to ascertain eating behaviours and their reporting [147–154] and to account for measurement errors in self-reported dietary measurements [155–159]. However, in the absence of an "ideal" reference instrument and in order to obtain "objective" observations of food consumption, the use of biomarkers emerged as a valuable research instrument. This motivated the collection of study subjects' biological material in population based studies [160]. Dietary biomarkers are biochemical indicators that can be viewed as an index of short to long-term dietary intake, of nutrient metabolism or markers of the biological consequences of food intake [161]. Biomarkers have been introduced in cancer epidemiology with the idea of relying on markers of relevant internal dose and markers of biologically effective dose to improve exposure assessment [162]. These markers are also known as "concentration" or "recovery" biomarkers

[139,163]. Other markers classified as "predictive" biomarkers are markers of early response/effect and are used to monitor early changes preceding disease occurrence [160,163]. Last, markers of susceptibility can be used in cancer epidemiology to identify subgroups in the population with greater susceptibility to cancer [139,161–163]. Biomarkers can be quantified in biological samples of serum, blood, plasma, urine. It is recognised that these quantities are also affected by random and systematic measurement errors, but these errors are assumed to be independent of errors associated with self-reported dietary assessments [139,160,163]. As such, they can also be used as a means of validation of dietary instruments to estimate the magnitude of systematic and random errors in questionnaires. Their use in calibration studies of diet/disease association has been advocated but seldom pursued [163]. A great extent of cancer research has developed around biomarkers with studies focusing on their validation [161,164–166], their methodological challenges [161,167], and their use in aetiological models [162,168–172]. The recent technological advancements in high-throughput technologies, particularly in the field of molecular biology, generated a slew of new round of metabolites, which can be acquired in biological samples collected in large-scale epidemiological studies [173,174]. Metabolomics is the branch of "- omics" concerned with the high-throughput identification and quantification of small molecule metabolites present in the human metabolome i.e. the ensemble of all metabolites [175,176]. It provides a complete picture of metabolic status and biochemical events happening within an organism [177]. These data have the potential to bring useful tools to improve our understanding of the role of diet in cancer research [175,178]. Biomarker research supports causal reasoning by linking exposures with disease via mechanisms. This is the premise on top of which the "meeting-in-the-middle" concept was proposed [162]. It aims to find overlap biomarkers that are indicative of a given exposure and that are, at the same time, predictive of disease outcome. This complementary approach sheds light on the mechanisms through which individual dietary (or more generally environmental) exposures diverge towards risk of cancer development by investigating life-course biological pathways using –omics technologies [162]. To achieve this, new statistical methodologies are being developed to provide holistic approaches for the combination of dietary questionnaires and biomarker data to be later used in aetiological models and to tackle the challenges brought on by the –omics data [179]. These data are characterised by high-dimensionality, a correlated

structure and a general lack of *a priori* biological hypotheses resulting in challenges for results interpretability [180]. Methodology that conceives a novel use of statistical tools, vastly relying on existing methods, has been developed to analyse this new wave of overwhelming and promising data [179,181–187]. These range from standard procedures of metabolome-wide association studies (MWAS) operated through adequate multiple statistical regression models coupled with multiple testing corrections to multivariate dimension reduction techniques and approaches for variable selection [179]. Some of these techniques are customised for supervised and unsupervised analyses of–omics data, in particular involving metabolomics [188–190]. Unsupervised learning methods' main aim is to explore, summarize and discover groups or trends that are entailed within the data, they need only a few prior assumptions and a little to no *a priori* knowledge [177]. These include techniques such as PCA, k-means clustering or hierarchical clustering. Supervised techniques are methods largely used in biomarker discovery, classification, and prediction and usually deal with sets of data with response variables. They mainly include partial least squares and support vector machine analyses and are now often used in metabolomics data analysis [177,179]. The use of mediation [191–195], pathway analyses [196,197], and approaches to model the "meeting-in-the-middle" concept are instrumental tools providing analytical solutions to fully exploit the multi-dimensional complexity of new generation nutritional epidemiological data.

The methodological work presented in this thesis will draw from already-existing or currently-developing statistical tools, notably multivariate factorial techniques, to explore the associations between diet and cancer. We take on a holistic approach making use of available dietary questionnaire exposures, lifestyle data as well as biomarker and –omics data to explore two cancer endpoints (breast and hepatocellular carcinoma) in an ideal setting to address challenges related to the multi-factorial complexities of dietary exposure.

These principles were applied in the European Prospective Investigation into Cancer and nutrition (EPIC), an on-going multicentre prospective cohort study, mainly designed to study the relationship between nutrition and cancer [198]. Over 521,000 participants, aged between 25 and 70 years, were recruited between 1992 and 2000 across 23 centres spanning 10 European countries including: France, Germany, Greece,

Italy, The Netherlands, Spain, the United Kingdom, Sweden, Denmark and Norway [199]. Dietary intake was assessed dependant on the local context using one of these three validated tools: extensive self-administrated quantitative dietary questionnaires (DQ), semi-quantitative food-frequency questionnaires (FFQ), or through combined dietary methods [199]. All these questionnaires were validated and country-specific, conceived to capture geographical specificity of diet. Indeed, the international multicentre setting of EPIC, combining study populations with different dietary habits, lifestyles and cancer incidences, aims to increase the overall statistical power providing a larger variability of dietary exposures and cancer outcomes. This heterogeneity across geographical regions raises methodological challenges, notably with regards to standardising dietary measurements, for a proper comparison on an absolute scale in all sub-cohorts [199–201]. To this end, in the EPIC calibration study a single 24 hour dietary recall (24-HDR) was collected by trained interviewers between 1995 and 2000 via the EPIC-Soft software (now called GLOBODIET, IARC, Lyon, France) from a random large stratified sample of roughly 8% of the cohort (approx. 37,000 subjects)[202] . The 24-HDR is used as a reference measurement and provides accurate mean estimates of nutrients and foods at the population level. Food portion sizes were estimated using a common picture book and other assessment methods (e.g. standard units and household measures)[200,202]. Foods were classified according to common food classification (88-266 foods) as described elsewhere [203] and individual intake of 25 priority nutrients, plus water, energy and more recently folate [204] were calculated using procedures standardized in the 'EPIC Nutrient DataBase' (ENDB) [203,205]. The calibration study and data harmonization ensured reliable comparisons of different intakes accounting for the heterogeneity of data when evaluating the association between nutritional exposures and disease outcome. Detailed baseline information including anthropometric measures, lifestyle habits (including history of tobacco smoking, alcohol consumption, physical activity, education level, etc.), history of previous illness and other relevant phenotypic information were collected by questionnaires or trained interviewers [202]. Additionally, biological samples were collected at baseline in 80% of the recruited cohort participants prior to cancer onset, providing invaluable biomarker measurements, as detailed in **Table 1**. Approval for this study was obtained from the ethical review boards of the International Agency for Research on Cancer and from all local institutions.

	Study subjects	
Country	**Questionnaire**	**Questionnaire + Blood**
France	74,524	28,083
Italy	47,749	47,725
Spain	41,440	39,579
U.K.	87,942	43,141
The Netherlands	40,072	36,318
Greece	28,555	28,483
Germany	53,091	50,678
Sweden	53,826	53,781
Denmark	57,054	56,131
Norway	37,215	31,000
Total	**521,468**	**414,889**

Table 1: Number of EPIC study subjects by country with questionnaires information and availability of blood samples.

The present thesis aims to investigate the applicability of multivariate statistical methods in the investigation of the relationship between nutrition and cancer, using questionnaires and biomarker data available from the EPIC study.

In a first study described in **Chapter 2**, we explored the applicability of a new dimension-reduction technique that has been recently introduced to the field of nutritional epidemiology: the Treelet Transform (TT). We investigated the relationship between the extracted nutrient patterns and risk of developing breast cancer overall and by hormonal-receptor status in the EPIC Study. Initially developed by Lee *et al.* [206], TT has been conceived as a statistical method aiming to reduce multidimensional datasets by harnessing features of PCA and combining them with those of hierarchical clustering. TT yields orthogonal components (eigenvectors of the correlation or covariance matrix of the data), that are linear projections of the starting variables while introducing sparsity in the component loadings, by making some of these loadings exactly equal to zero. In this way, TT produces components that are easier to interpret than in the well-established PCA [207], where findings' interpretation is complicated by the fact that all component loadings are nonzero. Additionally, TT returns a hierarchical tree reflecting the internal structure of the data. These elements make it a very promising technique in that respect as it allows for an easier interpretation of the findings and to spot the

variables that are mostly contributing to the high variability found within each factor. The **Chapter 2** paper compares nutrient patterns produced with the novel TT with those obtained via the classic PCA and then relates them to breast cancer (BC) outcomes. The use of TT can be extended to other high-dimensional datasets potentially characterized by highly correlated variables with redundant information and noise which may benefit from a method with a sparsity feature.

With the similar motivation for dimensionality reduction and extracting the lost relevant information, the paper presented in **Chapter 3** focused on two sets of data this time, one with untargeted metabolomics acquired through ^{1}H Nuclear Magnetic Resonance (NMR) protocols and the second set entailing a collection of lifestyle exposures. The objective from this work was to provide a practical implementation for the "Meeting-in-the-Middle" (MITM) principle, an idea conceived 10 years ago by Vineis and Perera [162] that relies on the identification of biomarkers that are both reflecting effects of exposures and also contributing to future disease risk. Our study conceptualized a statistical framework where such overlap biomarkers could be identified, first by disentangling the relationship between both sets followed by exploring their link with hepatocellular carcinoma (HCC) development in a nested-case control study within EPIC. This was done in a context characterized by challenges pertaining to the small sample size of the study at hand, making our study difficult to validate/replicate, and those pertaining to untargeted metabolomics in general (e.g. annotations). This first implementation was successful despite a small number of difficulties.

In **Chapter 4**, the statistical approach to model the MITM [162,208] is extended in another nested case-control study on HCC within EPIC and applied to targeted serum metabolomic data acquired through mass-spectrometry techniques. The work is refined by having a more restricted set of exposures from a modified healthy lifestyle index [86]. The statistical analyses are more comprehensive with Partial Least Squares (PLS) applied in turn to each exposure to yield exposure-specific signatures and extensive mediation analyses to investigate whether these specific biomarker profiles bridged their corresponding lifestyle exposures towards risk of HCC. This work allowed us to tackle statistical challenges related to the interpretation of parameters in a context characterized by confounding and various sets of potential mediators.

Finally, we were involved in another initiative studying the associations between biomarkers of fatty acids (FA) and breast cancer (BC) risk in EPIC. **Chapter 5** describes this study featuring measurements of 60 plasma phospholipid fatty acids from a large nested case-control study on BC, where for the first time it was possible to differentiate between *trans* fatty acids (TFA) coming from industrial products from those originating from animal sources. Univariate multivariable regressions were used to relate FA levels to BC risk, overall, by menopausal status and by hormonal receptor status. This work is a first step providing the background necessary for future and more sophisticated modelling, including FA patterns analyses and possibly another application of the MITM framework, hypothesis-driven this time around as opposed to the more agnostic exploratory implementations conducted thus far.

To conclude **Chapter 6** ensues with a general discussion on the findings and topics that were touched upon throughout this thesis.

CHAPTER II:

NUTRIENT PATTERNS AND BREAST CANCER IN EPIC

CONTEXT

Breast Cancer (BC) is the most frequent type of cancer affecting women worldwide; it is the most prevalent form of cancer in the world and the leading cause of mortality from cancer in women both in developed and developing countries [209]. Among modifiable risk factors, diet may account for up to 40% of preventable causes of cancer. In particular an estimated 50% of BC deaths are attributed to diet although despite substantial research, the relationship between diet and BC is still open to debate [2,6,210,211]. Usual approaches have often assessed the role of single dietary items i.e. micro/macronutrients, foods, energy and alcohol mostly through standard univariate analyses, and these have yielded significant results [145]. However, due to the complexity of diet and the potential interactions between different dietary components, approaches that focus on individual foods or restricted list of nutrients / dietary constituents may miss information on the role of diet in disease aetiology [71,73]. Dietary patterns have emerged as a tool of choice to depict a broader picture of the effects of overall diet. Conceptually, patterns are more akin to reflect reality than traditional approaches, as people usually consume a variety of foods often containing a complex combination of nutrients. Moreover, some nutrient effects may be too small to detect on their own, thus the cumulative effect of a pattern embracing multiple nutrients may be easier to identify [71,212]. In this study, nutrient patterns were obtained through two multivariate methods, the well-established Principal Component Analysis (PCA) [207] and the newly emerging Treelet Transform (TT) [129,213–215]. The association between the extracted nutrient patterns and BC was investigated within the EPIC study, a multicenter study with heterogeneous data, offering a vast playground to address methodological challenges.

OBJECTIVES

- To yield nutrient patterns within the women sub-cohort in EPIC by applying the TT, a new dimension reduction technique that has been recently introduced to the nutritional epidemiology landscape. To derive nutrient patterns using PCA, a more classic approach.
- To relate nutrient patterns to risk of BC in general, and by taking into account the heterogeneity of BC subtypes by integrating information on menopausal and hormone receptor status.

- To compare results from two multivariate dimension reduction techniques: PCA and TT.

APPROACH

The analyses focused on the women sub-cohort within EPIC (N=334,850) where 11,575 BC cases were ascertained across all centres. *A posteriori* nutrient patterns were obtained by applying multivariate methods (PCA and TT) to a covariance matrix of 23 log-transformed macro- and micronutrients obtained from dietary questionnaires. The aim of PCA is to reduce dimensionality by transforming a large set of correlated foods or nutrient items, into a smaller set of uncorrelated variables, called principal components that make up the nutrient patterns. TT additionally introduces sparsity in component loadings making some of them equal to zero, thus making the interpretation easier. TT also produces a hierarchical grouping of variables revealing intrinsic characteristics of data structure. Hazard ratios and 95% confidence intervals (HR, 95%CI) were estimated and quantified the association between the scores quintiles of the first two components and BC risk. The Cox proportional hazard models were stratified by age, centre, and adjusted for potential confounding factors including anthropometric measures, non-alcohol energy, lifestyle and reproductive variables.

MAIN FINDINGS

Two main patterns were retained in both TT and PCA analyses, and were consistent in terms of pattern identification and amount of total variability explained (over 50% of total observed variability). The first TT component (TC1) loaded highly on cholesterol, protein, retinol, vitamins B12 and D, while TC2 reflected a nutrient dense pattern with high contributions for β-carotene, riboflavin, thiamin, vitamins C and B6, fibre, Fe, Ca, K, Mg, P and folate (**Figure 1**). The TT components were highly correlated with those of PCA ($\rho_{TC1, PC1}$= 0.91, $\rho_{TC2, PC2}$= 0.86). The first pattern, that was akin to a Western diet, was associated with a non-significant increase of 5% in BC risk, whilst the second pattern was inversely associated with BC risk with HR=0.89(0.83, 0.95). This decrease was also significant for ER+, PR+, PR− and ER+/PR+ tumours.

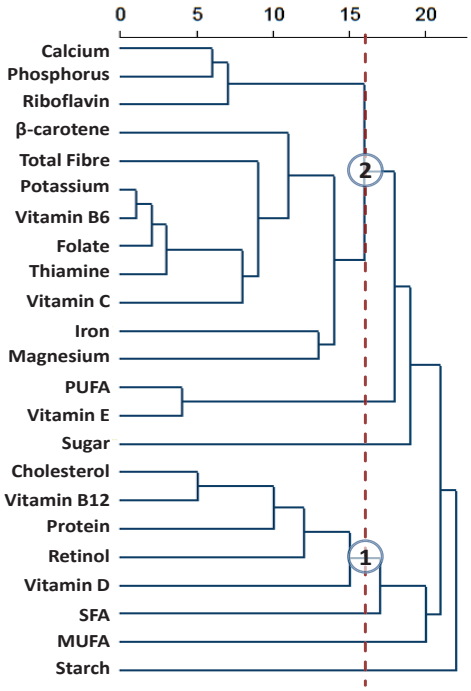


Figure 1: Cluster tree produced by the Treelet Transform.

Cut-level (red dashed line) was chosen after using a 10-fold cross-validation. Nutrients related to the treelet components (TC), indicated with numbered circles, have non-zero loadings on the given component.

CONCLUSION

This study investigated the association between nutrient patterns and BC in the international setting of the EPIC study using a new tool in nutritional epidemiology, the Treelet Transform. TT has the advantage of introducing sparsity in factor loadings thus leading to more easily interpretable patterns. When compared to a more standard approach, such as PCA, TT offers a complementary approach yielding comparable nutrient patterns accounting for similar amounts of variability. In essence, there is a sparsity trade-off: TC are easier to interpret but have a lower information resolution than PC, which may lead to disparities in some associations in models with TC scores vs. PC scores. The findings suggested a protective association for a diet rich in vitamins, minerals and β-carotene, indicating that a diet mostly plant-based decreased BC risk while a nutrient patterns characterized by a diet rich in macronutrients of animal origin, such as cholesterol or SFA, was related to an increase in BC risk, albeit non-significant.

PAPER

Contribution: First author, discussed the analytical strategy with the supervisor, conducted statistical analyses, wrote the first draft of the manuscript, submitted it to the journal and replied to reviewers' comments.

Public Health Nutrition: page 1 of 13 doi:10.1017/S1368980015000294

A treelet transform analysis to relate nutrient patterns to the risk of hormonal receptor-defined breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC)

Nada Assi[1,2], Aurelie Moskal[1], Nadia Slimani[1], Vivian Viallon[3,4,5], Veronique Chajes[1], Heinz Freisling[1], Stefano Monni[6], Sven Knueppel[7], Jana Förster[7], Elisabete Weiderpass[8,9,10,11], Leila Lujan-Barroso[12], Pilar Amiano[13,14], Eva Ardanaz[13,15], Esther Molina-Montes[13,16], Diego Salmerón[13,17,18], José Ramón Quirós[19], Anja Olsen[20], Anne Tjønneland[20], Christina C Dahm[21], Kim Overvad[21], Laure Dossus[22,23,24], Agnès Fournier[22,23,24], Laura Baglietto[25,26], Renee Turzanski Fortner[6], Rudolf Kaaks[6], Antonia Trichopoulou[27,28], Christina Bamia[29], Philippos Orfanos[29], Maria Santucci De Magistris[30], Giovanna Masala[31], Claudia Agnoli[32], Fulvio Ricceri[33], Rosario Tumino[34], H Bas Bueno de Mesquita[35,36,37], Marije F Bakker[38], Petra HM Peeters[38], Guri Skeie[8], Tonje Braaten[8], Anna Winkvist[39], Ingegerd Johansson[40], Kay-Tee Khaw[41], Nicholas J Wareham[42], Tim Key[43], Ruth Travis[43], Julie A Schmidt[43], Melissa A Merritt[37], Elio Riboli[37], Isabelle Romieu[1] and Pietro Ferrari[1,*]

[1]International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon Cedex 08, France: [2]Université Claude-Bernard Lyon 1, Villeurbanne, France: [3]Université de Lyon, Lyon, France: [4]Université Lyon 1, UMRESTTE, Lyon, France: [5]IFSTTAR, UMRESTTE, Bron, France: [6]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany: [7]Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany: [8]Department of Community Medicine, Faculty of Health Sciences, University of Tromsø, The Arctic University of Norway, Tromsø, Norway: [9]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden: [10]Cancer Registry of Norway, Oslo, Norway: [11]Department of Genetic Epidemiology, Folkhälsan Research Center, Helsinki, Finland: [12]Unit of Nutrition, Environment and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain: [13]CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain: [14]Public Health Division of Gipuzkoa, BioDonostia Research Institute, Health Department, San Sebastian, Spain: [15]Navarre Public Health Institute, Pamplona, Spain: [16]Escuela Andaluza de Salud Pública, Instituto de Investigación Biosanitaria de Granada (Granada.ibs), Granada, Spain: [17]Department of Epidemiology, Murcia Regional Health Council, Murcia, Spain: [18]Department of Health and Social Sciences, Universidad de Murcia, Murcia, Spain: [19]Public Health Directorate, Asturias, Oviedo, Spain: [20]Danish Cancer Society Research Center, Copenhagen, Denmark: [21]Section for Epidemiology, Department of Public Health, Aarhus University, Aarhus, Denmark: [22]Inserm, Centre for Research in Epidemiology and Population Health (CESP), Nutrition, Hormones and Women's Health Team, Villejuif, France: [23]Université Paris Sud, UMRS, Villejuif, France: [24]IGR, Villejuif, France: [25]Cancer Epidemiology Centre, Cancer Council of Victoria, Melbourne, Australia: [26]Centre for Epidemiology and Biostatistics, School of Population and Global Health, University of Melbourne, Melbourne, Australia: [27]Hellenic Health Foundation, Athens, Greece: [28]Bureau of Epidemiologic Research, Academy of Athens, Athens, Greece: [29]Department of Hygiene, Epidemiology and Medical Statistics, University of Athens Medical School, Athens, Greece: [30]Azienda Ospedaliera Universitaria (AOU) Federico II, Naples, Italy: [31]Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Institute – ISPO, Florence, Italy: [32]Epidemiology and Prevention Unit, Fondazione IRCCS, Istituto Nazionale dei Tumori, Milan, Italy: [33]Unit of Cancer Epidemiology – CERMS, Department of Medical Sciences, University of Turin and Città della Salute e della Scienza Hospital, Turin, Italy: [34]Cancer Registry and Histopathology Unit, 'Civile M.P. Arezzo' Hospital, Ragusa, Italy: [35]Department for Determinants of Chronic Diseases (DCD), National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands: [36]Department of Gastroenterology and Hepatology, University Medical Centre, Utrecht, The Netherlands: [37]Department of Epidemiology and Biostatistics, The School of Public Health, Imperial College London, London, UK: [38]Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands: [39]Department of Internal Medicine and Clinical Nutrition, The Sahlgrenska Academy, Göteborg, Sweden: [40]Department of Odontology, Umeå University, Umeå, Sweden: [41]Department of Public Health and Primary Care, University of Cambridge School of Clinical Medicine, Cambridge, UK: [42]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, UK: [43]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

Submitted 9 September 2014: Final revision received 12 January 2015: Accepted 20 January 2015

*Corresponding author: Email ferrarip@iarc.fr

Abstract

Objective: Pattern analysis has emerged as a tool to depict the role of multiple nutrients/foods in relation to health outcomes. The present study aimed at extracting nutrient patterns with respect to breast cancer (BC) aetiology.

Design: Nutrient patterns were derived with treelet transform (TT) and related to BC risk. TT was applied to twenty-three log-transformed nutrient densities from dietary questionnaires. Hazard ratios (HR) and 95 % confidence intervals computed using Cox proportional hazards models quantified the association between quintiles of nutrient pattern scores and risk of overall BC, and by hormonal receptor and menopausal status. Principal component analysis was applied for comparison.

Setting: The European Prospective Investigation into Cancer and Nutrition (EPIC).

Subjects: Women (*n* 334 850) from the EPIC study.

Results: The first TT component (TC1) highlighted a pattern rich in nutrients found in animal foods loading on cholesterol, protein, retinol, vitamins B_{12} and D, while the second TT component (TC2) reflected a diet rich in β-carotene, riboflavin, thiamin, vitamins C and B_6, fibre, Fe, Ca, K, Mg, P and folate. While TC1 was not associated with BC risk, TC2 was inversely associated with BC risk overall ($HR_{Q5\ v.\ Q1} = 0{\cdot}89$, 95 % CI 0·83, 0·95, $P_{trend} < 0{\cdot}01$) and showed a significantly lower risk in oestrogen receptor-positive ($HR_{Q5\ v.\ Q1} = 0{\cdot}89$, 95 % CI 0·81, 0·98, $P_{trend} = 0{\cdot}02$) and progesterone receptor-positive tumours ($HR_{Q5\ v.\ Q1} = 0{\cdot}87$, 95 % CI 0·77, 0·98, $P_{trend} < 0{\cdot}01$).

Conclusions: TT produces readily interpretable sparse components explaining similar amounts of variation as principal component analysis. Our results suggest that participants with a nutrient pattern high in micronutrients found in vegetables, fruits and cereals had a lower risk of BC.

Keywords
Nutrient patterns
Treelet transform
Breast cancer
European Prospective Investigation into Cancer and Nutrition
Principal component analysis

Breast cancer (BC) remains the highest incident cancer affecting women worldwide, with almost 1 670 000 cases registered in 2012. It is a major public health concern with mortality from BC accounting for over 522 000 deaths in 2012, including almost 198 000 deaths in Western countries and about 324 000 in less developed regions[1]. Established BC risk factors include age, genetic mutations, ethnicity, height, reproductive history, breast-feeding, hormone therapy and diabetes[2–6]. Besides these, a number of modifiable lifestyle factors are associated with BC such as smoking[7,8], body fat and obesity[9–11], physical inactivity[10,12,13], alcohol consumption[14–16] and diet[5,17,18]. Diet has been suggested to account for up to 25–40 % of preventable causes of cancers; in particular, 50 % of BC deaths are linked to diet, although the consensus around this estimate is not unanimous[12,19,20]. Standard approaches customarily evaluate the risk of BC associated with one or a group of dietary items, i.e. food(s) or nutrient(s). Nevertheless, associations between diet and disease might be missed when one parses the effect of a limited list of dietary constituents. Although this simplified approach of examining a single food or nutrient at a time has led to important results on the role of an individual dietary component in BC aetiology, such as fibre from vegetables, alcohol, tea consumption, folate and other micronutrients[12,14,18,20–23], research might benefit from a more comprehensive approach by exploring BC aetiology in terms of an integrated ensemble of dietary characteristics.

To capture the complexity of individuals' dietary habits, dietary pattern analysis has emerged as a complementary holistic methodology focusing on sets of dietary variables and addressing their inherent interrelations[24]. This approach is justified as components of dietary exposure are not independent[25,26] and because it allows to account for complex relationships between nutrients in biological pathways[25]. In addition, BC is a multifactorial disease[2–18], the aetiology of which possibly depends on more than a restricted list of dietary items.

Recent investigations carried out in Western populations[27–32] have consistently identified two main dietary patterns: the prudent/healthy and the Western/unhealthy[29,33]. While diet is related to cultural background, common nutrients are present in different combinations of foods; hence looking into diet–disease associations on the nutrient scale could lead to the identification of specific nutritional profiles relevant to BC aetiology.

In the present study, nutrient patterns within the European Investigation into Cancer and Nutrition (EPIC) were related to BC risk. Nutrient patterns were obtained by applying the treelet transform (TT) that has recently been introduced into nutritional epidemiology[34–36] and the well-known principal component analysis (PCA) was used for the sake of comparison[37]. TT yields sparse components and reveals the intrinsic structure of the data, thus simplifying interpretability. Aspects related to the application of TT to dietary data in the context of a multi-centre study are described and discussed. The association between nutrient

patterns and BC was evaluated using all BC cases and by taking into account the heterogeneity of BC subtypes by integrating information on menopausal and hormone receptor status.

Materials and methods

Study population and exclusion criteria

EPIC is a large prospective cohort of 521 330 healthy men and women designed to evaluate the relationships between dietary habits, nutrition, lifestyle factors and the incidence of cancer. The EPIC cohort includes participants from twenty-three centres in France, Germany, Denmark, Sweden, Norway, Greece, Italy, the Netherlands, Spain and the UK. In most centres, participants were recruited from the general population, the exceptions being France (women were enrolled from a national health insurance scheme covering teachers in the French education system employees), Italy (Turin and Ragusa: blood donors; Florence: screening programme participants), Spain (blood donors) and the Netherlands (Utrecht: women participating in BC screening). In Norway, only women from the general population were recruited and in the UK, one-half of the cohort (the Oxford sub-cohort) consisted of 'health-conscious' individuals from England, Wales, Scotland and Northern Ireland. The design of the study and its rationale along with the recruitment process have been described elsewhere[38].

Among the 521 330 EPIC participants, men were first removed (n 153 427). Women with prevalent cancers at any site at baseline (other than non-melanoma skin cancer; n 19 853) or lost to follow-up (n 2892) were excluded, as were women who did not complete any dietary questionnaire (n 3315) and those who did not complete a lifestyle questionnaire (n 26). To avoid including extreme values, participants in the top and bottom 1 % of the distribution of the ratio of reported total energy intake to energy requirement (n 6753) were excluded. After exclusion of non-first BC cases (n 2) the cohort included 335 062 women upon whom the dietary patterns were derived. An additional number of women (n 212) with missing information on BC status were excluded, which left 334 850 women retained for the statistical analyses.

Cancer assessment

Incident BC cases were identified through population cancer registries (Denmark, Italy, Netherlands, Norway, Spain, Sweden and UK) or through active follow-up (France, Germany, Naples and Greece), as detailed in Ferrari *et al.*[21]. Information on oestrogen receptor (ER) and progesterone receptor (PR) statuses was provided by each centre on the basis of pathology reports.

Dietary assessment

Long-term usual dietary intake was assessed at baseline using country-specific and validated dietary questionnaires (self-administered FFQ, semi-quantitative or interviewer-performed)[38–40]. In the validation studies, the dietary questionnaires were compared with a reference method which was in most centres 24 h dietary recalls, except in Sweden and the UK, where food records were used. Generally, the correlation coefficients were between 0·40 and 0·70 for all nutrients examined which was considered satisfactory[41]. Individual intakes of twenty-three nutrients and total energy were estimated using a common food composition database, the EPIC Nutrient Database (ENDB), which was compiled from national food composition databases of the ten countries represented in EPIC following standardized procedures[42,43].

Lifestyle questionnaires

Information on sociodemographic characteristics, including education, and lifestyle habits such as levels of physical activity, tobacco smoking, as well as consumption of alcohol and drinking habits, were collected using lifestyle questionnaires. In addition, anthropometric measures and past medical information were gathered at recruitment[38].

Nutrient pattern assessment

EPIC-wide nutrient patterns were derived among female participants in EPIC using TT in the main analysis and PCA in the sensitivity analysis. The sample covariance matrix of twenty-three log-transformed nutrient densities, computed using alcohol-free energy intake[44], was consistently used. The use of the sample covariance matrix allows variability to be informative in the pattern discovery phase. The distribution of nutrient consumption tends to be log-normal and may not be best described by the mean and variance on the original scale. Moreover micro- and macronutrients are expressed on different scales (micrograms, milligrams or grams). The nutrient densities were log-transformed to remove scale dependence and render their variance (or covariance) independent of the unit of measure. In line with previous work[28,45,46], alcohol intake was not included and was considered as a lifestyle factor. Total fat was divided into MUFA, PUFA and SFA, and total carbohydrates were broken down into starch and sugar. The micro- and macronutrients studied were Ca, β-carotene, cholesterol, MUFA, PUFA, SFA, Fe, fibre, K, Mg, P, protein, retinol, riboflavin, starch, sugar, thiamin, vitamins B_6, B_{12}, C, D, E and folate. The list of nutrients as well as the approach described for their handling is consistent with the nutrient patterns initiative within EPIC described by Moskal *et al.*[45].

Pattern extraction

The TT method used for pattern extraction is described in detail by Gorst-Rasmussen and co-workers[35,47]. Briefly, TT is a dimension reduction technique aimed at converting a set of observations of possibly correlated variables into orthogonal components. TT scores, corresponding to

the projection of data onto components, generally have a small degree of correlation, unlike PCA scores that are always uncorrelated. The number of retained components was based on the percentage of explained variance, scree plots and interpretability. The nutrient patterns were defined after the inspection of factor loadings, i.e. eigenvectors, expressing the contribution of nutrients to a given component. Score variables were determined for each component of TT and reflected adherence to a given type of diet/nutrient profile. TT combines the quantitative pattern extraction capabilities of PCA with interpretational advantages of hierarchical clustering of variables. In TT, the two variables displaying the highest correlation (or covariance) are identified, and a PCA is performed on them. The two variables are then replaced with the score of their first PCA component and a merge is indicated in the cluster tree. This operation is re-iterated until all variables have joined the cluster tree. In this way, TT produces a hierarchical grouping of variables which may reveal intrinsic characteristics of data structure. An important feature of TT is that it introduces sparsity into factors, making many factors loadings exactly equal to zero, potentially simplifying the interpretation. Alongside the cluster tree dendrogram produced by TT (as exemplified in Fig. 1), TT yields a coordinate system for the data at each level of the cluster tree. Selecting a cluster tree level (cut-level) for the TT cluster tree amounts to choosing the level of detail desired in the dimension reduction of data. More variation can be explained at the cost of factor sparsity when the cluster tree is cut near its 'root'. If the data have p variables, there are $p - 1$ possible cut-levels. After deciding on the number of components to retain, we performed a tenfold cross-validation to identify the optimal cut-level, i.e. the point at which increasing the cut-level does not substantially increase the variation of the retained patterns. We also performed a sensitivity analysis to assess the effect of different cut-levels[35,48].

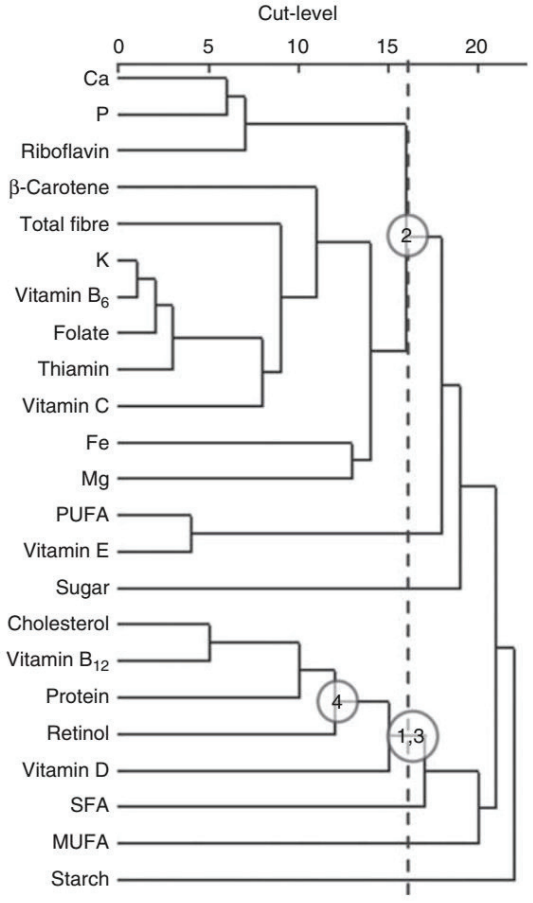


Fig. 1 Cluster tree produced by the treelet transform algorithm applied to twenty-three log-transformed nutrient densities for 335 062 women in the European Prospective Investigation into Cancer and Nutrition (EPIC). The dashed line indicates the chosen cut-level (16) to extract components. The highest-variance factors, i.e. treelet components at this level of the tree, are indicated with numbered circles. The nutrients related to these nodes have non-zero loadings on the given component. Components 1 and 3 share the same node but the variable loadings differ

Consistently, a PCA was also applied for the sake of comparison[37]. This technique yields orthogonal components that are invariant to the number of subsequent components retained. PCA identifies the best linear combination of the variables accounting for the most variance observed in the original data, producing components with uncorrelated scores. Results of TT analysis were compared with findings obtained with the more classic PCA method. To make the comparison easier, and because TT returns sparse vectors, only nutrients with absolute loadings greater than 0·2 were retained to identify a given pattern in PCA.

Patterns and breast cancer risk

The associations between nutrient patterns and risk of BC were investigated by using Cox proportional hazards regression models to estimate hazard ratios (HR) and 95 % confidence intervals. Breslow's method was adopted for handling time ties[49]. The time at entry was the age at recruitment and the time of exit was the age at cancer diagnosis, death, loss or end of follow-up, whichever happened first. Models were stratified by centre, to control for differences in questionnaire designs, follow-up procedures and other centre-specific effects, as well as for age at recruitment (1-year categories)[50]. Analyses were performed by considering the TT (and principal component (PC)) scores in quintiles to appreciate potential departure from linearity. Statistical analyses were adjusted for baseline menopausal status (premenopausal and perimenopausal (reference) or postmenopausal and women who underwent an ovariectomy), baseline alcohol intake (never drinkers (reference), former drinkers, drinkers only at recruitment, lifetime drinkers, unknown), height (continuous), BMI (below (reference) or above 25 kg/m^2), schooling level (none, primary (reference), technical/

professional/secondary, longer education, unknown/unspecified), age at first full-term pregnancy (nulliparous (reference), ≤21 years, 21–30 years, >30 years, unknown or missing), age at menarche (≤12 years (reference), 12–14 years, >14 years, missing), age at menopause (≤50 years (reference), >50 years, premenopausal or missing), use of hormone replacement therapy (never (reference), ever, unknown), level of physical activity (categorical, metabolic equivalents of task (MET)/h: inactive (reference), moderately inactive, moderately active, active, unknown) and alcohol-free energy (continuous). Use of oral contraceptive pills (never (reference), ever or unknown) and smoking status (never smokers (reference), ex-smokers, current smokers, unknown) were evaluated but not retained in the final models, due to limiting confounding exerted by these variables.

The overall significance of a score variable in categories was evaluated using the likelihood ratio test statistics (P_{LRT}) with df = 4. Additionally, P values for trend (P_{trend}) were computed by modelling a score variable with quintile-specific medians as continuous. The association between nutrient patterns and BC risk was evaluated in pre- and postmenopausal women and according to BC hormonal receptor status (ER/PR status). Interaction between menopausal status and pattern scores was explored. In addition, tests of heterogeneity of associations according to receptor status were performed using the data-augmentation method[51] by comparing the difference in the log likelihood between a model with receptor status-specific variable and a model with a single HR estimate for the two categories of receptor status to a χ^2 distribution with df = 1 ($P_{heterogeneity}$).

Departure from linearity was explored with restricted cubic splines[52], using five knots corresponding to the 1st and 99th percentiles and medians of the centred scores of quintiles 1, 3 and 5. Spline plots were produced by taking the median of the first quintile as reference. Departures from linearity were assessed via an evaluation of the joint significance of variables other than the linear one included in the model using Wald's test on df = 3. Associations between all of the PC and BC were investigated in a consistent way.

Statistical tests were two-sided, the per-test significance level was set to $\alpha = 0{\cdot}05$. All analyses were performed using the SAS statistical software package version 9·3; the 'tt' package in the STATA statistical software package release 12 was used to perform TT.

Results

A total of 11 576 BC cases were recorded in 11·5 years of median follow-up time and 3 670 439 person-years. Based on the information obtained at baseline, 2827 cases were premenopausal, 5872 were postmenopausal, 2548 were perimenopausal and 328 cases had a bilateral ovariectomy. Among incident cases, information on hormone receptor status for ER and PR was available only in 62 % and 52 % of total cancer cases, respectively, and was distributed as follows: 81 % ER^+ and 19 % ER^- tumours and 63 % PR^+ and 37 % PR^- tumours. Descriptive information of the study sample by EPIC country is available in Table 1.

Identification of nutrient patterns

Inspection of factor loadings allowed an initial identification of four nutrient patterns with TT, explaining 62 % of total nutrient intake variability within individuals. After a tenfold cross-validation along with a sensitivity analysis strategy and after evaluating the interpretability of each pattern, we chose to cut the cluster tree at level 16. Loadings of components 1 and 2 are shown in Table 2. TT yielded a dendrogram shown in Fig. 1, with numbered nodes indicating the four highest-variance factors, where factors 1 and 2 were identified as the first two components after setting the cut-level to 16 indicated by the dashed line. This dendrogram reveals the correlation structure of the log-transformed nutrient densities. The first treelet component (TC1) loaded on vitamin D, vitamin B_{12}, cholesterol, protein and retinol, suggesting a diet rich in animal products. The second treelet component (TC2)

Table 1 Numbers of women and breast cancer (BC) cases (first tumours only) in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort by country

Country	No. of women	Person-years	No. of BC cases	Follow-up time (years)*	Age at enrolment (years)*
France	67 356	699 216	3187	11·8	51·5
Italy	30 498	341 417	1047	11·7	50·9
Spain	24 846	299 575	495	12·6	47·7
UK general population	17 145	200 812	719	12·3	55·6
UK health-conscious	35 368	385353	761	11·3	41·5
Netherlands	26 839	315 554	916	12·2	52·7
Greece	15 224	148 594	198	10·7	53·6
Germany	27 390	272 011	834	10·9	48·4
Sweden	26 339	349 110	1095	13·9	50·6
Denmark	28 693	316 601	1340	11·6	56·3
Norway	35 152	342 195	984	10·1	48·0
Total	334 850	3 670 439	11 576	11·5	51·0

*Median is given for follow-up time and age at enrolment.

Table 2 Loadings of the first two components from treelet transform (TT; cut-level 16)

	TT 16 loadings	
Variable*	TC1	TC2
Ca		0·153
β-Carotene		0·721
Cholesterol	0·294	
MUFA		
PUFA		
SFA		
Fe		0·109
Fibre		0·183
K		0·157
Mg		0·144
P		0·074
Protein	0·086	
Retinol	0·679	
Riboflavin		0·141
Starch		
Sugar		
Thiamin		0·217
Vitamin B_6		0·185
Vitamin B_{12}	0·421	
Vitamin C		0·452
Vitamin D	0·517	
Vitamin E		
Folate		0·235
Explained variance	26 %	21 %

TC1, treelet component 1; TC2, treelet component 2.
*Log-transformed nutrient variables.

presented high positive loadings on β-carotene, thiamin, fibre, vitamin C and folate, and singled out some nutrients with mild loadings (<0·2), i.e. Fe, Ca, K, Mg and P (Table 2). TC2 may evoke a diet rich in vegetables, fruits and cereals. While the third treelet component (TC3) was largely driven by vitamin D, the fourth treelet component (TC4) was less straightforward to characterize, as displayed in the online supplementary material, Supplemental Table 1. Distributions of known risk factors for BC by quintiles of TT scores for the first two components are displayed in Table 3.

PC loadings are displayed in the online supplementary material, Supplemental Table 2. PCA produced patterns similar to TT with respect to the amount of variability explained and the nutrients contributing to the definition of each component: with PC1 displaying high loadings for cholesterol, retinol, vitamin B_{12} and vitamin D and negative loadings for vitamin C and β-carotene; and PC2 suggesting a micronutrient-dense pattern rich in fruits, vegetables, plant foods and dairy. The first two components (in TT and PCA) explained the most variability and were the most informative with respect to capturing meaningful nutrient patterns, and thus were further related to BC risk in disease models.

Nutrient patterns and breast cancer risk

Scores of nutrient patterns were related to BC risk. TC1 showed no statistically significant association with BC risk with $HR_{TC1\ Q5\ v.\ Q1} = 1{\cdot}05$ (95 % CI 0·98, 1·13, $P_{trend} = 0{\cdot}36$, $P_{LRT} = 0{\cdot}39$), while TC2 was significantly associated with BC risk with $HR_{TC2\ Q5\ v.\ Q1} = 0{\cdot}89$ (95 % CI 0·83, 0·95, $P_{trend} < 0{\cdot}001$, $P_{LRT} = 0{\cdot}02$), as shown in Table 4. The relationship between TT scores and BC risk was modelled through restricted cubic splines (RCS) and is presented in Fig. 2. Overall, there was a significant progressive decrease in BC risk for the second component. TC2 scores showed a linear decrease in BC risk (RCS_{TC2} $P_{trend} = 0{\cdot}02$). However, no departure from linearity was observed ($P_{Wald\ non\text{-}linearity} = 0{\cdot}94$ and 0·77, respectively, in TC1 and TC2; Fig. 2). Analyses of interaction between TC (or PC) scores and menopausal status were not statistically significant (results not shown).

Hormonal receptor status

In ER^- tumours, no significant association with BC risk was observed for TC1 and TC2 scores (Table 4). For ER^+ tumours there was a decrease in BC risk in the fourth and fifth quintiles of TC2 scores with $HR_{Q4\ v.\ Q1} = 0{\cdot}90$ (95 % CI 0·83, 0·99) and $HR_{Q5\ v.\ Q1} = 0{\cdot}89$ (95 % CI 0·81, 0·98, $P_{trend} = 0{\cdot}02$; Table 4). Regarding PR^- tumours (see online supplementary material, Supplemental Table 3), the second component TC2 showed a decreased BC risk with $HR_{Q5\ v.\ Q1} = 0{\cdot}84$ (95 % CI 0·72, 0·98). For PR^+ tumours, TC2 was linked with a decreased BC risk in participants in the fifth quintile with $HR_{Q5\ v.\ Q1} = 0{\cdot}87$ (95 % CI 0·77, 0·98). No significant association was seen for ER^-/PR^- tumours (Table 5). TC2 was linked with a decreased BC risk trend in ER^+/PR^+ tumours with $HR_{Q5\ v.\ Q1} = 0{\cdot}86$ (0·76, 0·98, $P_{trend} < 0{\cdot}01$; Table 5). Tests of heterogeneity yielded no significant results.

PCA derived components displayed a significant increase in BC risk for PC1 in participants in the highest quintile and a decreasing trend of BC risk for PC2, as shown in the online supplementary material, Supplemental Table 4 and Supplemental Fig. 1. Results of associations of PC with tumours by hormone receptor status are displayed in the online supplementary material, Supplemental Tables 4 and 5.

Discussion

In the present study, the role of nutrient patterns in the aetiology of BC was explored through the use of TT, a multivariate method recently introduced to the landscape of nutritional epidemiology[34–36]. The association was evaluated in the context of the EPIC study, characterized by large variability of dietary habits and by a large number of incident cancer cases across participating centres[38].

In recent years, dietary pattern analysis has emerged as a promising technique, complementary to methods focusing on individual foods or food components, to investigate the relationships between diet and risk of disease[25]. A systematic review and meta-analysis on dietary patterns in BC aetiology[33] selected eighteen

Table 3 Lifestyle and dietary baseline characteristics* according to the lowest, middle and highest quintiles of treelet transform (cut-level 16) scores for the first and second components among 334 850 women in the European Prospective Investigation into Cancer and Nutrition (EPIC)

	TC1						TC2					
	Q1		Q3		Q5		Q1		Q3		Q5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
No. of women	66 988		66 977		66 955		66 961		66 969		66 970	
Age (years)	50·2	11·8	50·8	9·5	52·0	8·1	49·6	9·3	51·1	9·5	52·2	10·9
Weight (kg)	63·0	11·6	64·8	11·8	65·0	11·9	64·0	11·9	64·0	11·7	63·8	11·5
Height (cm)	160·1	7·1	162·6	6·5	163·0	6·5	162·0	6·9	162·5	6·7	162·0	6·5
Non-alcohol energy (kJ/d)	7565	2280	7573	2171	7368	2121	8309	2406	7623	2138	6820	1929
Non-alcohol energy (kcal/d)	1808	545	1810	519	1761	507	1986	575	1822	511	1630	461
	%		%		%		%		%		%	
BMI class												
Below 25 kg/m^2	57		59		57		58		58		59	
Above 25 kg/m^2	43		41		43		42		42		41	
Schooling level												
None	11		3		2		5		5		4	
Primary	25		22		26		33		23		17	
Technical/professional/secondary	35		47		50		44		46		44	
Longer education	25		23		19		16		23		28	
Unspecified/unknown	4		5		3		2		3		8	
Use of hormone replacement therapy												
Never	82		68		60		71		68		69	
Ever	16		25		31		20		25		27	
Unknown	2		7		9		9		7		4	
Age at first term pregnancy												
Nulliparous	21		13		11		13		14		19	
≤21 years	16		18		24		20		18		17	
21–30 years	52		56		54		54		56		52	
>30 years	9		9		7		8		8		8	
Unknown	3		5		4		5		4		5	
Age at menarche												
≤12 years	38		35		33		33		35		39	
12–14	46		46		47		46		47		45	
>14 years	15		15		17		16		16		14	
Unknown	1		4		4		5		3		3	
Age at menopause												
≤50 years	19		16		18		17		17		18	
>50 years	19		18		19		16		18		19	
Unknown	63		66		63		67		65		62	
Menopausal status												
Pre and peri	55		55		49		60		53		49	
Post and ovariectomy	45		45		51		40		47		51	
Alcohol drinkers												
Never	16		6		4		8		8		9	
Former	6		3		2		4		3		4	
Only at recruitment	17		11		8		6		11		19	
Lifetime	51		56		46		44		54		57	
Unknown	10		22		40		38		24		11	
Physical activity												
Inactive	31		20		16		25		20		21	
Moderately inactive	33		33		28		30		31		33	
Moderately active	21		23		18		18		22		24	
Active	13		15		12		12		14		17	
Unknown	2		10		25		15		13		5	

TC1, treelet component 1; TC2, treelet component 2; Q1, quintile 1; Q3, quintile 3; Q5, quintile 5.
*Means and standard deviations are presented for continuous variables, and frequencies are presented for categorical variables.

relevant studies from case–control and cohort studies that used combinations of foods and micronutrients to identify dietary patterns[17,27,53–66]. Two *a posteriori* defined patterns emerged consistently: the Western/unhealthy (in seventeen studies) and the prudent/healthy (eighteen studies)[33]. In the aforementioned meta-analysis[33], the prudent/healthy dietary pattern, rich in intakes of vegetables, leafy vegetables, legumes and fish, was associated to decreased BC risk (relative risk comparing top *v.* bottom categories = 0·89, 95 % CI 0·82, 0·99), while the Western/

Table 4 Hazard ratios (HR) and 95 % confidence intervals for breast cancer (BC) by quintiles of pattern scores (first and second components of treelet transform, cut-level 16) for overall, oestrogen receptor-positive (ER^+) and oestrogen receptor-negative (ER^-) tumours in 334 850 women in the European Prospective Investigation into Cancer and Nutrition (EPIC)

	TC1						TC2					
Model*	Person-years	No. of BC cases	HR	95 % CI	P_{LRT}†	P_{trend}‡	Person-years	No. of BC cases	HR	95 % CI	P_{LRT}†	P_{trend}‡
Overall												
Q1	730 785	1784	1·00	Ref.	0·39	0·36	747 690	2317	1·00	Ref.	0·02	<0·001
Q2	738 136	2342	1·06	0·99, 1·13			736 718	2307	0·95	0·89, 1·00		
Q3	735 683	2376	1·04	0·97, 1·11			729 544	2365	0·95	0·89, 1·01		
Q4	737 533	2513	1·06	0·99, 1·14			725 903	2350	0·94	0·88, 1·00		
Q5	728 303	2561	1·05	0·98, 1·13			730 584	2237	0·89	0·83, 0·95		
ER^+												
Q1	725 634	885	1·00	Ref.	0·55	0·47	740 268	1133	1·00	Ref.	0·13	0·02
Q2	731 571	1214	1·07	0·98, 1·17			729 915	1140	0·92	0·84, 1·00		
Q3	728 782	1212	1·06	0·97, 1·16			722 467	1192	0·92	0·84, 1·00		
Q4	729 703	1247	1·08	0·98, 1·19			719 201	1193	0·90	0·83, 0·99		
Q5	720 422	1272	1·05	0·95, 1·16			724 261	1172	0·89	0·81, 0·98		
ER^-												
Q1	721 118	227	1·00	Ref.	0·94	0·43	734 469	287	1·00	Ref.	0·25	0·06
Q2	725 180	302	1·03	0·86, 1·23			724 168	318	1·06	0·90, 1·24		
Q3	722 496	301	0·99	0·82, 1·18			716 332	288	0·93	0·78, 1·10		
Q4	723 410	316	1·01	0·83, 1·22			713 221	288	0·93	0·78, 1·12		
Q5	714 166	292	0·95	0·78, 1·16			718 180	257	0·87	0·71, 1·05		
$P_{heterogeneity}$§					0·70						0·12	

TC1, treelet component 1; TC2, treelet component 2; Q1, quintile 1; Q2, quintile 2; Q3, quintile 3; Q4, quintile 4; Q5, quintile 5; Ref., reference category.
*Models were stratified by study centre and age in 1-year categories and adjusted for baseline menopausal status (premenopausal and perimenopausal (reference) or postmenopausal and women who underwent an ovariectomy), baseline alcohol intake (never drinkers (reference), former drinkers, drinkers only at recruitment, lifetime drinkers, unknown), height (continuous), BMI (below (reference) or above 25 kg/m^2), schooling level (none, primary (reference), technical/professional/secondary, longer education, unknown/unspecified), age at first full-term pregnancy (nulliparous (reference), ≤21 years, 21–30 years, >30 years, unknown or missing), age at menarche (≤12 years (reference), 12–14 years, >14 years, missing), age at menopause (≤50 years (reference), >50 years, pre-menopause or missing), use of hormone replacement therapy (never (reference), ever, unknown), level of physical activity (inactive (reference), moderately inactive, moderately active, active, unknown) and alcohol-free energy (continuous).
†P_{LRT}, *P* values for the likelihood ratio test (LRT) that was used to evaluate the overall significance of a score variable in quintile categories compared with a χ^2 distribution with df = 4.
‡P_{trend}, *P* values obtained by modelling score variables with quintile-specific medians as continuous variables.
§$P_{heterogeneity}$, *P* values for BC risks across ER status with df = 1 obtained using a data augmentation method.

unhealthy pattern, characterized by intakes of high-fat dairy products, red meat, processed meats and French fries, was not associated with BC risk. A recent study of the California Teachers Cohort identified a plant-based pattern, which was related to a reduction of BC risk[67]. In parallel, increasing evidence is accumulating that adherence to the *a priori* defined Mediterranean pattern is associated with a decreased BC risk[68–70], although results from these studies are not totally consistent, particularly for premenopausal women[70,71].

The dimension reduction techniques used herein were applied to nutrient densities. Nutrients are present in different combinations of foods, are less country-specific and are directly involved in biological reactions[72]. By exploring macro- and micronutrients, the present study aimed to provide an exhaustive representation of individuals' diet. Log-transformation was used to address scaling issues that can arise because macro- and micronutrients are expressed in different units. In this way, the variance and the components' decomposition are invariant to the unit of measure. Dietary normalization was achieved using equal energy, i.e. by dividing nutrient intakes by energy intake, minus energy from alcohol intake[44]. Most nutrients are associated with total energy because either they contribute to total energy directly or because people with higher energy values tend to display larger intakes of specific nutrients[44,73].

The first two patterns were retained as they were the most interpretable and depicted realistic nutrient patterns that could ultimately be linked with disease risk. The first pattern identified a diet characterized by animal products as opposed to a vegetarian diet, and was associated with a non-significant increase of 5 % in BC risk (TT). TC1 was quite comparable to a Western pattern. Two recent reviews on dietary patterns and BC[74,75] showed that diets rich in high-fat foods and processed meats were associated with an increased BC risk, although the findings described in both reviews have not been conclusive in this respect with most results reporting a positive association between Western-like dietary pattern and BC being not statistically significant[74,75]. In our study, the micronutrient-dense pattern characterized by a diet rich in vitamins and minerals, akin to a prudent pattern, was associated with an 11 % reduction in BC risk (TT), in line with previous findings[33,74,75]. The protective effect may come from the anti-carcinogenic properties of nutrients such as β-carotene, vitamins C and E, that may exert an antioxidant effect on oestrogen metabolism and reduce cell proliferation[75]. The TT components were

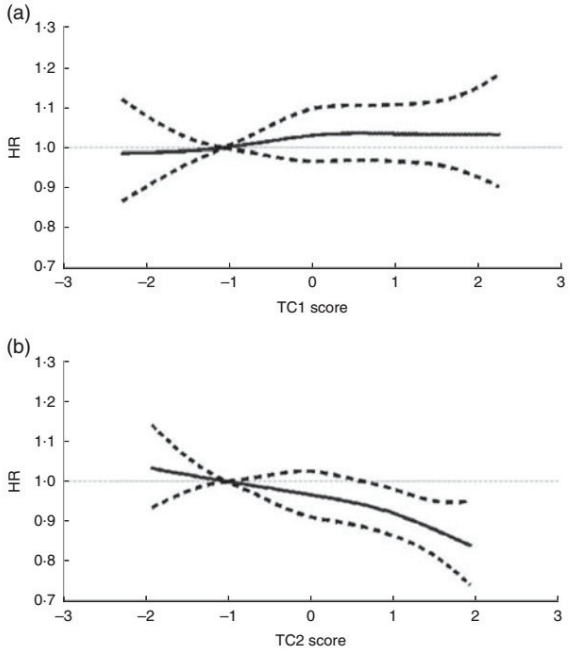


Fig. 2 Relationship between nutrient patterns derived from treelet transform and breast cancer risk (———, hazard ratio (HR); – – – – –, associated 95 % CI), obtained by using restrictive cubic splines with values of 1st and 99th percentiles and medians of quintiles 1, 3 and 5 used as knots, among 334 850 women in the European Prospective Investigation into Cancer and Nutrition (EPIC): (a) first treelet component (TC1), $P_{non\text{-}linearity} = 0{\cdot}94$, $P_{trend} = 0{\cdot}88$; (b) second treelet component (TC2), $P_{non\text{-}linearity} = 0{\cdot}77$, $P_{trend} = 0{\cdot}02$. Models were stratified by study centre and age in 1-year categories and adjusted for baseline menopausal status (premenopausal and perimenopausal (reference) or postmenopausal and women who underwent an ovariectomy), baseline alcohol intake (never drinkers (reference), former drinkers, drinkers only at recruitment, lifetime drinkers, unknown), height (continuous), BMI (below (reference) or above 25 kg/m^2), schooling level (none, primary (reference), technical/professional/secondary, longer education, unknown/unspecified), age at first full-term pregnancy (nulliparous (reference), ≤21 years, 21–30 years, >30 years, unknown or missing), age at menarche (≤12 years (reference), 12–14 years, >14 years, missing), age at menopause (≤50 years (reference), >50 years, pre-menopause or missing), use of hormone replacement therapy (never (reference), ever, unknown), level of physical activity (inactive (reference), moderately inactive, moderately active, active, unknown) and alcohol-free energy (continuous). P_{trend} was obtained by evaluating the joint significance of variables other than the linear one in the model by using Wald's test with df = 3

highly correlated with those of PCA ($\rho_{TC1,PC1} = 0{\cdot}91$, $\rho_{TC2,PC2} = 0{\cdot}86$). TT and PCA provided overall consistent findings in terms of pattern identification and amount of total variability explained. Further analyses were conducted by menopausal status at cohort enrolment, showing no differential association in pre- and postmenopausal women. Analyses carried out by hormonal receptor status showed that the second TT nutrient pattern was related to a significant decrease in BC risk for ER$^+$, PR$^+$, PR$^-$ and ER$^+$/PR$^+$ tumours. These results are complementary to previous literature findings on dietary patterns and hormonal defined risk of BC[58,67,70,75]. Indeed, Fung *et al.* found that a prudent dietary pattern was linked with decreased ER$^-$ risk (relative risk = 0·62, 95 % CI 0·45, 0·91)[76]. ER$^-$/PR$^-$ tumour risk was reduced in postmenopausal women among participants in the highest quintiles of a plant-based pattern and an *a priori* defined Mediterranean diet by 34 % and 20 %, respectively[67,70]. Results from the Pooling Project of Prospective Studies of Diet and Cancer found a protective association between total fruit or fruit and vegetable consumption in ER$^-$ tumours but not in ER$^+$ tumours or overall BC risk[77].

Whereas a large portion of the scientific literature on dietary patterns has used factor analysis or principal component factor analysis[74], the current paper promotes the use of TT. While PCA produces patterns that are eigenvectors of a covariance/correlation matrix of starting variables, TT is a multivariate technique that yields components by aggregating variables according to covariance/correlation[78], while at the same time exploring the clustering structure of variables, combining features of PCA with those of cluster analysis. Eventually, TT produces a cluster tree revealing the hierarchical grouping structure of variables. The dendrogram allows a visual inspection of the way different nutrients cluster, possibly easing interpretability of patterns. In addition, loadings are sparse, i.e. some of them are equal to zero as they do not pertain to the clustering node of the component so that a limited number of variables contributes to each treelet component.

In line with other clustering techniques[79], TT users are confronted with subjective decisions to select the appropriate cut-level for the cluster tree. Information on the grouping structure of variables that have joined (or not) the tree are specific to each level of the TT tree. By choosing a cut-level, the user decides on how much information to extract and the degree of sparsity of the components. If the tree is cut near the 'root', all nutrient variables join the tree. The information would be comparable to PCA output, i.e. all variables would contribute to treelet components. If the tree is cut closer to the 'leaves', i.e. when the cut-level is lower, loadings are sparse as many are equal to zero, possibly making the interpretation easier. By contrast, this may lead to components that do not capture dietary complexity and are therefore not informative. As pointed out by Meinshausen and Bühlmann, the use of TT leads to a trade-off between amount of variability explained and sparsity. The objective is to 'make the results as sparse as possible but not any sparser'[48]. To identify an optimal cut-level, cross-validation can be used. Once the cut-level is chosen, the loadings computed are invariant to the number of components to be retained; hence keeping n components is an *a priori* parameter to be specified in the cross-validation step.

The present study relied on dietary questionnaires to assess nutrient intakes, which are prone to measurement errors and may lack information on some relevant nutrients. Questionnaires were country-specific, potentially

Table 5 Hazard ratios (HR) and 95 % confidence intervals for breast cancer (BC) by quintiles of pattern scores (first and second components of treelet transform, cut-level 16) for oestrogen receptor-positive + progesterone receptor-positive (ER^+/PR^+) and oestrogen receptor-negative + progesterone receptor-negative (ER^-/PR^-) tumours in 334 850 women in the European Prospective Investigation into Cancer and Nutrition (EPIC)

	TC1						TC2					
Model*	Person-years	No. of BC cases	HR	95 % CI	P_{LRT}†	P_{trend}‡	Person-years	No. of BC cases	HR	95 % CI	P_{LRT}†	P_{trend}‡
ER^+/PR^+												
Q1	723 508	568	1·00	Ref.	0·16	0·26	737 812	753	1·00	Ref.	0·15	<0·01
Q2	728 884	811	1·15	1·03, 1·29			727 617	777	0·95	0·86, 1·05		
Q3	725 948	750	1·10	0·98, 1·23			719 931	777	0·94	0·84, 1·04		
Q4	726 667	751	1·11	0·98, 1·25			716 303	720	0·89	0·79, 0·99		
Q5	717 569	773	1·11	0·98, 1·26			720 914	626	0·86	0·76, 0·98		
ER^-/PR^-												
Q1	720 830	172	1·00	Ref.	0·60	0·31	734 117	218	1·00	Ref.	0·26	0·08
Q2	724 871	235	1·09	0·89, 1·33			723 844	241	1·05	0·87, 1·26		
Q3	722 003	207	0·93	0·75, 1·15			715 963	207	0·88	0·72, 1·08		
Q4	722 988	222	0·98	0·79, 1·23			712 804	210	0·93	0·76, 1·14		
Q5	713 798	214	0·97	0·77, 1·22			717 762	174	0·85	0·68, 1·06		
$P_{heterogeneity}$§					0·19						0·27	

TC1, treelet component 1; TC2, treelet component 2; Q1, quintile 1; Q2, quintile 2; Q3, quintile 3; Q4, quintile 4; Q5, quintile 5; Ref., reference category.
*Models were stratified by study centre and age in 1-year categories and adjusted for baseline menopausal status (premenopausal and perimenopausal (reference) or postmenopausal and women who underwent an ovariectomy), baseline alcohol intake (never drinkers (reference), former drinkers, drinkers only at recruitment, lifetime drinkers, unknown), height (continuous), BMI (below (reference) or above 25 kg/m^2), schooling level (none, primary (reference), technical/professional/secondary, longer education, unknown/unspecified), age at first full-term pregnancy (nulliparous (reference), ≤21 years, 21–30 years, >30 years, unknown or missing), age at menarche (≤12 years (reference), 12–14 years, >14 years, missing), age at menopause (≤50 years (reference), >50 years, pre-menopause or missing), use of hormone replacement therapy (never (reference), ever, unknown), level of physical activity (inactive (reference), moderately inactive, moderately active, active, unknown) and alcohol-free energy (continuous).
†P_{LRT}, *P* values for the likelihood ratio test (LRT) that was used to evaluate the overall significance of a score variable in quintile categories compared with a χ^2 distribution with df = 4.
‡P_{trend}, *P* values obtained by modelling score variables with quintile-specific medians as continuous variables.
§$P_{heterogeneity}$, *P* values for BC risks across ER/PR status with df = 1 obtained using a data augmentation method.

introducing systematic between-country differences in nutrient assessment. However, in the EPIC study, harmonized composition tables across European countries were used to translate food into nutrient intakes[42], thus sizeably improving the comparability of nutrient intakes.

One key element in pattern literature is reproducibility of patterns across populations. With twenty-three centres from ten countries, EPIC accounts for a wide heterogeneity in diet[80,81]. Previous findings in Moskal *et al.*'s study[45] on the EPIC data showed that more than 75 % of the variance that would be captured by centre-specific PC was captured by PC from overall PCA. This evidence suggested that overall PCA combining data from all EPIC centres allows capturing a good proportion of the variance explained by each EPIC centre. This motivated the choice of applying pattern decomposition on the overall data.

Conclusion

The current study presented results of a nutrient pattern analysis in an international setting using a new tool, TT, and subsequently related the patterns to risk of developing BC. TT is a complementary method to PCA in nutritional epidemiology as it produces readily interpretable sparse components. In the EPIC study, nutrient patterns characterized by a diet rich in macronutrients of animal origin, such as cholesterol or SFA, were associated with a non-significant increase in BC risk while a diet rich in vitamins, minerals and β-carotene, indicating a more plant-based diet, was associated with a significant decreased BC risk. This decrease was also significant for ER^+, PR^+, PR^- and ER^+/PR^+ tumours.

Acknowledgements

Acknowledgement: The authors thank Dr Anders Gorst-Rasmussen (Department of Cardiology, Aalborg University Hospital) for his critical input and useful discussions about the manuscript. *Financial support:* The coordination of the EPIC study is financially supported by the European Commission (Directorate General for Health and Consumer Affairs) and the International Agency for Research on Cancer (IARC). The national cohorts are supported by: the Health Research Fund (FIS) of the Spanish Ministry of Health RTICC 'Red Temática de Investigación Cooperativa en Cáncer (grant numbers Rd06/0020/0091 and Rd12/0036/0018), the Regional Governments of Andalucía, Asturias, Basque Country, Murcia (project 6236) and Navarra, and the 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, 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, Dutch Zorg Onderzoek Nederland, the World Cancer Research Fund and Statistics Netherlands (Netherlands); the European Research Council (2009-AdG 232997) and the Nordforsk, Nordic Centre of Excellence programme 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 (UK). The work undertaken by N.A. was supported by a Université de Lyon doctoral grant (EDISS doctoral school). *Conflict of interest:* None. *Authorship:* The authors' responsibilities were as follows. N.A. performed statistical analyses; N.A. and P.F. interpreted the findings and developed a first draft of the manuscript; A.M., N.S., V.V., V.C., H.F., S.M., S.K., J.F., E.W., L.L.-B. and I.R. contributed to the writing of the manuscript; P.A., E.A., E.M.-M., D.S., J.R.Q., A.O., A.Tj., C.C.D., K.O., L.D., A.F., L.B., R.T.F., R.K., A.Tr., C.B., P.O., M.S.D.M., G.M., C.A., F.R., R.Tu., H.B.B.d.M., M.F.B., P.H.M.P., G.S., T.B., A.W., I.J., K.-T.K., N.J.W., T.K., R.Tr., J.A.S., M.A.M. and E.R. substantially contributed to the interpretation of results and critically revised the content of the manuscript; and all authors contributed to the planning, execution and interpretation of the submitted manuscript, and read and approved the final manuscript. *Ethics of human subject participation:* This study was conducted according to the guidelines laid down in the Declaration of Helsinki and all procedures involving human subjects were approved by the IARC and the local ethical review committees. Written informed consent was obtained from all participants.

Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S1368980015000294

References

1. Bray F, Ren JS, Masuyer E *et al.* (2013) Estimates of global cancer prevalence for 27 sites in the adult population in 2008. *Int J Cancer* **132**, 1133–1145.
2. Key TJ, Verkasalo PK & Banks E (2001) Epidemiology of breast cancer. *Lancet Oncol* **2**, 133–140.
3. Collaborative Group on Hormonal Factors in Breast Cancer (2001) Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58 209 women with breast cancer and 101 986 women without the disease. *Lancet* **358**, 1389–1399.
4. Green J, Cairns BJ, Casabonne D *et al.* (2011) Height and cancer incidence in the Million Women Study: prospective cohort, and meta-analysis of prospective studies of height and total cancer risk. *Lancet Oncol* **12**, 785–794.
5. Chlebowski R (2007) Lifestyle change including dietary fat reduction and breast cancer outcome. *J Nutr* **137**, 1 Suppl., 233S–235S.
6. Anothaisintawee T, Wiratkapun C, Lerdsitthichai P *et al.* (2013) Risk factors of breast cancer: a systematic review and meta-analysis. *Asia Pac J Public Health* **25**, 368–387.
7. McKenzie F, Ellison-Loschmann L, Jeffreys M *et al.* (2013) Cigarette smoking and risk of breast cancer in a New Zealand multi-ethnic case–control study. *PLoS One* **8**, e63132.
8. Terry PD & Goodman M (2006) Is the association between cigarette smoking and breast cancer modified by genotype? A review of epidemiologic studies and meta-analysis. *Cancer Epidemiol Biomarkers Prev* **15**, 602–611.
9. Rohan TE, Heo M, Choi L *et al.* (2013) Body fat and breast cancer risk in postmenopausal women: a longitudinal study. *J Cancer Epidemiol* **2013**, 754815.
10. McCullough LE, Eng SM, Bradshaw PT *et al.* (2012) Fat or fit: the joint effects of physical activity, weight gain, and body size on breast cancer risk. *Cancer* **118**, 4860–4568.
11. Amadou A, Hainaut P & Romieu I (2013) Role of obesity in the risk of breast cancer: lessons from anthropometry. *J Oncol* **2013**, 906495.
12. World Cancer Research Fund/American Institute for Cancer Research (2010) Continuous Update Project Report. Food, Nutrition, Physical Activity, and the Prevention of Breast Cancer. http://www.dietandcancerreport.org/cancer_resource_center/downloads/cu/Breast-Cancer-2010-Report.pdf
13. Monninkhof EM, Elias SG, Vlems FA *et al.* (2007) Physical activity and breast cancer: a systematic review. *Epidemiology* **18**, 137–157.
14. Fagherazzi G, Vilier A, Boutron-Ruault M-C *et al.* (2014) Alcohol consumption and breast cancer risk subtypes in the E3N-EPIC cohort. *Eur J Cancer Prev* (Epublication ahead of print version).
15. Tjønneland A, Christensen J, Olsen A *et al.* (2007) Alcohol intake and breast cancer risk: the European Prospective Investigation into Cancer and Nutrition (EPIC). *Cancer Causes Control* **18**, 361–373.
16. Zhang SM, Lee I-M, Manson JE *et al.* (2007) Alcohol consumption and breast cancer risk in the Women's Health Study. *Am J Epidemiol* **165**, 667–676.
17. Cui X, Dai Q, Tseng M *et al.* (2007) Dietary patterns and breast cancer risk in the Shanghai breast cancer study. *Cancer Epidemiol Biomarkers Prev* **16**, 1443–1448.
18. Levi F, Pasche C, Lucchini F *et al.* (2001) Dietary intake of selected nutrients and breast-cancer risk. *Int J Cancer* **91**, 260–263.
19. Doll R (1992) The lessons of life: keynote address to the nutrition and cancer conference. *Cancer Res* **52**, 7 Suppl., 2024S–2029S.
20. Anand P, Kunnumakkara AB, Kunnumakara AB *et al.* (2008) Cancer is a preventable disease that requires major lifestyle changes. *Pharm Res* **25**, 2097–2116.
21. Ferrari P, Rinaldi S, Jenab M *et al.* (2013) Dietary fiber intake and risk of hormonal receptor-defined breast cancer in the European Prospective Investigation into Cancer and Nutrition study. *Am J Clin Nutr* **97**, 344–353.
22. Wu AH, Yu MC, Tseng C-C *et al.* (2003) Green tea and risk of breast cancer in Asian Americans. *Int J Cancer* **106**, 574–579.
23. Shrubsole MJ, Jin F, Dai Q *et al.* (2001) Dietary folate intake and breast cancer risk: results from the Shanghai Breast Cancer Study. *Cancer Res* **61**, 7136–7141.

24. Jacques PF & Tucker KL (2001) Are dietary patterns useful for understanding the role of diet in chronic diseases? *Am J Clin Nutr* **73**, 1–2.
25. Hu FB (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol* **13**, 3–9.
26. Jacobs DR & Steffen LM (2003) Nutrients, foods, and dietary patterns as exposures in research: a framework for food synergy. *Am J Clin Nutr* **78**, 3 Suppl., 508S–513S.
27. De Stefani E, Deneo-Pellegrini H, Boffetta P *et al.* (2009) Dietary patterns and risk of cancer: a factor analysis in Uruguay. *Int J Cancer* **124**, 1391–1397.
28. Edefonti V, Bravi F, Garavello W *et al.* (2010) Nutrient-based dietary patterns and laryngeal cancer: evidence from an exploratory factor analysis. *Cancer Epidemiol Biomarkers Prev* **19**, 18–27.
29. Nkondjock A, Krewski D, Johnson KC *et al.* (2005) Dietary patterns and risk of pancreatic cancer. *Int J Cancer* **114**, 817–823.
30. Schulze MB, Hoffmann K, Kroke A *et al.* (2001) Dietary patterns and their association with food and nutrient intake in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study. *Br J Nutr* **85**, 363–373.
31. Vrieling A, Buck K, Seibold P *et al.* (2013) Dietary patterns and survival in German postmenopausal breast cancer survivors. *Br J Cancer* **108**, 188–192.
32. De Stefani E, Boffetta P, Ronco AL *et al.* (2008) Nutrient patterns and risk of lung cancer: a factor analysis in Uruguayan men. *Lung Cancer* **61**, 283–291.
33. Brennan SF, Cantwell MM, Cardwell CR *et al.* (2010) Dietary patterns and breast cancer risk: a systematic review and meta-analysis. *Am J Clin Nutr* **91**, 1294–1302.
34. Dahm CC, Gorst-Rasmussen A, Crowe FL *et al.* (2012) Fatty acid patterns and risk of prostate cancer in a case–control study nested within the European Prospective Investigation into Cancer and Nutrition. *Am J Clin Nutr* **96**, 1354–1361.
35. Gorst-Rasmussen A, Dahm CC, Dethlefsen C *et al.* (2011) Exploring dietary patterns by using the treelet transform. *Am J Epidemiol* **173**, 1097–1104.
36. Schoenaker DAJM Dobson AJ, Soedamah-Muthu SS *et al.* (2013) Factor analysis is more appropriate to identify overall dietary patterns associated with diabetes when compared with Treelet transform analysis. *J Nutr* **143**, 392–398.
37. Jolliffe IT (2002) *Principal Component Analysis*, 2nd ed. New York: Springer-Verlag.
38. Riboli E, Hunt KJ, Slimani N *et al.* (2002) European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* **5**, 1113–1124.
39. Riboli E & Kaaks R (1997) The EPIC project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol* **26**, Suppl. 1, S6–S14.
40. Kaaks R, Slimani N & Riboli E (1997) Pilot phase studies on the accuracy of dietary intake measurements in the EPIC project: overall evaluation of results. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol* **26**, Suppl. 1, S26–S36.
41. Margetts B & Pietinen P (1997) European Prospective Investigation into Cancer and Nutrition: validity studies on dietary assessment methods. *Int J Epidemiol* **26**, Suppl. 1, S1–S5.
42. Slimani N, Deharveng G, Unwin I *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. *Eur J Clin Nutr* **61**, 1037–1056.
43. Bouckaert KP, Slimani N, Nicolas G *et al.* (2011) Critical evaluation of folate data in European and international databases: recommendations for standardization in international nutritional studies. *Mol Nutr Food Res* **55**, 166–180.
44. Willett WC, Howe GR & Kushi LH (1997) Adjustment for total energy intake in epidemiologic studies. *Am J Clin Nutr* **65**, 4 Suppl., 1220S–1228S.
45. Moskal A, Pisa P, Ferrari P *et al.* (2014) Nutrient patterns and their food sources in an international study setting: report from the EPIC study. *PLoS One* **9**, e98647.
46. Imamura F, Lichtenstein AH, Dallal GE *et al.* (2009) Confounding by dietary patterns of the inverse association between alcohol consumption and type 2 diabetes risk. *Am J Epidemiol* **170**, 37–45.
47. Gorst-Rasmussen A (2011) tt: treelet transform with Stata. *Stata J* **12**, 130–146.
48. Meinshausen N & Bühlmann P (2008) Discussion of: treelets – an adaptive multi-scale basis for sparse unordered data. *Ann Appl Stat* **2**, 478–481.
49. Thiébaut ACM & Bénichou J (2004) Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Stat Med* **23**, 3803–3820.
50. Ferrari P, Day NE, Boshuizen HC *et al.* (2008) The evaluation of the diet/disease relation in the EPIC study: considerations for the calibration and the disease models. *Int J Epidemiol* **37**, 368–378.
51. Lunn M & McNeil D (1995) Applying Cox regression to competing risks. *Biometrics* **51**, 524–532.
52. Heinzel H & Kaider A (1997) Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Comput Methods Programs Biomed* **54**, 201–208.
53. Männistö S, Dixon LB, Balder HF *et al.* (2005) Dietary patterns and breast cancer risk: results from three cohort studies in the DIETSCAN project. *Cancer Causes Control* **16**, 725–733.
54. Agurs-Collins T, Rosenberg L, Makambi K *et al.* (2009) Dietary patterns and breast cancer risk in women participating in the Black Women's Health Study. *Am J Clin Nutr* **90**, 621–628.
55. Terry P, Suzuki R & Hu FB (2001) A prospective study of major dietary patterns and the risk of breast cancer. *Cancer Epidemiol Biomarkers Prev* **10**, 1281–1285.
56. Sieri S, Krogh V, Pala V *et al.* (2004) Dietary patterns and risk of breast cancer in the ORDET Cohort. *Cancer Epidemiol Biomarkers Prev* **13**, 567–572.
57. Adebamowo CA, Hu FB, Cho E *et al.* (2005) Dietary patterns and the risk of breast cancer. *Ann Epidemiol* **15**, 789–795.
58. Fung TT, Hu FB, Holmes MD *et al.* (2005) Dietary patterns and the risk of postmenopausal breast cancer. *Int J Cancer* **116**, 116–121.
59. Nkondjock A & Ghadirian P (2005) Associated nutritional risk of breast and colon cancers: a population-based case–control study in Montreal, Canada. *Cancer Lett* **223**, 85–91.
60. Velie EM, Schairer C, Flood A *et al.* (2005) Empirically derived dietary patterns and risk of postmenopausal breast cancer in a large prospective cohort study. *Am J Clin Nutr* **82**, 1308–1319.
61. Hirose K, Matsuo K, Iwata H *et al.* (2007) Dietary patterns and the risk of breast cancer in Japanese women. *Cancer Sci* **98**, 1431–1438.
62. Murtaugh MA, Sweeney C, Giuliano AR *et al.* (2008) Diet patterns and breast cancer risk in Hispanic and non-Hispanic white women: the Four-Corners Breast Cancer Study. *Am J Clin Nutr* **87**, 978–984.
63. Wu AH, Yu MC, Tseng C *et al.* (2009) Dietary patterns and breast cancer risk in Asian American women. *Am J Clin Nutr* **89**, 1145–1154.
64. Cottet V, Touvier M, Fournier A *et al.* (2009) Postmenopausal breast cancer risk and dietary patterns in the E3N-EPIC prospective cohort study. *Am J Epidemiol* **170**, 1257–1267.
65. Ronco AL, de Stefani E, Aune D *et al.* (2010) Nutrient patterns and risk of breast cancer in Uruguay. *Asian Pac J Cancer Prev* **11**, 519–524.

66. Edefonti V, Decarli A, La Vecchia C *et al.* (2008) Nutrient dietary patterns and the risk of breast and ovarian cancers. *Int J Cancer* **122**, 609–613.
67. Link LB, Canchola AJ, Bernstein L *et al.* (2013) Dietary patterns and breast cancer risk in the California Teachers Study cohort. *Am J Clin Nutr* **98**, 1524–1532.
68. Trichopoulou A, Bamia C, Lagiou P *et al.* (2010) Conformity to traditional Mediterranean diet and breast cancer risk in the Greek EPIC (European Prospective Investigation into Cancer and Nutrition) cohort. *Am J Clin Nutr* **92**, 620–625.
69. Demetriou CA, Hadjisavvas A, Loizidou MA *et al.* (2012) The Mediterranean dietary pattern and breast cancer risk in Greek-Cypriot women: a case–control study. *BMC Cancer* **12**, 113.
70. Buckland G, Travier N, Cottet V *et al.* (2013) Adherence to the Mediterranean diet and risk of breast cancer in the European prospective investigation into cancer and nutrition cohort study. *Int J Cancer* **132**, 2918–2927.
71. Couto E, Sandin S, Lo M *et al.* (2013) Mediterranean dietary pattern and risk of breast cancer. *PLoS One* **8**, e55374.
72. Edefonti V, Hashibe M, Ambrogi F *et al.* (2012) Nutrient-based dietary patterns and the risk of head and neck cancer: a pooled analysis in the International Head and Neck Cancer Epidemiology consortium. *Ann Oncol* **23**, 1869–1880.
73. Freedman LS, Hartman AM, Kipnis V *et al.* (1997) Comments on: adjustment for total energy intake in epidemiologic studies. *Am J Clin Nutr* **65**, 1229–1231.
74. Edefonti V, Randi G, La Vecchia C *et al.* (2009) Dietary patterns and breast cancer: a review with focus on methodological issues. *Nutr Rev* **67**, 297–314.
75. Albuquerque RCR, Baltar VT & Marchioni DML (2014) Breast cancer and dietary patterns: a systematic review. *Nutr Rev* **72**, 1–17.
76. Fung TT, Rimm EB, Spiegelman D *et al.* (2001) Association between dietary patterns and plasma biomarkers of obesity and cardiovascular disease risk. *Am J Clin Nutr* **73**, 61–67.
77. Jung S, Spiegelman D, Baglietto L *et al.* (2013) Fruit and vegetable intake and risk of breast cancer by hormone receptor status. *J Natl Cancer Inst* **105**, 219–236.
78. Gorst-Rasmussen A, Dahm CC, Dethlefsen C *et al.* (2011) Gorst-Rasmussen et al. respond to 'Dietary pattern analysis'. *Am J Epidemiol* **173**, 1109–1110.
79. Krzanowski WJ (2000) *Principles of Multivariate Analysis: A User's Perpective*, 2nd ed. New York: Oxford University Press Inc.
80. Freisling H, Fahey MT, Moskal A *et al.* (2010) Region-specific nutrient intake patterns exhibit a geographical gradient within and between European countries. *J Nutr* **140**, 1280–1286.
81. Slimani N & Margetts B (2009) Nutrient intakes and patterns in the EPIC cohorts from 10 European countries. *Eur J Clin Nutr* **63**, Suppl. 4, S1–S274.

Online Supplementary Material

A treelet transform analysis to relate nutrient patterns to the risk of hormonal receptor–defined breast cancer in the European Prospective Investigation into Cancer and Nutrition study.

Supplementary Table 1: TT (cut-level 16) loadings of the third and fourth components.

Variables *	TT 16 loadings	
	TC3	TC4
Calcium (Ca)		
β-Carotene		
Cholesterol	-0.178	0.448
MUFA		
PUFA		
SFA		
Iron (Fe)		
Fibre		
Potassium (K)		
Magnesium (Mg)		
Phosphorus (P)		
Protein	-0.052	0.132
Retinol	-0.410	-0.609
Riboflavin		
Starch		
Sugar		
Thiamin		
Vitamin B_6		
Vitamin B_{12}	-0.254	0.641
Vitamin C		
Vitamin D	0.856	
Vitamin E		
Folate		
Explained variance	9%	6%

TC3, treelet component 3. TC4, treelet component 4.
* log-transformed nutrient variables.

Supplementary Table 2: PCA loadings of the 4 derived components.

Variables *†	PCA loadings			
	PC1	PC2	PC3	PC4
Calcium (Ca)	-0.024	0.12	-0.136	**0.314**
β-Carotene	**-0.275**	**0.601**	-0.121	**-0.495**
Cholesterol	**0.276**	0.07	-0.172	0.064
MUFA	0.018	-0.043	-0.123	-0.148
PUFA	-0.006	0.102	0.131	**-0.211**
SFA	0.119	-0.031	-0.155	-0.105
Iron (Fe)	-0.054	0.102	-0.019	0.048
Fibre	-0.131	0.145	0.136	0.006
Potassium (K)	-0.065	0.174	0.065	0.169
Magnesium (Mg)	-0.045	0.142	0.042	0.115
Phosphorus (P)	0.003	0.108	0.01	0.19
Protein	0.042	0.077	-0.003	0.159
Retinol	**0.601**	**0.271**	**-0.295**	-0.275
Riboflavin	0.004	**0.206**	-0.131	**0.322**
Starch	-0.004	-0.112	0.137	-0.068
Sugar	-0.098	0.073	0.01	0.175
Thiamin	-0.076	0.174	0.133	0.183
Vitamin B_6	-0.075	0.177	0.072	0.189
Vitamin B_{12}	**0.362**	**0.254**	**-0.266**	**0.306**
Vitamin C	**-0.276**	**0.316**	-0.033	0.126
Vitamin D	**0.431**	**0.25**	**0.796**	0.006
Vitamin E	-0.098	0.153	0.068	**-0.256**
Folate	-0.141	**0.249**	-0.014	0.105
Explained variance	28%	22%	10%	8%

PC1, principal component 1. PC2, principal component 2. PC3, principal component 3. PC4, principal component.

* log-transformed nutrient variables

† In bold are PCA loadings >0.20

Supplementary Table 3: HRs (95%CI) for BC by quintiles of pattern scores (1st and 2nd components of TT cut-level 16) for PR positive and PR negative tumours in EPIC women.

Model*		First component					Second component				
		PY	BC cases	HR (95% CI)	P-LRT[a]	P-trend[b]	PY	BC cases	HR (95% CI)	P-LRT[a]	P-trend[b]
PR Positive											
	Q1	723,730	611	1.00 (ref)			738,063	801	1.00 (ref)		
	Q2	729,055	850	1.12 (1.01,1.25)			727,815	823	0.96 (0.86,1.06)		
	Q3	726,226	805	1.10 (0.98,1.22)	0.31	0.28	720,137	827	0.95 (0.85,1.05)	0.17	<0.01
	Q4	726,869	800	1.10 (0.98,1.23)			716,542	766	0.90 (0.81,1.00)		
	Q5	717,755	812	1.10 (0.97,1.24)			721,078	661	0.87 (0.77,0.98)		
PR Negative											
	Q1	722,296	386	1.00 (ref)			735,796	467	1.00 (ref)		
	Q2	726,449	468	0.98 (0.86,1.13)			725,303	449	0.89 (0.78,1.02)		
	Q3	723,483	433	0.91 (0.79,1.06)	0.46	0.10	717,455	434	0.84 (0.73,0.96)	0.10	0.03
	Q4	724,668	468	0.99 (0.85,1.15)			714,395	454	0.90 (0.78,1.03)		
	Q5	715,243	435	0.90 (0.77,1.06)			719,189	386	0.84 (0.72,0.98)		
P- heterogeneity[c]					0.07					0.36	

HR: hazard ratio. 95%CI, 95% confidence interval. BC, breast cancer. PR, progesterone receptor. PY, person-years.

[a] P-LRT, p-values for the likelihood ratio test (LRT), that was used to evaluate overall significance of a score variable in quintile categories compared with a chi-square distribution with 4 df.

[b] P-trend values were obtained by modelling score variables with quintile-specific medians as continuous variables.

[c] P-heterogeneity values for BC risks across PR status on 1 df were obtained using a data augmentation method.

*Models were stratified by study centre and age in 1-y categories and adjusted for baseline menopausal status (premenopausal and perimenopausal [reference] or postmenopausal and women who underwent an ovariectomy), baseline alcohol intake (never drinkers [reference], former drinkers, drinkers only at recruitment, lifetime drinkers, unknown), height (continuous), BMI (below [reference] or above 25), schooling level (none, primary [reference], technical/professional/secondary, longer education, unknown /unspecified), age at first full-term pregnancy (nulliparous [reference], ≤ 21years, 21-30 years, > 30 years, unknown or missing), age at menarche (≤ 12 years [reference], 12-14 years, >14 years, missing), age at menopause (≤50 years [reference], > 50 years, pre-menopause or missing), use of hormones (never[reference], ever, unknown), levels of physical activity (inactive [reference], moderately inactive, moderately active, active, unknown) and alcohol-free energy(continuous).

Supplementary Table 4: HRs (95%CI) for BC by quintiles of pattern scores (1st and 2nd components of PCA) for overall, ER positive and ER negative tumours in EPIC women.

Model*		First component					Second component				
		PY	BC cases	HR (95% CI)	P-LRT[a]	P-trend[b]	PY	BC cases	HR (95% CI)	P-LRT[a]	P-trend[b]
Overall											
	Q1	729,222	1,843	1.00 (ref)			748,437	2,143	1.00 (ref)		
	Q2	736,877	2,292	1.03 (0.96,1.09)			737,177	2,339	1.03 (0.97,1.10)		
	Q3	734,382	2,445	1.06 (1.00,1.13)	0.29	0.07	732,009	2,280	0.98 (0.92,1.04)	0.15	0.046
	Q4	735,659	2,478	1.06 (1.00,1.13)			727,730	2,354	0.98 (0.99,1.05)		
	Q5	734,300	2,509	1.07 (1.00,1.15)			725,087	2,460	0.96 (0.89,1.02)		
ER Positive											
	Q1	723,700	882	1.00 (ref)			741,994	1,087	1.00 (ref)		
	Q2	730,480	1,201	1.07 (0.98,1.17)			730,010	1,142	1.00 (0.92,1.09)		
	Q3	727,426	1,260	1.09 (0.99,1.19)	0.27	0.09	725,034	1,113	0.94 (0.86,1.03)	0.46	0.10
	Q4	728,361	1,286	1.11 (1.01,1.22)			720,800	1,173	0.94 (0.86,1.03)		
	Q5	726,145	1,201	1.09 (0.99,1.21)			718,273	1,315	0.95 (0.86,1.04)		
ER Negative											
	Q1	719,177	215	1.00 (ref)			736,399	280	1.00 (ref)		
	Q2	724,194	287	1.01 (0.85,1.22)			724,298	312	1.10 (0.93,1.30)		
	Q3	720,958	333	1.13 (0.94,1.35)	0.56	0.91	719,335	301	1.05 (0.88,1.25)	0.02	0.11
	Q4	721,850	306	1.01 (0.83,1.22)			714,609	245	0.83 (0.69,1.00)		
	Q5	720,190	297	1.04 (0.85,1.27)			711,728	300	0.96 (0.80,1.16)		
P- heterogeneity[c]					0.80					0.13	

HR: hazard ratio. 95%CI, 95% confidence interval. BC, breast cancer. ER, estrogen receptor. PY, person-years.

[a] P-LRT, p-values for the likelihood ratio test (LRT), that was used to evaluate overall significance of a score variable in quintile categories compared with a chi-square distribution with 4 df.

[b] P-trend values were obtained by modelling score variables with quintile-specific medians as continuous variables.

[c] P-heterogeneity values for BC risks across ER status on 1 df were obtained using a data augmentation method.

*Models were stratified by study centre and age in 1-y categories and adjusted for baseline menopausal status (premenopausal and perimenopausal [reference] or postmenopausal and women who underwent an ovariectomy), baseline alcohol intake (never drinkers [reference], former drinkers, drinkers only at recruitment, lifetime drinkers, unknown), height (continuous), BMI (below [reference] or above 25), schooling level (none, primary [reference], technical/professional/secondary, longer education, unknown /unspecified), age at first full-term pregnancy (nulliparous [reference], ≤ 21years, 21-30 years, > 30 years, unknown or missing), age at menarche (≤ 12 years [reference], 12-14 years, >14 years, missing), age at menopause (≤50 years [reference], > 50 years, pre-menopause or missing), use of hormones (never[reference], ever, unknown), levels of physical activity (inactive [reference], moderately inactive, moderately active, active, unknown) and alcohol-free energy(continuous).

Supplementary Table 5: HRs (95%CI) for BC by quintiles of pattern scores (1st and 2nd components of PCA) for ER & PR positive and ER & PR negative tumours in EPIC women.

Model*	First component					Second component				
	PY	BC cases	HR (95% CI)	P-LRT[a]	P-trend[b]	PY	BC cases	HR (95% CI)	P-LRT[a]	P-trend[b]
ER and PR Positive										
Q1	721,384	525	1.00 (ref)			718,901	161	1.00 (ref)		
Q2	727,780	775	1.15 (1.03,1.29)			723,803	211	1.00 (0.81,1.23)		
Q3	724,554	805	1.16 (1.03,1.31)	0.07	0.04	720,508	242	1.09 (0.89,1.35)	0.77	0.65
Q4	725,315	790	1.16 (1.03,1.31)			721,445	224	0.98 (0.79,1.23)		
Q5	723,543	758	1.17 (1.03,1.33)			719,832	212	0.99 (0.78,1.25)		
ER and PR Negative										
Q1	739,692	743	1.00 (ref)			736,067	215	1.00 (ref)		
Q2	727,688	774	1.03 (0.93,1.14)			723,975	241	1.10 (0.91,1.32)		
Q3	722,601	720	0.96 (0.86,1.07)	0.38	0.09	718,949	214	0.97 (0.80,1.19)	0.06	<0.05
Q4	717,804	694	0.94 (0.84,1.05)			714,277	180	0.82 (0.66,1.02)		
Q5	714,791	722	0.94 (0.84,1.06)			711,222	200	0.90 (0.72,1.12)		
P- heterogeneity[c]				0.45						0.12

HR: hazard ratio. 95%CI, 95% confidence interval. BC, breast cancer. ER, estrogen receptor. PR, progesterone receptor. PY, person-years.

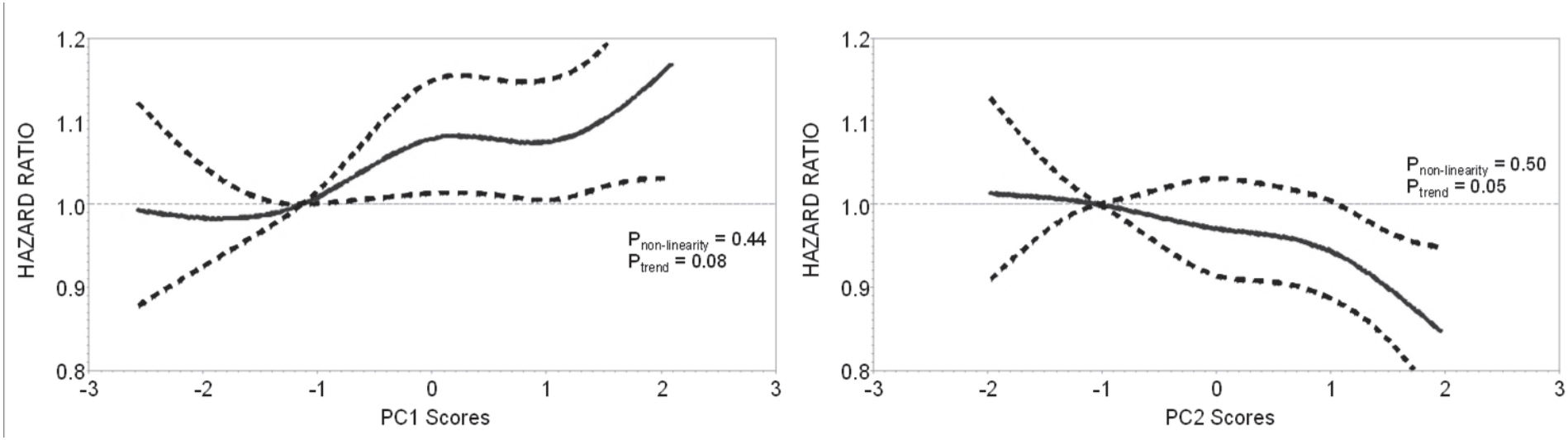
[a] P-LRT, p-values for the likelihood ratio test (LRT), that was used to evaluate overall significance of a score variable in quintile categories compared with a chi-square distribution with 4 df.

[b] P-trend values were obtained by modelling score variables with quintile-specific medians as continuous variables.

[c] P-heterogeneity values for BC risks across ER\PR status on 1 df were obtained using a data augmentation method.

*Models were stratified by study centre and age in 1-y categories and adjusted for baseline menopausal status (premenopausal and perimenopausal [reference] or postmenopausal and women who underwent an ovariectomy), baseline alcohol intake (never drinkers [reference], former drinkers, drinkers only at recruitment, lifetime drinkers, unknown), height (continuous), BMI (below [reference] or above 25), schooling level (none, primary [reference], technical/professional/secondary, longer education, unknown /unspecified), age at first full-term pregnancy (nulliparous [reference], ≤ 21years, 21-30 years, > 30 years, unknown or missing), age at menarche (≤ 12 years [reference], 12-14 years, >14 years, missing), age at menopause (≤50 years [reference], > 50 years, pre-menopause or missing), use of hormones (never[reference], ever, unknown), levels of physical activity (inactive [reference], moderately inactive, moderately active, active, unknown) and alcohol-free energy(continuous).

Supplementary Figure 1: Relations between PCA nutrient patterns and BC risk (and associated 95%CI) obtained by using restrictive cubic splines with values of 1st and 99th percentile and medians of quintiles 1, 3 and 5 used as knots.



Models were stratified by study centre and age in 1-y categories and adjusted for baseline menopausal status (premenopausal and perimenopausal [reference] or postmenopausal and women who underwent an ovariectomy), baseline alcohol intake (never drinkers [reference], former drinkers, drinkers only at recruitment, lifetime drinkers, unknown), height (continuous), BMI (below [reference] or above 25), schooling level (none, primary [reference], technical/ professional/ secondary, longer education, unknown / unspecified), age at first full-term pregnancy (nulliparous [reference], ≤ 21years, 21-30 years, > 30 years, unknown or missing), age at menarche (≤ 12 years [reference], 12-14 years, >14 years, missing), age at menopause (≤50 years [reference], > 50 years, pre-menopause or missing), use of hormones (never[reference], ever, unknown), levels of physical activity (inactive [reference], moderately inactive, moderately active, active, unknown) and alcohol-free energy(continuous). P-linearity was obtained by evaluating the joint significance of variables other than the linear one in the model by using Wald's test with 3 df.

CHAPTER III:

A STATISTICAL FRAMEWORK FOR THE "MEETING-IN-THE-MIDDLE" APPLIED TO UNTARGETED METABOLOMIC DATA

CONTEXT

Biosciences in the era of Big Data have undergone a profound change in the way research is focused, structured and executed. Particularly, recent technological advances in the fields of molecular biology and spectrometry resulted in an increased availability of ever-complex high-dimensional –omics datasets. Such data pose logistical challenges pertaining to their storage, their processing but also to analytical approaches to fully exploit them [173]. Aside from the well-established genomics, -omics also encompass a variety of other fields including transcriptomics, epigenomics, proteomics and metabolomics, an opportunity to examine the "exposome" (i.e., the entirety of life-course environmental exposures) in a comprehensive manner [216]. Unlike the genome, the "exposome" is modifiable, and can be explored through exposure-biomarker approach. One such approach has emerged through the "Meeting-In-The-Middle" (MITM) principle, a research strategy that can potentially reveal exposure-specific biomarkers that are at the same time predictive of morbid conditions [162,217] by looking at associations between exposures, intermediate markers and disease, particularly in settings using metabolomics. This is best investigated in prospective studies which are especially well-tailored for this purpose as they rely on biological samples collected before disease onset, often at recruitment, and therefore are marginally influenced by metabolic changes that arise in the disease-development process.

OBJECTIVES

- To conceive a statistical framework for the MITM approach whose aim is to identify biomarkers that are related to specific exposures and that are, at the same time, predictive of disease outcome.
- To include multivariate techniques in the analytical framework for dimensionality reduction and relating different sets of data.
- To apply the analytical strategy within the European Prospective Investigation into Cancer and nutrition (EPIC) where biological samples were collected at baseline in disease-free participants. Untargeted metabolomic data was acquired using NMR techniques from subjects in a nested case-control study on hepatocellular carcinoma (HCC), for which information on lifestyle and dietary exposures was available.

APPROACH

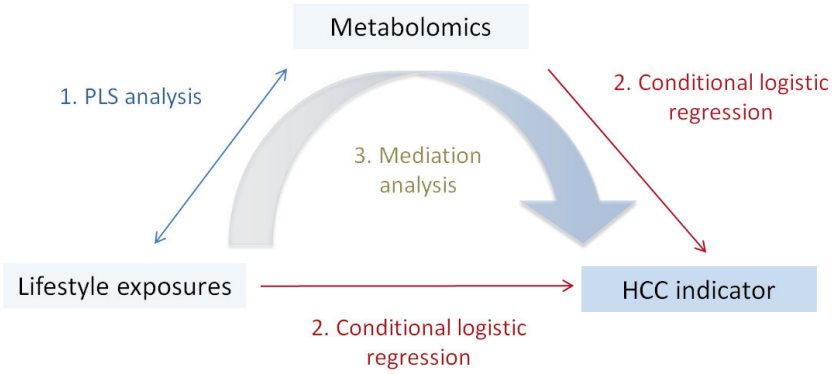


Figure 2: General original scheme to model the MITM principle.

The analytical strategy for the MITM was applied towards an analysis of the dietary and lifestyle determinants of HCC. In a case-control study on HCC nested within EPIC, serum ^{1}H NMR spectra (800 MHz) were acquired for 114 cases and 222 matched controls, and resulted in 285 metabolic variables (the "responses"). These made up the metabolomics set that was related to a set of 21 lifestyle variables (the "predictors", including information on diet, anthropometry and clinical attributes) through Partial Least Squares (PLS) (**Figure 2**). PLS is most suitable for this purpose, as it generalizes features of Principal Component Analysis (PCA) and Multiple Linear Regression (MLR), by iteratively extracting components that maximize the covariance between two sets of variables [218,219]. This resulted on the one hand in extracting the bulk of information explaining the most variability, and on the other hand in retaining a restricted number of factors, achieving dimensionality reduction. The derived scores were related to HCC risk in conditional logistic regressions, and odds ratios and their corresponding 95% confidence intervals were computed (OR, 95%CI). Finally, the mediating role of the metabolomic signatures between the lifestyle profiles and risk of developing HCC was assessed in mediation analyses [208].

MAIN FINDINGS

PLS allowed the simultaneous identification of relevant lifestyle and metabolic factors whose link can be predictive in the aetiology of chronic diseases. Three PLS factors reflected in a lifestyle and metabolic components were selected. A first lifestyle factor characterized by a healthy pattern with negative loadings for diabetes status, smoking

status and lifetime alcohol intake was not associated with HCC risk, neither was its metabolomics counterpart. The lifestyle component of the second PLS factor reflected a 'higher-risk exposures' lifestyle pattern, and showed a significant 54% increase in HCC risk. Likewise, its associated metabolic component displayed a significant HCC risk rise by 11%. The third PLS lifestyle factor included participants with lower vegetables intake, elevated lifetime alcohol consumption, more likely to be ever smokers and have a hepatitis infection; one standard deviation increase of this component was associated with a statistically significant 37% increase in HCC risk. Similarly, its metabolic counterpart characterised by positive signals of ethanol and myoinositol and negative loadings for glucose displayed a 22% significant increase in HCC risk.

CONCLUSION

This integrated framework allowed the use of all potentially informative aspects of high-dimensional data including untargeted metabolomics, dietary and lifestyle exposures and disease outcome resulting in intermediate biomarker signatures discovery. This study devised a way to bridge lifestyle variables to HCC risk through NMR metabolomics data possibly highlighting the intersection of relevant markers of exposure with predictive markers of disease outcome. This implementation of the MITM was applied towards the investigation of HCC determinants; it can be easily extended to similar aetiological contexts and to settings characterized by high-dimensional data, increasingly frequent in the –omics generation.

PAPER

Contribution: First author, discussed the analytical strategy with the supervisor, conducted statistical analyses, wrote the first draft of the manuscript, submitted it to the journal and replied to reviewers' comments.

Mutagenesis Advance Access published July 7, 2015

Mutagenesis, 2015, 1–11
doi:10.1093/mutage/gev045
Original Manuscript

Original Manuscript

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[2,19], 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,†] and 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, UK, [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]Department of Hygiene, Epidemiology and Medical Statistics, WHO Collaborating Center for Food and Nutrition Policies, 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, Umeå, 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 [19]Present address: Chemical Physics Department, Weizmann Institute of Science, Rehovot, Israel.

*To whom correspondence should be addressed. Tel: +33 472 73 8031; Fax: +33 472 73 8361. E-mail: ferrarip@iarc.fr
†These authors contributed equally to this work.

Received 10 April 2015; Revised 18 May 2015; Accepted 3 June 2015.

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 nuclear magnetic resonance (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 1SD 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 characterised by high-dimensional data, increasingly frequent in the omics generation.

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 aetiological 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 characterising 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 and2000, over 520 000 participants aged 20–85 years, were recruited from 23 centres in 10 Western European countries including Denmark, France, Germany, Greece, Italy, Norway, Spain, Sweden, the Netherlands and UK (11). The design, rationale and methods of the EPIC study including information on dietary assessment methodology, blood collection protocols and follow-up procedures were discussed previously (11).

Between 1992 and 1998, standardised 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 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 h) and fasting status at blood collection (<3, 3–6, >6 h); 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 8500 bins of 0.001 ppm width over the chemical shift range of 0.5–9 ppm. Spectra were normalised to total intensity, centred and Pareto scaled, and additionally normalised for batch effects using the batch profiling calibration method (16). After removal of the structured noise (characterised by a specific mean and standard deviation) located in a well-known noise region (8.5–9 ppm) 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 Supplementary Appendix, available at *Mutagenesis* Online. This permitted a reduction of the number of NMR variables from 8500 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, available at *Mutagenesis* Online.

Statistical analyses

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) intercorrelation 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 h). 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 generalises 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 maximise their covariance (22). The factors extracted from

the predictors' set are orthogonal. Computational details of PLS are described in the Supplementary Appendix, available at *Mutagenesis* Online. As a standard step for the PLS algorithm, the X- and Y-sets were centred and standardised for the analysis and a simple expectation–maximisation (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 7-fold cross validation analysis was carried out to select the number h of significant PLS factors to retain (8) (see Supplementary Appendix, available at *Mutagenesis* Online). This was achieved by splitting the data into seven groups of observations. In turn, each group of observations was considered as the test set, while 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 minimising 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 characterise 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 10th percentile and larger than the 90th percentile for the X variables, and lower than the 5th and larger than the 95th 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 (1SD) 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 summarises 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 ($\pi = 0.0004$), 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. 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.* (33) For each one of the 1000 drawn bootstrap samples, a 10-fold cross-validation was performed, repeated 10 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 2.5th and 97.5th percentile values made up the 95% confidence intervals.

Sensitivity analysis

A sensitivity analysis was performed by running PLS on data excluding sets where cases were diagnosed within the first 2 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 2 years from enrollment 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 Supplementary Appendix, available at *Mutagenesis* Online).

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 (34). Briefly, mediation was computed using a Baron and Kenny approach adapted to dichotomous outcomes (35), 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) (34). 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 (35). 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 centre and age at blood collection for sake of consistency with previous steps of the analysis.

Statistical analyses were performed using R (36) and SAS (37) in general, with the following packages for specific purposes: PROC PLS in SAS 9.4 for PLS analyses, 'paramed' in Stata 12 (38) 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%. PLS analysis was carried controlling for this variable.

After a 7-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 characterising 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 0.859 and 0.853, respectively. An increase in the resampled cross-validated AUC values was also observed for all three X- and Y-scores, albeit smaller, with 0.836 and 0.827, respectively. Results from the sensitivity analysis conducted on data excluding sets where cases were diagnosed within the first 2 years of follow-up, showed similarities in terms of lifestyle variables' and metabolites' loadings on the PLS factors (Supplementary Table 2, available at *Mutagenesis* Online). 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, body mass index (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, available at *Mutagenesis* Online). 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, available at *Mutagenesis* Online). 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

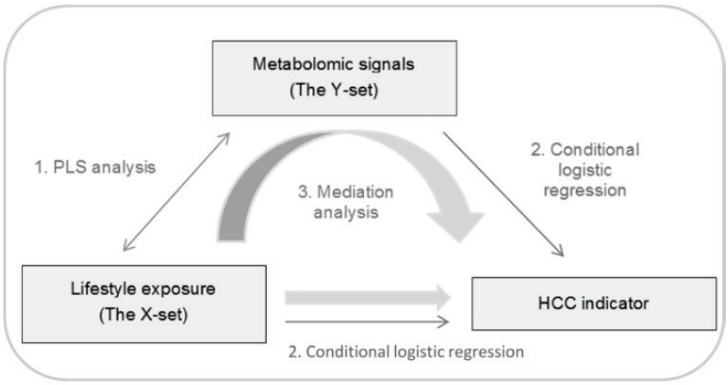


Figure 1. General scheme of the analytical framework developed in the study. A PC-PR2 analysis is carried out beforehand to identify relevant sources of variation. In the PLS model, the X- and Y-sets are related to each other, and scores are computed (1). X- and Y-scores are, in turn, associated to a 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).

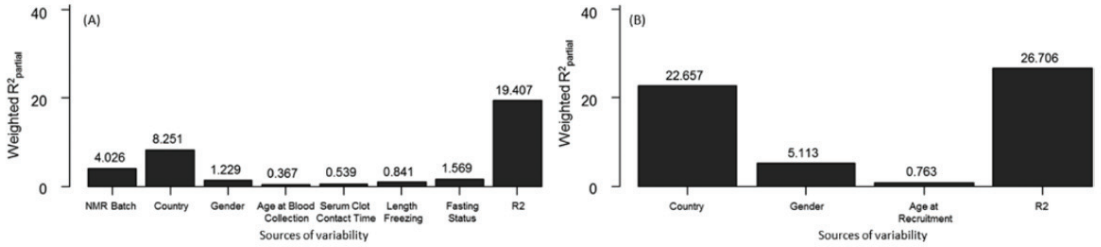


Figure 2. PC-PR2 analysis results* identifying the sources of variability in the NMR data (**A**) and in the lifestyle data (**B**).
* 17 and 14 components were retained to account for 80% (threshold used) of total NMR (A) and lifestyle variability (B), respectively. The R2 value represents the amount of variability in NMR/lifestyle variable explained by the ensemble of investigated predictors.

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

factor for the Y-set. ROC-associated statistics for different models are presented in Supplementary Table 3, available at *Mutagenesis* Online. 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 (A) and for the sensitivity model (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 NIE 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 subcohort 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 covariability in sets of predictors and response variables (8,39). 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 characterisation of metabolomic signatures. In addition, diabetes is the main metabolic risk factor for HCC alongside with fatty liver disease (40,41), and chronic infection with hepatitis B (HBV) and particularly hepatitis C (HCV) viruses were classified as class I carcinogens for HCC by IARC (42).

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 (26,43), 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 summarising data variability through a limited number of factors. Based on a 7-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 characterising 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

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[a]	Loading value	CS (ppm)[a,b]	Metabolite[c]	Loading value
1	Dairy products	0.28	5.22	Glucose	−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		−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–CH_2–COOC bond of lipids + acetone	−0.04
			1.86	Lysine	0.09
			1.87		0.10
			1.53	CH_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	CH=CH 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 + myoinositol	−0.12
			4.05	Creatinine	−0.14
			3.88	Glucose	0.15
			3.82		0.16
			3.76		0.15
			3.71		0.15
			3.54		0.15
			3.50		0.16
			3.48		0.16
			3.44	Acetoacetate	0.16
			3.23	Choline + glycerophosphocholine	0.15
			2.80	Aspartate	−0.12
			2.22	CH_2–CH_2–COOC bond of lipids + acetone	−0.11
			2.19	CH_2–CH_2–COOC bond of lipids	−0.15
			2.02	Proline + glutamate + CH2=C bonds of lipids	−0.13
			1.53	CH_2–CH_2–COOC bond of lipids	−0.13
			1.25	CH_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–CH_2	0.11
	Hepatitis	0.26	3.88	Glucose	−0.11
			3.82		−0.11
			3.76		−0.12
			3.71		−0.11
			3.69		−0.11
			3.63	Myoinositol	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	Myoinositol	0.12
			3.23	Choline + glycerophosphocholine	−0.12
			2.80	Aspartate	−0.13
			2.76	part of =CH–CH_2–CH= bond of lipids	−0.13

Table 2. *Continued*

PLS factor	Lifestyle variable[a]	Loading value	CS (ppm)[a,b]	Metabolite[c]	Loading value
			2.35	Proline + glutamate	0.12
			2.33		0.13
			1.20	3-hydroxybutyrate + CH_2 bond of lipids	0.11
			1.16	Ethanol	0.15
			0.66	Cholesterol	0.11

[a]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.

[b]CS: ^{1}H chemical shift (ppm) of the cluster (centre value).

[c]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[a] 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[b] (95% CI)	P-Wald[c]	Factor	OR[b] (95% CI)	P-Wald[c]
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

[a]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 h), fasting status at blood collection (<3/3–6/>6 h); among women, additional matching criteria included menopausal status (pre-/peri-/postmenopausal) and hormone replacement therapy use at time of blood collection (yes/no).

[b]ORs expressing the change in HCC risk associated to 1SD increase in the score.

[c]Wald's test was for continuous exposure compared with a Chi-square distribution with 1 degree of freedom (dof).

risk was observed for its metabolic counterpart, characterised by positive signals of ethanol and myoinositol, 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 (34). 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 (44). In a setting with low HCC prevalence (π = 0.0004), in line with Western populations (45), extremely low PPV estimates were observed. In the absence of a very specific test, many false positive tests arise from disease-free individuals (44), thus leading to a dilution of PPV.

A sensitivity analysis was carried out excluding the first 2 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 (46,47). In such strategies, only one set of variables is multidimensional 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 (48–50). In addition, canonical correlation analysis (CCA) (51) is a method applied to identify the optimum structure or dimensionality of each variable set that maximises the relationship between two sets of multidimensional variables. The main difference between CCA and PLS regression is that CCA maximises the correlation between the two new dimensions, i.e. extracted factors, whereas PLS maximises their covariance. PLS can be considered as a trade-off between CCA and PCA, since maximising the covariance corresponds to maximising 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 SRV (17,52), 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

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[b]	Sensitivity	Specificity	Accuracy	PPV
Adjustment covariates (ADJ)[a]	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

[a]The model is run on the ADJ including the C-reactive protein concentration, alpha-fetoprotein concentration and a composite score for liver damage.

[b]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[a] and 95% confidence intervals (OR, 95%CI) associated with the lifestyle (X-set) and the NMR clusters (Y-set) PLS scores in the sensitivity analysis (N=271, 92 cases, 179 controls)

PLS lifestyle variables X-scores			PLS NMR variables Y-scores		
Factor	OR[b] (95% CI)	P-Wald[c]	Factor	OR[b] (95% CI)	P-Wald[c]
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

The sensitivity analysis was conducted excluding sets where cases were diagnosed within the first 2 years of follow-up (X-set = 21, Y-set = 285).

[a]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 h), fasting status at blood collection (<3/3–6/>6 h); among women, additional matching criteria included menopausal status (pre-/peri-/postmenopausal) and hormone replacement therapy use at time of blood collection (yes/no).

[b]ORs expressing the change in HCC risk associated to 1SD increase in the score.

[c]Wald's test was for continuous exposure compared with a Chi-square distribution with 1 degree of freedom (dof).

time (53), and the Human Metabolome Database (HMDB) and other related resources (18,54), 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 (54).

Throughout the different steps of this work, the scaling problem was first tackled by normalising spectra to total intensity. NMR data were also centred 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

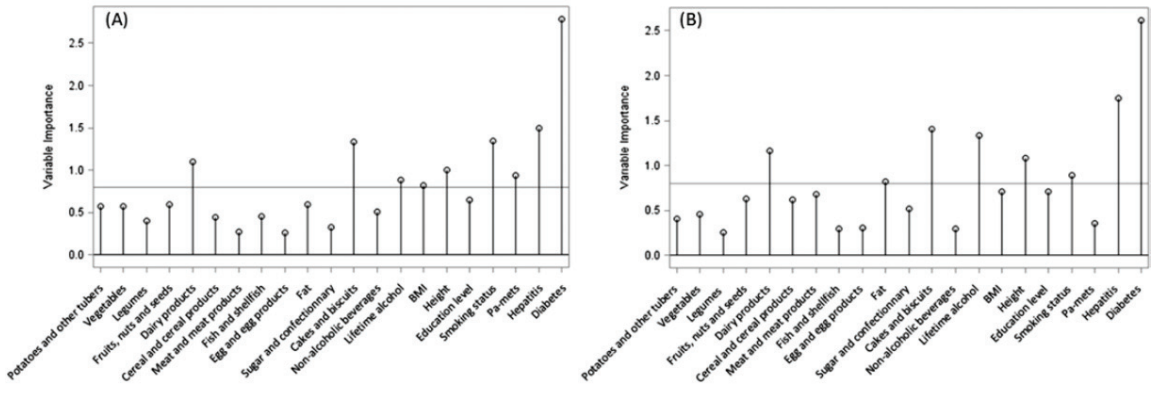


Figure 3. Variable importance plot (VIP) displaying the variable importance for projection statistic of the predictor variables for the PLS analyses. (**A**) Results from the main PLS model run on all observations (*N* = 336, X-set = 21, Y-set = 285). (**B**) Results from the PLS sensitivity analysis run on a subsample (*N* = 271, 92 cases, 179 controls) excluding sets where cases were diagnosed within the first 2 years of follow-up (X-set = 21, Y-set = 285). The horizontal line corresponds to Wold's criterion (0.8), the threshold used to rule if a variable has an important contribution to the construction of the Y variables (see Supplementary Appendix, available at *Mutagenesis* Online for further details).

Table 6. Results from the mediation analysis ($N = 336$, X-set = 21, Y-set = 285): natural indirect effect (NIE) and 95%CI[a]

Model[b]				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

[a]The standard errors used to compute the 95% CI were obtained using the delta method.

[b]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.

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 centre.

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.

Supplementary data

Supplementary Tables 1–3 and Appendix are available at *Mutagenesis* Online.

Funding

The coordination of EPIC is financially supported by the European Commission (DG-SANCO) and the International Agency for Research on Cancer. The national cohorts are supported by Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); Deutsche Krebshilfe, Deutsches Krebsforschungszentrum and Federal Ministry of Education and Research (Germany); the Hellenic Health Foundation (Greece); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); Nordic Centre of Excellence programme on Food, Nutrition and Health. (Norway); Health Research Fund (FIS), PI13/00061 to Granada), Regional Governments of Andalucía, Asturias, Basque Country, Murcia (no. 6236) and Navarra, ISCIII RETIC (RD06/0020) (Spain); Swedish Cancer Society, Swedish Scientific Council and County Councils of Skåne and Västerbotten (Sweden); Cancer Research UK (14136 to EPIC-Norfolk; C570/A16491 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk) (United Kingdom).

This work was supported by the French National Cancer Institute (L'Institut National du Cancer; INCA) [grant number 2009-139; PI: M. Jenab]. The work undertaken by N Assi was supported by the Université de Lyon I through a doctoral fellowship awarded by the EDISS doctoral school.

Acknowledgements

We would like to acknowledge the assistance of Dr Elodie Jobard from the ISA-CRMN in obtaining the annotation of the NMR data.
Conflict of interest statement: None declared.

References

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. (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. *et al.* (2014) Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study. *Int. J. Obes. (Lond).*, 38, 1388–1396.
4. Trushina, E. and Mielke, M. M. (2014) 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. *et al.* (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. (2013) High-field NMR metabonomics for investigation of cancer in human populations and metabolic perturbations in model systems. PhD Thesis, Ecole Normale Supérieure de Lyon.
11. Riboli, E., Hunt, K.J., Slimani, N. *et al.* (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. *Int. J. Epidemiol*., 26, 26–36.
13. Trichopoulos, D., Bamia, C., Lagiou, P. *et al.* (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., 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. *et al.* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res*., 37, D603–D610.
19. Cui, Q., Lewis, I. A., Hegeman, A. D. *et al.* (2008) Metabolite identification via the Madison Metabolomics Consortium Database. *Nat. Biotechnol*., 26, 162–164.
20. Slimani, N., Deharveng, G., Unwin, I. *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*. Duxbury Press, Belmont, CA.
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. *et al.* (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. *et al.* (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. (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. and Moons, K. G. (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. 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.
34. 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.
35. Vanderweele, T. J. and Vansteelandt, S. (2010) Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol*., 172, 1339–1348.
36. R Foundation for Statistical Computing and R Core Team. (2013) *R: A language and environment for statistical computing*.
37. *Base SAS® 9.4 Procedures Guide*. (2012) SAS Institute Inc., Cary, NC.
38. *Stata Statistical Software: Release 12*. (2011) StataCorp. College Station, TX.
39. Abdi, H. (2010) Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdiscip. Rev. Comput. Stat*., 2, 97–106.
40. 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.
41. Gomaa, A.-I. (2008) Hepatocellular carcinoma: Epidemiology, risk factors and pathogenesis. *World J. Gastroenterol*., 14, 4300–4308.
42. Cogliano, V.J., Baan, R., Straif, K. *et al.* (2011) Preventable exposures associated with human cancers. *J. Natl. Cancer Inst*., 103, 1827–1839.
43. Bialecki, E. S. and Di Bisceglie, A. M. (2005) Diagnosis of hepatocellular carcinoma. *HPB (Oxford)*., 7, 26–34.
44. 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.
45. Leong, T. Y. and Leong, A. S. (2005) Epidemiology and carcinogenesis of hepatocellular carcinoma. *HPB (Oxford)*., 7, 5–15.
46. Rothwell, J., Fillâtre, Y., Martin, J.-F. *et al.* (2014) New biomarkers of coffee consumption identified by the non-targeted metabolomic profiling of cohort study subjects. *PLoS One*, 9, e93474.
47. Guo, M., Zhao, B., Liu, H. *et al.* (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.
48. Anderson, T. W. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Stat*., 22, 327–351.
49. Izenman, A. J. (1975) Reduced-rank regression for the multivariate linear model. *J. Multivar. Anal*., 5, 248–264.
50. Aldrin, M. (2002) Reduced-rank regression. In El-Shaarawi, A. H., Piegorsch, W. W. (ed.), *Encyclopedia of Environmetrics*. John Wiley & Sons Ltd, Chichester, pp. 1724–1728.
51. Hotelling, H. (1936) Relations between two sets of variates. *Biometrika*, 28, 321–377.
52. Navratil, V., Pontoizeau, C., Billoir, E. and Blaise, B. J. (2013) SRV: an open-source toolbox to accelerate the recovery of metabolic biomarkers and correlations from metabolic phenotyping datasets. *Bioinformatics*, 29, 1348–1349.
53. Wishart, D. S. (2008) Quantitative metabolomics using NMR. *TrAC Trends Anal. Chem*., 27, 228–237.
54. Psychogios, N., Hau, D. D., Peng, J. *et al.* (2011) The human serum metabolome. *PLoS One*, 6, e16957.

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		-0.07
	Lifetime Alcohol Consumption	-0.34	3.88		-0.06
	Smoking Status	-0.26	3.82		-0.07
	Diabetes	-0.59	3.76	Glucose	-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.

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 $i = 1, \ldots, r$:

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

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.

CHAPTER IV:

A REFINEMENT OF THE "MEETING-IN-THE MIDDLE" FRAMEWORK WITH AN APPLICATION IN TARGETED METABOLOMICS

CONTEXT

The MITM principle [162,180] was used as a research strategy to identify biomarkers that are related to specific exposures and that are also predictive of disease outcome, by looking at associations between exposures, contender intermediate markers and disease. This strategy is particularly of interest in epidemiological studies with metabolomic data. A first implementation of the MITM principle was presented as a proof of concept [220], it explored intermediate biomarkers separately, relating them to nutrient variables and to colon and breast cancer in a nested case-control study. In our first MITM paper [208], we set-up a single statistical framework by integrating multivariate methods, namely PLS, and mediation analyses, to fully exploit data originating from different high-dimensional sets. Building on these previous implementations of the MITM, and using targeted metabolomic data, we further refined and developed the analytical scheme by focusing on a restricted set of exposures and by adapting the mediation analysis to matched case-control study designs. The application looked yet again into determinants of HCC, the most common form of liver cancer, which ranks as the 2nd most frequent cause of cancer death worldwide [209]. HCC being a multi-factorial disease strongly associated with lifestyle factors and with dietary habits [221], components of a modified Healthy Lifestyle Index (HLI) scores' link with serum metabolites are jointly investigated to possibly identify modifiable lifestyle exposure patterns and metabolite signatures related to HCC that may ultimately lead to the identification of targeted cancer prevention schemes.

OBJECTIVES

- To apply the MITM approach in order to explore the components from a modified HLI with respect to serum metabolites in a nested case-control study on HCC within the EPIC cohort. Targeted metabolites were acquired through the BiocratesKit from pre-diagnostic sera samples.
- To further establish and tune the analytical framework previously developed to yield exposure-specific metabolomics profiles through multiple PLS.
- To develop and adapt the mediation analysis structure to accommodate the matched nested case-control design.

APPROACH

Following a similar scheme as in the previous MITM implementation, for 147 HCC cases and their matched controls, 132 metabolites levels were acquired from pre-diagnostic serum samples using standard targeted metabolite profiling protocols (BiocratesKit). Through PLS analysis, this metabolomics set, including an additional liver damage score, was linked to a set of 7 lifestyle variables corresponding to components of HLI, including diet, Body Mass Index (BMI) (kg/m^2), physical activity (hourly Metabolic Equivalent of Task Met-h/week), lifetime alcohol consumption (g/day), smoking, diabetes at baseline and hepatitis infection. A series of multiple PLS was further applied using each HLI variable separately to yield metabolite patterns that are specific to each exposure under scrutiny. Mediation analyses were then performed to assess the mediating role of the metabolomic profiles in the relationships between the overall lifestyle profile first, then for each individual HLI component in turn and HCC. Estimates of the Natural Direct Effect (NDE) and Natural Indirect Effect (NIE) were computed by adapting formulae from VanderWeele & Vansteelandt (AJE, 2010) [192], to accommodate conditional logistic regressions for the matched design. Total effects were also presented. Statistical significance controlled for multiple testing through False Discovery Rate (FDR) in the multiple PLS results.

MAIN FINDINGS

In the overall analysis, the lifestyle PLS factor scored high for study subjects characterised, on average, by low propensity towards smoking, alcohol drinking and obesity. Its metabolic counterpart was positively related to sphingolipids with hydroxyl group including SM(OH) C14:1, SM(OH) C16:1 and SM(OH) C22:2, and negatively with glutamic acid, hexoses, PC aaC32:1 and liver damage score. Both components displayed decreased HCC risks quantified with total effects through with odds ratios (OR) equal to 0.53[95% CI: 0.39, 0.71] and mediator effects adjusted for the exposure OR=0.30[0.19, 0.47] per 1-SD change in components' scores, respectively. There was evidence of mediation between this overall "healthy" pattern and HCC through its metabolic counterpart with NIE=0.62 [0.50, 0.77]. Results from multiple PLS, showed that specific metabolic signatures of BMI, alcohol intake, diet, smoking and diabetes were found to be mediators of the relationship between corresponding HLI variables and HCC risk. Their

respective NIE was equal to 1.56[1.24, 2.96], 1.09[1.03, 1.15], 0.85[0.74, 0.97], 1.22[1.04, 1.44] and 5.11[1.99, 13.14].

CONCLUSION

Using a multiple PLS scheme within a MITM framework, we were able to yield lifestyle-specific metabolomic signatures. These metabolic profiles bridged healthy behaviours to HCC risk through mediation analyses. The models were fine-tuned and metabolomic signals specific to BMI, alcohol intake, diet, smoking and diabetes were found to be mediators on the pathway between each of these exposures and risk of developing HCC. Future studies applying the MITM should utilize larger sample sizes for improved power. Nevertheless, the present work clearly offers the utility of the MITM in exploring environment-disease associations in an integrated setting with highly-dimensional data.

PAPER

Contribution: First author, discussed the analytical strategy with the supervisor, conducted statistical analyses, wrote the first draft of the manuscript, submitted it to peer-review journals.

The manuscript is currently in draft format. It has been circulated to the writing group, and to EPIC collaborators. After a first unsuccessful submission to the Journal of the National Cancer Institute (JNCI), it has been now submitted and is under consideration at the International Journal of Epidemiology (IJE).

The meeting-in-the-middle framework using metabolomics in the relationship between lifestyle factors and hepatocellular carcinoma risk

Nada Assi[1], Duncan Thomas[2], Michael Leitzmann[3], Magdalena Stepien[1], Véronique Chajès[1], Thierry Philip[4], Paolo Vineis[5], Christina Bamia[6,7], Marie-Christine Boutron-Ruault[8,9], Torkjel M Sandanger[10], Amaia Molinuevo[11,12], Hendriek Boshuizen[13], Anneli Sundkvist[14], Tilman Kühn[15], Ruth Travis[16], Kim Overvad[17], Elio Riboli[5], Marc Gunter[1], Augustin Scalbert[1], Mazda Jenab[1], Pietro Ferrari[1*], Vivian Viallon [18]

[1] Section of Nutrition and Metabolism, International Agency for Research on Cancer (IARC), Lyon, France. [2] University of Southern California, Los Angeles, CA, USA. [3] Department of Epidemiology and Preventive Medicine, Regensburg University, Regensburg, Germany. [4]Unité Cancer et Environnement, Centre Léon Bérard, 28 rue Laennec, 69373, Lyon 08 Cedex, France. [5] Department of Epidemiology and Biostatistics, MRC-HPA Centre for Environment and Health, School of Public Health, Imperial College London, Norfolk Place W2 1PG London, UK. [6] Hellenic Health Foundation, Athens, Greece. [7] WHO Collaborating Center for Nutrition and Health, Unit of Nutritional Epidemiology and Nutrition in Public Health, Dept. of Hygiene, Epidemiology and Medical Statistics, University of Athens Medical School, Greece. [8] Université Paris-Saclay, Université Paris-Sud, UVSQ, CESP, INSERM, Villejuif, France. [9] Gustave Roussy, F-94805, Villejuif, France. [10] Department of Community Medicine, UiT the Arctic University of Norway, Tromsø, Norway. [11] Public Health Division of Gipuzkoa, Regional Government of the Basque Country, Spain. [12] CIBER of Epidemiology and Public Health (CIBERESP), Spain. [13] National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands. [14] Department of Radiation Sciences Oncology, Umeå University 901 87 Umeå, Sweden. [15] Division of Cancer Epidemiology, German Cancer

Research Center (DKFZ), Heidelberg, Germany. [16] Cancer Epidemiology Unit, University of Oxford, Oxford OX3 7LF, UK. [17] The Department of Epidemiology, School of Public Health, Aarhus University, Aarhus, Denmark. [18] Université de Lyon, Université Claude Bernard Lyon1, Ifsttar, UMRESTTE, UMR T_9405, F- 69373, LYON

*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

Background: The “meeting-in-the-middle” (MITM) is a principle to identify exposure biomarkers that are also predictors of disease. The MITM statistical framework was applied in a nested case-control study on hepatocellular carcinoma (HCC) within the EPIC cohort where the components of a modified healthy lifestyle index (HLI) were related to serum metabolites.

Methods: Lifestyle and targeted metabolomic data were available from 147 HCC cases and 147 matched controls. Partial Least Squares (PLS) analysis related 7 modified HLI variables (diet, BMI, physical activity, lifetime alcohol, smoking, diabetes, hepatitis) to 132 serum-measured metabolites, and a liver function score. Exposure-specific signatures were also extracted with PLS models. Mediation analysis evaluated the role of metabolomic PLS scores in the relationship between the modified HLI and HCC risk.

Results: The overall PLS factor's lifestyle component was negatively associated with lifetime alcohol, BMI, smoking, diabetes and positively associated with physical activity. Its metabolic counterpart was positively related to SM(OH) C14:1, C16:1 and C22:2, and negatively to glutamate, hexoses, and PC aaC32:1. The lifestyle and metabolomics components were inversely related to HCC risk. The PLS scores expressing metabolic signatures mediated the association between smoking and lifetime alcohol and HCC with Natural Indirect Effects respectively equal to 1.22(95% confidence interval [CI]=1.04 to 1.44) and 1.09(95%CI=1.03 to 1.15).

Conclusions: This study refined the analytical framework of the MITM principle as a way to investigate the relations between lifestyle factors and disease risk using metabolomics. Relevant metabolomic signatures were identified as mediators in the relationship between specific lifestyle exposures and HCC.

Keywords: Meeting-in-the-middle, mediation analysis, partial least squares, hepatocellular carcinoma, targeted metabolomics, healthy lifestyle index, EPIC.

Key Messages:

- This work presents a flexible analytical framework for the "meeting-in-the-middle" principle, a promising tool to potentially identify causal pathways. The statistical strategy relied on an integrative approach to relate exposures to a wide array of metabolomics data in relation to hepatocellular carcinoma outcome.
- Using an individual Partial Least Square approach, exposure-specific metabolic signatures were identified and were shown to be predictive for disease outcome. This was especially noteworthy for BMI, alcohol, smoking as well as diabetes- specific metabolic profiles.
- The approach can be further extended to similar aetiological contexts and/or using other types of –Omics data.

Introduction

Metabolomics have become a focal point in epidemiological studies, as a result of large scale collection of biological samples and technological advances in the fields of molecular biology and chemometrics[1–4]. Metabolomics offers a broad spectrum of potential biomarkers to explore in search of causal and mechanistic pathways in disease development and aetiology. Such endeavours have revealed a number of mechanistic insights in the understanding of disease progression at metabolic levels and led to biomarker discovery[5].

Metabolomic datasets raise challenges from the processing of complex high-dimensional data, to the analytical approaches to fully exploit them[1]. New statistical methodologies are increasingly sought to address the multivariate nature of metabolomic data[6] and to discover relevant pathological processes that metabolomics may help investigate. In this scenario, the "meeting-in-the-middle" (MITM) principle[7,8] is used as a research strategy to identify biomarkers that are related to specific exposures and that are, at the same time, predictive of the outcome.

The MITM has been previously implemented in a nested case-control study where intermediate biomarkers were related to nutrients and to colon and breast cancer indicators[9]. The implementation to multivariate modelling was further extended in a Partial Least Squares (PLS) analysis to integrate a set of 21 lifestyle variables and 285 metabolic variables from ^{1}H NMR spectra in relation to hepatocellular carcinoma (HCC) risk[10].

Since HCC is a multi-factorial disease strongly associated with lifestyle factors[11], the MITM was applied to identify metabolite signatures related to HCC. The lifestyle components of a *modified* healthy lifestyle index (HLI)[12,13] were related to specific metabolic signals.

In this study an in-depth proof of concept of the MITM is revisited with a focused strategy to explore the mediating role of metabolic signatures on the path from exposure to disease in a HCC case-control study nested within the European Prospective Investigation into Cancer and nutrition (EPIC) using targeted metabolomic data.

Material and Methods

The nested case-control design

Within a nested case-control study of HCC[14,15] in EPIC, this study focused on 147 cases and 147 matched controls with available biological samples identified in the period between subjects' recruitment into the cohort (1993-1998) and 2010[15,16]. Cases of HCC originated from all participating EPIC centres except for Norway and France that were not a part of this study. All subjects were cancer-free at the time of blood collection. Information on population, data collection of dietary and lifestyle data, follow-up, case ascertainment and matching criteria can be read in **Supplementary Methods**.

The lifestyle variables (X-set of predictors)

The lifestyle variables were the predictors, referred to as the X-set, and included body mass index (BMI) (continuous, kg/m²), average lifetime alcohol intake (continuous, g/day), the diet score (continuous) described in the **Supplementary Methods**, physical activity (continuous metabolic equivalents of task in MET-h/week), smoking (never, ex-smokers quit>10 years, ex-smokers quit <=10 y, current smokers <=15 cig/day, current smokers > 15 cig/day), hepatitis infection (yes/no) and self-reported diabetes at baseline (yes/no). These are the components of a healthy lifestyle index (HLI)[12,13], hereby modified to include hepatitis and diabetes status, as detailed in **Supplementary Methods**.

The metabolites set (M-set of responses)

Metabolomic data

Metabolic biomarkers from serum samples were measured by tandem mass spectrometry at IARC, Lyon, France, using the BIOCRATES AbsoluteIDQ p180 Kit (Biocrates,

Innsbruck, Austria). Details of the sample preparation and mass spectrometry analyses are provided elsewhere[15,17]. Out of 145 metabolites measured in serum, this study included 132 metabolites with at most 40% of missing values. Metabolite nomenclature has been previously described[18] and can be found in **Supplementary Methods**. Measurements that were below the limit of detection were set to half that value and those below limit of quantification were set to half that limit (applicable to a total of 16 metabolites for 0.3% to 29.3% of participants). Additionally, measurements that were above the highest concentration calibration standards were set to the highest values.

Liver function score

A composite score indicative of liver function identifying the number of abnormal values for six circulating liver blood biomarker tests indicating possible underlying liver dysfunction[10,14,15] was included in the set of metabolites, the M-set, as detailed in **Supplementary Methods**. These biomarkers were acquired at the same time as the metabolites from the pre-diagnostic blood samples collected at recruitment.

Statistical analyses

Modified HLI and HCC risk

The association between the modified HLI and HCC risk was evaluated through conditional logistic regression models. Odds ratios, and their 95% confidence intervals (OR, 95%CI) were computed to express a change in HCC risk reflecting one standard deviation (1-SD) increase in the index. Unadjusted and liver function score adjusted ORs were estimated.

Principal Component Partial R-squared (PC-PR2) analyses

Sources of systematic variability within the X-set of HLI variables and the M-set of metabolites were identified and quantified through the PC-PR2 method[10,19] as described in **Supplementary Methods**. For both X- and M-sets, residuals on country and batch (M-set only) were computed in univariate linear regression models and used in the PLS analyses.

Primary PLS analyses: overall and individual PLS

Exposure variables were related to metabolomic data through PLS analysis that extracts linear combinations, referred to as PLS factors, of predictors (the X-set of lifestyle variables) and responses (the M-set of metabolites), allowing a simultaneous decomposition of both sets with the aim of maximizing their covariance[20,21]. An overall PLS was conducted using the entire X-set, then a series of individual PLS analyses was further applied using each HLI variable separately as the predictor to yield exposure-specific metabolomics signatures. In an attempt to yield even more specific metabolic signatures, sensitivity PLS analyses using mutually adjusted lifestyle residuals and country for the X-set and with country and batch residuals for the M-set were computed and presented in **Supplementary Tables**. More details on the process are provided in **Supplementary Material**.

Mediation analyses

Mediation analysis assessed whether the metabolic profiles mediated the relation between individual lifestyle factors and HCC risk. For the overall and individual PLS analyses, mediating effects were computed for each extracted pair of lifestyle variable and M-score, adapting the formulae from VanderWeele and Vansteelandt[22] to accommodate continuous exposures and mediators and conditional logistic regression for our matched setting. For each examined lifestyle variable, estimates of the natural direct effect (NDE),

the natural indirect effect (NIE), and the total effect (TE) were obtained, along with the effect of the corresponding M-score adjusted for its counterpart lifestyle exposure and for confounding variables and referred to as the mediator effect (ME). For more details, see **Supplementary Methods**.

All statistical tests were two-sided and p-values < 0.05 were considered statistically significant. Statistical analyses were performed using PROC PLS in SAS[23] for PLS analyses and the R Software[24] for linear and conditional logistic regressions and mediation analyses.

Results

Study population characteristics by case-control status are presented in **Table 1**. One PLS factor was retained after 7-fold cross validation for PLS analysis. The lifestyle PLS factor identified a 'healthy' behavior profile with positive loadings for physical activity, negative loadings for BMI, lifetime alcohol consumption and smoking (**Table 2**). The corresponding metabolomics PLS factor was characterized by glutamic acid, hexoses and sphingomyelins. The PLS lifestyle factor was inversely associated with HCC risk, with TE=0.53 (95%CI=0.39-0.71, P_{value}=2.64E-05) **(Table 4),** whereas the HLI score was not related to HCC with OR=0.93, 95%CI=0.84 to 1.02, P_{value}=0.117 (results not shown). The PLS metabolic profile showed a strong inverse association with HCC risk, with ME (Mediator Effect) equal to 0.30 (95%CI=0.19 to 0.47, P_{value}=1.94E-07). The association of the lifestyle factor with HCC risk was mediated by the metabolic profile, with NIE=0.62 (0.50 to 0.77, P_{value}=2.12E-05), with an estimated mediated proportion of 52% (**Table 4**).

Individual PLS analyses yielded metabolite signatures for each component of the modified HLI (**Table 3**). For lifetime alcohol, the signature was negatively related to SM C16:1, SM C18:1, SM(OH) C14:1, SM(OH) C16:1 and SM(OH) C22:2 and positively related to glutamic acid and PC aaC32:1. Metabolites associated with smoking included SM C16:1 and C18:1, SM(OH) C14:1 and C22:2, LysoPC aC28:1 and PC aeC30:2 with negative loadings and hexoses with positive loadings. In the sensitivity analysis, smoking was negatively associated with serine, lysine and biogenic taurine and positively with PC aaC36:1 and aaC40:3 (**Supplementary Table 3**). Different phosphatidylcholines characterized the metabolic signature related to diet. The metabolic profile of BMI included glutamic acid, tyrosine, PC aaC38:3, the liver function score with positive loadings and glutamine, LysoPC aC17:0 and

LysoPC aC18:2 with negative values. Hexoses and amino acids valine, isoleucine and phenylalanine were positively associated with diabetes status.

All PLS metabolic signatures, with the exception of physical activity and hepatitis infection, were associated with HCC risk, with strong evidence of mediation (**Table 4**). In particular, for both diabetes and BMI, a positive association for the NIE, equal to 5.11 (1.99 to 13.14, P_{value}=6.99E-04) and 1.56 (1.24 to 1.96, P_{value}=1.72E-04), respectively, was observed, together with a lack of association for the NDE, thus suggesting that the relationship between these two variables and HCC risk was fully mediated by the corresponding metabolic signatures. As for smoking, diet and lifetime alcohol, the mediated proportions were 56%, 38% and 24%, respectively, with NIE equal to 1.22 (1.04 to 1.44, P_{value}=0.018), 0.85 (0.74 to 0.97, P_{value}=0.025) and 1.09 (1.03 to 1.15, P_{value}=0.002), respectively. Noteworthy, the NIE estimate for smoking in the sensitivity analysis was 1.98 (1.34 to 2.92, P_{value}=5.65E-04), and the relation between smoking and HCC was fully mediated by the M-score (**Supplementary Table 4**).

The TE estimates showed strong associations for lifetime alcohol (1.40, 95%CI=1.14 to 1.72, P_{value}=1.40E-03), diet score (0.66, 0.47 to 0.92, P_{value}=0.014) and hepatitis infection (16.70, 4.82 to 57.84, P_{value}=8.92E-06) (**Table 4**). Most of these associations remained statistically significant after FDR correction. With the exception of smoking and, to a lesser extent, lifetime alcohol, the PLS metabolic profiles and estimated associations were virtually unchanged in the sensitivity analysis (**Supplementary Tables 3 and 4**).

256 **Discussion**

257 This study extended the statistical framework of the MITM[10] with a focused
258 strategy to comprehensively explore the mediating role of metabolite signatures in the
259 relationship between HLI and HCC.

260 In a previous implementation of the MITM[10], 21 lifestyle variables were related to
261 285 metabolic variables acquired from pre-diagnostic sera ^{1}H NMR spectra. In this study ,
262 the X-set of predictors was restricted to the original components of the HLI, most of which
263 have been previously associated with HCC risk[11,25–34]. Variables from the existing
264 index[12,13] were complemented by indicators of hepatitis infection and diabetes status at
265 baseline, which are well-known HCC risk factors[25,26,35]. Alcohol use at recruitment was
266 replaced by lifetime alcohol intake, mainly to address reverse causality. A more focused
267 methodology was further developed building on a similar analytical structure.

268 PLS analysis was used to relate the sets of HLI variables to metabolites. Preliminarily,
269 an overall factor depicted a lifestyle pattern characterized by low propensity towards
270 smoking, alcohol drinking and obesity, low prevalence of baseline diabetes or hepatitis
271 infection and high levels of physical activity. Mediation analyses indicated the metabolite
272 signature mediated 52% of the association between the healthy lifestyle factor and risk of
273 HCC. In a second phase, individual PLS models were related to specific components of the
274 HLI. The specific metabolite signatures were found to mediate the relation with HCC risk for
275 BMI, lifetime alcohol consumption, smoking, diabetes and diet, with a proportion mediated
276 of 100, 24, 56, 100 and 38%, respectively. These findings suggested that varying proportion
277 of the total effect on HCC is exerted via the metabolite signatures, possibly through specific
278 underlying mechanisms by which the exposure is acting.

279 Specifically, a recent IARC handbook evaluation on body fatness and obesity
280 reported a positive relationship between BMI and risk of liver cancer[36]. Our study
281 suggests that the increase in HCC risk is entirely mediated by a BMI-specific metabolic
282 signature characterized by phosphatidylcholines (LysoPC aC18:2, LysoPC aC17:0 and PC
283 aeC36:2) and tyrosine. PCs are required for lipoprotein assembly and secretion; in particular
284 acyl-alkyl-PCs were correlated with high-density cholesterol[37,38]. Tyrosine levels
285 imbalance has been previously related to insulin resistance and type 2 diabetes[39–41].
286 Correlation studies conducted in the EPIC-Potsdam cohort exploring the association
287 between lifestyle factors and blood metabolite levels, acquired with the same targeted
288 technology showed similar findings, with serum acyl-alkyl-phosphatidylcholines (PC ae),
289 LysoPC aC17:0, aC18:2 and PC aeC36:2 negatively associated with obesity and BMI whereas
290 tyrosine was positively related to BMI[42–44].

291 The metabolic signature fully mediated the association between diabetes, a well-
292 established HCC risk factor[11], and HCC. The contributing metabolites were hexoses,
293 phenylalanine and LysoPCs, consistently with previous studies based on targeted[41] and
294 untargeted[45] approaches. These metabolites were further linked with insulin resistance
295 and involved in glycolysis and gluconeogenesis, and their metabolic alterations was
296 associated with an increased diabetes risk[41].
297 The metabolomics signature of lifetime alcohol intake was negatively associated with
298 sphingomyelins and positively associated to phosphatidylcholines. Similar metabolites
299 patterns were observed in a study that focused on alcohol-dependent patients [46]. As
300 ethanol has been hypothesised to induce lipogenesis in the liver tissues[47], alcohol can

301 lead to hepatic injuries causing a disruption of the metabolism of fatty acids and
302 phospholipids[48].
303 The identification of specific metabolic signatures for alcohol and smoking was particularly
304 challenging in our study, as these two factors are strongly correlated[49–51]. An overlap
305 between the smoking and alcohol-specific metabolite signatures was observed in the
306 preliminary analysis, where four common sphingomyelins , i.e. SM C16:1, SM C18:1, SM(OH)
307 C14:1 and SM(OH) C22:2,were identified. In the sensitivity analysis, the different lifestyle
308 exposures were mutually adjusted for prior to PLS, thus leading to a new list of metabolites
309 associated with smoking which included serine, SM(OH) C22:2 and PC aaC36:1, consistently
310 to what was reported in the KORA study[52]. As a result, the estimated proportion of
311 mediation increased from 57 to 100 %, resulting in a metabolic signature capturing smoking-
312 related metabolic features that is more predictive of HCC.

313 The application of mediation analysis in this study was another challenging aspect of
314 the analytical framework. A temporal sequence among, in turn, lifestyle exposures,
315 metabolites and outcome is required[53,54] for the NDE and NIE to have a causal
316 interpretation. In our study, while cancer occurrence was assessed during follow-up,
317 lifestyle exposures were assessed at baseline, at the same time of the collection of biological
318 samples that provided metabolomics data. In this respect it is worth noticing that lifestyle
319 and metabolomics reflect different exposure windows. The metabolites likely reflect
320 exogenous and endogenous exposures in a limited timeframe, i.e. between weeks and a few
321 months as the reliability studies that of serum metabolomics data seem to
322 indicate[17,18,55]. The diet score was derived from questionnaires that covered the dietary
323 habits of participants over the past 12 months prior to baseline[56,57]. While lifetime

alcohol reflected the history of exposure across adult life, other exposures such as BMI, smoking, physical activity, hepatitis infection and diabetes status were the result of one point in time assessment at recruitment. Our analytical framework study consistently relied on the hypothesis that lifestyle factors were stable over time in the middle-age study populations recruited in EPIC.

Another key aspect of mediation analysis is what is referred to as the 'cross-world assumption', whereby NDE and NIE cannot be identified in the presence of a mediator-outcome confounder that is affected by the exposure[58]. In our study the composite liver function score, an index compiled from measures of circulating biomarkers of hepatic function indicating underlying liver impairment[14] was likely affected by lifestyle exposure, and was, in turn, likely influencing metabolite levels and HCC risk. The use of weighting-based estimation methods to look at joint mediators to compute randomized interventional effects has been proposed as a solution in the presence of mediator-outcome confounder[58]. In this study the liver function was added to the list of mediators. In this way, the metabolic signatures comprised of relevant information on the liver function, and the link with relevant lifestyle factors was evaluated.

This study was characterized by limited sample size, a direct consequence of the fact that HCC is a rare disease. Findings from this comprehensive approach suggested that certain exposure-specific metabolite profiles are intermediate biomarkers on the metabolic pathway towards hepatocellular carcinogenesis, but replication of these findings in an independent setting is warranted.

This study further refined an endeavor for high-throughput data to integrate metabolomics, lifestyle exposures together with disease indicators. Metabolomics lends

347 itself as a promising tool to identify metabolites bridging the link between exposure(s) and
348 disease, as advocated by the MITM principle[7,8]. The framework we developed allows the
349 identification of informative metabolic signatures, which are useful to elucidate the
350 underlying biological mechanisms in the relationship between lifestyle exposure to risk of
351 cancer risk[59].

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 EPIC is financially supported by the European Commission (DG-SANCO) and the International Agency for Research on Cancer. The national cohorts are supported by Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); Deutsche Krebshilfe, Deutsches Krebsforschungszentrum and Federal Ministry of Education and Research (Germany); the Hellenic Health Foundation (Greece); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); Nordic Centre of Excellence programme on Food, Nutrition and Health. (Norway); Health Research Fund (FIS), PI13/00061 to Granada), Regional Governments of Andalucía, Asturias, Basque Country, Murcia (no. 6236) and Navarra, ISCIII RETIC (RD06/0020) (Spain); Swedish Cancer Society, Swedish Scientific Council and County Councils of Skåne and Västerbotten (Sweden); Cancer Research UK (14136 to EPIC-Norfolk; C570/A16491 and C8221/A19170 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk and MR/M012190/1 to EPIC-Oxford) (United Kingdom). The work undertaken by N Assi was supported by the Université de Lyon I through a doctoral fellowship awarded by the EDISS doctoral school.

"For information on how to submit an application for gaining access to EPIC data and/or biospecimens, please follow the instructions at http://epic.iarc.fr/access/index.php"

Conflict of Interest

None to declare.

Acknowledgements:

The authors wish to thank Dr Joshua Sampson from the Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA, for useful discussions and insightful comments on this work. The authors would like to extend their thanks to Mr Bertrand Hémon and Ms Carine Biessy from the International Agency for Research on Cancer for their kind help with issues related to data management.

References

1. Baker M. The ’Omes Puzzle. Nature. 2013;494:416–9.

2. Nicholson JK, Holmes E, Elliott P. The metabolome-wide association study: a new look at human disease risk factors. J Proteome Res. 2008 Sep;7(9):3637–8.

3. Wild CP, Scalbert A, Herceg Z. Measuring the Exposome: A Powerful Basis for Evaluating Environmental Exposures and Cancer Risk. Environ Mol Mutagen. 2013;54(3):480:499.

4. Wild CP. Complementing the Genome with an “‘ Exposome ’”: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. Cancer Epidemiol biomarkers Prev. 2005;14(August):1847–51.

5. Abu Bakar MH, Sarmidi MR, Cheng K-K, Ali Khan A, Suan CL, Zaman Huri H, et al. Metabolomics - the complementary field in systems biology: a review on obesity and type 2 diabetes. Mol Biosyst [Internet]. 2015 Jul [cited 2016 Jul 28];11(7):1742–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25919044

6. Chadeau-Hyam M, Campanella G, Jombart T, Bottolo L, Portengen L, Vineis P, et al. Deciphering the Complex: Methodological Overview of Statistical Models to Derive OMICS-Based Biomarkers. Environ Mol Mutagen. 2013;54:542–57.

7. Vineis P, Perera F. Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. Cancer Epidemiol Biomarkers Prev. 2007 Oct;16(10):1954–65.

8. Vineis P, van Veldhoven K, Chadeau-Hyam M, Athersuch TJ. Advancing the application of omics-based biomarkers in environmental epidemiology. Environ Mol Mutagen [Internet]. 2013 Aug [cited 2016 Jul 28];54(7):461–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23519765

9. Chadeau-Hyam M, Athersuch TJ, Keun HC, De Iorio M, Ebbels TMD, Jenab M, et al. Meeting-

in-the-middle using metabolic profiling - a strategy for the identification of intermediate biomarkers in cohort studies. Biomarkers. 2011 Feb;16(1):83–8.

10. Assi N, Fages A, Vineis P, Chadeau-hyam M, Stepien M, Duarte-salles T, et al. A statistical framework to model the meeting-in-the-middle principle using metabolomic data : application to hepatocellular carcinoma in the EPIC study. Mutagenesis. 2015;30(6):743–53.

11. Gomaa A-I. Hepatocellular carcinoma: Epidemiology, risk factors and pathogenesis. World J Gastroenterol. 2008;14(27):4300–8.

12. McKenzie F, Biessy C, Ferrari P, Freisling H, Rinaldi S, Chajès V, et al. Healthy Lifestyle and Risk of Cancer in the European Prospective Investigation Into Cancer and Nutrition Cohort Study. Medicine (Baltimore) [Internet]. 2016 Apr [cited 2016 Jul 28];95(16):e2850. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27100409

13. McKenzie F, Ferrari P, Freisling H, Chajès V, Rinaldi S, de Batlle J, et al. Healthy lifestyle and risk of breast cancer among postmenopausal women in the European Prospective Investigation into Cancer and Nutrition cohort study. Int J cancer [Internet]. 2015 Jun 1 [cited 2016 Jul 28];136(11):2640–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25379993

14. Fedirko V, Trichopolou A, Bamia C, Duarte-Salles T, Trepo E, Aleksandrova K, et al. Consumption of fish and meats and risk of hepatocellular carcinoma: the European Prospective Investigation into Cancer and Nutrition (EPIC). Ann Oncol. 2013 Aug;24(8):2166–73.

15. Stepien M, Duarte-Salles T, Fedirko V, Floegel A, Kumar-Barupal D, Rinaldi S, et al. Alteration of Amino Acid and Biogenic Amine Metabolism in Hepatobiliary Cancers: Findings from a Prospective Cohort Study. Submitt to Am J Gastroenterol. 2015;

16. Trichopoulos D, Bamia C, Lagiou P, Fedirko V, Trepo E, Jenab M, et al. Hepatocellular carcinoma risk factors and disease burden in a European cohort: a nested case-control study. J Natl Cancer Inst. 2011 Nov 16;103(22):1686–95.

17. Carayol M, Licaj I, Achaintre D, Sacerdote C, Vineis P, Key TJ, et al. Reliability of serum metabolites over a two-year period: A targeted metabolomic approach in fasting and non-fasting samples from EPIC. PLoS One [Internet]. 2015;10(8):1–10. Available from: http://dx.doi.org/10.1371/journal.pone.0135437

18. Floegel A, Drogan D, Wang-Sattler R, Prehn C, Illig T, Adamski J, et al. Reliability of serum metabolite concentrations over a 4-month period using a targeted metabolomic approach. PLoS One. 2011;6(6).

19. Fages A, Ferrari P, Monni S, Dossus L, Floegel A, Mode N, et al. Investigating sources of variability in metabolomic data in the EPIC study: the Principal Component Partial R-square (PC-PR2) method. Metabolomics. 2014;10(6):1074–83.

20. Abdi H. Partial least squares regression and projection on latent structure regression (PLS Regression). Wiley Interdiscip Rev Comput Stat. 2010;2(1):97–106.

21. Tenenhaus M. La régression PLS. Technip. Paris; 1998.

22. Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. Am J Epidemiol. 2010 Dec 15;172(12):1339–48.

23. SAS Institute Inc., Cary N. Base SAS® 9.4 Procedures Guide. 2012.

24. R Foundation for Statistical Computing, R Core Team. R: A language and environment for statistical computing. Vienna, Austria.; 2013.

25. Akuta N, Suzuki F, Kobayashi M, Hara T, Sezaki H, Suzuki Y, et al. Correlation Between Hepatitis B Virus Surface antigen Level and Alpha-Fetoprotein in Patients Free of

Hepatocellular Carcinoma or Severe Hepatitis. J Med Virol. 2014;86:131–8.

26. Yang W-S, Va P, Bray F, Gao S, Gao J, Li H-L, et al. The role of pre-existing diabetes mellitus on hepatocellular carcinoma occurrence and prognosis: a meta-analysis of prospective cohort studies. PLoS One. 2011 Jan;6(12):e27326.

27. Berzigotti A, Saran U, Dufour J-F. Physical Activity and Liver Diseases. Hepatology. 2016;63(3):1026–40.

28. Liu X, Xu J. Body Mass Index and Waistline are Predictors of Survival for Hepatocellular Carcinoma After Hepatectomy. Med Sci Monit [Internet]. 2015;21:2203–9. Available from: http://www.medscimonit.com/abstract/index/idArt/894202

29. Niu J, Lin Y, Guo Z, Niu M, Su C. The Epidemiological Investigation on the Risk Factors of Hepatocellular Carcinoma: A Case-Control Study in Southeast China. Medicine (Baltimore) [Internet]. 2016;95(6):e2758. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26871825

30. Raffetti E, Portolani N, Molfino S, Baiocchi GL, Limina RM, Caccamo G, et al. Role of aetiology, diabetes, tobacco smoking and hypertension in hepatocellular carcinoma survival. Dig Liver Dis [Internet]. 2015;47(11):950–6. Available from: http://dx.doi.org/10.1016/j.dld.2015.07.010

31. Rong X, Wei F, Geng Q, Ruan J, Shen H, Li A, et al. The Association Between Body Mass Index and the Prognosis and Postoperative Complications of Hepatocellular Carcinoma: A Meta-Analysis. Medicine (Baltimore) [Internet]. 2015;94:e1269. Available from: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=ovftq&AN=00005792-201508010-00026

32. Testino G, Leone S, Borro P. Alcohol and hepatocellular carcinoma: A review and a point of

view. World J Gastroenterol. 2014;20(43):15943–54.

33. Turati F, Trichopoulos D, Polesel J, Bravi F, Rossi M, Talamini R, et al. Mediterranean diet and hepatocellular carcinoma. J Hepatol [Internet]. 2014;60(3):606–11. Available from: http://dx.doi.org/10.1016/j.jhep.2013.10.034

34. Chiang C-H, Lu C-W, Han H-C, Hung S-H, Lee Y-H, Yang K-C, et al. The Relationship of Diabetes and Smoking Status to Hepatocellular Carcinoma Mortality. Medicine (Baltimore) [Internet]. 2016;95(6):e2699. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00005792-201602090-00046

35. Trichopoulos D, Bamia C, Lagiou P, Fedirko V, Trepo E, Jenab M, et al. Hepatocellular carcinoma risk factors and disease burden in a European cohort: a nested case-control study. J Natl Cancer Inst. 2011 Nov 16;103(22):1686–95.

36. Lauby-Secretan B, Scoccianti C, Loomis D, Grosse Y, Bianchini F, Straif K. Body Fatness and Cancer — Viewpoint of the IARC Working Group. N Engl J Med [Internet]. 2016 Aug 25 [cited 2016 Sep 28];375(8):794–8. Available from: http://www.nejm.org/doi/10.1056/NEJMsr1606602

37. Floegel A, Stefan N, Yu Z, Muhlenbruch K, Drogan D, Joost H-G, et al. Identification of Serum Metabolites Associated With Risk of Type 2 Diabetes Using a Targeted Metabolomic Approach. Diabetes [Internet]. 2013 Feb 1 [cited 2016 Sep 23];62(2):639–48. Available from: http://diabetes.diabetesjournals.org/cgi/doi/10.2337/db12-0495

38. Cole LK, Vance JE, Vance DE. Phosphatidylcholine biosynthesis and lipoprotein metabolism. Biochim Biophys Acta - Mol Cell Biol Lipids [Internet]. 2012 May [cited 2016 Nov 16];1821(5):754–61. Available from: http://linkinghub.elsevier.com/retrieve/pii/S138819811100179X

39. Kawanaka M, Nishino K, Oka T, Urata N, Nakamura J, Suehiro M, et al. Tyrosine levels are associated with insulin resistance in patients with nonalcoholic fatty liver disease. Hepatic Med Evid Res. 2015;7:29–35.

40. Hellmuth C, Kirchberg FF, Lass N, Harder U, Peissner W, Koletzko B, et al. Tyrosine Is Associated with Insulin Resistance in Longitudinal Metabolomic Profiling of Obese Children. J Diabetes Res [Internet]. 2016 [cited 2016 Nov 16];2016:1–10. Available from: http://www.hindawi.com/journals/jdr/2016/2108909/

41. Floegel A, Stefan N, Yu Z, M??hlenbruch K, Drogan D, Joost HG, et al. Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. Diabetes. 2013;62(2):639–48.

42. Floegel A, Wientzek A, Bachlechner U, Jacobs S, Drogan D, Prehn C, et al. Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study. Int J Obes (Lond). 2014;(February):1–9.

43. Bachlechner U, Floegel A, Steffen A, Prehn C, Adamski J, Pischon T, et al. Associations of anthropometric markers with serum metabolites using a targeted metabolomics approach: results of the EPIC-potsdam study. Nutr Diabetes [Internet]. 2016 [cited 2016 Sep 23];6(6):e215. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27348203

44. Kim JY, Park JY, Kim OY, Ham BM, Kim H-J, Kwon DY, et al. Metabolic Profiling of Plasma in Overweight/Obese and Lean Men using Ultra Performance Liquid Chromatography and Q-TOF Mass Spectrometry (UPLC−Q-TOF MS). J Proteome Res [Internet]. 2010 Sep 3 [cited 2016 Nov 16];9(9):4368–75. Available from: http://pubs.acs.org/doi/abs/10.1021/pr100101p

45. Drogan D, Dunn WB, Lin W, Buijsse B, Schulze MB, Langenberg C, et al. Untargeted Metabolic Profiling Identifies Altered Serum Metabolites of Type 2 Diabetes Mellitus in a Prospective, Nested Case Control Study. Clin Chem [Internet]. 2015 Mar 1 [cited 2016 Sep 23];61(3):487–

97. Available from: http://www.clinchem.org/cgi/doi/10.1373/clinchem.2014.228965

46. Reichel M, Hönig S, Liebisch G, Lüth A, Kleuser B, Gulbins E, et al. Alterations of plasma glycerophospholipid and sphingolipid species in male alcohol-dependent patients. Biochim Biophys Acta - Mol Cell Biol Lipids. 2015;1851(11):1501–10.

47. You M, Fischer M, Deeg MA, Crabb DW. Ethanol Induces Fatty Acid Synthesis Pathways by Activation of Sterol Regulatory Element-binding Protein (SREBP). J Biol Chem. 2002;277(32):29342–7.

48. Glen I, Skinner F, Glen E, Mbch B, Macdonell L, Al GET. The Role of Essential Fatty Acids in Alcohol Dependence and Tissue Damage. Alcohol Clin Exp Res. 1987;11:37–41.

49. Gubner NR, Delucchi KL, Ramo DE. Associations between binge drinking frequency and tobacco use among young adults. Addict Behav [Internet]. 2016;60:191–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27156220

50. Barrett SP, Tichauer M, Leyton M, Pihl RO. Nicotine increases alcohol self-administration in non-dependent male smokers. Drug Alcohol Depend. 2006;81(2):197–204.

51. Kuper H, Tzonou A, Kaklamani E, Hsieh C-C, Lagiou P, Adami H-O, et al. Tobacco Smoking , Alcohol Consumption and Their Interaction in the Causation of Hepatocellular Carcinoma. Int J Cancer. 2000;502:498–502.

52. Xu T, Holzapfel C, Dong X, Bader E, Yu Z, Prehn C, et al. Effects of smoking and smoking cessation on human serum metabolite profile: results from the KORA cohort study. BMC Med [Internet]. 2013 [cited 2016 Sep 28];11:60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23497222

53. Valeri L, Vanderweele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS

macros. Psychol Methods. 2013;18(2):137–50.

54. Gelfand LA, Mensinger JL, Tenhave T. Mediation Analysis: A Retrospective Snapshot of Practice and More Recent Directions. J Gen Psychol. 2009;136(2):153–76.

55. Sampson JN, Boca SM, Shu XO, Stolzenberg-Solomon RZ, Matthews CE, Hsing AW, et al. Metabolomics in Epidemiology: Sources of Variability in Metabolite Measurements and Implications. Cancer Epidemiol Biomarkers Prev [Internet]. 2013 Apr 1 [cited 2016 Sep 12];22(4):631–40. Available from: http://cebp.aacrjournals.org/cgi/doi/10.1158/1055-9965.EPI-12-1109

56. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. 2002 Dec;5(6B):1113–24.

57. Kaaks R, Slimani N, Riboli E. Pilot Phase Studies on the Accuracy of Dietary Intake Measurements in the EPIC Project : Overall Evaluation of Results. 1997;26(1):26–36.

58. Vanderweele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. Epidemiology. 2014;25(2):300–6.

59. Bro R, Kamstrup-Nielsen MH, Engelsen SB, Savorani F, Rasmussen MA, Hansen L, et al. Forecasting individual breast cancer risk using plasma metabolomics and biocontours. Metabolomics [Internet]. 2015;11(5):1376–80. Available from: http://dx.doi.org/10.1007/s11306-015-0793-8

Tables and Figures

Table 1: Baseline characteristics of the study population of the EPIC nested case-control study on hepatocellular carcinoma.

	Cases *(N=147)*	**Controls** *(N=147)*
Characteristics	*Mean (sd) or Frequency*	
Sex		
Male	102	102
Female	45	45
Age at blood collection (y)	60.08 (7.15)	60.06 (7.17)
Height (cm)	167.70 (10.31)	169.30 (9.91)
Weight (kg)	79.78 (17.04)	78.28 (12.88)
BMI (kg/m2)	28.24 (4.74)	27.33 (4.10)
Total energy (kcal/d)	2260.84 (1001.13)	2276.57 (640.07)
Alcohol at recruitment (g/d)	21.56 (34.25)	14.73 (18.92)
Physical Activity (Met-h/week)	77.87 (53.44)	83.27 (52.23)
Education Level		
None or Primary School completed	79	77
Technical/Professional School	33	33
Secondary School	6	12
Longer Education (incl. university degree)	22	25
Unknown	7	0
Lifetime Alcohol Consumption (g/d)*	31.59 (46.32)	18.13 (18.81)
Dietscore*	25.69 (6.69)	27.35 (6.16)
Hepatitis Infection*		
Yes	41	5
No	106	142
Diabetes at Baseline*		
Yes	19	10
No	128	137
Smoking Status*		
Current > 15 cigarettes/d	25	23
Current <= 15 cigarettes/d	34	10
Ex-smokers quit <=10y	17	25
Ex-smokers quit >10y	29	29
Never	42	60

*Missing values were imputed with the EM algorithm. See also frequencies in **Supplementary Table 1**.

Table 2: Exposure variables of the modified HLI and corresponding metabolites contributing to the first PLS factor (N=294, X-set= 7, M-set=133). Results from the overall analysis using residuals based on country (X- and M-sets) and batch (M-set only).

Exposure Variable	Loadings	Metabolites	Loadings*
BMI	-0.385	Glutamic Acid	-0.192
Lifetime Alcohol	-0.695	Hexoses	-0.191
Diet score	-0.058	SM(OH) C14:1	0.196
Physical activity	0.297	SM(OH) C16:1	0.179
Smoking	-0.409	SM(OH) C22:2	0.214
Hepatitis Infection	-0.176	PC aaC32:1	-0.184
Diabetes	-0.282	Liver function score	-0.186

* Metabolite variables contributing to each PLS factor were selected based on extreme loading values, i.e. below or above the 2.5th and 97.5th percentiles.

Table 3: Metabolites contributing to the PLS factor of each HLI component (N=294, X-set=1, M-set=133)*. Results from multiple PLS models performed using residuals based on country (X- and M-sets) and batch residuals (M-set only).

	Metabolite	Loadings		Metabolite	Loadings		Metabolite	Loadings
BMI			**Lifetime alcohol**			**Diet score**		
	Glutamine	-0.186		Glutamic Acid	0.170		PC aaC36:1	-0.178
	Glutamic Acid	0.230		SM C16:1	-0.171		PC aaC38:0	0.195
	Tyrosine	0.243		SM C18:1	-0.167		PC aaC38:6	0.230
	LysoPC aC17:0	-0.218		SM(OH) C14:1	-0.180		PC aaC40:6	0.204
	LysoPC aC18:2	-0.236		SM(OH) C16:1	-0.184		PC aaC42:2	0.263
	PC aeC36:2	-0.203		SM(OH) C22:2	-0.211		PC aeC34:1	-0.195
	Liver function score	0.191		PC aaC32:1	0.211		PC aeC40:6	0.167
Physical activity			**Smoking**			**Hepatitis infection**		
	Biogenic Creatinine	-0.199		Hexoses	0.136		SM C20:2	-0.179
	Biogenic Taurine	-0.181		SM C16:1	-0.238		SM(OH) C16:1	-0.178
	Glutamic Acid	-0.212		SM C18:1	-0.194		PC aaC32:2	0.188
	PC aaC34:2	-0.188		SM(OH) C14:1	-0.214		PC aaC34:1	0.184
	PC aeC34:2	0.209		SM(OH) C22:2	-0.182		PC aaC34:3	0.180
	PC aeC34:3	0.176		LysoPC aC28:1	-0.204		PC aaC34:4	0.197
	PC aeC36:3	0.193		PC aeC30:2	-0.264		PC aaC36:5	0.189
Diabetes status								
	Biogenic Alpha AAA	0.236						
	Isoleucine	0.168						
	Phenylalanine	0.158						
	Valine	0.211						
	Hexoses	0.551						
	Lyso PC aC16:1	-0.145						
	Liver function score	0.226						

* Metabolite variables contributing to each PLS factor were selected based on extreme loading values, i.e. below or above the 2.5th and 97.5th percentiles.

Table 4: Results from the mediation analyses, with natural direct effect (NDE), natural indirect effect (NIE), total effects (TE), mediator effects (ME) and their associated 95% confidence intervals, using residuals based on country (X- and M-sets) and batch (M-set only).

Models*	NDE	P value	FDR	NIE	P value	FDR	Total Effect	P value	FDR	ME	P value	FDR	% mediated
Overall, 7 components	**0.64 (0.44,0.92)**	**0.015**	-	**0.62 (0.50, 0.77)**	**2.12E-05**	-	**0.53 (0.39,0.71)**	**2.64E-05**	-	**0.30 (0.19,0.47)**	**1.94E-07**	-	52
BMI	0.85 (0.60,1.20)	3.44E-01	4.81E-01	**1.56 (1.24,1.96)**	**1.72E-04**	**1.20E-03**	1.23 (0.93,1.62)	1.49E-01	1.74E-01	**4.04 (2.22,7.36)**	**4.77E-06**	**3.17E-05**	100
Lifetime Alcohol	**1.31 (1.06,1.61)**	**1.20E-02**	**4.20E-02**	**1.09 (1.03,1.15)**	**2.40E-03**	**4.67E-03**	**1.40 (1.14,1.72)**	**1.40E-03**	**3.50E-03**	**2.50 (1.57,3.97)**	**1.00E-04**	**2.48E-04**	24
Diet score	0.77 (0.54,1.11)	1.68E-01	3.92E-01	**0.85 (0.74,0.97)**	**1.80E-02**	**2.52E-02**	**0.66 (0.47,0.92)**	**1.40E-02**	**3.27E-02**	**0.61 (0.41,0.89)**	**1.10E-02**	**1.54E-02**	38
Physical activity	0.98 (0.72,1.35)	9.18E-01	9.18E-01	0.97 (0.87,1.09)	6.17E-01	6.17E-01	0.98 (0.71,1.34)	8.84E-01	8.84E-01	0.90 (0.60,1.35)	6.15E-01	6.15E-01	**
Smoking	1.17 (0.77,1.77)	4.59E-01	5.36E-01	**1.22 (1.04,1.44)**	**1.77E-02**	**2.52E-02**	1.42 (0.99,2.03)	5.82E-02	1.02E-01	**3.33 (1.96,5.66)**	**9.04E-06**	**3.17E-05**	57
Hepatitis Infection	**17.99 (5.15,62.80)**	**5.87E-06**	**4.11E-05**	0.94 (0.83, 1.06)	2.98E-01	3.48E-01	**16.70 (4.82,57.84)**	**8.92E-06**	**6.24E-05**	1.22 (0.88,1.69)	2.23E-01	2.60E-01	0
Diabetes	0.46 (0.11,1.93)	2.87E-01	4.82E-01	**5.11 (1.99,13.14)**	**6.99E-04**	**2.45E-03**	2.45 (0.84,7.18)	1.01E-01	1.41E-01	**2.75 (1.59,4.78)**	**3.17E-04**	**5.55E-04**	100

* Models were mutually adjusted for all HLI variables with the exception of the overall model including all 7 components of the HLI in PLS analysis. Cases and controls were matched on age at blood collection (±1 year), sex, study centre, date (± 2 months) and time of day at blood collection (± 3h), fasting status at blood collection (<3/3-6/>6h); women were additionally matched on menopausal status (pre/peri/postmenopausal) and hormone replacement therapy. The mediator models were linear. The outcome models were computed through conditional logistic regressions. In the mediation analysis, the exposure was the original modified HLI lifestyle factor (for the overall model the exposure was the X-score), the mediator was the associated M-score (metabolic profile) and the outcome was HCC. ** As the associations were null for direct and indirect effects, the proportion mediated was not computed. NDE and NIE and their 95%CI computed from formulae detailed in **Supplementary Methods**.

Supplementary material:

Supplementary Table 1: Descriptive statistics of the different components of the modified healthy lifestyle index (HLI) and its scoring, in the current nested case-control study on HCC (Cases=147, Controls=147).

HLI variable	Scoring details	Frequency	Missing	Frequency after EM
BMI (kg/m^2)			0	
5th quintile (>30)	0	76		
4th quintile (26-29.9)	1	107		
3rd quintile (24-25.9)	2	52		
2nd quintile (22-23.9)	3	34		
1st quintile (<22)	4	25		
Lifetime alcohol consumption (g/day)			42	
m: >30 ; w: >20	0	65		85
m: 15-30 ; w: 10-20	1	55		76
m: 5-15 ; w: 5-10	2	54		55
0.1-5	3	59		59
Never	4	19		19
Diet score			12	
1st quintile (6-21)	0	65		65
2nd quintile (22-25)	1	55		65
3rd quintile (26-28)	2	52		54
4th quintile (29-33)	3	77		77
5th quintile (34-46)	4	33		33
Physical activity (METs-h/week)			0	
1st quintile (<45)	0	51		
2nd quintile (46-69)	1	59		
3rd quintile (70-96)	2	44		
4th quintile (97-133)	3	60		
5th quintile (>=134)	4	80		
Smoking			7	
Current > 15 cigarettes/day	0	48		48
Current <= 15 cigarettes/day	1	43		44
Ex smokers quit <= 10-years	2	36		42
Ex smokers quit > 10 years	3	58		58
Never	4	102		102
Hepatitis Infection			76	
Yes	0	41		46
No	4	177		248
Diabetes at baseline			29	
Yes	0	29		29
No	4	236		265

Supplementary Table 2: PC-PR2 results* identifying the sources of variability in the modified HLI variables and in the Metabolomic data.

Modified Healthy Lifestyle Index - 7 original variables					
Country	*Age at recruitment*	*Sex*	*R^2*		
6,165	0,645	3,602	10,697		
Metabolomic data - 132 metabolites					
Country	*Age at blood collection*	*Batch*	*Sex*	*BMI*	*Diet score*
13,146	0,539	**7,103**	4,028	1,263	0,667
Physical Activity	*Alcohol at recruitment*	*Smoking*	*Hepatitis*	*Diabetes*	*R^2*
0,555	2,498	0,312	2,664	0,969	29,458

* 6 and 21 components were retained to account for 80% (threshold used) of total modified HLI and metabolites variables' variability, respectively. The R^2 value represents the amount of variability in modified HLI/metabolites variable explained by the ensemble of investigated predictors.

Supplementary Table 3: Metabolites contributing* to two selected modified HLI variable-specific PLS factors: smoking and lifetime alcohol (N=294, X-set= 1 in turn, M-set=133) – Results reported from the primary analysis, using residuals based on country (X- and M-sets) and batch (M-set only), and from the sensitivity analysis, using mutually adjusted lifestyle residuals as well as residuals for country and batch (the latter only in the M-set).

Primary Analysis			
Lifetime Alcohol		**Smoking**	
Metabolites	**Loadings**	**Metabolites**	**Loadings**
SM C16:1	-0,173	Lysine	-0,173
SM C18:1	-0,175	SM C16:1	-0,218
SM(OH) C14:1	-0,205	SM C18:1	-0,176
SM(OH) C16:1	-0,193	SM(OH) C14:1	-0,196
SM(OH) C22:2	-0,212	SM(OH) C22:2	-0,171
LysoPC aC28:1	-0,170	LysoPC aC28:1	-0,170
PC aeC30:2	-0,177	PC aeC30:2	-0,235
Sensitivity Analysis			
Lifetime Alcohol		**Smoking**	
Metabolites	**Loadings**	**Metabolites**	**Loadings**
SM C18:1	-0.161	Biogenic Taurine	-0.201
SM(OH) C16:1	-0.168	Lysine	-0.211
SM(OH) C22:1	-0.168	Serine	-0.189
SM(OH) C22:2	-0.203	SM(OH) C14:1	-0.195
LysoPC aC16:1	0.162	PC aaC36:1	0.23
PC aaC32:1	0.234	PC aaC40:3	0.202
Acylcarnitine C2	0.152	PC aeC30:2	-0.206

* Metabolite variables contributing to each PLS factor were selected based on extreme loading values, i.e. below or above the 2.5th and 97.5th percentiles.

Supplementary Table 4: Results from the mediation analyses, with natural direct effect (NDE), natural indirect effect (NIE), total effects (TE), mediator effects (ME) and their associated 95% confidence intervals in the sensitivity analysis, using mutually adjusted lifestyle residuals as well as residuals for country and batch (the latter only in the M-set).

Models*	NDE	P value	FDR	NIE	P value	FDR	TE	P value	FDR	ME	P value	FDR	% mediated
Overall - 7 components[1]	0.81 (0.57,1.15)	2.37E-01	-	**0.49 (0.36,0.67)**	**6.11E-06**	-	**0.55 (0.40,0.74)**	**4.58E-05**	-	**0.17 (0.10,0.32)**	**1.59E-08**	-	77
BMI	0.84 (0.58,1.22)	3.53E-01	4.12E-01	**1.52 (1.22,1.89)**	**1.72E-04**	**1.20E-03**	1.23 (0.93,1.62)	1.49E-01	1.74E-01	**3.22 (1.90,5.46)**	1.35E-05	4.73E-05	100
Lifetime Alcohol	**1.30 (1.05,1.60)**	**1.50E-02**	**5.25E-02**	**1.10 (1.04,1.17)**	**1.68E-03**	**2.94E-03**	**1.40 (1.14,1.72)**	1.43E-03	5.00E-03	**2.60 (1.62,4.18)**	8.28E-05	1.93E-04	27
Diet score	0.79 (0.54,1.15)	2.16E-01	4.12E-01	**0.85 (0.75,0.97)**	**1.40E-02**	**1.96E-02**	**0.66 (0.47,0.92)**	1.39E-02	3.24E-02	**0.57 (0.38,0.85)**	5.74E-03	8.04E-03	41
Physical activity	0.94 (0.68,1.32)	7.30E-01	7.30E-01	0.85 (0.74,0.98)	2.48E-02	2.89E-02	0.98 (0.71,1.34)	8.84E-01	8.84E-01	0.57 (0.36,0.88)	1.21E-02	1.41E-02	72
Smoking	0.71 (0.36,1.39)	3.15E-01	4.12E-01	**1.98 (1.34,2.92)**	**5.65E-04**	**1.32E-04**	1.42 (0.99,2.03)	5.82E-02	1.02E-01	**11.63 (4.33,31.28)**	1.16E-06	8.12E-06	100
Hepatitis Infection	**15.56 (4.44,54.54)**	**1.78E-05**	**1.25E-04**	1.07 (0.91,1.26)	4.33E-01	4.33E-01	**16.70 (4.82, 57.84)**	8.92E-06	6.24E-05	0.88 (0.64,1.21)	4.17E-01	4.17E-01	2
Diabetes	0.43 (0.10,1.87)	2.61E-01	4.12E-01	**5.39 (2.07,14.0)**	**5.53E-04**	**1.32E-04**	**2.45 (1.50,3.88)**	1.01E-01	1.41E-01	**2.83 (1.63,4.94)**	2.39E-04	4.18E-04	100

* Models were mutually adjusted for all HLI variables with the exception of the overall model including all 7 components of the HLI in PLS analysis. Cases and controls were matched on age at blood collection (±1 year), sex, study centre, date (± 2 months) and time of day at blood collection (± 3h), fasting status at blood collection (<3/3-6/>6h); women were additionally matched on menopausal status (pre/peri/postmenopausal) and hormone replacement therapy. The mediator models were linear. The outcome models were computed through conditional logistic regressions. In the mediation analysis, the exposure was in turn the original HLI lifestyle factor (for the overall model the exposure was the X-score), the mediator was the associated M-score (metabolic profile) and the outcome was HCC. NDE and NIE and their 95%CI computed from formulae detailed in **Supplementary Methods**.

Supplementary Methods

Material and Methods

The EPIC Study

EPIC is a multicentre prospective study designed to investigate the link between diet, lifestyle and environmental factors with cancer incidence and other chronic disease outcomes. Over 520,000 healthy men and women aged 25-85 were enrolled between 1992 and 2000 across 23 EPIC administrative centres in 10 European countries including Denmark, France, Germany, Greece, Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom[1]. In most of EPIC centers, participants were recruited amongst the general population with the following exceptions: for France, women were enrolled from a health insurance scheme for school and university employees; in Utrecht, The Netherlands and in Florence, Italy, participants came from breast cancer screening programs; some centers in Italy (Turin and Ragusa) and Spain recruited blood donors; and the Oxford sub-cohort (United Kingdom) included mostly health-conscious individuals recruited throughout the UK. Finally, the French, Norwegian and Naples (Italy) cohorts comprised only women. Extensive details of the study design and recruitment methods have been previously published[1,2].

Data collection of dietary and lifestyle data

During the enrolment period, participants gave informed consent and completed questionnaires on diet, lifestyle and medical history. Approval for this study was obtained from the ethical review boards of the participating institutions and the International Agency for Research on Cancer (IARC). Biological samples were collected for approximately 80% of the cohort prior to disease onset. Serum samples were stored at IARC, Lyon, France in -196°C liquid nitrogen for all countries, with the exception of samples originating from Sweden (-80°C freezers) and Denmark (-150°C nitrogen vapour). Usual diet over the previous 12 months was assessed for each individual through validated country-specific dietary questionnaires (DQs)[1]. Nutrient intakes were then estimated using a common harmonized food composition database across EPIC countries (EPIC Nutrient Database, ENDB)[3,4]. Information on sociodemographic data including education, smoking and alcohol drinking histories as well as physical activity were gathered in lifestyle questionnaires. Anthropometric characteristics were directly measured by trained study personnel for most of the participants[1], but were self-reported in baseline questionnaires for a subset of

participants from the EPIC-Oxford sub-cohort, although the accuracy of these self-reported data have been validated[5].

Follow-up and case ascertainment in the nested case-control study

Follow-up started at date of entry to the study and finished at date of diagnosis, death or last completed follow-up (from December 2004 up to June 2010), whichever came first. Cancer incidence was determined through population cancer registries or through active follow-up as detailed elsewhere[6]. Incident HCC cases were defined as first primary invasive tumours and identified through the 10th Revision of International Statistical Classification of Diseases, Injury and Causes of Death (ICD10) as C22.0 with morphology codes ICD-O-2 "8170/3"and "8180/3". Metastatic cases and other types of primary liver cancer were excluded.

Matching criteria for the nested case-control study

For each HCC case, one control (n=147) was selected by incidence density sampling[7] from all cohort members alive and free of cancer (except for non-melanoma skin cancer), and matched by age at blood collection (±1 year), sex, study centre, time of the day at blood collection (±3 hours), fasting status at blood collection (<3, 3-6,and >6 hours); among women, the pair was additionally matched by menopausal status (pre-, peri-, and postmenopausal), and hormone replacement therapy use at time of blood collection (yes/no).

Modified Healthy Lifestyle Index (HLI) construction

The overall HLI had five initial components and was determined for the entire EPIC cohort as an unweighted sum of the scores of its individual components, each assigned scores of 0 to 4, where a higher score indicated a healthier behaviour[8,9]. This study utilized a modified version of the HLI and included smoking, quintiles of physical activity, BMI, quintiles of the diet score and lifetime alcohol consumption instead of alcohol at recruitment to avoid reverse causality with respect to HCC outcome. In addition, two components reflecting two major risk factors of liver cancer[10–12] were added to the modified index to make it more HCC-specific: diabetes at baseline (No=4, Yes=0)[11]; and hepatitis infection (No=4, Yes=0, assessed from biomarker measures of hepatitis B and hepatitis C viruses' (HBV, HCV)

seropositivity [ARCHITECT HBsAg and anti-HCV chemiluminescent microparticle immunoassays; Abbott Diagnostics, France])[12]. To some extent hepatitis infection can reflect certain lifestyle exposures and behaviours. Missing values in some of the index components were imputed by an expectation-maximization (EM) algorithm that preserved the variance-covariance structure of the data[13]. Descriptive and scoring details on the modified HLI components can be viewed in **Supplementary Table 1**.

Metabolomic data nomenclature

Fatty acids side chains are labelled "Cx:y", where x and y are the numbers of carbon atoms and double bonds, respectively. Measured metabolites included 12 acylcarnitines (abbreviated according to the fatty acid side chain), 21 amino acids and 6 biogenic amines (labelled with their full name), 78 phosphatidylcholines (PC) of which there were 11 "LysoPC a" (PCs having one fatty acid side chain with an acyl bound), 34 "PC aa" and 33 "PC ae" (PCs having respectively two acyl side chains [diacyl] and one acyl and one alkyl side chains), a total of 14 sphingomyelins "SM" of which 5 had a hydroxyl group "SM(OH)" (additionally labelled according to the fatty acid side chain) and finally 1 sum of hexoses (including glucose, fructose and galactose). PCs were separated by type of bond and number of fatty acids side chains.

Liver function score construction

This score includes the following 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 (Centre de Biologie République, Lyon, France) based on assay specifications as previously described[6,14].

The diet score (included in the X-set, continuous and in the modified HLI, categorical)

An *a priori* score for diet was proposed within EPIC based on dietary components that have been posited to affect risk of cancer[9,8]. The diet score combined six dietary items including cereal fiber, red and processed meats, ratio of polyunsaturated to saturated fatty acids, margarine (used as a surrogate marker for trans-fat from industrial sources), glycaemic load,

and fruits and vegetables. Details of the diet score computation are provided elsewhere[9]. The resulting continuous variable was included in the X-set as previously mentioned.

Statistical Analyses

Principal Component Partial R2 (PC-PR2)

PC-PR2 combines aspects of PCA with the partial R^2 statistic in multiple linear regression models. Briefly, the set under scrutiny is reduced through PCA and a number of components explaining an amount of total variability above a designated threshold (here, 80%), is retained. Multiple linear models are then fitted where each component's variability is explained by regressing it on a list of relevant covariates, yielding an R^2 statistic for each of the latter. The R^2 quantifies the amount of variability each independent variable explains, conditional on all other covariates included in the model. Finally, an overall partial R^2 is computed as a weighed mean for each covariate, using the eigenvalues as components' weights.

In this study, PC-PR2 was applied to the X-set of 7 exposure variables where the covariates explored for systematic variability were country, age at recruitment and sex. With the similar objective of identifying sources of variability in the metabolite data, another PC-PR2 analysis was run on the M-set and the examined covariates included country, age at blood collection, batch, sex, BMI, diet score, physical activity, alcohol at recruitment, smoking, hepatitis and diabetes at baseline.

After running PC-PR2, a total of 6 and 21 principal components were retained explaining around 80% of total variability among the modified HLI original variables and the metabolites set, respectively. The ensemble of explanatory systematic variables accounted for 10.7 and 29.5% of total variance within the X- and M- sets, respectively. "Country of origin" was the highest contributor with consistently 6.2 and 13.1% in the X- and M-sets, followed by "Batch" with 7.1% in the M-set (**Supplementary Table 2**). PLS analyses were carried out controlling for these two variables in the respective sets. Sensitivity analyses were also conducted where mutually adjusted lifestyle residuals and country residuals were used in the X-set. Country and batch residuals were used in the M-set **(Supplementary Tables 3-4)**.

Details on the PLS procedure

PLS is a multivariate method that generalizes features of PCA with those of multiple linear regression[15,16]. Mathematical and computational details of the PLS method and its applicability within the MITM framework have been thoroughly described previously[17]. Missing values in the M-set were imputed through a simple EM algorithm[18,19] consisting of the two following steps. First, the missing values were replaced by the average of the non-missing values for each related variable and a PLS model is run. In a second step, the missing data are assigned their predicted values based on the first model and PLS is then rerun. The number of PLS factors to retain was selected after carrying a 7-fold cross-validation to minimize the predicted residual sum of squares (PRESS) statistic, a measure of PLS performance. Details of the process can be found elsewhere[17]. PLS factor loadings, i.e. the coefficients quantifying how much each original variable contributes to the PLS factor, characterize each extracted HLI and metabolomics profile. As the M-set was particularly dense in metabolite variables, the interpretation of the metabolomics profile mainly focused on those most significantly contributing to the PLS component, reporting variables with loading values lower than the 5th and larger than the 95th percentiles. One PLS factor was retained in each one of the individual PLS analyses. All lifestyle and metabolomic components of PLS factors where mirrored in their respective PLS-scores (X- and M-scores).

Details on the mediation analyses

The NDE and NIE were produced through two main models: a linear mediator model and a conditional logistic outcome model. HCC being a rare outcome, direct and indirect effects were estimated taking into account the nested case-control design. This is done by running the mediator regression only for the controls[20]. After testing, there was no exposure-mediator interaction, the models can then be simply written as follows:
Let x be the exposure, m the mediator, c a set of different confounders, y HCC and j the pair number ranging among the set {1,…, n=147}:

$$E[M|x,c] = \beta_0 + \beta_1 x + \beta_2' c$$

$$logit[P(Y=1|x,m,c,j)] = \theta_{0,j} + \theta_1 x + \theta_2 m + \theta_3' c$$

Thus, NDE and NIE are given as follows for a one standard deviation increase in x and m:

$$NDE_{x|c} \approx \exp(\theta_1 sd(x))$$

$$NIE_{x|c} = \exp(\theta_2\ \beta_1\ sd(x))$$

95% CI for NDE and NIE were computed through the following formulae:

$$95\%CI\left(NDE_{x|c}\right) = \exp\left(\log\left(\widehat{NDE}_{x|c}\right) \pm 1.96 * \widehat{sd(x)} * \sqrt{\hat{\sigma}_{11}^{\theta}}\right)$$

$$95\%CI\left(NIE_{x|c}\right) = \exp\left(\log\left(\widehat{NIE}_{x|c}\right) \pm 1.96 * \widehat{sd(x)} * \sqrt{\hat{\theta}_2^2\hat{\sigma}_{11}^{\beta} + \hat{\beta}_1^2\hat{\sigma}_{22}^{\theta}}\right),$$

where $\hat{\sigma}_{11}^{\theta}, \hat{\sigma}_{22}^{\theta}$ and $\hat{\sigma}_{11}^{\beta}$ are the estimated variances of the coefficients $\widehat{\theta_1}, \widehat{\theta_2}$ and $\widehat{\beta_1}$, respectively.

The total effect of X (TE) was computed from the following conditional logistic regressions:

$$logit[P(Y = 1|x, c, j)] = \gamma_{0,j} + \gamma_1 x + \gamma_2' c$$

with TE given by:

$$TE_{x|c} = \exp(\gamma_1\ sd(x))$$

Usually TE can be written as the product of NDE and NIE. However, in our setting employing conditional logistic regression, this is no longer the case because discordant pairs in the model adjusted for the mediator are not the same as the model not including the mediator (TE).

The mediator effect (ME), corresponding to the "independent" effect of the M-score adjusted for its counterpart lifestyle exposure and for confounding variables was given by:

$$ME_{m|x,c} = \exp(\theta_2\ sd(m))$$

To control for potential confounding, mediation analyses models were adjusted for the modified HLI variables except the one under scrutiny (multiple PLS), with the exception of the models from the overall PLS. P-values for NDE and NIE were inferred from the 95%CI, whereas for the ME and TE, p-values associated with Wald's test for continuous exposure compared with a chi-square distribution with 1 degree of freedom are reported. The false discovery rate (FDR) correction[21] was applied to mediation results stemming from the multiple PLS analyses.

For each mediation analysis the estimates for the NDE, NIE, TE and ME were reported for an increase in the exposure as follows: an increase of 1-SD for the overall PLS analysis and for smoking, an increase of 1-SD among the controls for BMI, physical activity and the diet score, an increase of 1 unit (0 to 1) for diabetes and hepatitis, and finally an increase of 12 g/day (corresponding to one alcohol unit) for lifetime alcohol.

Since TE=NDE*NIE does not hold in our setting, the mediated proportion was computed using the following formula:

$$\%\, mediated = min\left(max\left(0, \left(\frac{\log(NIE)}{\log(NDE) + \log(NIE)}\right) * 100\right), 100\right)$$

Indeed, the proportion mediated makes real sense only when NDE and NIE have the same direction of association and is bounded between 0% and 100%. In this case our formula reduces to:

$$\%mediated = \frac{\log(NIE)}{\log(\mathrm{NDE}) + \log(NIE)}.$$

When NDE and NIE have opposite directions, the mediated proportion is not well-defined. For example, if $NDE = 0.5$ and $NIE = 2$ so that $TE = 1$, it is not clear what the mediated proportion would be. In our results, NDE and NIE always had the same direction when they were both statistically significant. For example, in our analyses for diabetes (or equivalently for BMI), the NIE is significantly associated with an increased risk of HCC and the NDE was not significant and had the opposite direction of association. This suggested that TE=NIE and using our first formula above we get the appropriate value of 100%.

References

1. Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr*. 2002;5(6B):1113-1124.

2. Kaaks R, Slimani N, Riboli E. Pilot Phase Studies on the Accuracy of Dietary Intake Measurements in the EPIC Project : Overall Evaluation of Results. 1997;26(1):26-36.

3. Slimani N, Deharveng G, Unwin I, et al. The EPIC nutrient database project (ENDB): a first attempt to standardize nutrient databases across the 10 European countries participating in the EPIC study. *Eur J Clin Nutr*. 2007;61(9):1037-1056. doi:10.1038/sj.ejcn.1602679.

4. Nicolas G, Witthöft CM, Vignat J, et al. Compilation of a standardised international folate database for EPIC. *Food Chem*. 2016;193:134-140.

doi:10.1016/j.foodchem.2014.11.044.

5. Spencer E a, Roddam AW, Key TJ. Accuracy of self-reported waist and hip measurements in 4492 EPIC-Oxford participants. *Public Health Nutr*. 2004;7(6):723-727. doi:10.1079/PHN2004600.

6. Stepien M, Duarte-Salles T, Fedirko V, et al. Alteration of Amino Acid and Biogenic Amine Metabolism in Hepatobiliary Cancers: Findings from a Prospective Cohort Study. *Submitt to Am J Gastroenterol*. 2015.

7. Pearce N. Incidence density matching with a simple SAS computer program. *Int J Epidemiol*. 1989;18(4):981-984. http://www.ncbi.nlm.nih.gov/pubmed/2621036. Accessed July 28, 2016.

8. McKenzie F, Ferrari P, Freisling H, et al. Healthy lifestyle and risk of breast cancer among postmenopausal women in the European Prospective Investigation into Cancer and Nutrition cohort study. *Int J cancer*. 2015;136(11):2640-2648. doi:10.1002/ijc.29315.

9. McKenzie F, Biessy C, Ferrari P, et al. Healthy Lifestyle and Risk of Cancer in the European Prospective Investigation Into Cancer and Nutrition Cohort Study. *Medicine (Baltimore)*. 2016;95(16):e2850. doi:10.1097/MD.0000000000002850.

10. Chiang C-H, Lu C-W, Han H-C, et al. The Relationship of Diabetes and Smoking Status to Hepatocellular Carcinoma Mortality. *Medicine (Baltimore)*. 2016;95(6):e2699. doi:10.1097/MD.0000000000002699.

11. Yang W-S, Va P, Bray F, et al. The role of pre-existing diabetes mellitus on hepatocellular carcinoma occurrence and prognosis: a meta-analysis of prospective cohort studies. *PLoS One*. 2011;6(12):e27326.

12. Akuta N, Suzuki F, Kobayashi M, et al. Correlation Between Hepatitis B Virus Surface antigen Level and Alpha-Fetoprotein in Patients Free of Hepatocellular Carcinoma or Severe Hepatitis. *J Med Virol*. 2014;86:131-138.

13. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the

EM Algorithm. *J R Stat Soc Ser B*. 1977;53:1-38. doi:10.1017/CBO9781107415324.004.

14. Trichopoulos D, Bamia C, Lagiou P, et al. Hepatocellular carcinoma risk factors and disease burden in a European cohort: a nested case-control study. *J Natl Cancer Inst*. 2011;103(22):1686-1695.

15. Abdi H. Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdiscip Rev Comput Stat*. 2010;2(1):97-106.

16. Tenenhaus M. *La Régression PLS*. Technip. Paris; 1998.

17. Assi N, Fages A, Vineis P, et al. A statistical framework to model the meeting-in-the-middle principle using metabolomic data : application to hepatocellular carcinoma in the EPIC study. *Mutagenesis*. 2015;30(6):743-753. doi:10.1093/mutage/gev045.

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

19. Bastien P. Régression PLS et Données Censurées. 2008.

20. VanderWeele TJ, Tchetgen Tchetgen EJ. Mediation Analysis With Matched Case-Control Study Designs. *Am J Epidemiol*. 2016;183(1):1-2. doi:10.1093/aje/kww038.

21. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B*. 1995;57(1):289-300.

CHAPTER V:

FATTY ACIDS AND BREAST CANCER IN EPIC

CONTEXT

Breast cancer (BC) is the most frequent cancer affecting women as one in five new cancer cases detected in women is BC, and it is the main cause of cancer death in women worldwide. BC incidence is on the rise and is expected to keep rising as the world population ages [209,222]. BC is a multifactorial disease whose aetiology embraces environmental, lifestyle and dietary risk factors [13,20,22,25,81,88,223–228]. Diet can account for about 40% of causes of cancer although there is no consensus around this estimate [2,229]. Nonetheless, intakes of some fatty acids (FA) have been suggested to affect BC risk. While a high dietary intake of ω-3 polyunsaturated FA (PUFA) from marine origins have been hypothesized to decrease BC risk [230], effects of *trans* FA (TFA) have been postulated to increase the development of many non-communicable diseases (NCDs) and cancers, including BC, due to a high ratio of *cis* monounsaturated to saturated FA (MUFA to SFA) [231–233]. Many studies were conducted investigating the relation between TFA and BC [234], but results from epidemiological data based on dietary questionnaires were inconsistent. This is due to the lack of reliable data on FA in food composition tables, expressly for TFA, hence biomarkers offer a promising objective measure [231]. An investigation into the French arm of EPIC – the E3N sub-cohort – in a nested case-control study with FA biomarker data showed a statistically significant link between industrially produced TFA and increased risk of BC [35]. The following work aims to confirm the findings from the latter study by extending the analysis to a larger nested case-control sample including subjects from all EPIC countries, providing a wider geographical gradient of FA intake.

OBJECTIVES

- To assess the association between biomarkers of dietary FA intake and risk of BC within a large nested case-control study in EPIC.
- To investigate this association by different hormonal receptor status (different BC subtypes) and by menopausal status.
- To confirm the findings from the French arm of EPIC – E3N – where evidence showed the detrimental effects of total *trans* monounsaturated FA, *trans*

palmitoleic and *trans* elaidic acids on BC risk, using a larger sample size from the whole EPIC cohort with more variability.

- To provide the necessary evidence on the effects of individual FA, particularly TFAs, prior to moving to more complex frameworks exploring the lipidome in multivariate and pathway analyses.

APPROACH

Within a nested case-control study on BC within EPIC, including 2,982 cases and as many matched controls, sixty fatty acids levels were measured by gas chromatography in pre-diagnostic plasma. For each plasma phospholipid FA, conditional logistic regressions were applied to estimate the odds ratios and associated 95% confidence interval (OR, 95%CI). The models were adjusted for date at blood collection, education level, BMI, height, menopausal hormone use at baseline, alcohol, age at first birth and parity combined, energy intake, and family history of BC. This univariate multivariable approach was additionally used in subgroup analyses where the relationships between FA were investigated by menopausal status and by oestrogen receptor (ER) and progesterone receptor (PR) status in tumours.

MAIN FINDINGS

After controlling for multiple testing through the FDR correction, evidence of an increased overall BC risk was found associated with high levels of palmitoleic acid with OR=1.37 (1.14, 1.64, p-trend<.001, q-value=0.004) comparing the highest quartile with the lowest. High levels of the desaturation index DI_{16} (16:1n-7/16:0) which is a biomarker of endogenous hepatic synthesis of MUFA, were associated with a statistically significant increase in BC risk by 28%. Contrariwise, high levels of plasma phospholipid n-6 PUFA were associated with a decrease in BC risk with OR=0.81 (0.69, 0.96, p-trend=0.035) but this association did not withstand FDR correction. In subgroup analyses by menopausal status, the results did not markedly differ, whereas specific associations emerged by hormonal receptor status. Specifically, ER- BC cases significantly arose by two-fold in participants with high levels of industrial TFA. This increase was not however present in ER+, PR- and PR+ subtypes.

CONCLUSION

Findings from this study carried out on data from all EPIC participating sub-cohorts showed that an early increase in endogenous synthesis of MUFA might increase BC risk. This confirmed early findings from E3N, where specific MUFA were linked with an increased BC risk. These results were consistent and independent from menopausal and hormonal receptor status. Dietary industrially-produced TFA increased ER- BC risk. These results may contribute to issue guidelines for BC prevention, by considerably lowering or eliminating TFA in industrially processed foods. This latter measure would likewise benefit the ER- BC subtype that has one of the highest mortality rates. This analysis is a first stepping stone looking into the associations between FA and BC. Future analyses will look into the complex lipid interactions at the heart of the lipidome, and disentangle these associations when considering the common metabolic pathways shared by numerous FA, with the scope of looking at BC outcome.

PAPER

Contribution: Co-author, contributed to statistical analyses and to finalizing the report, participated to the addressing reviewer's comments, read and approved the final report.

The manuscript is currently in draft format. It has been circulated to the writing group, and to EPIC collaborators. After a first unsuccessful submission to the Journal of the National Cancer Institute (JNCI), it will be shortly submitted to another target journal.

A prospective evaluation of plasma phospholipid fatty acids and breast cancer risk in the EPIC study

Véronique Chajès[1*], Nada Assi[1], Carine Biessy[1], Pietro Ferrari[1], Sabina Rinaldi[1], Nadia Slimani[1], Gilbert M. Lenoir[2], Laura Baglietto[2,3], Mathilde His[2,3], Marie-Christine Boutron-Ruault[2,3], Antonia Trichopoulou[4,5], Pagona Lagiou[4,5,6], Michail Katsoulis[4], Rudolf Kaaks[7], Tilman Kühn[7], Salvatore Panico[8], Valeria Pala[9], Giovanna Masala[10], H.B(as). Bueno-de-Mesquita[11,12,13], Petra H. Peeters[14], Carla van Gils[14], Anette Hjartåker[15], Karina Standahl Olsen[16], Runa Borgund Barnung[16], Aurelio Barricarte[17,18,22], Daniel Redondo-Sanchez[19,22], Virginia Menéndez[20], Pilar Amiano[21,22], Maria Wennberg[23], Tim Key[24], Kay-Tee Khaw[25], Melissa A. Merritt[12], Elio Riboli[12], Marc J. Gunter[1], Isabelle Romieu[1]

[1]Nutrition and Metabolism Section, International Agency for Research on Cancer, Lyon, France

[2]Institut Gustave Roussy, Villejuif, France

[3]Université Paris-Saclay, Université Paris-Sud, UVSQ, CESP, INSERM, Villejuif, France

[4]Hellenic Health Foundation, Athens, Greece

[5]WHO Collaborating Center for Nutrition and Health, Unit of Nutritional Epidemiology and Nutrition in Public Health, Department of Hygiene, Epidemiology and Medical Statistics, University of Athens Medical School, Athens, Greece

[6]Department of Epidemiology, Harvard School of Public Health, Boston, USA

[7]The German Cancer Research Center (DKFZ), Heildelberg, Germany

[8]Dipartimento Di Medicina Clinica E Chirurgia, Federico II University, Naples, Italy

[9]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

[10]Cancer Risk Factors and Life-Style Epidemiology Unit, Cancer Research and Prevention Institute – ISPO, Florence, Italy

[11]Department for Determinants of Chronic Diseases (DCD), National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands

[12]Department of Epidemiology and Biostatistics, The School of Public Health, Imperial College London, London, United Kingdom

[13]Department of Social & Preventive Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia

[14]Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, The Netherlands

[15]Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway

[16]Department of Community Medicine, University of Tromsø-UiT The Artic University of Norway, Tromsø, Norway

[17]Navarra Public Health Institute, Pamplona, Spain

[18]Navarra Institute for Health Research (IdiSNA), Pamplona, Spain

[19]Escuela Andaluza de Salud Pública, Instituto de Investigación Biosanitaria ibs Granada, Hospitales Universitarios de Granada/Universidad de Granada, Granada, Spain

[20]Public Health Directorate, Asturias, Spain (VM);

[21]Public Health Division of Gipuzkoa, Health Department, Basque Region, San Sebastian, Spain

[22]CIBER Epidemiology and Public Health CIBERESP, Spain

[23]Public Health and Clinical Medicine, Nutritional Research, Umeå University, Umeå, Sweden

[24]The Cancer Epidemiology Unit, University of Oxford, Oxford, United Kingdom

[25]University of Cambridge School of Clinical Medicine, Cambridge, United Kingdom

*Corresponding author Dr. Véronique Chajès at Nutrition and Metabolism, International Agency for Research on Cancer, 150 cours Albert Thomas, 69373 Lyon cedex 08, France; Tel: +33 4 72738014; chajesv@iarc.fr

Key words: fatty acids, biomarkers, breast cancer, epidemiology, EPIC

Running title: plasma phospholipid fatty acids and breast cancer

Conflict of interest: The authors have declared no potential conflict of interest.

Financial support

This work was supported by World Cancer Research Funds, and the Institut National du Cancer (INCA).

The coordination of EPIC is financially supported by the European Commission (DG-SANCO) and the International Agency for Research on Cancer. The national cohorts are supported by Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); Deutsche Krebshilfe, Deutsches Krebsforschungszentrum and Federal Ministry of Education and Research (Germany); the Hellenic Health Foundation (Greece); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); UiT The Arctic University of Norway; Health Research Fund (FIS), PI13/00061 to Granada), Regional Governments of Andalucía, Asturias, Basque Country, Murcia (no. 6236) and Navarra, ISCIII RETIC (RD06/0020) (Spain); Swedish Cancer Society, Swedish Scientific Council and County Councils of Skåne and Västerbotten (Sweden); Cancer Research UK (14136 to EPIC-Norfolk; C570/A16491 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk) (United Kingdom).

Abstract

Intakes of specific fatty acids have been postulated to impact breast cancer risk but epidemiological data based on dietary questionnaires remain conflicting. We assessed the association between plasma phospholipid fatty acids and breast cancer risk in a case-control study nested within the European Prospective Investigation into Cancer and Nutrition (EPIC) study. Sixty fatty acids were measured by gas chromatography in pre-diagnostic plasma phospholipids from 2,982 incident breast cancer cases matched to 2,982 controls. Conditional logistic regression models were used to estimate relative risk of breast cancer by fatty acid level. The false discovery rate (q-values) was computed to control for multiple comparisons. Subgroup analyses were performed by estrogen receptor (ER) and progesterone receptor (PR) expression in the tumours. A high level of palmitoleic acid (odds ratio, OR for the highest quartile compared with the lowest OR[Q4-Q1]=1.37; 95%CI=1.14-1.64; p for trend=0.0001, q-value=0.004) as well as a high desaturation index (DI_{16}) (16:1n-7/16:0) (OR[Q4-Q1]=1.28; 95%CI=1.07-1.54; p for trend=0.002, q-value=0.037), as biomarkers of endogenous synthesis of monounsaturated fatty acids, were significantly associated with increased risk of breast cancer. Levels of industrial trans-fatty acids were positively associated with ER-negative tumors (OR for the highest tertile compared with the lowest [T3-T1]=2.01; 95% CI=1.03-3.90; p for trend=0.047), while no association was found for ER-positive tumors (P-heterogeneity =0.01). These findings suggest that increased endogenous synthesis of palmitoleic acid estimated many years prior to diagnosis is associated with higher breast cancer risk. Dietary trans fatty acids derived from industrial processes may specifically increase ER-negative breast cancer risk.

Introduction

Breast cancer is the most frequently diagnosed cancer among women worldwide with an estimated 1.8 million new cancer cases diagnosed in 2013 (25% of all cancers) (1). While multiple risk factors for breast cancer such as family history, obesity, alcohol, breastfeeding, and reproductive history, are well established, very few additional modifiable risk factors have been identified.

Variation in diet has been suggested to account for up to 25-40% of preventable causes of cancers (2). A potential link between dietary fat and breast cancer has been a focus of intense research; however, overall findings to date are conflicting (3-5). Epidemiological studies indicate that, rather than total fat intake, subtypes of fatty acids could diversely affect breast cancer risk. A high dietary intake of *cis* monounsaturated fat (MUFA) (6), or long-chain n-3 polyunsaturated fatty acids (PUFA) from marine sources (7), may reduce breast cancer risk. Conversely, a positive association has been reported between dietary intake of saturated fatty acids (SFA) and ER-positive breast cancer (8). Finally, a high estimated intake of industrial *trans* fatty acids (ITFA) derived from industrially-produced hydrogenated vegetable oils may increase the risk of postmenopausal breast cancer (9). However, overall data on specific fatty acids are still discrepant.

Epidemiological data on biomarkers of exposure to fatty acids and breast cancer risk are also limited. Meta-analyses of prospective and/or case-control studies have suggested a protective effect of n-3 PUFA on breast cancer risk (7), while some SFA and MUFA have been associated with an increased risk of breast cancer (10). One prospective study showed a significant association between high blood levels of ITFA and increased risk of breast cancer (11). However, in general prospective studies have not shown clear associations between patterns of fatty acids and risk of breast cancer, overall and by hormonal receptor status (12). More epidemiological prospective studies that integrate reliable biomarkers of exposure to fatty acids are needed to further investigate the contribution of different types of fatty acids to the etiology of breast cancer, overall and by hormone receptor subtypes.

The purpose of the current study was to investigate associations between plasma phospholipid fatty acids and risk of breast cancer, overall and by hormonal receptor status, in a large case-control study nested within the prospective EPIC cohort.

Materials and Methods

The EPIC STUDY

The EPIC study includes 519,978 participants in 10 European countries: Denmark, France, Germany, Greece**,** Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom. Participants gave informed consent and completed questionnaires on diet, lifestyle, and medical history. In most centers, participants were recruited from the general population. Exceptions were the French cohort (women of the health insurance scheme covering teachers), the Utrecht cohort (women attending breast cancer screening), the Ragusa cohort (blood donors and their spouses), and one-half of the Oxford cohort (vegetarians and health-conscious volunteers). Following a standardized protocol, blood samples were collected (1993-2002), aliquoted into plasma, serum, white blood cells and erythrocytes, and stored in liquid nitrogen.

Outcome assessment

Incident breast cancer cases were identified through population cancer registries or by active follow up using health insurance records, cancer and pathology registries, and contacts with participant. Subjects were followed up until cancer diagnosis (except non-melanoma skin cancer), death, emigration, or the end of the follow-up period.

Nested case-control study

Of 367,993 women, the present analysis excluded women with prevalent cancers at any site (n=19,853), missing diagnosis or censoring date (n=2,892), missing dietary or lifestyle information (n=3,339), in the top or bottom 1% of the ratio of energy intake to energy requirement (n=6,753), and non-first breast cancer cases (n=217), which left 334,939 women. Within this group, 11,576 women with invasive breast cancer were identified after a median follow-up of 11.5 years. We designed a case-control study nested among those who provided a blood sample. Within this subgroup, 3,858 women with invasive breast cancer were identified. Due to flooding that occurred in

the Danish Biobank, samples from Denmark were not included, leading to a total of 2,982 cases. For each case, one matched control was chosen randomly among cohort women without breast cancer. Controls were matched to cases by center, age at blood donation (± 3 months), menopausal status (pre; surgical post; natural post), time of the day at blood collection (± 1 hour), fasting status (< 3hrs; 3-6 hrs.; >6 hrs.) and phase of the menstrual cycle (early follicular; late follicular; peri-ovulatory; midluteal; other luteal).

The EPIC study was approved by the Ethical Committee of the International Agency for Research on Cancer and individual EPIC centers.

Fatty acid analyses

Fatty acids measured in plasma and erythrocyte membrane phospholipids are highly correlated, and exhibit similar coefficient correlations with dietary fatty acids estimated through questionnaires (13), suggesting that both matrices can be used as biomarkers of habitual intake. In the present study, fatty acid concentrations were determined in plasma phospholipids, as our previous cross-sectional study within the EPIC study showed that some specific fatty acids measured in this fraction are reliable biomarkers of specific food intakes (14,15).

As previously described (11), total lipids were extracted from plasma samples (200 µl) with chloroform-methanol 2:1 (v/v) containing antioxidant butylated hydroxytoluene and L-A-phosphatidylcholine-dimyristoyl-d_{54} as an internal standard. Phospholipids were purified by adsorption chromatography. Fatty acid methyl esters were formed by transmethylation. Analyses were carried out on 7890A gas chromatographs (7890N GC Agilent Technologies). Samples form cases and controls were processed in the same batch, and laboratory staff was blinded to any participant characteristics. Human plasma were used as quality control samples and included in each batch. Fatty acids were identified by their retention times compared with those of commercial standards. The relative concentration of each fatty acid, expressed as percent of total fatty acids, was quantified by integrating the area under the peak and dividing the result by the total area. Fatty acids

were also expressed as absolute concentrations in plasma (μmol/liter) based on the quantity of the methyl deuterated internal standard.

Coefficients of variation for fatty acids ranged from 1.81% for large peaks to 9.75% for the smallest peaks.

We calculated the percentage of the following groups: saturated fatty acids (SFA), cis-monounsaturated fatty acids (cis-MUFA), ruminant trans fatty acids, industrial trans fatty acids, cis-n-6 polyunsaturated fatty acids (cis-n-6 PUFA), long-chain n-6 PUFA (20:2n-6, 20:3n-6, 20:4n-6, 22:4n-6, 22:5n-6), n-3 PUFA, long-chain n-3 PUFA (20:3n-3, 20:4n-3, 20:5n-3, 22:5n-3, 24:5n-3, 24:6n-3, 22:6n-3), and ratio of long-chain n-6/long-chain n-3 PUFA. We also determined the desaturation indexes (DI) as the ratio of product to substrate, either oleic acid to stearic acid (DI_{18}) or the ratio of palmitoleic acid to palmitic acid (DI_{16}), as biomarkers of endogenous lipogenesis (16).

Hormonal receptor status

Information on estrogen receptor (ER) expression was available for 2,047 cases (1,649 ER-positive, 398 ER-negative), and on progesterone receptor (PR) expression for 1,729 cases (1,150 PR-positive, 579 PR-negative). Immunohistochemical measurement of ER and PR expression was performed in each EPIC centre. To standardize the quantification of the receptor status, the following criteria were applied for a positive receptor status: ≥10% cells stained, any 'plus system' description, ≥20fmo/mg, an Allred score of ≥3, an IRS ≥2, or an H-score≥10.

Statistical analyses

Baseline characteristics of cases and controls were compared using paired t-tests for continuous variables. For categorical variables, the statistical significance of case – control differences was tested using a chi-square test. All missing values were excluded from calculations.

In order to evaluate the association between fatty acids and breast cancer risk (overall and specific breast cancer subtypes by receptor status), odds ratios (OR) and their 95% confidence intervals (CI) were estimated using conditional logistic regression models. Plasma fatty acids were categorized into

quartiles (overall cancer risk; cancer by time between blood collection and breast cancer diagnosis or by menopausal status at the time of blood collection) or tertiles (analyses by hormonal receptor subtypes) based on the distribution of plasma levels in controls.

Multivariable models included potential confounding factors related to fatty acids and breast cancer risk: date of blood collection, body mass index (BMI, kg/m^2) (as a continuous variable), years of education (low; medium; high), height (as a continuous variable), menopausal hormone use at baseline (ever; never), alcohol intake at recruitment (as a continuous variable), age at first birth and parity combined (nulliparous; first birth before age 30y, 1-2 children; first birth before age 30y, $\geq$3 children; first birth after age 30y), energy intake (as a continuous variable), and family history of breast cancer (yes; no). Tests for trend were computed using the quartile-or tertile-specific means of each fatty acid.

Additionally, a forward selection procedure was run on all fatty acids including groupings, to select fatty acids that mostly contribute to the aetiological model. Adjustment variables mentioned above were forced into the model and fatty acids considered as explanatory effects are tested. Chi-Square statistic was computed for each variable not in the model, if it is significant at the entry level=0.05, the corresponding fatty acid was then added to the model. The procedure was repeated until none of the remaining variables meets with the entry criterion.

Sub-analyses were conducted according to hormonal receptor status (ER-positive, ER-negative, PR-positive, PR-negative), and tests of heterogeneity of associations were performed. Formal tests of heterogeneity were based on chi-square statistics, calculated as the deviations of logistic beta-coefficients observed in each of the subgroups relative to the overall beta-coefficient.

The false discovery rate (FDR, q-values) was computed for results from the multivariable models from the main analysis using the Benjamini-Hochberg correction to control for multiple comparisons (17).

Statistical tests were 2-sided, and $P<0.05$ was considered significant. All analyses were performed with the SAS 9.2 software (SAS Institute Inc., Cary N. Base SAS® 9.3 Procedures Guide. 2011).

Results

Characteristics of participants

Baseline characteristics of cases and controls are presented in Table 1. Cases had a significantly higher BMI, adult height, a lower number of full term pregnancies and an older age at first full term pregnancy.

Plasma phospholipid fatty acids in cases and controls

Mean plasma phospholipid fatty acid levels in cases and controls are provided in Table 2. Palmitic acid is the main SFA, oleic acid the main cis-MUFA, and linoleic acid the main n-6 PUFA, with a ratio of n-6 to n-3 PUFA higher than 2. Elaidic acid, the main ITFA, represents a higher percentage than vaccenic acid, the natural trans fatty acid.

Plasma phospholipid fatty acids and overall breast cancer risk

Table 3 presents OR and 95% CI of overall breast cancer according to quartiles of fatty acids, expressed as percent of total fatty acids. SFA were not statistically significantly associated with breast cancer risk. Higher levels of cis-MUFA were associated with increased risk of breast cancer (OR for the highest quartile compared with the lowest [Q4-Q1]=1.17; 95%CI=0.98-1.39; p for trend=0.042, q-value=0.259). Only palmitoleic acid remained statistically significantly related to breast cancer risk after FDR correction (OR [Q4-Q1]=1.37; 95%CI=1.14-1.64; p for trend=0.0001, q-value=0.004). Consistently, palmitoleic acid (16:1n-7) was the only fatty acid retained by the forward selection procedure (data not shown).

No significant association was found between overall breast cancer and levels of trans-MUFA or trans PUFA from natural ruminant sources or industrial sources (Table 3).

Levels of individual cis n-6 or n-3 PUFAs were not significantly associated with breast cancer incidence (Table 3). However, levels of total cis n-6 PUFA were inversely associated with breast

cancer risk (OR [Q4-Q1]=0.81; 95%CI=0.69-0.96; p for trend=0.035), while no further association was detected with total cis n-3 PUFA. However, the association with n-6 PUFA did not withstand correction for multiple testing (q-value=0.259). Further, the ratio of n-6 to n-3 PUFA was not associated with breast cancer development (Table 3).

A higher DI_{18} was positively associated with breast cancer (OR [Q4-Q1]=1.16; 95%CI=0.97-1.40; p for trend=0.031, q-value=0.259). Particularly, increased risk of breast cancer was associated with a high DI_{16}, even after controlling for multiple testing (OR for the highest quartile compared with the lowest [Q4-Q1]=1.28; 95%CI=1.07-1.54; p for trend=0.002, q-value=0.037).

Plasma phospholipid fatty acids and breast cancer risk by hormonal receptor status

Table 4 presents OR and 95% of breast cancer according to fatty acid groupings, presented by subgroup of hormonal receptor expression. Although not statistically significant, the positive association between breast cancer risk and DI_{16} remained irrespective of hormonal receptor status. Increased risk of ER-negative breast cancer was specifically associated with high levels of ITFA (OR for the highest tertile compared with the lowest [T3-T1]=2.01; 95%CI=1.03-3.90; p for trend=0.047), while no significant association was found with ER-positive breast cancer (p for heterogeneity=0.015).

Discussion

In this large prospective study, we found evidence that higher levels of MUFA, particularly palmitoleic acid, as well as higher DI_{16}, were associated with increased risk of breast cancer. In addition, higher levels of ITFA were specifically associated with ER-negative breast cancer.

Nutritional epidemiology has been limited by the assessment of dietary fatty acids through dietary assessment methodologies, prone to substantial measurement error. Measurement of plasma phospholipid fatty acid offer specific biomarkers of past dietary intakes of fatty acids that cannot be endogenously synthesized, irrespective of the source and quality of food (14,15). In contrast, weaker associations were found between dietary intakes and SFA, and MUFA because of endogenous synthesis and complex fatty acid metabolism (15).

Accumulating evidence supports a role of early increased *de novo* synthesis of MUFA in the development of breast cancer (16, 18). Stearoyl-CoA desaturase-1 (SCD-1) is the key enzyme in the synthesis of MUFA from SFA, suggesting the implication of SCD-1 activity in the biological alterations of breast cancer (16, 18). In agreement with our findings, some epidemiological studies reported a significant association between increased risk of breast cancer and increasing levels of plasma/serum phospholipid or erythrocyte membrane MUFA (palmitoleic acid and/or oleic acid) (19-21). Lipid imaging and profiling for tissue samples from different types of cancer reported abundant amounts of MUFA relative to PUFA in the cancer microenvironment compared with the adjacent normal tissue, leading to decreased in membrane fluidity, which, in turns, influences many crucial membrane-associated functions (22). MUFA can serve as mediators of signal transduction and cellular differentiation, and unbalanced levels of these mediators have been also implicated in carcinogenesis (16,18). On the other hand, data available from epidemiological studies have generally shown a negative association between estimated dietary intake of MUFA with breast cancer risk, at least in Mediterranean countries (23,24), suggesting the role of endogenously synthesized MUFA in breast cancer development, rather than exogenous dietary MUFA. Thus, these

data support the hypothesis that increased endogenous synthesis of MUFA, rather than exogenous dietary MUFA, may stimulate breast cancer development, and might represent a specific target for breast cancer prevention.

There are limited data on the impact of SFA and MUFA in the DI measured in plasma phospholipids. In a controlled cross-over study, a high dietary intake of SFA has been shown to increase the DI_{16} measured in blood cholesterol esters and phospholipids (25). As a consequence, a high DI_{16} in plasma phospholipids that is positively associated with breast cancer risk may be the result of a diet rich in SFA, with concomitant increased hepatic desaturation of dietary SFA to MUFA. In a large cross sectional study within EPIC, a weak correlation was found between dietary intake of oleic acid, the main dietary MUFA, and plasma phospholipid DI_{18}, suggesting that dietary MUFA may not be a strong determinant in the DI_{18} compared with endogenous synthesis from stearic acid. These data may suggest the effect of dietary SFA rather than dietary MUFA in high DI measured in plasma phospholipids.

We found no significant association between breast cancer risk overall or by hormonal receptor status and levels of n-3 PUFA from marine sources. In contrast, prospective studies conducted in Asian populations consistently reported an inverse association between breast cancer risk and dietary intake or biomarkers of n-3 PUFA (7). Because n-3 PUFA intake in Asian populations is higher compared to Western populations, it was suggested that n-3 PUFA intake from fish might be too low in the EPIC population to reveal a possible protective effect on breast cancer (11). However, in a prospective study conducted in Japan with high intakes of n-3 PUFA, no significant inverse association was found between n-3 PUFA and breast cancer risk, while a negative trend was reported between EPA and ER+PR+ breast cancer (26). Because of the competition between n-3 PUFA and n-6 PUFA for eicosanoids production as an underlying mechanism, ratio of n-3/n-6 PUFA in diet and blood phospholipids has been suggested to play a determinant role in breast cancer risk. Indeed, data from a meta-analysis of prospective studies reported a decreased risk of breast cancer

associated with increasing ratio of n-3/n-6 PUFA measured in diet or in serum phospholipids (27). However, no significant association remained among European populations (27). In agreement with this latter finding, we failed to report a significant inverse association between n-3/n-6 ratio and breast cancer risk within the EPIC study. In a prospective study conducted in a French population, breast cancer risk was not related to any dietary PUFA overall (28); however, opposite associations were seen according to food sources of PUFA (28), emphasizing the importance of considering food sources of PUFA. If long chain n-3 PUFA originates mainly from fish sources, we cannot distinguish the contribution of different food sources (vegetable oils, meat, processed foods) to n-6 PUFA levels in plasma phospholipids. This high level of heterogeneity between epidemiological studies may suggest that other micronutrients and biochemical pathways may modulate the relationship between PUFA and breast cancer. In support of this hypothesis, one prospective study showed that antioxidant supplementation modified the association between PUFA and breast cancer risk (29). Further epidemiological studies should incorporate markers of micronutrient intake and other metabolic factors linked to breast cancer (e.g. insulin, inflammatory markers).

Trans fatty acids are classified as natural or industrially produced. Natural trans fatty acids are produced by the gut bacteria of ruminant animals and are found in small amounts in the food products from these animals. ITFA are formed when fats and oils are partially hydrogenated during industrial processing techniques, and these fatty acids are found in fast foods, industrially-produced products, snack, deep-fried foods, and baked goods. There is evidence that ITFA significantly increases the risk of coronary heart disease more than any other dietary component (30). The average intake of ITFA in many European countries is now relatively low; however, as the majority of the European countries still do not limit the content of ITFA in food, a large number of products containing high levels of ITFA are still available in Europe (31).

Some epidemiological studies have reported a positive association between intake of ITFA and risk of breast cancer (11), ovarian cancer (32), colon cancer, and prostate cancer (33). In the current study,

we confirm and refine our previous data on breast cancer (11) by reporting a positive association between plasma phospholipid ITFA isomers and breast cancer risk restricted to the subtype of ER-negative tumours. Few mechanistic data on the effect of ITFA on cancer development are available. One study showed that elaidic acid, the main ITFA, induces hepatic *de novo* fatty acid synthesis in vitro through upregulating the SREBP-1 pathway, while cis MUFA and SFA did not show an effect (34). In contrast to ITFA, we found no significant association between natural trans fatty acids and breast cancer risk, overall or by hormonal receptor status.

This study had several strengths including its prospective design, based on a very large number of incident breast cancer cases with detailed clinical and epidemiologic data. Additionally, we were able to separate trans fatty acid isomers from natural and industrial processes. The major limitation of the study is the single collection of blood samples at baseline. Finally, given the longer lifespan of fatty acids in adipose tissue and erythrocytes compared with plasma, it might be suggested that fatty acids measured in these matrices offer a better measure of longer-term intake than fatty acids measured in plasma phospholipids. However, there are data suggesting that plasma fatty acids are correlated with erythrocyte levels (13).

These findings suggest that increased endogenous synthesis of MUFA estimated several years prior to diagnosis may be associated with breast cancer development. The identification of dietary/lifestyle factors as potential regulators of endogenous MUFA synthesis could provide new strategies for breast cancer prevention. ITFA may also specifically increase ER-negative breast cancer risk. The poor prognosis and high burden of ER-negative breast cancer mortality make this subtype a priority for prevention. Eliminating ITFA in industrial processes and in foods could offer a relatively straightforward public health action for reducing non-communicable disease risk.

Acknowledgments

The authors gratefully acknowledge Mrs Mélanie Collin and Mrs Anne-Sophie Gross within the Lipidomic Platform (Plateforme de Biologie Intégrée) at Institut Gustave Roussy, and Mrs Béatrice Vozar at the International Agency for Research on Cancer, for their outstanding assistance with laboratory measurements of plasma fatty acids.

References

1. Fitzmaurice C, Dicker D, Pain A, Hamavid H, Moradi-Lakeh M, MacIntyre MF, et al. The global burden of cancer 2013. JAMA Oncol 2015;1(4):505-27.

2. Ezzati M, Riboli E. Can noncommunicable diseases be prevented? Lessons from studies of populations and individuals. Science 2012;337(6101):1482-87.

3. Turner LB. A meta-analysis of fat intake, reproduction, and breast cancer risk: an evolutionary perspective. Am J Hum Biol 2011;23(5):601-8.

4. Thiébaut A, Kipnis V, Chang SC, Subar AF, Thompson FE, Rosenberg PS, et al. Dietary fat and postmenopausal invasive breast cancer in the National Institutes of Health-AARP Diet and Health Study Cohort. J Natl Cancer Inst 2007;99(6):451-62.

5. Prentice RL, Caan B, Chlebowski RT, Patterson R, Kuller LH, Ockene JK, et al. Low-fat dietary pattern and risk of invasive breast cancer: the Women's Health Initiative Randomized Controlled Dietary Modification Trial. JAMA 2006;295(6):629-42.

6. Voorrips LE, Brants HAM, Kardinaal AFM, Hiddink GJ, van den Brandt PA, Goldbohm RA. Intake of conjugated linoleic acid, fat, and other fatty acids in relation to postmenopausal breast cancer: the Netherlands Cohort Study on Diet and Cancer. Am J Clin Nutr 2002; 76(4):873-82.

7. Zheng JS, Hu XJ, Zhao YM, Yang J, Li D. Intake of fish and marine n-3 polyunsaturated fatty acids and risk of breast cancer: meta-analysis of data from 21 independent prospective cohort studies. BMJ 2013 ;346:f3706.

8. Sieri S, Chiodini P, Agnoli C, Pala V, Berrino F, Trichopoulou A, et al. Dietary Fat Intake and Development of Specific Breast Cancer Subtypes. J Natl Cancer Inst 2014;106(5):106-14.

9. Kim EH, Willett WC, Colditz GA, Hankinson SE, Stampfer MJ, Hunter DJ, et al. Dietary fat and risk of postmenopausal breast cancer in a 20-year follow-up. Am J Epidemiol 2006;164(10):990-7.

10. Saadatian-Elahi M, Norat T, Goudable J, Riboli E Biomarkers of dietary fatty acid intake and the risk of breast cancer: a meta-analysis. Int J Cancer 2004;111(4):584-91.

11. Chajès V, Thiébaut ACM, Rotival M, Gauthier E, Maillard V, Boutron-Ruault MC, et al. Association between serum trans-monounsaturated fatty acids and breast cancer risk in the E3N-EPIC study. Am J Epidemiol 2008;167(11):1312-20.

12. Schmidt JA, Gorst-Rasmussen A, Nyström PW, Christensen JH, Schmidt EB, Dethlefsen C, et al. Baseline patterns of adipose tissue fatty acids and long-term risk of breast cancer: a case-cohort study in the Danish cohort Diet, Cancer, and Health. Eur J Clin Nutr 2014;68(10):1088-94.

13. Patel PS, Sharp SJ, Jansen E, Luben RN, Khaw KT, Wareham NJ, et al. Fatty acids measured in plasma and erythrocyte-membrane phospholipids and derived by food-frequency questionnaire and the risk of new-onset type 2 diabetes: a pilot study in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Norfolk cohort. Am J Clin Nutr 2010;92(5):1214-22.

14. Chajès V, Biessy C, Byrnes G, Deharveng G, Saadatian-Elahi M, Jenab M, et al. Ecological-level associations between highly processed food intakes and plasma phospholipid elaidic acid concentrations: results from a cross-sectional study within the European Prospective Investigation into Cancer and Nutrition (EPIC). Nutr Cancer 2011;63(8):1235-50.

15. Saadatian-Elahi M, Slimani N, Chajès V, Jenab M, Goudable J, Biessy C, et al. Plasma phospholipid fatty acid profiles and their association with food intakes: results from a cross-

sectional study within the European Prospective Investigation into Cancer and Nutrition. Am J Clin Nutr 2009;89(1):331-46.

16. Chajès V, Joulin V, Clavel-Chapelon F. The fatty acid desaturation index of blood lipids, as a biomarker of hepatic stearoyl-CoA desaturase expression, is a predictive factor of breast cancer risk. Curr Opin Lipidol 2011;22(1):6-10.

17. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Ser B 1995;57:289-300.

18. Igal, RA. Stearoyl CoA desaturase-1 : New insights into a central regulator of cancer metabolism. Biochim Biophys Acta 2016;1861(12 Pt A):1865-80.

19. Chajès V, Hultén K, Van Kappel AL, Winkvist A, Kaaks R, Hallmans G, et al. Fatty acid composition in serum phospholipids and risk of breast cancer : an incident case-control study in Sweden. Int J Cancer 1999;83(5):585-90.

20. Pala V, Krogh V, Muti P, Chajès V, Riboli E, Micheli A, et al. Erythrocyte membrane fatty acids and subsequent breast cancer: a prospective Italian study. J Natl Cancer Inst 2001;93(14):1088-95.

21. Shannon J, King IB, Moshofsky R, Lampe JW, Gao DL, Ray RM, et al. Erythrocyte fatty acids and breast cancer risk: a case-control study in Shanghai, China. Am J Clin Nutr 2007;85(4):1090-7.

22. Guo S, Wang Y, Zhou D, Li Z. Significantly increased monounsaturated lipids relative to polyunsaturated lipids in six types of cancer microenvironment are observed by mass spectrometry imaging. Sci Rep 2014;4:5959.

23. Bosetti C, Pelucchi C, La Vecchia C. Diet and cancer in Mediterranean countries: carbohydrates and fats. Public Health Nutr 2009;12(9A):1595-600.

24. Escrich E, Moral R, Solanas M. Olive oil, an essential component of the Mediterranean diet, and breast cancer. Public Health Nutr 2011;14(12A):2323-32.

25. Warensjö E, Risérus U, Gustafsson IB, Mohsen R, Cederholm T, Vessby B. Effects of saturated and unsaturated fatty acids on estimated desaturase activities during a controlled dietary intervention. Nutr Metab Cardiovasc Dis 2008;18(10): 683-90.

26. Kiyabu GY, Inoue M, Saito E, Abe SK, Sawada N, Ishihara J, et al. Fish, n-3 polyunsaturated fatty acids and n-6 polyunsaturated fatty acids intake and breast cancer risk: the Japan Public Health Center-based prospective study. Int J Cancer 2015;137(12):2915-26.

27. Yang B, Ren XL, Fu YQ, Gao JL, Li D. Ratio of n-3/n-6 PUFAs and risk of breast cancer: a meta-analysis of 274135 adult females from 11 independent prospective studies. BMC Cancer 2014;14:105.

28. Thiébault ACM, Chajès V, Gerber M, Boutron-Ruault MC, Joulin V, Lenoir G, et al. Dietary intakes of ω-6 and ω-3 polyunsaturated fatty acids and the risk of breast cancer. Int J Cancer 2009;124(4):924-31.

29. Pouchieu C, Chajès V, Laporte F, Kesse-Guyot E, Galan P, Hercberg S, et al. Prospective associations between plasma saturated, monounsaturated and polyunsaturated fatty acids and overall and breast cancer risk – Modulation by antioxidants: a nested case-control study. Plos One 2014;9(2):e90442.

30. Mozaffarian D, Katan MB, Ascherio A, Stampfer MJ, Willett WC. Trans fatty acids and cardiovascular disease. New Engl J Med 2006;354(15):1601-13.

31. Stender S, Dyerberg J, Bysted A, Leth T, Astrup A. A trans world journey. Atheroscler Suppl 2006;7(2):47-52.

32. Merritt MA, Cramer DW, Missmer SA, Vitonis AF, Titus LJ, Terry KL. Dietary fat intake and risk of epithelial ovarian cancer by tumour histology. Br J Cancer 2014;110(5):1392-401.

33. Hu J, La Vecchia C, de Groh M, Negri E, Morrison H, Mery L. Dietary trans fatty acids and cancer risk. Eur J Cancer Prev 2011;20(6):530-8.

34. Shao F, Ford DA. Elaidic acid increases hepatic lipogenesis by mediating Sterol Regulatory Element Binding Protein-1c activity in HuH-7 cells. Lipids 2014;49(5):403-13.

Table 1. Baseline characteristics of control and cancer subjects in the EPIC Study

Characteristic	Controls (n=2,982)	Cases (n=2,982)	P[1] value
Mean age (years)	53.95 (8.17)	53.94 (8.17)	-
Mean Body Mass Index (kg/m^2)	25.30 (4.23)	25.53 (4.47)	0.03
Mean adult height (cm)	161.23 (6.51)	161.58 (6.55)	0.02
Age at menarche (years)	12.98 (1.56)	12.95 (1.51)	0.34
Age (years) at menarche (%)			0.76
<12	491 (16.7)	473 (16.2)	
12-13	649 (22.1)	669 (22.9)	
≥13	1802 (61.2)	1782 (60.9)	
Full term pregnancy (%)	2553 (87.4)	2500 (85.7)	0.04
Age at first full term pregnancy (years) – among parous women	25.24 (4.25)	25.62 (4.32)	0.03
Number of full term pregnancy – among parous women	2.32 (1.05)	2.24 (0.98)	0.006
Combined age at first birth and parity (%)			0.05
Nulliparous	368 (12.9)	418 (14.7)	
First birth before age 30 years, 1-2 children	1360 (47.8)	1309 (46.0)	
First birth before age 30 years, ≥3 children	741 (26.1)	705 (24.7)	
First birth after age 30 years	375 (13.2)	416 (14.6)	
Age (years) **at menopause** (%)			0.49
Pre-menopausal	753 (25.2)	753 (25.3)	
<45	47 (1.6)	49 (1.6)	
45-54	826 (27.7)	821 (27.5)	
55+	1356 (45.5)	1359 (45.6)	
Ever use of menopausal hormones (%)	888 (31.2)	897 (31.4)	0.96

Years of education (%)			0.66
Low	998 (44.9)	979 (44.7)	
Medium	606 (27.2)	581 (26.6)	
High	620 (27.9)	627 (28.7)	
Family history of breast cancer (%)	152 (11.0)	183 (13.2)	0.34
Smoking status (%)			0.63
Never	1689 (57.9)	1653 (56.7)	
Former	705 (24.2)	727 (24.9)	
Smoker	522 (17.9)	535 (18.4)	
Mean Physical activity (at work and leisure expressed as	103.28	101.20	0.11
Metabolic Equivalent Tasks (METS) units)	(53.18)	(53.28)	
Physical activity (%)			0.13
Inactive	356 (12.5)	406 (14.3)	
Moderately inactive	903 (31.6)	907 (31.9)	
Moderately active	1313 (45.9)	1279 (44.9)	
Active	286 (10.0)	255 (8.9)	
Mean alcohol intake (g/d)	8.34 (12.07)	8.62 (12.31)	0.35
Mean alcohol intake – consumers only (g/d)	10.09 (12.59)	10.50 (12.84)	0.43
Mean energy intake (kcal/d)	1949.66	1973.61	0.07
	(544.34)	(535.32)	

Data are presented as means (SD) or percentages. All missing values were excluded from calculations.

[1]Baseline characteristics of cases and controls were compared using paired t-tests for continuous variables. For categorical variables, the statistical significance of case – control differences was tested using a chi-square test. - No p-value was computed for comparing mean ages between cases and controls because control subjects were matched to cases by age at blood donation.

Table 2. Mean plasma phospholipid fatty acids at baseline among control and case subjects in the EPIC Study

Fatty acids (% of total fatty acids)	Controls (n=2,982) Mean (SD)	Cases (n=2,982) Mean (SD)
14:0 (myristic acid)	0.27 (0.09)	0.27 (0.09)
15:0 (pentanoic acid)	0.18 (0.06)	0.18 (0.06)
16:0 (palmitic acid)	25.53 (2.23)	25.62 (2.23)
17:0 (heptanoic acid)	0.39 (0.08)	0.39 (0.08)
18:0 (stearic acid)	14.09 (1.64)	14.03 (1.55)
16:1n-7 (palmitoleic acid)	0.64 (0.25)	0.66 (0.27)
18:1n-5	0.16 (0.12)	0.16 (0.13)
18:1n-7 (cis-vaccenic acid)	1.50 (0.39)	1.52 (0.34)
18:1n-9 (oleic acid)	10.32 (2.09)	10.42 (2.10)
16:1n-7/9 (palmitelaidic acid)	0.44 (0.47)	0.43 (0.44)
18:1n-9/12 (elaidic acid)	0.36 (0.24)	0.36 (0.22)
18:1n-7 (vaccenic acid)	0.30 (0.15)	0.29 (0.14)
18:2n-6 (linoleic acid)	22.10 (3.22)	21.97 (3.25)
18:3n-6 (γ-linolenic acid)	0.10 (0.05)	0.11 (0.47)
20:3n-6 (di-homo-γ-linolenic acid)	3.34 (0.83)	3.34 (0.84)
20:4n-6 (arachidonic acid)	10.97 (2.21)	10.98 (2.17)
22:4n-6 (adrenic acid)	0.37 (0.54)	0.38 (0.57)
22:5n-6 (osbond acid)	0.25 (0.10)	0.26 (0.11)
CLA9cis,11trans (conjugated linoleic acid)	0.22 (0.38)	0.22 (0.36)
18:2ct, 18:2tc, 18:2tt (trans linoleic acid)	0.18 (0.11)	0.18 (0.10)

18:3n-3ccc (α-linolenic acid)	0.20 (0.09)	0.20 (0.09)
20:5n-3 (eicosapentaenoic acid, EPA)	1.18 (0.77)	1.16 (0.73)
22:5n-3 (docosapentaenoic acid, DPA)	1.00 (0.28)	1.00 (0.31)
22:6n-3 (docosahexaenoic acid, DHA)	4.73 (1.47)	4.73 (1.47)
18:3n-3cct, ctt, ttt (trans α-linolenic acid)	0.03 (0.03)	0.03 (0.03)
20:3n-9 (mead acid)	0.19 (0.14)	0.20 (0.14)
Total SFA	40.54 (2.24)	40.56 (2.08)
(10:0, 12:0, 14:0, 15:0, 16:0, 17:0, 18:0, 20:0, 22:0, 24:0)		
Total cis-MUFA	13.00 (2.37)	13.13 (2.39)
(14:1, 15:1, 16:1n-7, 17:1, 18:1n-5, 18:1n-7, 18:1n-9, 20:1, 22:1, 24:1)		
Total trans ruminant fatty acids	0.94 (0.65)	0.93 (0.61)
(trans 16:1n-7/9, trans 18:1n-7, CLA)		
Total trans industrial fatty acids	0.57 (0.30)	0.57 (0.28)
(18 :1n-9/12, trans 18:2n-6, trans 18:3n-3)		
Total cis-n-6 PUFA	37.50 (3.23)	37.39 (3.20)
(18:2, 18:3, 20:2, 20:3, 20:4, 22:4, 22:5)		
Total long-chain n-6 PUFA	15.30 (2.53)	15.32 (2.48)
(20:2, 20:3, 20:4, 22:4, 22:5)		
Total cis-n-3 PUFA	7.19 (2.17)	7.17 (2.14)
(18:3, 18:4, 20:4, 20:5, 22:5, 24:5, 24:6, 22:6)		
Total long-chain n-3 PUFA	6.98 (2.16)	6.97 (2.14)
(20:4, 20:5, 22:5, 24:5, 24:6, 22:6)		
Long-chain n-6/n-3 PUFA	2.39 (0.80)	2.40 (0.83)
Desaturation index$_{18}$ (18:1n-9cis/18:0)	0.75 (0.20)	0.76 (0.20)
Desaturation index$_{16}$ (16:1n-7/9cis/16:0)	0.03 (0.01)	0.03 (0.01)

Table 3. Odds ratios (OR) and 95% confidence intervals (CI) of breast cancer by quartiles of plasma phospholipid fatty acids (percentage of total fatty acids)

Plasma phospholipid fatty acids	Quartiles of plasma phospholipid fatty acids					
	1 (reference)	2	3	4	P for trend on means††	Q-value (FDR corrected p-values)
Saturated fatty acids (SFA)						
14:0 (myristic acid)						
Range (%)	< 0.20	[0.20-0.26[	[0.26-0.32[	≥ 0.32		
Cases/controls (n)	726/746	807/804	696/727	753/705		
OR; 95% CI	1.00	1.05;0.90-1.23	1.00;0.85-1.18	1.13;0.95-1.34	0.279	0.794
15:0 (pentanoic acid)						
Range (%)	< 0.15	[0.15-0.18[	[0.18-0.21[	≥ 0.21		
Cases/controls (n)	854/816	778/772	663/684	687/710		

OR; 95% CI	1.00	0.98;0.85-1.13	0.94;0.81-1.11	0.95;0.80-1.12	0.471	0.934
16:0 (palmitic acid)						
Range (%)	< 24.26	[24.26-25.57[	[25.57-26.77[	≥ 26.77		
Cases/controls (n)	679/746	807/749	711/742	785/745		
OR; 95% CI	1.00	1.21;1.03-1.42	1.08;0.91-1.29	1.20;0.99-1.45	0.146	0.491
17:0 (heptanoic acid)						
Range (%)	< 0.34	[0.34-0.39[	[0.39-0.43[	≥ 0.43		
Cases/controls (n)	852/779	709/714	745/744	676/745		
OR; 95% CI	1.00	0.94;0.81-1.10	0.93;0.79-1.01	0.83;0.69-0.99	0.054	0.273
18:0 (stearic acid)						
Range (%)	< 13.15	[13.15-14.05[	[14.05-14.98[	≥ 14.98		
Cases/controls (n)	763/746	776/745	744/749	699/742		
OR; 95% CI	1.00	1.01;0.86-1.18	0.94;0.80-1.10	0.86;0.72-1.03	0.086	0.354
Cis-monounsaturated						
fatty acids (MUFA)						
16:1n-7 (palmitoleic						

acid)						
Range (%)	< 0.47	[0.47-0.59[	[0.59-0.75[	≥ 0.75		
Cases/controls (n)	693/769	650/722	803/746	836/745		
OR; 95% CI	1.00	1.02;0.87-1.19	1.29;1.09-1.52	1.37;1.14-1.64	0.0001	0.004
18:1n-5						
Range (%)	< 0.10	[0.10-0.13[	[0.13-0.18[	≥ 0.18		
Cases/controls (n)	772/747	832/775	673/715	705/745		
OR; 95% CI	1.00	1.05;0.90-1.22	0.94;0.80-1.10	0.96;0.80-1.14	0.422	0.935
18:1n-7 (cis-vaccenic						
acid)						
Range (%)	< 1.30	[1.30-1.47[	[1.47-1.66[	≥ 1.66		
Cases/controls (n)	700/750	757/745	715/742	810/745		
OR; 95% CI	1.00	1.13;0.98-1.32	1.09;0.94-1.28	1.23;1.05-1.45	0.024	0.259
18:1n-9 (oleic acid)						
Range (%)	< 8.89	[8.89-10.05[	[10.05-11.50[	≥ 11.50		
Cases/controls (n)	735/750	684/741	766/746	797/745		

OR; 95% CI	1.00	0.96;0.83-1.12	1.08;0.93-1.27	1.16;0.97-1.39	0.059	0.273
Trans-monounsaturated fatty acids						
16:1n-7/9 (palmitelaidic acid)						
Range (%)	< 0.16	[0.16-0.28[	[0.28-0.48[	≥ 0.48		
Cases/controls (n)	739/746	767/745	743/747	733/744		
OR; 95% CI	1.00	1.07;0.86-1.34	1.03;0.78-1.37	0.92;0.63-1.34	0.753	0.935
18:1n-9/12 (elaidic acid)						
Range (%)	< 0.20	[0.20-0.29[	[0.29-0.46[	≥ 0.46		
Cases/controls (n)	753/768	768/725	784/758	677/731		
OR; 95% CI	1.00	1.11;0.95-1.30	1.09;0.92-1.30	1.00;0.82-1.24	0.963	0.991
18:1n-7 (vaccenic acid)						

Range (%)	< 0.19	[0.19-0.28[	[0.28-0.38[	≥ 0.38		
Cases/controls (n)	756/746	873/770	683/724	670/742		
OR; 95% CI	1.00	1.10;0.93-1.30	0.92;0.76-1.12	0.88;0.71-1.09	0.161	0.496
Cis-n-6 polyunsaturated fatty acids (PUFA)						
18:2n-6 (linoleic acid)						
Range (%)	< 20.03	[20.03-22.08[	[22.08-24.11[	≥ 24.11		
Cases/controls (n)	814/746	723/745	705/746	740/745		
OR; 95% CI	1.00	0.88;0.76-1.02	0.87;0.75-1.02	0.93;0.79-1.08	0.312	0.825
18:3n-6 (γ-linolenic acid)						
Range (%)	< 0.06	[0.06-0.09[	[0.09-0.12[	≥ 0.12		
Cases/controls (n)	749/777	684/714	831/793	718/698		
OR; 95% CI	1.00	0.98;0.84-1.14	1.06;0.91-1.24	1.03;0.87-1.21	0.583	0.935
20:3n-6 (di-homo-γ-						

linolenic acid)						
Range (%)	< 2.78	[2.78-3.29[	[3.29-3.86[	≥ 3.86		
Cases/controls (n)	711/746	794/747	736/744	741/745		
OR; 95% CI	1.00	1.10;0.95-1.28	1.00;0.85-1.17	0.97;0.82-1.14	0.545	0.935
20:4n-6 (arachidonic						
acid)						
Range (%)	< 9.55	[9.55-10.95[	[10.95-12.38[	≥ 12.38		
Cases/controls (n)	726/746	785/745	719/749	752/742		
OR; 95% CI	1.00	1.04;0.89-1.22	0.98;0.83-1.16	1.01;0.85-1.20	0.948	0.991
22:4n-6 (adrenic acid)						
Range (%)	< 0.28	[0.28-0.33[	[0.33-0.40[	≥ 0.40		
Cases/controls (n)	771/748	700/743	807/791	704/700		
OR; 95% CI	1.00	0.90;0.77-1.05	0.95;0.81-1.13	0.95;0.80-1.14	0.737	0.935
22:5n-6 (osbond acid)						
Range (%)	< 0.18	[0.18-0.24[	[0.24-0.31[	≥ 0.31		
Cases/controls (n)	747/758	793/752	714/737	728/735		

OR; 95% CI	1.00	1.05;0.90-1.23	0.95;0.80-1.13	0.98;0.82-1.18	0.643	0.935
Trans-n-6 PUFA						
Conjugated linoleic						
acid (CLA)						
CLA9cis,11trans						
Range (%)	< 0.11	[0.11-0.18[	[0.18-0.26[	≥ 0.26		
Cases/controls (n)	747/739	721/752	732/725	754/738		
OR; 95% CI	1.00	0.96;0.76-1.21	1.05;0.81-1.35	1.12;0.85-1.47	0.558	0.935
Non conjugated trans-						
n-6 PUFA						
18:2ct, 18:2tc, 18:2tt						
(trans linoleic acid)						
Range (%)	< 0.10	[0.10-0.15[	[0.15-0.24[	≥ 0.24		
Cases/controls (n)	827/824	707/688	738/725	710/745		
OR; 95% CI	1.00	1.06;0.88-1.27	1.01;0.80-1.29	0.92;0.69-1.23	0.649	0.935
Cis-n-9 PUFA						

20:3n-9 (mead acid)						
Range (%)	< 0.13	[0.13-0.18[	[0.18-0.24[	≥ 0.24		
Cases/controls (n)	721/746	835/819	694/712	732/705		
OR; 95% CI	1.00	1.08;0.93-1.25	1.02;0.86-1.20	1.08;0.91-1.28	0.515	0.935
Cis-n-3 PUFA						
18:3n-3ccc (α-linolenic acid)						
Range (%)	< 0.14	[0.14-0.18[	[0.18-0.24[	≥ 0.24		
Cases/controls (n)	786/798	711/693	763/746	722/745		
OR; 95% CI	1.00	1.05;0.91-1.22	1.06;0.90-1.23	0.97;0.82-1.15	0.768	0.935
20:5n-3 (eicosapentaenoic acid, EPA)						
Range (%)	< 0.71	[0.71-0.98[	[0.98-1.40[	≥ 1.40		
Cases/controls (n)	757/746	746/745	751/746	728/745		
OR; 95% CI	1.00	1.00;0.86-1.16	1.00;0.86-1.16	0.93;0.79-1.09	0.355	0.876

22:5n-3 (docosapentaenoic acid, DPA)						
Range (%)	< 0.80	[0.80-0.99[	[0.99-1.17[	≥ 1.17		
Cases/controls (n)	735/752	753/762	775/732	719/736		
OR; 95% CI	1.00	1.05;0.89-1.24	1.12;0.93-1.36	1.04;0.85-1.27	0.615	0.935
22:6n-3 (docosahexaenoic acid, DHA)						
Range (%)	< 3.72	[3.72-4.57[	[4.57-5.59[	≥ 5.59		
Cases/controls (n)	757/746	731/745	748/748	746/743		
OR; 95% CI	1.00	0.96;0.83-1.12	0.97;0.83-1.14	1.00;0.84-1.19	0.987	0.991
Trans-n-3 PUFA						
18:3n-3cct, ctt, ttt (trans α-linolenic acid)						
Range (%)	< 0.01	[0.01-0.02[	[0.02-0.04[	≥ 0.04		

Cases/controls (n)	770/743	528/554	625/625	580/581		
OR; 95% CI	1.00	0.91;0.76-1.10	0.99;0.80-1.22	1.01;0.77-1.33	0.991	0.991
Grouping						
Total SFA*						
Range (%)	< 39.49	[39.49-40.47[	[40.47-41.40[	≥ 41.40		
Cases/controls (n)	717/746	727/748	724/746	814/742		
OR; 95% CI	1.00	1.00;0.85-1.18	1.03;0.87-1.23	1.16;0.96-1.42	0.125	0.463
Total cis-MUFA†						
Range (%)	< 11.40	[11.40-12.71[	[12.71-14.34[	≥ 14.34		
Cases/controls (n)	740/746	667/745	765/746	810/745		
OR; 95% CI	1.00	0.93;0.80-1.08	1.08;0.92-1.26	1.17;0.98-1.39	0.042	0.259
Total trans ruminant						
fatty acids‡						
Range (%)	< 0.56	[0.56-0.77[	[0.77-1.11[	≥ 1.11		
Cases/controls (n)	754/746	749/747	742/744	737/745		
OR; 95% CI	1.00	0.99;0.84-1.17	1.00;0.81-1.23	1.03;0.77-1.36	0.874	0.935

Total trans industrial fatty acids§						
Range (%)	< 0.37	[0.37-0.48[	[0.48-0.69[	≥ 0.69		
Cases/controls (n)	780/746	770/746	722/745	710/745		
OR; 95% CI	1.00	0.99;0.85-1.16	0.93;0.78-1.11	0.97;0.76-1.16	0.514	0.935
Total cis n-6 PUFA‖						
Range (%)	< 35.81	[35.81-37.64[	[37.64-39.60[	≥ 39.60		
Cases/controls (n)	833/746	679/745	781/746	689/745		
OR; 95% CI	1.00	0.81;0.70-0.94	0.93;0.79-1.08	0.81;0.69-0.96	0.035	0.259
Total long-chain n-6 PUFA¶						
Range (%)	< 13.61	[13.61-15.30[	[15.30-17.04[	≥ 17.04		
Cases/controls (n)	718/746	747/745	818/752	699/739		
OR; 95% CI	1.00	1.02;0.87-1.19	1.11;0.94-1.30	0.94;0.78-1.12	0.709	0.935
Total cis n-3 PUFA#						
Range (%)	< 5.70	[5.70-6.80[	[6.80-8.32[	≥ 8.32		

Cases/controls (n)	776/747	694/744	789/748	723/743		
OR; 95% CI	1.00	0.90;0.77-1.05	0.99;0.85-1.17	0.93;0.78-1.11	0.661	0.935
Total long-chain n-3						
PUFA**						
Range (%)	< 5.48	[5.48-6.59[	[6.59-8.11[	≥ 8.11		
Cases/controls (n)	754/746	716/746	787/748	725/743		
OR; 95% CI	1.00	0.95;0.81-1.11	1.02;0.87-1.20	0.97;0.81-1.15	0.888	0.991
Long-chain n-6/n-3						
PUFA						
Range	< 1.82	[1.82-2.33[	[2.33-2.89[	≥ 2.89		
Cases/controls (n)	741/746	736/745	748/746	757/745		
OR; 95% CI	1.00	0.99;0.85-1.16	1.03;0.88-1.21	1.02;0.85-1.21	0.783	0.935
Desaturation indexes						
Desaturation index$_{18}$						
(18:1n-9cis/18:0)						
Range	< 0.62	[0.62-0.72[	[0.72-0.86[	≥ 0.86		

Cases/controls (n)	702/746	691/745	835/746	754/745		
OR; 95% CI	1.00	1.01;0.87-1.18	1.25;1.07-1.48	1.16;0.97-1.40	0.031	0.259
Desaturation index$_{16}$						
(16:1n-7cis/16:0)						
Range	< 0.018	[0.018-0.023[	[0.023-0.029[	≥ 0.029		
Cases/controls (n)	684/745	675/745	802/745	819/745		
OR; 95% CI	1.00	1.01;0.87-1.19	1.26;1.07-1.49	1.28;1.07-1.54	0.002	0.037

*Includes 10:0, 12:0, 14:0, 15:0, 16:0, 17:0, 18:0, 20:0, 22:0, 24:0; †Includes 14:1, 15:1, 16:1n-7, 17:1, 18:1n-5, 18:1n-7, 18:1n-9, 20:1, 22:1, 24:1; ‡Includes trans 16:1n-7/9, trans 18:1n-7, CLA; §Includes 18 :1n-9/12, trans 18:2n-6, trans 18:3n-3; ‖Includes 18:2, 18:3, 20:2, 20:3, 20:4, 22:4, 22:5; ¶Includes 20:2, 20:3, 20:4, 22:4, 22:5; #Includes 18:3, 18:4, 20:4, 20:5, 22:5, 24:5, 24:6, 22:6; **Includes 20:4, 20:5, 22:5, 24:5, 24:6, 22:6. †† Conditional logistic regression adjusted for date at blood collection, years of education, Body Mass Index, height, menopausal hormone use at baseline, alcohol at baseline, age at first birth and parity combined, energy intake, family history of breast cancer. ‡‡ FDR: false discovery rate

Table 4. Odds ratios (OR) and 95% confidence intervals (CI) for the highest tertile compared to the lowest of breast cancer by plasma phospholipid fatty acid groupings according to hormonal receptor status

Fatty acids (% of total fatty acids)	**ER-positive** (n=1,649 cases)		**ER-negative** (n=398 cases)			**PR-positive** (n=1,150 cases)		**PR-negative** (n=579 cases)		
	OR 95% CI	P trend on means††	**OR** 95% CI	P trend on means††	P for heterogeneity	**OR** 95%CI	P trend on means††	**OR** 95% CI	P trend on means††	P for heterogeneity
SFA*	**1.13** 0.90-1.43	0.299	**0.76** 0.45-1.29	0.315	0.182	**0.96** 0.72-1.29	0.795	**0.79** 0.51-1.22	0.279	0.449
cis MUFA†	**1.04** 0.83-1.29	0.749	**1.24** 0.80-1.92	0.331	0.460	**0.99** 0.77-1.30	0.970	**1.07** 0.75-1.53	0.779	0.822
cis n-6 PUFA‡	**1.01** 0.83-1.24	0.928	**0.93** 0.62-1.39	0.682	0.835	**0.94** 0.75-1.19	0.603	**1.14** 0.81-1.60	0.457	0.451
Long-chain cis n-6 PUFA§	**1.17** 0.95-1.44	0.140	**0.66** 0.42-1.01	0.067	0.006	**1.29** 1.00-1.64	0.062	**0.79** 0.55-1.14	0.223	0.013
cis n-3 PUFA‖	**0.99** 0.81-1.22	0.979	**0.82** 0.54-1.26	0.356	0.473	**1.06** 0.83-1.36	0.641	**0.76** 0.52-1.09	0.183	0.208

Long-chain cis n-3	**1.01**	0.838	**0.86**	0.456	0.577	**1.10**	0.496	**0.76**	0.199	0.189
PUFA¶	0.82-1.25		0.56-1.34			0.86-1.40		0.52-1.11		
Long-chain n-6/n-	**1.15**	0.214	**1.11**	0.638	0.416	**1.05**	0.715	**1.32**	0.131	0.521
3 PUFA	0.94-1.42		0.72-1.71			0.82-1.34		0.93-1.87		
Ruminant trans	**1.04**	0.755	**0.71**	0.410	0.446	**1.05**	0.601	**0.89**	0.704	0.473
fatty acids#	0.76-1.43		0.34-1.47			0.69-1.61		0.47-1.69		
Industrial trans	**0.82**	0.102	**2.01**	0.047	0.015	**0.98**	0.963	**0.82**	0.467	0.545
fatty acids**	0.62-1.10		1.03-3.90			0.68-1.41		0.47-1.42		
DI_{16}	**1.18**	0.136	**1.47**	0.086	0.583	**1.21**	0.138	**1.25**	0.234	0.494
	0.95-1.47		0.95-2.29			0.94-1.56		0.87-1.79		
DI_{18}	**0.95**	0.649	**1.18**	0.490	0.686	**0.92**	0.549	**1.08**	0.679	0.429
	0.76-1.18		0.75-1.84			0.71-1.21		0.74-1.58		

*Includes 10:0, 12:0, 14:0, 15:0, 16:0, 17:0, 18:0, 20:0, 22:0, 24:0; †Includes 14:1, 15:1, 16:1n-7, 17:1, 18:1n-5, 18:1n-7, 18:1n-9, 20:1, 22:1, 24:1; ‡Includes 18:2, 18:3, 20:2, 20:3, 20:4, 22:4, 22:5; §Includes 20:2, 20:3, 20:4, 22:4, 22:5; ‖Includes 18:3, 18:4, 20:4, 20:5, 22:5, 24:5, 24:6, 22:6; ¶Includes 20:4, 20:5, 22:5, 24:5, 24:6, 22:6; #Includes trans 16:1n-7/9, trans 18:1n-7, CLA; **Includes 18 :1n-9/12, trans 18:2n-6, trans 18:3n-3; ††Conditional logistic regression adjusted for date at blood collection, years of education, Body Mass Index, height, menopausal hormone use at baseline, alcohol, age at first birth and parity combined, energy intake, family history of breast cancer

CHAPTER VI:

GENERAL DISCUSSION

In this work we have explored aspects of nutritional epidemiology by combining self-reported dietary and lifestyle information together with biomarker measurements to deeply investigate features of the diet and cancer association. Our main objective was to develop novel statistical frameworks for the application of multivariate statistical techniques. This work was made possible by exploiting the availability of data and the unique features of the European Prospective Investigation into Cancer and nutrition study. Different themes were tackled, ranging from nutrient patterns to use of metabolomics and fatty acids, different endpoints, including carcinomas of the breast and the liver. This thesis focused on the use of multivariate analytical solutions to make full use of available exposure data, thus extracting relevant information that could improve our understanding of cancer aetiology in the field of nutritional epidemiology. Our approach progressively moved from conventional statistical modelling harbouring multivariable regressions coupled with multiple testing corrections, towards a more holistic scheme embracing multivariate contexts, using increasingly complex mathematical techniques. Evaluations primarily focused on nutrients and cancer association and then moved towards integration of dietary biomarkers, of features of untargeted and targeted metabolomics. These different features were evaluated together with lifestyle exposures, the common denominator of all investigations carried out throughout this thesis, using a methodological challenging integrative strategy to fully exploit a large amount of epidemiological information.

In this chapter, we will discuss different aspects pertaining to the data from different sources exploited within this thesis, addressing some strengths and weaknesses of questionnaire, biomarker and metabolomic data. Advances in lab technology, the importance of the validation of the findings, the necessity of replication as well as the rationale and evolution of the statistical framework that has been developed will be touched upon. Features of mediation analysis, that holds a central part in our MITM implementation, are extensively explained. Finally, future perspectives are evoked whereby the tools investigating the diet-cancer relation can be further extended to embrace Mendelian randomisation or through more complex pathway analyses.

A large part of the evidence assessed in this thesis relied on dietary information originating from validated questionnaire data, whereby nutrients and total energy were estimated from harmonised food composition tables, the ENDB, compiled from national

databases of the ten EPIC countries following standardized procedures [203,235,236]. Thus, analysis described in **Chapter 2** featured 23 nutrients and total energy, the predictors set in the MITM implementation outlined in **Chapter 3** included 13 main EPIC food groups, and the diet score used in the study presented in **Chapter 4** was constructed based on six dietary items known to be related to cancer risks [86]. In addition the variables for alcohol consumption (e.g. alcohol at baseline and lifetime alcohol intake), used either as part of the main exposures (**Chapters 3** and **4**) or as adjustment confounders (**Chapters 2** and **5**) were also appraised from lifestyle questionnaires [199].

Standard dietary assessment methods, like food frequency questionnaires are feasible and cost-effective to be administered in large epidemiological studies, but are prone to exposure misclassification [133]. Measurement error may account for some of the lack of consistency that has been pointed out in findings within and across studies relying on data from FFQs examining diet and cancer risk [143]. Some of the early results found in large cohorts were not confirmed with long-term follow-up [237] and many strong findings on the nutrition-cancer relationship unveiled in case-control studies could not be replicated in clinical trials [238,239] or in cohort studies [240]. Questionnaires are nonetheless a valuable tool for large-scale dietary assessment and remain the standard measure for diet in epidemiologic research [5,143]. Much research is taking on the challenge of evaluating FFQs and enhancing the quality of their reporting [143,241].

Regardless, new strategies are sought to move from traditional nutritional epidemiology that focuses on self-reported dietary and lifestyle factors towards ways to investigate the aetiology of diseases not relying on study participants' capacity to recall previous habits, yet exploiting objective measures to assess exposure [143]. Biomarkers measured in biological specimens are increasingly being used for this scope [139,163]. Dietary biomarkers and –omics technologies provide a very promising means to quantify dietary and other environmental exposures [242].

The work developed in this thesis utilized biomarker measurements, either to estimate the diet-disease risk associations, or as a complementary tool to combine evidence from different sources. In **Chapters 3** and **4** analytical frameworks that integrated, respectively, untargeted NMR and targeted MS data with dietary and lifestyle questionnaire data are described. Metabolic profiles were identified that were the

overlap signals in the MITM principle: biomarker signatures that are related to specific exposures and are predictive of cancer risk at the same time. The evaluation outlined in **Chapter 5** used biomarker data as the primary exposure of interest where 60 plasma fatty acids concentrations were examined in relation with breast cancer risk. These were quantified through an improved gas chromatography procedure that allowed a good separation of *trans* fatty acids. Combining questionnaire with biomarker data provided us with an unprecedented opportunity to deeply investigate the complex relationships between diet and the risk of cancer, using increasingly sophisticated statistical techniques.

An interesting property of dietary biomarkers measured in biological samples is that some of them reflect a great number of endogenous factors influencing foods and nutrients (e.g. involvement in metabolic pathways, genetic characteristics, excretion, tissue turnover, absorption effects, etc.) that affect the correlation of a biomarker with its corresponding dietary exposure [139]. Additionally, they also reflect more closely the dietary compound's bioavailable dose, the latter being the relevant parameter in any metabolic process they are involved in [243]. With all this in mind, valuable additional information of dietary exposure can be obtained through biomarker assessment.

Different classes of dietary biomarkers can be identified: the "recovery" biomarkers provide unbiased estimates of absolute dietary intakes and are therefore suitable to be used as reference measurements to assess the accuracy of dietary assessments [165]. These markers often reflect the short-term nutritional status and display moderate correlation values with estimates of dietary intake [139,163]. However, only a few recovery biomarkers are available, i.e. urinary doubly labelled water for total energy intake, and urinary nitrogen and potassium for dietary protein and potassium intakes, respectively [244]. Blood samples are usually collected in cohort studies at recruitment, largely because collecting many replicates of biosamples requires considerable resources. This may not be sufficient to describe the evolution of long-term dietary exposure using biomarker measurements. A repeated sampling of biospecimens would be a valuable asset to monitor changes in diet overtime in prospective designs and to better depict dietary intake / nutrient state at baseline and during follow-up [5]. In addition, the potential for bias in biosamples collected in nested case-control studies within prospective design is reduced but not absent. While these samples are collected before diagnosis, the impact of preclinical conditions may impact the biochemical

parameters, thus causing spurious associations [63]. Concentration values of dietary biomarkers may be difficult to compare across different studies, mainly due to heterogeneity in laboratory processes that may introduce systematic bias affecting the biomarker measurement [139,242]. These include the type of biological specimens obtained, the differences in sample handling (e.g. procedures of collection, storage, thawing), the methodologies employed to measure the biomarker (machinery, precision, limits of detection and quantification, day-to-day drifts, etc.) [139].

With recent advances in technology, many elements related to the laboratory settings have improved [245–247]. For example, a method (the group-batch profile – GBP method) has been developed to adjust NMR data for systematic variations introduced by sample work-up prior to spectral data acquisition [248]. The PC-PR2 method has been conceived to identify and quantify the contribution of relevant sources of variation in metabolomics data prior to investigation of etiological hypotheses [183]. This technique has been used in studies described in **Chapters 3** and **4**. Considerable efforts are currently underway to harmonize metabolomics data in order to allow pooling data together from different studies, to ensure a better comparability of results in international settings. Such harmonisation efforts have started in international collaborations such as the The COnsortium of METabolomics Studies (COMETS), a partnership among prospective cohort studies involved in the acquirement of metabolomics profiling. International consortia face the need to provide interdisciplinary solutions to investigate complex data, at a time when epidemiologic investigations are accumulating –omics data [249].

The unique attributes of metabolomics data and the increase in the amount of information they bring make them an appealing opportunity to take on the challenge brought by highly dimensional, collinear, nonlinear and non-normal data. With such overwhelming sets of data to process, there is an increased demand for statistical methodologies and modelling approaches that are needed for better analysis of data.

After pre-processing and exploratory steps, data analyses of metabolomics currently rely mostly on regression-based methods including multivariable regression models, multiple testing correction procedures, use of multivariate dimension reduction techniques, and to a lesser extent variable selection approaches [179,242]. Univariate approaches are employed in the first instance to uncover simple associations between

metabolites and exposure or response variables or alternately with disease outcomes. Multivariate techniques of dimension reduction applied to large metabolomics sets mainly aim to summarize information into a restricted number of latent variables known as the principal components. PCA and its derivatives are the most widely used methods, while Discriminant Analysis (DA) partitions observations with respect to the investigated outcome by maximising the ratio of intergroup to intragroup variation. PLS-based multivariate approaches combine PCA and MLR to identify latent factors capturing as much variation in predictors and responses by extracting linear combinations maximising the covariance of the latter sets. Variable selection techniques entail a penalisation introduced in regression approaches to ensure sparsity by shrinking the values of some of the regression coefficient estimates towards zero. These are known as regularized linear regressions and mainly comprise ridge regression, Lasso and its variants as well as Elastic Net. These methods are progressively being applied to –omics data. In particular, multivariate approaches are subject to over-fitting making validation a mandatory step for analytical strategies employing these methods. Cross-validation techniques that do not call for the appraisal of additional independent samples are typically used to internally validate the findings. In this procedure, the data is randomly partitioned into a training set used to build a given model and a test set that is removed, usually with a 90%-10% fold proportion. The process is then iterated until each sample has served as a test set once. It is a model validation technique evaluating the accurate predictive performance of the model in practice and its robustness in face of data perturbations [242,250,251]. Yet cross-validation does not guarantee good performance across different populations and may even lead to an overestimation of the discriminatory classifier performance likely due to biases introduced in the process [251,252]. The direction is now in favour of an external independent validation of results that would produce more conservative results, but alas even such external validations can possibly be subject to some biases, selective reporting and optimism causing them to be inflated [251,253]. Validation has become an issue of special concern with the exponential growth of –omics that powered expectations for a cutting-edge era of personalized medicine. The current recommendation is to adopt routine external validation of biomarkers and metabolites, preferably in much larger studies than in current practice, and if possible by different teams [252]. Given the inherent complexity of biomarker data, it is essential to differentiate true signals from false positives and

assess the generalizability of metabolic signatures that arise from analyses [251,252]. In **Chapter 3**, an internal cross validation procedure was performed to evaluate the predictive performance of the PLS models. The receiver operating characteristic (ROC) curve and the associated area under the curve (AUC) were determined from conditional logistic models including progressively the PLS scores, separately for lifestyle and metabolomic signatures. The AUC unavoidably increases with the number of covariates added to the conditional logistic model. A resampling scheme was devised to compute objective unbiased estimates of the AUC inspired from the work of Uno et al [254]. For each one of the 1000 drawn bootstrap samples, a 10-fold cross-validation was performed, repeated 10 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. A replication of these findings in independent studies is needed.

Another motivation for a replication of our findings in external studies or using larger samples is the small sample size we had at hand. In the nested case-control studies on hepatocellular carcinoma presented in **Chapter 3** and **Chapter 4**, the sample sizes were very modest with 114 cases and 222 controls, and 147 cases and 147 controls, respectively. We made a rather opportunistic use of the available data that were at our disposal within different nested case-control studies in EPIC where metabolomic data was accessible to investigate the diet-cancer associations or to implement statistical strategies in proof-of-concept designs. In **Chapter 5**, we looked into associations between levels of 60 plasma phospholipids fatty acids in one of the largest nested case-control studies to date to ascertain fatty acids from biomarkers collected within a prospective study. Due to a flooding that occurred in the Danish Biobank, samples from Denmark were not included, when these will be added to the fold, there will be possibly more power to detect associations that did not withstand multiple correction testing.

Throughout this thesis, we moved from a multivariate problem with dietary data (**Chapter 2**) to a higher-level multivariate problem integrating biomarkers (**Chapter 3**) and then onto a more specific and more tightly defined problem (**Chapter 4**). We first employed TT, a dimension reduction technique to take on one set of nutrients (**Chapter 2**), then made use of PLS to best summarise information from two sets of data and then

applied a multiple PLS scheme in a more carefully controlled context. We also improved on our usage of mediation analysis from a generic use to evaluate the mediating role played by the extracted metabolic signatures (**Chapter 3**) to a more refined use adapted to our study design (**Chapter 4**). More specifically, in the different stages of the development of the statistical framework for the MITM implementation different factors and exposures were considered. We first embraced a multitude of exposures in the first application of the MITM, with 13 main EPIC food groups out of 21 diverse lifestyle exposures in **Chapter 3**. In the next exercise presented in **Chapter 4**, we simplified the exposure to diet by using a diet score constructed based on 6 dietary items, this may have been a simplification but it reduced the dilution / dispersion of information by having one factor for diet, and possibly resulted in a more specific metabolic factor in relation to dietary exposure. The framework developed is flexible and can accommodate other statistical methods that can fit like block parts and replace those in use (e.g. sparse-PLS or canonical correlation analysis instead of PLS) and can be tailored to be used with other –omics datasets and disease endpoints. This stems from the conceptual strength of the MITM [162] sustaining that any past exposure may leave alterations, either metabolic, genetic, epigenetic *inter alia*, that are only expressed far later in time, depending on subsequent exposures. The MITM sets the challenge to first identify these changes that can be recognised as overlap biomarkers mirroring previous exposures and related to pathophysiological conditions, and then to monitor those complex changes at the molecular level and relate them and interpret their effects with respect to the mechanisms of carcinogenesis. These will ultimately lead to a better understanding of the underlying ecology of cancer development in an attempt to connect the external exposures to the palette of internal biochemical modifications.

In our evaluation of whether the metabolic signals mediated the association between a given exposure or a lifestyle profile and HCC risk, we resorted to mediation analysis (**Chapters 3** and **4**). Mediation analysis is an increasingly utilised technique, widely used across many disciplines, to explore various causal pathways, beyond the estimation of simple associations. Mediation analysis investigates the mechanisms that underlie an observed relationship between an exposure variable and an outcome variable and examines how they relate to a third intermediate variable, the mediator [195]. Rather than hypothesizing only a direct causal relationship between the independent variable and the dependent variable, a mediational model hypothesizes

that the exposure variable causes the mediator variable, which in turn causes the outcome variable. The direct and the indirect (through the mediator) levels of association levels are then estimated from the outcome and mediator models [193]. Although mediation analysis has become very popular in social sciences, its use remains challenging. Over simplistic regression models, the possibly greatest merit of mediation analysis is that it allows the synergistic structure of the relationship between exposure, mediator and outcome variables to be captured and quantified. By introducing more complex functional relationships between variables, thus mimicking features of pathway analysis, the interpretation of model parameters needs to account for the large amount of underlying hypotheses subjacent each mediation model. Very strong assumptions are required for such an ambitious causal endeavour, they must be met and confounders must be accounted for in order to have a causal interpretation of the findings [192,193]. We were faced with some of these challenges that we addressed especially in **Chapter 4**. One such example relates to temporality; the exposure must precede the mediator that in turn precedes the outcome to satisfy the chronological ordering assumption. In EPIC and most observational epidemiology settings, most variables of interest, including the exposures and mediators under study, were simultaneously assessed at baseline, together with the collection of biological samples. Yet, lifestyle and metabolomics reflect exposure windows of different nature and time length, thus our working assumption was to consider these factors as relatively stable in EPIC. A number of issues still require further investigation including intermediate confounders, multiple mediators and their inter-correlations and mediator-outcome confounders that are affected by the exposure to mention a few. These scenarios may not be trivial to handle, and current research is focusing on such challenging aspects and solutions are emerging [192–194,255–257]. Nonetheless, mediation analysis remains a tightly controlled environment where every variable entering the DAG and every association arrow that is drawn has to comply with strict hypotheses [258].

To overcome challenges related to confounding and reverse causality in aetiological models, a Mendelian randomization (MR) method was developed as a way to use genetic variants as an instrumental variable for the exposure of interest [259,260]. The rationale is that, due to the random heritability of genetic traits brought by the random assortment of alleles at the time of gamete formation [260], if a genetic variant alters some dietary or lifestyle exposure, including the level of a biomarker, then

the direct association of the variant with cancer risk would strongly suggest that the biomarker–cancer relationship is not confounded by other factors, and that the primary link between the exposure of interest and cancer is causal [261]. Aside from establishing causal associations, MR provides estimates of the magnitude of effect between exposure and outcome [259]. MR could be used in the diet-biomarker-cancer relationship by including information on genetic variations upstream (for instance, with single-nucleotide polymorphisms). The current knowledge on how genetic variations influence dietary habits, nutrient metabolism or how they affect mechanism, bioavailability, adsorption or biotransformation of nutrients is progressively growing [262]. It is noteworthy to remember that MR, similarly to mediation analysis, also embraces a series of assumptions to account for in order to be implemented. Bias can arise when the genetic variant targets an exposure that is different from the one of interest [259]. In this case the instrumental variable is invalid, either because 1) the variant is not predictive of the exposure, 2) is also related to confounding factors of the exposure-outcome association or 3) is also indirectly related to the outcome, conditional to the exposure and confounders. The latter assumptions refer to pleiotropy (multiple effects of a single gene), which in essence requires that the genetic variant be strictly linked to the exposure of interest, and nothing else [260,263]. Current MR developments are striving to fill the methodological gap in order to obtain causal estimates and to evaluate MR performance when using invalid instruments [264]. New research is also joining efforts between mediation analysis and MR to focus into causal pathways, by investigating more complex networks of relationships between variables, through the integration of regression-based methods and structural equation models along with the use of genetic variants as instrumental variables [263]. In the context of MR this new development allows to estimate the direct and indirect effects even in the presence of unmeasured confounding. Both mediation and MR analyses tackle causality with different approaches but both are rigorous concepts limiting variables amongst them, and where a set of assumptions on the exposure, mediator, instrument and outcome are required for mediation effects to be interpreted as causal irrespective of the statistical models used [193,195].

Alternatively, pathway analysis has been suggested as a valuable way to investigate etiological mechanisms [197,265]. Pathway analysis employs what is referred to as mixed-method research to search for mechanisms, exploiting the principle

is that quantitative and qualitative studies have complementary strengths that can be used to explore underlying relationships between some explanatory variable and an outcome, controlling for other factors [265]. A critical aspect of pathway analysis is the need for an *a priori* knowledge of the expected relationship between the exposure and the outcome, the nature of the outcome, and the state of knowledge about causal pathways, which is often limited and uncertain. Another degree of complexity is that mechanisms in the context of pathway analysis are treated analogously to mediators or intermediate variables in standard mediation approaches [266–268], i.e. that the mechanism is caused by the exposure and causes the outcome [265]. The implementation of pathway analysis is not straightforward and many approaches are being developed to adequately apply it [269–271]. A number of metabolic pathway analysis tools which includes pathway enrichment analysis [272] can reveal underlying complex biological processes and connectivities, and are now used for metabolomics data [273,274].

Statistical innovations and new methodologies to analyse increasingly high-dimensional, biologically complex data will be key to pursue the investigation of the diet-disease relationship, a relation that evolves in time and crystallizes many already-established components, but that will inevitably pick up new contributing factors along the way.

REFERENCES

1. Doll, R. and Peto, R. (1981) The Causes of Cancer : Quantitative Estimates of Avoidable Risks of Cancer in the United States Today. *J. Natl. Cancer Inst.*, **66**,1191–1308.

2. Doll, R. (1992) The lessons of life: keynote address to the nutrition and cancer conference. *Cancer Res.*, **52**,2024s–2029s.

3. Willett, W.C. (2000) Diet and Cancer. *Oncologist*, **5**,393–404.

4. Willett, W.C. (2014) Webcast of diet and cancer: status report in 2014. Proceedings of the Annual Meeting of the American Association for Cancer Research. San Diego: AACR. p. April 5-9.

5. Miller, A.B. and Linseisen, J. (2010) Achievements and future of nutritional cancer epidemiology. *Int. J. Cancer*, **126**,1531–1537. doi:10.1002/ijc.25006.

6. Anand, P., Kunnumakkara, A.B., Kunnumakara, A.B., et al. (2008) Cancer is a preventable disease that requires major lifestyle changes. *Pharm. Res.*, **25**,2097–2116.

7. Stepien, M., Chajes, V., and Romieu, I. (2016) The role of diet in cancer : the epidemiologic link. *Salud Publica Mex.*, **58**,261–273.

8. Potter, J.D. (2015) Nutritional Epidemiology — There â€™s Life in the Old Dog Yet ! *Cancer Epidemiol. biomarkers Prev.*, **24**,323–330. doi:10.1158/1055-9965.EPI-14-1327.

9. Allen, N.E., Beral, V., Casabonne, D., Kan, S.W., Reeves, G.K., and Brown, A. (2009) Moderate Alcohol Intake and Cancer Incidence in Women. *J. Natl. Cancer Inst.*, **101**,296–305. doi:10.1093/jnci/djn514.

10. Baan, R., Straif, K., Grosse, Y., El Ghissassi, F., Bouvard, V., Altieri, A., Cogliano, V.J., and WHO International Agency for Research on Cancer Monograph Working Group. (2007) Carcinogenicity of alcoholic beverages. *Lancet Oncol.*, **8**,292–293.

11. Boyle, P. and Boffetta, P. (2009) Alcohol consumption and breast cancer risk. *Breast Cancer Res.*, **11**,S3. doi:10.1186/bcr2422.

12. Chen, W.Y., Rosner, B., Hankinson, S.E., Colditz, G.A., and Willett, W.C. (2011) Moderate alcohol consumption during adult life, drining patterns, and breast cancer. *JAMA*, **306**,1884–1890. doi:10.1001/jama.2011.1590.Moderate.

13. Fagherazzi, G., Vilier, A., Boutron-Ruault, M.-C., Mesrine, S., and Clavel-Chapelon, F. (2014) Alcohol consumption and breast cancer risk subtypes in the E3N-EPIC cohort. *Eur. J. Cancer Prev.*, **epub ahead**,1–6. Accessed 5 December 2014.

14. Imamura, F., Lichtenstein, A.H., Dallal, G.E., Meigs, J.B., and Jacques, P.F. (2009) Confounding by dietary patterns of the inverse association between alcohol consumption and type 2 diabetes risk. *Am. J. Epidemiol.*, **170**,37–45. doi:10.1093/aje/kwp096.

15. Romieu, I., Scoccianti, C., Chajès, V., et al. (2015) Alcohol intake and breast cancer in the European prospective investigation into cancer and nutrition. *Int. J. Cancer*, **137**,1921–1930. doi:10.1002/ijc.29469.

16. Schütze, M., Boeing, H., Pischon, T., et al. (2011) Alcohol attributable burden of incidence of cancer in eight European countries based on results from prospective cohort study.

BMJ, **342**,d1584.

17. Seitz, H.K., Pelucchi, C., Bagnardi, V., and Vecchia, C. La. (2012) Epidemiology and Pathophysiology of Alcohol and Breast Cancer : Update 2012. *Alcohol Alcohol.*, **47**,204–212. doi:10.1093/alcalc/ags011.

18. Sieri, S., Agudo, A., Kesse, E., et al. (2002) Patterns of alcohol consumption in 10 European countries participating in the European Prospective Investigation into Cancer and Nutrition (EPIC) project. *Public Health Nutr.*, **5**,1287–1296. doi:10.1079/PHN2002405.

19. Smith-Warner, S.A., Spiegelman, D., Yaun, S.-S., et al. (1998) Alcohol and Breast Cancer in Women A Pooled Analysis of Cohort Studies. *JAMA*, **279**,535–540.

20. Tjønneland, A., Christensen, J., Olsen, A., et al. (2007) Alcohol intake and breast cancer risk: the European Prospective Investigation into Cancer and Nutrition (EPIC). *Cancer Causes Control*, **18**,361–373. doi:10.1007/s10552-006-0112-9.

21. Trichopoulos, D., Bamia, C., Lagiou, P., et al. (2011) Hepatocellular carcinoma risk factors and disease burden in a European cohort: a nested case-control study. *J. Natl. Cancer Inst.*, **103**,1686–1695.

22. Zhang, S.M., Lee, I.-M., Manson, J.E., Cook, N.R., Willett, W.C., and Buring, J.E. (2007) Alcohol consumption and breast cancer risk in the Women's Health Study. *Am. J. Epidemiol.*, **165**,667–676. doi:10.1093/aje/kwk054.

23. Bagnardi, V., Rota, M., Botteri, E., et al. (2013) Light alcohol drinking and cancer : a meta-analysis. *Ann. Oncol.*, **24**,301–308. doi:10.1093/annonc/mds337.

24. Amadou, A., Hainaut, P., and Romieu, I. (2013) Role of Obesity in the Risk of Breast Cancer : Lessons from Anthropometry. *J. Oncol.*, **2013**.

25. McCullough, L.E., Eng, S.M., Bradshaw, P.T., Cleveland, R.J., Teitelbaum, S.L., Neugut, A.I., and Gammon, M.D. (2012) Fat or fit: the joint effects of physical activity, weight gain, and body size on breast cancer risk. *Cancer*, **118**,4860–4868. doi:10.1002/cncr.27433.

26. Wang, Y.C., Mcpherson, K., Marsh, T., Gortmaker, S.L., and Brown, M. (2011) Health and economic burden of the projected obesity trends in the USA and the UK. *Lancet*, **378**,815–825. doi:10.1016/S0140-6736(11)60814-3.

27. Swinburn, B.A., Sacks, G., Hall, K.D., Mcpherson, K., Finegood, D.T., Moodie, M.L., and Gortmaker, S.L. (2011) The global obesity pandemic : shaped by global drivers and local environments. *Lancet*, **378**,804–814. doi:10.1016/S0140-6736(11)60813-1.

28. Eliassen, A.H., Colditz, G.A., Rosner, B., Willett, W.C., and Hankinson, S.E. (2006) Adult Weight Change and Risk of Postmenopausal Breast Cancer. *JAMA*, **296**,193–201.

29. Willett, W.C., Stampfer, M.J., Colditz, G.A., Rosner, B.A., and Speizer, F.E. (1990) Relation of Meat, Fat, and Fiber Intake to the Risk of Colon Cancer in a Prospective Study among Women. *N. Engl. J. Med.*, **323**,1664–1672.

30. Giovannucci, E., Rimm, E.B., Stampfer, M.J., and Colditz, A. (1994) Intake of Fat , Meat , and Fiber in Relation to Risk of Colon Cancer in Men Intake of Fat , Meat , and Fiber in Relation to Risk of Colon Cancer in Men ',2390–2397.

31. Kuriki, K., Hirose, K., Wakai, K., et al. (2007) Breast cancer risk and erythrocyte compositions of n-3 highly unsaturated fatty acids in Japanese. *Int. J. Cancer*, **121**,377–385. doi:10.1002/ijc.22682.

32. Terry, P.D., Rohan, T.E., and Wolk, A. (2003) Intakes of fish and marine fatty acids and the risks of cancers of the breast and prostate and of other hormone-related cancers : a review of the epidemiologic evidence. *Am. J. Clin. Nutr.*, **77**,532–543.

33. Mozaffarian, D., Aro, A., and Willett, W.C. (2009) Health effects of trans-fatty acids: experimental and observational evidence. *Eur. J. Clin. Nutr.*, **63 Suppl 2**,S5-21. doi:10.1038/sj.ejcn.1602973.

34. Vinikoor, L.C., Schroeder, J.C., Millikan, R.C., Satia, J. a., Martin, C.F., Ibrahim, J., Galanko, J. a., and Sandler, R.S. (2008) Consumption of trans-fatty acid and its association with colorectal adenomas. *Am. J. Epidemiol.*, **168**,289–297. doi:10.1093/aje/kwn134.

35. Chajès, V., Thiébaut, A.C.M., Rotival, M., et al. (2008) Association between serum trans-monounsaturated fatty acids and breast cancer risk in the E3N-EPIC Study. *Am. J. Epidemiol.*, **167**,1312–1320. doi:10.1093/aje/kwn069.

36. Voorrips, L.E., Brants, H. a M., Kardinaal, A.F.M., Hiddink, G.J., van den Brandt, P. a., and Goldbohm, R.A. (2002) Intake of conjugated linoleic acid, fat, and other fatty acids in relation to postmenopausal breast cancer: the Netherlands Cohort Study on Diet and Cancer. *Am. J. Clin. Nutr.*, **76**,873–882.

37. Willett, W.C. (1997) Specific fatty acids and risks of breast cancer : dietary intake. *Am. J. Clin. Nutr.*, **66**,1557S–63S.

38. Richman, E.L., Kenfield, S.A., Chavarro, J.E., Stampfer, M.J., Giovannucci, E.L., Willett, W.C., and Chan, J.M. (2014) Fat intake after diagnosis and risk of lethal prostate cancer and all-cause mortality. *JAMA Intern. Med.*, **173**,1318–1326. doi:10.1001/jamainternmed.2013.6536.Fat.

39. Cho, E., Spiegelman, D., Hunter, D.J., Chen, W.Y., Stampfer, M.J., Colditz, G.A., and Willett, W.C. (2003) Premenopausal Fat Intake and Risk of Breast Cancer. *J. Natl. Cancer Inst.*, **95**,1079–1085.

40. Le Marchand, L., Hankin, J.H., Pierce, L.M., et al. (2002) Well-done red meat, metabolic phenotypes and colorectal cancer in Hawaii. *Mutat. Res.*, **506–507**,205–214.

41. Norat, T., Lukanova, A., Ferrari, P., and Riboli, E. (2002) Meat Consumption And Colrectal Cancer Risk : Dose-Response Meta-Analysis Of Epidemiological Studies. *Int. J. Cancer*, **256**,241–256. doi:10.1002/ijc.10126.

42. Norat, T., Bingham, S., Ferrari, P., et al. (2005) Meat, fish, and colorectal cancer risk: the European Prospective Investigation into cancer and nutrition. *J. Natl. Cancer Inst.*, **97**,906–916. doi:10.1093/jnci/dji164.

43. English, D.R., Macinnis, R.J., Hodge, A.M., Hopper, J.L., Haydon, A.M., and Giles, G.G. (2004) Red Meat , Chicken , and Fish Consumption and Risk of Colorectal Cancer Red Meat , Chicken , and Fish Consumption and Risk of Colorectal Cancer. *Cancer Epidemiol. biomarkers Prev.*, **13**,1509–1514.

44. Sinha, R., Chow, W.H., Kulldorff, M., Denobile, J., Butler, J., Weil, R., Hoover, R.N., and Rothman, N. (1999) Well-done , Grilled Red Meat Increases the Risk of Colorectal Adenomas Well-done , Grilled Red Meat Increases the Risk of Colorectal Adenomas. *Am. Assoc. Cancer Res.*, **59**,4320–4324.

45. Fedirko, V., Trichopolou, A., Bamia, C., et al. (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.

46. Linos, E., Willett, W.C., Cho, E., Colditz, G., and Frazier, L.A. (2008) Red Meat Consumption during Adolescence among Premenopausal Women and Risk of Breast Cancer. *Cancer Epidemiol. biomarkers Prev.*, **17**,2146–2151. doi:10.1158/1055-9965.EPI-08-0037.Red.

47. Choi, Y., Song, S., Song, Y., and Lee, J.E. (2013) Consumption of red and processed meat and esophageal cancer risk : Meta-analysis. *World J. Gastroenterol.*, **19**,1020–1029. doi:10.3748/wjg.v19.i7.1020.

48. Alexander, D.D., Weed, D.L., Miller, P.E., Mohamed, M.A., Colorado, D.D.A., Arbor, A., Michigan, D.D.A., and Consulting, D.L.W. (2015) Red Meat and Colorectal Cancer : A Quantitative Update on the State of the Epidemiologic Science. *J. Am. Coll. Nutr.*, **34**,521–543.

49. Jung, S., Spiegelman, D., Baglietto, L., et al. (2013) Fruit and vegetable intake and risk of breast cancer by hormone receptor status. *J. Natl. Cancer Inst.*, **105**,219–236. doi:10.1093/jnci/djs635.

50. Boggs, D.A., Palmer, J.R., Wise, L.A., Spiegelman, D., Stampfer, M.J., and Rosenberg, L. (2010) Original Contribution Fruit and Vegetable Intake in Relation to Risk of Breast Cancer in the Black Women's Health Study. *Am. J. Epidemiol.*, **172**,1268–1279. doi:10.1093/aje/kwq293.

51. Fung, T.T., Chiuve, S.E., Willett, W.C., Hankinson, S.E., Hu, F.B., and Holmes, M.D. (2013) Intake of specific fruits and vegetables in relation to risk of estrogen receptor-negative breast cancer among postmenopausal women. *Breast Cancer Res. Treat.*, **138**,925–930. doi:10.1007/s10549-013-2484-3.Intake.

52. Liu, J., Wang, J., Leng, Y., and Lv, C. (2012) Intake of fruit and vegetables and risk of esophageal squamous cell carcinoma : A meta-analysis of observational studies. *Int. J. Cancer*, **133**,473–486. doi:10.1002/ijc.28024.

53. Mulholland, H.G., Murray, L.J., Cardwell, C.R., and Cantwell, M.M. (2009) Glycemic index, glycemic load, and risk of digestive tract neoplasms : a systematic review and meta-analysis. *Am. J. Clin. Nutr.*, **89**,568–576. doi:10.3945/ajcn.2008.26823.1.

54. Aune, D., Chan, D.S.M., Lau, R., Vieira, R., Greenwood, D.C., Kampman, E., and Norat, T. (2012) Carbohydrates, glycemic index, glycemic load, and colorectal cancer risk: a systematic review and meta-analysis of cohort studies. *Cancer Causes Control*, **23**,521–535. doi:10.1007/s10552-012-9918-9.

55. Sang, L., Chang, B., Li, X., and Jiang, M. (2013) Consumption of coffee associated with reduced risk of liver cancer : a meta-analysis. *BMC Gastroenterol.*, **13**.

56. Bravi, F., Scotti, L., Bosetti, C., Gallus, S., Negri, E., Vecchia, C. La., and Tavani, A. (2009) Coffee drinking and endometrial cancer risk : a meta-analysis of observational studies. *Am. J. Obstet. Gyncology*, **200**,130–135. doi:10.1016/j.ajog.2008.10.032.

57. Je, Y. and Giovannucci, E. (2012) Coffee consumption and risk of endometrial cancer: findings from a large up-to-date meta-analysis. *Int. J. Cancer*, **131**,1700–1710. doi:10.1002/ijc.27408.

58. Chow, W., Blot, W.J., Vaughan, T.L., et al. (1998) Body Mass Index and Risk of Adenocarcinomas of the Esophagus and Gastric Cardia. *J. Natl. Can*, **90**,150–155.

59. Samanic, C., Gridley, G., Chow, W., Lubin, J., Hoover, R.N., and Fraumeni, J.F. (2004) Obesity and cancer risk among white and black United States veterans. *Cancer Causes Control*, **15**,35–43.

60. Lee, E., Levine, E.A., Franco, V.I., Allen, G.O., Gong, F., Zhang, Y., and Hu, J.J. (2014) Combined Genetic and Nutritional Risk Models of Triple Negative Breast Cancer. *Nutr. Cancer*, **66**,955–963. doi:10.1080/01635581.2014.932397.

61. Potter, J.D., Cerhan, J.R., Sellers, T.A., McGovern, P.G., Drinkard, C., Kushi, L.R., and Folsom, A.R. (1995) Progesterone and Estrogen Receptors in the Iowa Women's Health Kinds of Breast Cancer and Mammary Neoplasia Study : How Many Are There ? *Cancer Epidemiol. biomarkers Prev.*, **4**,319–326.

62. Giovannucci, E.L., Rimm, E.B., Stampfer, M.J., Colditz, G.A., Ascheiro, A., and Willett, W.C. (1994) Intake of Fat , Meat , and Fiber in Relation to Risk of Colon Cancer in Men '. *Cancer Res.*, **54**,2390–2397.

63. Willett, W. (1987) Nutritional Epidemiology : Issues and Challenges. *Int. J. Epidemiol.*, **16**,312–317.

64. Arija, V., Abellana, R., Ribot, B., and Ramón, J.M. (2015) Biases and adjustments in nutritional assessments from dietary questionnaires. *Nutr. Hosp.*, **31**,113–118. doi:10.3305/nh.2015.31.sup3.8759.

65. Hatfield, D.L. and Gladyshev, V.N. (2009) The Outcome of Selenium and Vitamin E Cancer Prevention Trial (SELECT) reveals the need for better understanding of selenium biology. *Mol. Interv.*, **9**,18–21.

66. National Research Council - Committee on Diet and Health. (1989) Diet and health: implications for reducing chronic disease risk. Washington, DC: National Academy Press.

67. van Dam, R.M. (2005) New approaches to the study of dietary patterns. *Br. J. Cancer*, **93**,573–574. doi:10.1079/BJN20051453.

68. Jacobs, D.R. and Steffen, L.M. (2003) Nutrients, foods, and dietary patterns as exposures in research: a framework for food synergy. *Am. J. Clin. Nutr.*, **78**,508S–513S.

69. Ferrari, P., Roddam, A., Fahey, M.T., et al. (2009) A bivariate measurement error model for nitrogen and potassium intakes to evaluate the performance of regression calibration in the European Prospective Investigation into Cancer and Nutrition study. *Eur. J. Clin. Nutr.*, **63**,179–187. doi:10.1038/ejcn.2009.80.

70. Willett, W.C., Howe, G.R., and Kushi, L.H. (1997) Adjustment for total energy intake in epidemiologic studies. *Am. J. Clin. Nutr.*, **65**,1220S–1228S; discussion 1229S–1231S.

71. Hu, F.B. (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr. Opin. Lipidol.*, **13**,3–9.

72. Edefonti, V., Randi, G., La Vecchia, C., Ferraroni, M., and Decarli, A. (2009) Dietary patterns and breast cancer: a review with focus on methodological issues. *Nutr. Rev.*, **67**,297–314. doi:10.1111/j.1753-4887.2009.00203.x.

73. Jacques, P.F. and Tucker, K.L. (2001) Are dietary patterns useful for understanding the role of diet in chronic diseases? *Am. J. Clin. Nutr.*, **73**,1–2.

74. Bamia, C., Orfanos, P., Ferrari, P., et al. (2005) Dietary patterns among older Europeans : the EPIC-Elderly study. *Br. J. Nutr.*, **94**,100–113. doi:10.1079/BJN20051456.

75. Bamia, C., Trichopoulos, D., Ferrari, P., et al. (2007) Dietary patterns and survival of older Europeans : The EPIC-Elderly Study (European Prospective Investigation into Cancer and Nutrition). *Public Health Nutr.*, **10**,590–598. doi:10.1017/S1368980007382487.

76. Romaguera, D., Vergnaud, A., Peeters, P.H., et al. (2012) Is concordance with World Cancer Research Fund / American Institute for Cancer Research guidelines for cancer prevention related to subsequent risk of cancer ? Results from the EPIC study. *Am. J. Clin. Nutr.*, **96**,150–163. doi:10.3945/ajcn.111.031674.

77. Guenther, P., Casavale, K., Reedy, J., Kirkpatrick, S., Hiza, H., Kuczynski, K., Kahle, L., and Krebs-Smith, S. (2014) Update of the Healthy Eating Index: HEI-2010. *J. Acad. Nutr. Diet.*, **113**,569–580. doi:10.1016/j.jand.2012.12.016.Update.

78. Buckland, G., Travier, N., Cottet, V., et al. (2013) Adherence to the mediterranean diet and risk of breast cancer in the European prospective investigation into cancer and nutrition cohort study. *Int. J. Cancer*, **132**,2918–2927. doi:10.1002/ijc.27958.

79. Agnoli, C., Grioni, S., Sieri, S., et al. (2013) Italian Mediterranean Index and risk of colorectal cancer in the Italian section of the EPIC cohort. *Int. J. Cancer*, **132**,1404–1411. doi:10.1002/ijc.27740.

80. Couto, E., Boffetta, P., Lagiou, P., et al. (2011) Mediterranean dietary pattern and cancer risk in the EPIC cohort. *Br. J. Cancer*, **104**,1493–1499. Accessed 2 April 2013.

81. Couto, E., Sandin, S., Lo, M., Ursin, G., and Adami, H. (2013) Mediterranean Dietary Pattern and Risk of Breast Cancer. *PLoS One*, **8**,e55374. doi:10.1371/Citation.

82. Turati, F., Trichopoulos, D., Polesel, J., et al. (2014) Mediterranean diet and hepatocellular carcinoma. *J. Hepatol.*, **60**,606–611.

83. de Lorgeril, M., Salen, P., Martin, J.-L., Monjaud, I., Boucher, P., and Mamelle, N. (1998) Mediterranean Dietary Pattern in a Randomized Trial: prolonged survival and possible reduced cancer rate. *Arch. Intern. Med.*, **158**,1181–1187.

84. Trichopoulou, A., Bamia, C., Lagiou, P., and Trichopoulos, D. (2010) Conformity to traditional Mediterranean diet and breast cancer risk in the Greek EPIC (European Prospective Investigation into Cancer and Nutrition) cohort. *Am. J. Clin. Nutr.*, **92**,620–625.

85. McKenzie, F., Ferrari, P., Freisling, H., et al. (2015) Healthy lifestyle and risk of breast cancer among postmenopausal women in the European Prospective Investigation into Cancer and Nutrition cohort study. *Int. J. cancer*, **136**,2640–2648. Accessed 28 July 2016.

86. McKenzie, F., Biessy, C., Ferrari, P., et al. (2016) Healthy Lifestyle and Risk of Cancer in the European Prospective Investigation Into Cancer and Nutrition Cohort Study. *Medicine (Baltimore).*, **95**,e2850. Accessed 28 July 2016.

87. Schulze, M.B., Hoffmann, K., Kroke, A., and Boeing, H. (2007) Dietary patterns and their association with food and nutrient intake in the European Prospective Investigation into Cancer and Nutrition (EPIC)–Potsdam study. *Br. J. Nutr.*, **85**,363. Accessed 1 March 2013.

88. Männistö, S., Dixon, L.B., Balder, H.F., et al. (2005) Dietary patterns and breast cancer risk: results from three cohort studies in the DIETSCAN project. *Cancer Causes Control*, **16**,725–733. doi:10.1007/s10552-005-1763-7.

89. De Stefani, E., Ronco, A.L., Boffetta, P., Deneo-Pellegrini, H., Correa, P., Acosta, G., and Mendilaharsu, M. (2012) Nutrient-derived dietary patterns and risk of colorectal cancer: a factor analysis in Uruguay. *Asian Pac. J. Cancer Prev.*, **13**,231–235.

90. Fung, T.T., Hu, F.B., Holmes, M.D., Rosner, B.A., Hunter, D.J., Colditz, G.A., and Willett, W.C. (2005) Dietary patterns and the risk of postmenopausal breast cancer. *Int. J. Cancer*,

116,116–121. doi:10.1002/ijc.20999.

91. Buck, K., Vrieling, A., Flesch-Janys, D., and Chang-Claude, J. (2011) Dietary patterns and the risk of postmenopausal breast cancer in a German case-control study. *Cancer Causes Control*, **22**,273–282. doi:10.1007/s10552-010-9695-2.

92. Palli, D., Russo, A., and Decarli, A. (2001) Dietary patterns, nutrient intake and gastric cancer in a high-risk area of Italy. *Cancer Causes Control*, **12**,163–172.

93. Turati, F., Edefonti, V., Bravi, F., et al. (2011) Nutrient-based dietary patterns, family history, and colorectal cancer. *Eur. J. Cancer Prev.*, **20**,456–461. Accessed 4 December 2013.

94. Kesse, E., Clavel-Chapelon, F., and Boutron-Ruault, M.C. (2006) Dietary patterns and risk of colorectal tumors: a cohort of French women of the National Education System (E3N). *Am. J. Epidemiol.*, **164**,1085–1093. doi:10.1093/aje/kwj324.

95. Bosetti, C., Bravi, F., Turati, F., et al. (2013) Nutrient-based dietary patterns and pancreatic cancer risk. *Ann. Epidemiol.*, **23**,124–128.

96. Satia, J. a., Tseng, M., Galanko, J. a., Martin, C., and Sandler, R.S. (2009) Dietary patterns and colon cancer risk in Whites and African Americans in the North Carolina Colon Cancer Study. *Nutr. Cancer*, **61**,179–193. doi:10.1080/01635580802419806.

97. Bravi, F., Edefonti, V., Bosetti, C., et al. (2010) Nutrient dietary patterns and the risk of colorectal cancer: a case-control study from Italy. *Cancer Causes Control*, **21**,1911–1918. doi:10.1007/s10552-010-9619-1.

98. Magalhães, B., Peleteiro, B., and Lunet, N. (2012) Dietary patterns and colorectal cancer: systematic review and meta-analysis. *Eur. J. Cancer Prev.*, **21**,15–23. doi:10.1097/CEJ.0b013e3283472241.

99. Wirfält, E., Hedblad, B., Gullberg, B., Mattisson, I., Andrén, C., Rosander, U., Janzon, L., and Berglund, G. (2001) Food patterns and components of the metabolic syndrome in men and women: a cross-sectional study within the Malmö Diet and Cancer cohort. *Am. J. Epidemiol.*, **154**,1150–1159.

100. Chan, J.M., Gong, Z., Holly, E.A., and Bracci, P.M. (2013) Dietary patterns and risk of pancreatic cancer in a large population-based case-control study in the san francisco bay area. *Nutr. Cancer*, **65**,157–164. doi:10.1080/01635581.2012.725502.

101. Hirose, K., Matsuo, K., Iwata, H., and Tajima, K. (2007) Dietary patterns and the risk of breast cancer in Japanese women. *Cancer Sci.*, **98**,1431–1438. doi:10.1111/j.1349-7006.2007.00540.x.

102. Agurs-collins, T., Rosenberg, L., Makambi, K., Palmer, J.R., Adams-campbell, L., and Al, A.E.T. (2009) Dietary patterns and breast cancer risk in women participating in the Black Women ' s Health Study. *Am. J. Clin. Nutr.*, **90**,621–628. doi:10.3945/ajcn.2009.27666.1.

103. Zhu, Y., Wu, H., Wang, P.P., et al. (2013) Dietary patterns and colorectal cancer recurrence and survival: a cohort study. *BMJ Open*, **3**. doi:10.1136/bmjopen-2012-002270.

104. Baglietto, L., Krishnan, K., Severi, G., et al. (2011) Dietary patterns and risk of breast cancer. *Br. J. Cancer*, **104**,524–531. doi:10.1038/sj.bjc.6606044.

105. Deneo-Pellegrini, H., Boffetta, P., De Stefani, E., et al. (2013) Nutrient-based dietary patterns of head and neck squamous cell cancer: a factor analysis in Uruguay. *Cancer*

Causes Control, **24**,1167–1174. doi:10.1007/s10552-013-0196-y.

106. Bertuccio, P., Edefonti, V., Bravi, F., Ferraroni, M., Pelucchi, C., Negri, E., Decarli, A., and La Vecchia, C. (2009) Nutrient dietary patterns and gastric cancer risk in Italy. *Cancer Epidemiol. Biomarkers Prev.*, **18**,2882–2886. doi:10.1158/1055-9965.EPI-09-0782.

107. Edefonti, V., Hashibe, M., Ambrogi, F., et al. (2012) Nutrient-based dietary patterns and the risk of head and neck cancer: a pooled analysis in the International Head and Neck Cancer Epidemiology consortium. *Ann. Oncol.*, **23**,1869–1880. doi:10.1093/annonc/mdr548.

108. Terry, P., Suzuki, R., and Hu, F.B. (2001) A Prospective Study of Major Dietary Patterns and the Risk of Breast Cancer A Prospective Study of Major Dietary Patterns and the Risk of Breast Cancer 1. *Cancer, Epidemiol. Biomarkers Prev.*, **10**,1281–1285.

109. Williams, C.D., Satia, J. a., Adair, L.S., Stevens, J., Galanko, J., Keku, T.O., and Sandler, R.S. (2009) Dietary patterns, food groups, and rectal cancer risk in Whites and African-Americans. *Cancer Epidemiol. Biomarkers Prev.*, **18**,1552–1561. doi:10.1158/1055-9965.EPI-08-1146.

110. De Stefani, E., Boffetta, P., Fagundes, R.B., Deneo-Pellegrini, H., Ronco, A.L., Acosta, G., and Mendilaharsu, M. (n.d.) Nutrient patterns and risk of squamous cell carcinoma of the esophagus: a factor analysis in uruguay. *Anticancer Res.*, **28**,2499–2506.

111. Butler, L.M., Wang, R., Koh, W.-P., and Yu, M.C. (2008) Prospective study of dietary patterns and colorectal cancer among Singapore Chinese. *Br. J. Cancer*, **99**,1511–1516. doi:10.1038/sj.bjc.6604678.

112. Nkondjock, A., Krewski, D., Johnson, K.C., and Ghadirian, P. (2005) Dietary patterns and risk of pancreatic cancer. *Int. J. Cancer*, **114**,817–823. doi:10.1002/ijc.20800.

113. Albuquerque, R.C.R., Baltar, V.T., and Marchioni, D.M.L. (2014) Breast cancer and dietary patterns: a systematic review. *Nutr. Rev.*, **72**,1–17. doi:10.1111/nure.12083.

114. Sieri, S., Krogh, V., Pala, V., Muti, P., Micheli, A., Evangelista, A., Tagliabue, G., and Berrino, F. (2004) Dietary Patterns and Risk of Breast Cancer in the ORDET Cohort. *Cancer Epidemiol. biomarkers Prev.*, **13**,567–572.

115. Edefonti, V., Bravi, F., Garavello, W., et al. (2010) Nutrient-based dietary patterns and laryngeal cancer: evidence from an exploratory factor analysis. *Cancer Epidemiol. Biomarkers Prev.*, **19**,18–27. doi:10.1158/1055-9965.EPI-09-0900.

116. Appel, L.J., Moore, T.J., Obarzanek, E., et al. (1997) A Clinical Trial of the Effects of Dietary Patterns on Blood Pressure. *N. Engl. J. Med.*, **336**,1117–1124.

117. Terry, P., Hu, F.B., Hansen, H., and Wolk, A. (2001) Prospective study of major dietary patterns and colorectal cancer risk in women. *Am. J. Epidemiol.*, **154**,1143–1149.

118. De Stefani, E., Deneo-Pellegrini, H., Boffetta, P., et al. (2009) Dietary patterns and risk of cancer: a factor analysis in Uruguay. *Int. J. Cancer*, **124**,1391–1397. doi:10.1002/ijc.24035.

119. Turati, F., Edefonti, V., Bravi, F., et al. (2011) Nutrient-based dietary patterns, family history, and colorectal cancer. *Eur. J. Cancer Prev.*, **20**,456–461. doi:10.1097/CEJ.0b013e328348fc0f.

120. Edefonti, V., Decarli, A., La Vecchia, C., Bosetti, C., Randi, G., Franceschi, S., Dal Maso, L., and

Ferraroni, M. (2008) Nutrient dietary patterns and the risk of breast and ovarian cancers. *Int. J. Cancer*, **122**,609–613. doi:10.1002/ijc.23064.

121. Wu, A.H., Yu, M.C., Tseng, C., Stanczyk, F.Z., and Pike, M.C. (2009) Dietary patterns and breast cancer risk in Asian American women. *Am. J. Clin. Nutr.*, **89**,1145–1154. doi:10.3945/ajcn.2008.26915.1.

122. Adebamowo, C.A., Hu, F.B., Cho, E., Spiegelman, D., Holmes, M.D., and Willett, W.C. (2005) Dietary patterns and the risk of breast cancer. *Ann. Epidemiol.*, **15**,789–795. doi:10.1016/j.annepidem.2005.01.008.

123. Vrieling, A., Buck, K., Seibold, P., Heinz, J., Obi, N., Flesch-Janys, D., and Chang-Claude, J. (2013) Dietary patterns and survival in German postmenopausal breast cancer survivors. *Br. J. Cancer*, **108**,188–192. doi:10.1038/bjc.2012.521.

124. Flood, A., Rastogi, T., Wirfält, E., et al. (2008) Dietary patterns as identified by factor analysis and colorectal cancer among middle-aged Americans. *Am. J. Clin. Nutr.*, **88**,176–184.

125. Millen, B.E., Quatromoni, P.A., Gagnon, D.R., Cupples, L.A., Franz, M.M., and D'Agostino, R.B. (1996) Dietary Patterns of men and women suggest targets for health promotion: the Farmingham Nutrition Studies. *Am. J. Heal. Promot.*, **11**,42–52.

126. Cottet, V., Touvier, M., Fournier, A., Touillaud, M.S., Lafay, L., Clavel-Chapelon, F., and Boutron-Ruault, M.-C. (2009) Postmenopausal breast cancer risk and dietary patterns in the E3N-EPIC prospective cohort study. *Am. J. Epidemiol.*, **170**,1257–1267. doi:10.1093/aje/kwp257.

127. Nkondjock, A. and Ghadirian, P. (2005) Associated nutritional risk of breast and colon cancers: a population-based case-control study in Montreal, Canada. *Cancer Lett.*, **223**,85–91. doi:10.1016/j.canlet.2004.11.034.

128. Murtaugh, M.A., Sweeney, C., Giuliano, A.R., Herrick, J.S., Hines, L., Byers, T., Baumgartner, K.B., and Slattery, M.L. (2008) Diet patterns and breast cancer risk in Hispanic and non-Hispanic white women: the Four-Corners Breast Cancer Study. *Am. J. Clin. Nutr.*, **87**,978–984.

129. Gorst-Rasmussen, A., Dahm, C.C., Dethlefsen, C., Scheike, T., and Overvad, K. (2011) Exploring dietary patterns by using the treelet transform. *Am. J. Epidemiol.*, **173**,1097–1104. doi:10.1093/aje/kwr060.

130. Hoffmann, K. (2004) Application of a New Statistical Method to Derive Dietary Patterns in Nutritional Epidemiology. *Am. J. Epidemiol.*, **159**,935–944. doi:10.1093/aje/kwh134.

131. Martinez, M.E., Marshall, J.R., and Sechrest, L. (1998) Factor Analysis and the Search for Objectivity. *Am. J. Epidemiol.*, **148**,17–19.

132. Rothman, K.J. and Greenland, S. (1998) Modern Epidemiology. Second Edi. Philadelphia, PA, USA: Lippincott-Raven publishers. 738 p.

133. Day, N. and Ferrari, P. (2002) Some methodological issues in nutritional epidemiology. *IARC Sci. Publ.*, **156**,5–10.

134. Boffetta, P. (2016) Reflections on nutritional cancer epidemiology. *Am. J. Clin. Nutr.*, **103**,3–4. doi:10.3945/ajcn.115.126508.Am.

135. Willett, W.C. (2013) Nutritional Epidemiology. Third Edit. New York USA: Oxford

University Press Inc. 529 p.

136. Belanger, C.F., Hennekens, C.H., Rosner, B., and Speizer, F.E. (1978) The nurses' health study. *Am. J. Nurs.*, **78**,1039–1040.

137. 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.

138. Riboli, E., Hunt, K.J., Slimani, N., et al. (2002) European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.*, **5**,1113–1124. doi:10.1079/PHN2002394.

139. Jenab, M., Slimani, N., Bictash, M., Ferrari, P., and Bingham, S.A. (2009) Biomarkers in nutritional epidemiology : applications , needs and new horizons. *Hum. Genet.*, **125**,507–525. doi:10.1007/s00439-009-0662-5.

140. Schatzkin, A. and Kipnis, V. (2004) Could exposure assessment problems give us wrong answers to nutrition and cancer questions? *J. Natl. Cancer Inst.*, **96**,1564–1565. doi:10.1093/jnci/djh329.

141. Freedman, L.S., Hartman, A.M., Kipnis, V., and Brown, C. (1997) Comments on: Adjustment for total energyintakein epidemiologic. *Am. J. Clin. Nutr.*, **65**,1229–1231.

142. Kipnis, V., Subar, A.F., Midthune, D., et al. (2003) Structure of Dietary Measurement Error : Results of the OPEN Biomarker Study. *Am. J. Epidemiol.*, **158**,14–21. doi:10.1093/aje/kwg091.

143. Kristal, A.R., Peters, U., and Potter, J.D. (2005) Is It Time to Abandon the Food Frequency Questionnaire ? *Cancer Epidemiol. biomarkers Prev.*, **14**,2826–2829. doi:10.1158/1055-9965.EPI-editorial.

144. IARC Working Group on the Evaluation of Cancer Preventive Strategies. (2003) Fruit and Vegetables. *IARC Handbooks Cancer Prev.*, **8**.

145. World Cancer Research Fund / American Institute for Cancer Research. AICR. (2007) Food, Nutrition, Physical Activity and the Prevention of Cancer: a Global Perspective. AICR, editor Washington, DC.

146. World Cancer Research Fund / American Institute for Cancer Research. AICR. (1997) Food, Nutrition and the Prevention of Cancer: A Global Perspective. Washington, DC.

147. Millen, A.E., Midthune, D., Thompson, F.E., Kipnis, V., and Subar, A.F. (2006) The National Cancer Institute Diet History Questionnaire : Validation of Pyramid Food Servings. *Am. J. Epidemiol.*, **163**,279–288. doi:10.1093/aje/kwj031.

148. Tooze, J.A., Kipnis, V., Buckman, D.W., et al. (2010) A mixed-effects model approach for estimating the distribution of usual intake of nutrients: The NCI method. *Stat. Med.*, **29**,2857–2868. doi:10.1002/sim.4063.A.

149. Carroll, R.J., Midthune, D., Subar, A.F., Shumakovich, M., Freedman, L.S., Thompson, F.E., and Kipnis, V. (2012) Taking Advantage of the Strengths of 2 Different Dietary Assessment Instruments to Improve Intake Estimates for Nutritional Epidemiology. *Am. J. Epidemiol.*, **175**,340–347. doi:10.1093/aje/kwr317.

150. Subar, A.F., Midthune, D., Tasevska, N., Kipnis, V., and Freedman, L.S. (2013) Checking for completeness of 24-h urine collection using para-amino benzoic acid not necessary in the Observing Protein and Energy Nutrition study. *Eur. J. Clin. Nutr.*, **67**,863–867.

doi:10.1038/ejcn.2013.62.

151. Illner, A.-K., Freisling, H., Boeing, H., Huybrechts, I., Crispim, S.P., and Slimani, N. (2012) Review and evaluation of innovative technologies for measuring diet in nutritional epidemiology. *Int. J. Epidemiol.*, **41**,1187–1203. doi:10.1093/ije/dys105.

152. Prentice, R.L., Sugar, E., Wang, C.Y., Neuhouser, M., and Patterson, R. (2002) Research strategies and the use of nutrient biomarkers in studies of diet and chronic disease. *Public Health Nutr.*, **5**,977–984. doi:10.1079/PHN2002382.

153. Schutz, Y., Weinsier, R.L., and Hunter, G.R. (2001) Assessment of Free-Living Physical Activity in Humans : An Overview of Currently Available and Proposed New Measures. *Obes. Res.*, **9**,368–379.

154. Hankin, J.H., Rhoads, G.G., and Glober, G.A. (1975) A dietary method for study of gastrointestinal cancer. *Am. J. Clin. Nutr.*, **28**,1055–1061.

155. Lissner, L., Troiano, R.P., Midthune, D., Heitmann, B.L., Kipnis, V., Subar, A.F., and Potischman, N. (2007) OPEN about obesity : recovery biomarkers , dietary reporting errors and BMI. *Int. J. Obes.*, **31**,956–961. doi:10.1038/sj.ijo.0803527.

156. Schatzkin, A., Subar, A.F., Moore, S., et al. (2009) Observational Epidemiologic Studies of Nutrition and Cancer: The Next Generation (with Better Observation). *Cancer Epidemiol. biomarkers Prev.*, **18**,1026–1032. doi:10.1158/1055-9965.EPI-08-1129.Observational.

157. Zheng, C., Beresford, S.A., Horn, L. Van., et al. (2014) Simultaneous Association of Total Energy Consumption and Activity-Related Energy Expenditure With Risks of Cardiovascular Disease , Cancer , and Diabetes Among Postmenopausal Women. *Am. J. Epidemiol.*, **180**,526–535. doi:10.1093/aje/kwu152.

158. Prentice, R.L. (1996) Measurement Error and Results From Analytic Epidemiology : Dietary Fat and Breast Cancer. *J. Natl. Cancer Inst.*, **88**,1738–1747.

159. Freedman, L.S., Schatzkin, A., and Wax, Y. (1990) The Impact of Dietary Measurement Error on Planning Sample Size Required in a Cohort Study. *Am. J. Epidemiol.*, **132**,1185–1195.

160. Potischman, N. (2003) Biomarkers of Nutritional Exposure and Nutritional Status Biologic and Methodologic Issues for Nutritional Biomarkers. *J. Nutr.*, **133**,875S–880S.

161. Potischman, N. and Freudenheim, J.L. (2003) Biomarkers of Nutritional Exposure and Nutritional Status Biomarkers of Nutritional Exposure and Nutritional Status : An Overview. *J. Nutr.*, **133**,873–874.

162. 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.

163. Kaaks, R., Riboli, E., and Sinha, R. (1997) Biochemical markers of dietary intake. *IARC Sci. Publ.*,103–126.

164. Tasevska, N., Runswick, S.A., Mctaggart, A., and Bingham, S.A. (2005) Urinary Sucrose and Fructose as Biomarkers for Sugar Consumption. *Cancer Epidemiol. biomarkers Prev.*, **14**,1287–1295.

165. Bingham, S.A. (2002) Biomarkers in nutritional epidemiology. *Public Health Nutr.*, **5**,821–827. doi:10.1079/PHN2002368.

166. Bingham, S.A., Hughes, R., and Cross, A.J. (2002) Effect of White Versus Red Meat on Endogenous N-Nitrosation in the Human Colon and Further Evidence of a Dose Response. *J. Nutr.*, **132**,3522S–3525S.

167. Schatzkin, A., Kipnis, V., Carroll, R.J., et al. (2003) A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study : results from the biomarker-based Observing Protein and Energy Nutrition (OPEN) study. *Int. J. Epidemiol.*, **32**,1054–1062. doi:10.1093/ije/dyg264.

168. Bougnoux, P., Hajjaji, N., and Couet, C. (2008) The lipidome as a composite biomarker of the modifiable part of the risk of breast cancer. *Prostaglandins. Leukot. Essent. Fatty Acids*, **79**,93–96. doi:10.1016/j.plefa.2008.09.004.

169. Chajès, V., Joulin, V., and Clavel-Chapelon, F. (2011) The fatty acid desaturation index of blood lipids, as a biomarker of hepatic stearoyl-CoA desaturase expression, is a predictive factor of breast cancer risk. *Curr Opin Lipidol*, **22**,6–10.

170. van Duijnhoven, F.J.B., Bueno-De-Mesquita, H.B., Calligaro, M., et al. (2011) Blood lipid and lipoprotein concentrations and colorectal cancer risk in the European Prospective Investigation into Cancer and Nutrition. *Gut*, **60**,1094–1102. doi:10.1136/gut.2010.225011.

171. Cottet, V., Collin, M., Gross, A.-S., Boutron-Ruault, M.C., Morois, S., Clavel-Chapelon, F., and Chajès, V. (2013) Erythrocyte membrane phospholipid fatty acid concentrations and risk of colorectal adenomas : a case-control nested in the French E3N-EPIC cohort study. *Cancer Epidemiol. Biomarkers Prev.*,. doi:10.1158/1055-9965.EPI-13-0168.

172. Eliassen, A.H., Hendrickson, S.J., Brinton, L.A., et al. (2012) Circulating Carotenoids and Risk of Breast Cancer : Pooled Analysis of Eight Prospective Studies,1905–1916. doi:10.1093/jnci/djs461.

173. Baker, M. (2013) The 'Omes Puzzle. *Nature*, **494**,416–419.

174. Trifonova, O.P., Il'in, V.A., Kolker, E. V., and Lisitsa, A. V. (2013) Big Data in Biology and Medicine. *Acta Naturae*, **5**,13–16.

175. Psychogios, N., Hau, D.D., Peng, J., et al. (2011) The human serum metabolome. *PLoS One*, **6**,e16957.

176. Hollywood, K., Brison, D.R., and Goodacre, R. (2006) Metabolomics : Current technologies and future trends. *Proteomics*, **6**,4716–4723. doi:10.1002/pmic.200600106.

177. Ren, S., Hinzman, A.A., Kang, E.L., Szczesniak, R.D., and Lu, L.J. (2015) Computational and statistical analysis of metabolomics data. *Metabolomics*, **11**,1492–1513. doi:10.1007/s11306-015-0823-6.

178. 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.

179. 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.

180. Vineis, P., van Veldhoven, K., Chadeau-Hyam, M., and Athersuch, T.J. (2013) Advancing the application of omics-based biomarkers in environmental epidemiology. *Environ. Mol. Mutagen.*, **54**,461–467. Accessed 28 July 2016.

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

182. Bro, R., Kamstrup-Nielsen, M.H., Engelsen, S.B., et al. (2015) Forecasting individual breast cancer risk using plasma metabolomics and biocontours. *Metabolomics*, **11**,1376–1380.

183. 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.

184. Chadeau-Hyam, M., Athersuch, T.J., Keun, H.C., et al. (2011) Meeting-in-the-middle using metabolic profiling - a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers*, **16**,83–88.

185. Kühn, T., Floegel, A., Sookthai, D., et al. (2016) Higher plasma levels of lysophosphatidylcholine 18:0 are related to a lower risk of common cancers in a prospective metabolomics study. *BMC Med.*, **14**,13.

186. Stepien, M., Duarte-Salles, T., Fedirko, V., et al. (2015) Alteration of Amino Acid and Biogenic Amine Metabolism in Hepatobiliary Cancers: Findings from a Prospective Cohort Study. *Submitt. to Am. J. Gastroenterol.,*.

187. Drogan, D., Dunn, W.B., Lin, W., et al. (2015) Untargeted metabolic profiling identifies altered serum metabolites of type 2 diabetes mellitus in a prospective, nested case control study. *Clin. Chem.*, **61**,487–497. doi:10.1373/clinchem.2014.228965.

188. Floegel, A., Drogan, D., Wang-Sattler, R., et al. (2011) Reliability of serum metabolite concentrations over a 4-month period using a targeted metabolomic approach. *PLoS One*, **6**. doi:10.1371/journal.pone.0021103.

189. Floegel, A., Stefan, N., Yu, Z., et al. (2013) Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes*, **62**,639–648. doi:10.2337/db12-0495.

190. Bathe, O.F., Shaykhutdinov, R., Kopciuk, K., et al. (2011) Feasibility of identifying pancreatic cancer based on serum metabolomics. *Cancer Epidemiol. Biomarkers Prev.*, **20**,140–147.

191. Nguyen, Q.C., Osypuk, T.L., Schmidt, N.M., Glymour, M.M., and Tchetgen, E.J.T. (2015) Practical Guidance for Conducting Mediation Analysis With Multiple Mediators Using Inverse Odds Ratio Weighting. *Am. J. Epidemiol.*, **181**,349–356. doi:10.1093/aje/kwu278.

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

193. 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.

194. Vanderweele, T.J. and Vansteelandt, S. (2014) Mediation Analysis with Multiple Mediators. *Epidemiol. Method.*, **2**,95–115. doi:10.1515/em-2012-0010.Mediation.

195. Mackinnon, D.P., Fairchild, A.J., and Fritz, M.S. (2007) Mediation Analysis. *Annu. Rev. Psychol.*, **5**,593–614. doi:10.1146/annurev.psych.58.110405.085542.

196. Brown, D.G., Rao, S., Weir, T.L., Malia, J.O., Bazan, M., Brown, R.J., and Ryan, E.P. (2016) Metabolomics and metabolic pathway networks from human colorectal cancers , adjacent

mucosa , and stool. *Cancer Metab.*, **4**,11 eCollection. doi:10.1186/s40170-016-0151-y.

197. Bai, Y., Zhang, H., Sun, X., Sun, C., and Ren, L. (2014) Biomarker identification and pathway analysis by serum metabolomics of childhood acute lymphoblastic leukemia. *Clin. Chim. Acta*, **436**,207–216. doi:10.1016/j.cca.2014.05.022.

198. Riboli, E. and Kaaks, R. (1997) The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.*, **26 Suppl 1**,S6-14.

199. Riboli, E., Hunt, K.J., Slimani, N., et al. (2002) European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.*, **5**,1113–1124.

200. Ferrari, P., Day, N.E., Boshuizen, H.C., et al. (2008) The evaluation of the diet/disease relation in the EPIC study: considerations for the calibration and the disease models. *Int. J. Epidemiol.*, **37**,368–378. doi:10.1093/ije/dym242.

201. Freisling, H., Fahey, M.T., Moskal, A., et al. (2010) Region-Specific Nutrient Intake Patterns Exhibit a Geographical Gradient within and between European Countries. *J. Nutr.*, **140**,1280–1286. doi:10.3945/jn.110.121152.or.

202. Slimani, N., Kaaks, R., Ferrari, P., Casagrande, C., Lotze, G., and Mattisson, I. (2002) European Prospective Investigation into Cancer and Nutrition (EPIC) calibration study : rationale , design and population characteristics. *Public Health Nutr.*, **5**,1125–1145. doi:10.1079/PHN2002395.

203. Slimani, N., Deharveng, G., Unwin, I., 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. *Eur. J. Clin. Nutr.*, **61**,1037–1056. doi:10.1038/sj.ejcn.1602679.

204. Nicolas, G., Witthöft, C.M., Vignat, J., Knaze, V., Huybrechts, I., Roe, M., Finglas, P., and Slimani, N. (2016) Compilation of a standardised international folate database for EPIC. *Food Chem.*, **193**,134–140. doi:10.1016/j.foodchem.2014.11.044.

205. Slimani, N., Ferrari, P., Ocké, M., et al. (2000) Standardization of the 24-hour diet recall calibration method used in the european prospective investigation into cancer and nutrition (EPIC): general concepts and preliminary results. *Eur. J. Clin. Nutr.*, **54**,900–917.

206. Lee, A.B., Nadler, B., and Wasserman, L. (2008) Treelets—An adaptive multi-scale basis for sparse unordered data. *Ann. Appl. Stat.*, **2**,435–471. doi:10.1214/07-AOAS137.

207. Jolliffe, I.T. (2002) Principal Component Analysis, Second Edition. 2nd editio. Springer, editor 489 p.

208. Assi, N., Fages, A., Vineis, P., et al. (2015) A statistical framework to model the meeting-in-the-middle principle using metabolomic data : application to hepatocellular carcinoma in the EPIC study. *Mutagenesis*, **30**,743–753. doi:10.1093/mutage/gev045.

209. Ferlay, J., Soerjomataram, I., Ervik, M., et al. (2013) GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. *Lyon, Fr. Int. Agency Res. Cancer*,.

210. Willett, W.C. (2005) Diet and Cancer: An Evolving Picture. *JAMA*, **293**,233–234.

211. World Cancer Research Fund / American Institute for Cancer Research. (2010) Continuous Update Project Report . Food, Nutrition, Physical Activity, and the Prevention

of Breast Cancer.

212. Lee, C.N., Reed, D.M., MacLean, C.J., Yano, K., and Chiu, D. (1988) Dietary potassium and stroke. *N. Engl. J. Med.*, **318**,995–996. doi:10.1056/NEJM198804143181516.

213. Gorst-rasmussen, A. (2011) tt:Treelet Transform with Stata. *Stata J.*, **12**,130–146.

214. Dahm, C.C., Gorst-Rasmussen, A., Crowe, F.L., et al. (2012) Fatty acid patterns and risk of prostate cancer in a case-control study nested within the European Prospective Investigation into Cancer and Nutrition. *Am. J. Clin. Nutr.*, **96**,1354–1361. doi:10.3945/ajcn.112.034157.

215. Schoenaker, D.A.J.M., Dobson, A.J., Soedamah-Muthu, S.S., and Mishra, G.D. (2013) Factor Analysis Is More Appropriate to Identify Overall Dietary Patterns Associated with Diabetes When Compared with Treelet Transform Analysis. *J. Nutr.*,392–398. doi:10.3945/jn.112.169011.TABLE.

216. Wild, C.P. (2005) Complementing the Genome with an "' Exposome '": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiol. biomarkers Prev.*, **14**,1847–1851. doi:10.1158/1055-9965.EPI-05-0456.

217. Vineis, P. and Chadeau-Hyam, M. (2011) Integrating biomarkers into molecular epidemiological studies. *Curr. Opin. Oncol.*, **23**,100–105. doi:10.1097/CCO.0b013e3283412de0.

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

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

220. Chadeau-Hyam, M., Athersuch, T.J., Keun, H.C., et al. (2011) Meeting-in-the-middle using metabolic profiling - a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers*, **16**,83–88.

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

222. Bray, F., Ren, J.S., Masuyer, E., and Ferlay, J. (2013) Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int. J. Cancer*, **132**,1133–1145. doi:10.1002/ijc.27711.

223. Anothaisintawee, T., Wiratkapun, C., Lerdsitthichai, P., et al. (2013) Risk factors of breast cancer: a systematic review and meta-analysis. *Asia. Pac. J. Public Health*, **25**,368–387. doi:10.1177/1010539513488795.

224. Chlebowski, R. (2007) Lifestyle Change Including Dietary Fat Reduction and Breast Cancer Outcome. *J. Nutr.*, **137**,233S–235S.

225. Collaborative Group on Hormonal Factors in Breast Cancer. (2001) Familial breast cancer : collaborative reanalysis of individual data from 52 epidemiological studies including 58 209 women with breast cancer and 101 986 women without the disease. *Lancet*, **358**,1389–1399.

226. Cui, X., Dai, Q., Tseng, M., Shu, X.-O., Gao, Y.-T., and Zheng, W. (2007) Dietary patterns and breast cancer risk in the shanghai breast cancer study. *Cancer Epidemiol. Biomarkers Prev.*, **16**,1443–1448. doi:10.1158/1055-9965.EPI-07-0059.

227. Levi, F., Pasche, C., Lucchini, F., and La Vecchia, C. (2001) Dietary Intake of Selected Nutrients and Breast-Cancer Risk. *Int J Cancer*, **91**,260–263.

228. McKenzie, F., Ellison-Loschmann, L., Jeffreys, M., Firestone, R., Pearce, N., and Romieu, I. (2013) Cigarette smoking and risk of breast cancer in a New Zealand multi-ethnic case-control study. *PLoS One*, **8**,e63132. Accessed 5 December 2014.

229. Bougnoux, P., Giraudeau, B., and Couet, C. (2006) Diet, cancer, and the lipidome. *Cancer Epidemiol. Biomarkers Prev.*, **15**,416–421. doi:10.1158/1055-9965.EPI-05-0546.

230. Terry, P.D. and Goodman, M. (2006) Is the association between cigarette smoking and breast cancer modified by genotype? A review of epidemiologic studies and meta-analysis. *Cancer Epidemiol. Biomarkers Prev.*, **15**,602–611. Accessed 5 December 2014.

231. Willett, W. and Mozaffarian, D. (2008) Ruminant or industrial sources of trans fatty acids : public health issue or food label skirmish ? *Am. J. Clin. Nutr.*, **87**,515–516.

232. Molin, M., Odden, N., Ha, L., Henriksen, T.N., Frazier, K.S., Strand, M.F., and Westerberg, A.C. (2013) Trans fat can be a hidden health risk. *Tidsskr Nor Legeforen*, **133**,1844–1847.

233. Nishida, C. and Uauy, R. (2009) WHO Scientific Update on health consequences of trans fatty acids : introduction. *Eur. J. Clin.*, **63**,S1–S4. doi:10.1038/ejcn.2009.13.

234. Smith, B.K., Robinson, L.E., Nam, R., and Ma, D.W.L. (2009) Trans-fatty acids and cancer: a mini-review. *Br. J. Nutr.*, **102**,1254–1266. doi:10.1017/S0007114509991437.

235. Bouckaert, K.P., Slimani, N., Nicolas, G., Vignat, J., Wright, A.J.A., Roe, M., Witthöft, C.M., and Finglas, P.M. (2011) Critical evaluation of folate data in European and international databases: recommendations for standardization in international nutritional studies. *Mol. Nutr. Food Res.*, **55**,166–180. doi:10.1002/mnfr.201000391.

236. Margetts, B. and Pietinen, P. (1997) European Prospective Investigation into Cancer and Nutrition: validity studies on dietary assessment methods. *Int. J. Epidemiol.*, **26**,Suppl 1:S1-S5.

237. Potter, J.D. (2005) Vegetables, fruit, and cancer. *Lancet*, **366**,527–530.

238. Schatzkin, A., Lanza, E., Corle, D., et al. (2000) Lack of effect of a low-fat, high-fiber diet on the recurrence of colorectal adenomas. *N. Engl. J. Med.*, **342**,1149–1155.

239. Alberts, D.S., Martinez, M.E., Roe, D.J., et al. (2000) Lack of ffect of a high-fiber cereal supplement on the recurrence of colorectal adenomas. *N. Engl. J. Med.*, **342**,1156–1162.

240. Voorrips, L.E., Goldbohm, R.A., van Poppel, G., Sturmans, F., Hermus, R.J.J., and van den Brandt, P.A. (2000) Vegetable and Fruit Consumption and Risks of Colon and Rectal Cancer in a Prospective Cohort Study. *Am. J. Epidemiol.*, **152**,1081–1092.

241. Smith, A.J., Jobe, J.B., and Mingay, D.J. (1995) Retrieval from memory of dietary information. *Appl. Cogn. Psychol.*, **5**,269–296.

242. Tzoulaki, I., Ebbels, T.M.D., Valdes, A., Elliott, P., and Ioannidis, J.P.A. (2014) Design and Analysis of Metabolomics Studies in Epidemiologic Research : A Primer on -Omic Technologies. *Am. J. Epidemiol.*, **180**,129–139. doi:10.1093/aje/kwu143.

243. Holst, B. and Williamson, G. (2008) Nutrients and phytochemicals : from bioavailability to bioefficacy beyond antioxidants. *Curr. Opin. Biotechnol.*, **19**,73–82. doi:10.1016/j.copbio.2008.03.003.

244. Kaaks, R., Ferrari, P., Ciampi, A., Plummer, M., and Riboli, E. (2002) Uses and limitations of statistical accounting for random error correlations , in the validation of dietary questionnaire assessments. *Public Health Nutr.*, **5**,969–976. doi:10.1079/PHN2002380.

245. Bictash, M., Ebbels, T.M., Chan, Q., et al. (2010) Opening up the " Black Box ": Metabolic phenotyping and metabolome-wide association studies in epidemiology. *J. Clin. Epidemiol.*, **63**,970–979. doi:10.1016/j.jclinepi.2009.10.001.

246. Verma, M., Khoury, M., and Ioannidis, J.P. (2013) Opportunities and Challenges for Selected Emerging Technologies in Cancer Epidemiology: Mitochondrial, Epigenomic, Metabolomic, and Telomerase Profiling. *Cancer Epidemiol. biomarkers Prev.*, **22**,189–200. doi:10.1158/1055-9965.EPI-12-1263.Opportunities.

247. Nicholson, J.K., Holmes, E., Kinross, J.M., Darzi, A.W., Takats, Z., and Lindon, J.C. (2012) Metabolic phenotyping in clinical and surgical environments. *Nature*, **491**,384–392. doi:10.1038/nature11708.

248. 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.

249. Khoury, M.J., Lam, T.K., Ioannidis, J.P.A., et al. (2013) Transforming Epidemiology for 21st Century Medicine and Public Health. *Cancer Epidemiol. biomarkers Prev.*, **22**,508–517. doi:10.1158/1055-9965.EPI-13-0146.

250. Kohavi, R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proc. Fourteenth Int. Jt. Conf. Artif. Intell.*, **2**,1137–1143.

251. Ioannidis, J.P.A. and Khoury, M.J. (2011) Improving Validation Practices in "Omics" Research. *Science (80-.).*, **334**,1230–1232.

252. Castaldi, P.J., Dahabreh, I.J., John, P., and Ioannidis, A. (2011) An empirical assessment of validation practices for molecular classifiers. *Brienfings Bioinforma.*, **12**,189–202. doi:10.1093/bib/bbq073.

253. Ioannidis, J.P.A. (2010) Expectations, validity, and reality in omics. *J. Clin. Epidemiol.*, **63**,945–949. doi:10.1016/j.jclinepi.2010.04.002.

254. 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.

255. Daniel, R.M., Stavola, B.L. De., Cousens, S.N., and Vansteelandt, S. (2015) Causal Mediation Analysis with Multiple Mediators. *Biometrics*, **71**,1–14. doi:10.1111/biom.12248.

256. Vanderweele, T.J., Vansteelandt, S., and Robins, J.M. (2014) Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, **25**,300–306. doi:10.3851/IMP2701.Changes.

257. Vanderweele, T.J. (2016) Mediation Analysis With Matched Case-Control Study Designs. *Am. J. Epidemiol.*, **183**,869–870. doi:10.1093/aje/kww038.

258. Pearl, J. (2012) The Mediation Formula : A guide to the assessment of causal pathways in nonlinear models Mediation : Direct and Indirect Effects. In: Berzuini C, Dawid P, Bernardinelli L, editors. Causality: Statistical Perspectives and Applications. Chichester, UK: John Wiley and Sons, Ltd. pp. 151–179.

259. Taylor, A.E., Davies, N.M., Ware, J.J., Vanderweele, T., Davey, G., and Munafo, M.R. (2014) Mendelian randomization in health research : Using appropriate genetic variants and avoiding biased estimates. *Econ. Hum. Biol.*, **13**,99–106. doi:10.1016/j.ehb.2013.12.002.

260. Smith, G.D. and Ebrahim, S. (2004) Mendelian randomization : prospects , potentials , and limitations. *Int. J. Epidemiol.*, **33**,30–42. doi:10.1093/ije/dyh132.

261. Hunter, D.J. (2006) The Influence of Genetic Polymorphism. *J. Nutr.*, **136**,2711–2713.

262. Kaput, J. (2008) Nutrigenomics research for personalized nutrition and medicine. *Curr. Opin. Biotechnol.*, **19**,110–120. doi:10.1016/j.copbio.2008.02.005.

263. Burgess, S., Daniel, R.M., Butterworth, A.S., Thompson, S.G., and Consortium, E. (2015) Mendelian Randomization Methodology Network Mendelian randomization : using genetic variants as instrumental variables to investigate mediation in causal pathways,484–495. doi:10.1093/ije/dyu176.

264. Bowden, J., Smith, G.D., Haycock, P.C., and Burgess, S. (2016) Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.*, **40**,304–314. doi:10.1002/gepi.21965.

265. Weller, N. and Barnes, J. (2014) Pathway Analysis and the Search for Causal Mechanisms. *Sociol. Methods Res.*, **45**,424–457. doi:10.1177/0049124114544420.

266. Tingley, D., Yamamoto, T., Keele, L., and Imai, K. (2013) mediation: R Package for Causal Mediation Analysis.

267. Baron, R. and Kenny, D. (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Personnal. Soc. Psychol.*, **51**,1173–1182.

268. Gelfand, L.A. and Tenhave, T. (2009) Mediation Analysis: A Retrospective Snapshot of Practice and More Recent Directions. **136**,1–22.

269. Aslibekyan, S., Almeida, M., and Tintle, N. (2015) Pathway analysis approaches for rare and common variants: Insights from GAW18. *Genet. Epidemiol.*, **38**,S86–S91. doi:10.1002/gepi.21831.Pathway.

270. Zhou, Y.-H. (2016) Pathway Analysis for RNA-Seq Data Using a Score-Based Approach. *Biometrics*, **72**,165–174. doi:10.1111/biom.12372.

271. Huang, Y.-T., Hsu, T., and Christiani, D.C. (2014) TEGS-CN: A Statistical Method for Pathway Analysis of Genome-wide Copy Number Profile. *Cancer Inform.*, **13**,15–23. doi:10.4137/CIN.S13978.Received.

272. Kankainen, M., Gopalacharyulu, P., Holm, L., and Oreši, M. (2011) MPEA — metabolite pathway enrichment analysis. *Bioinfomatics*, **27**,1878–1879. doi:10.1093/bioinformatics/btr278.

273. Ma, X., Chi, Y., Niu, M., et al. (2016) Metabolomics Coupled with Multivariate Data and Pathway Analysis on Potential Biomarkers in Cholestasis and Intervention Effect of Paeonia lactiflora Pall. *Front. Pharmacol.*, **7**,14 eCollection. doi:10.3389/fphar.2016.00014.

274. Tianjiao, L., Shuai, W., Xiansheng, M., Yongrui, B., Shanshan, G., Bo, L., and Lu, C. (2014) Metabolomics Coupled with Multivariate Data and Pathway Analysis on Potential Biomarkers in Gastric Ulcer and Intervention Effects of Corydalis yanhusuo Alkaloid. *PLoS*

One, **9**,e82499. doi:10.1371/journal.pone.0082499.