

Journal of the National Cancer Institute

Novel biomarkers of habitual alcohol intake and associations with risk of pancreatic and liver cancers and liver disease mortality

--Manuscript Draft--

Manuscript Number:	JNCI-20-1776R1
Full Title:	Novel biomarkers of habitual alcohol intake and associations with risk of pancreatic and liver cancers and liver disease mortality
Article Type:	Article
Corresponding Author:	Pietro Ferrari, PhD International Agency for Research on Cancer Lyon, Fr FRANCE
Order of Authors Secondary Information:	
Keywords:	alcohol intake; untargeted metabolomics; 2-hydroxy-3-methylbutyric acid; biomarkers; EPIC; ATBC
Section/Category:	Epidemiology
Manuscript Classifications:	Cancer Site (AEs, Reviewers, and Authors please choose at least one); Liver Disease and Liver Cancer; Pancreatic Cancer; Study Design (Authors and Statistical Reviewers choose at least one); Preclinical/Experimental Therapeutics; Biomarker(s)
Author Comments:	Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer /World Health Organization.
Additional Information:	
Question	Response
Choose one statement regarding the planning and drafting of this manuscript:	All authors of this research paper have directly participated in the planning, execution, or analysis of the study.
Enter your initials in the box to confirm the following statement: All authors of this paper have read and approved the final version submitted.	PF
Enter your initials in the box to confirm the following statement: The contents of this manuscript have not been copyrighted or published previously.	PF
Enter your initials in the box to confirm the following statement: The contents of this manuscript are not now under consideration for publication elsewhere.	PF
Enter your initials in the box to confirm the following statement: The contents of this manuscript will not be copyrighted, submitted, or published elsewhere while acceptance by the Journal is under consideration.	PF
Does the research described in this manuscript meet ethical guidelines,	Yes

including adherence to the legal requirements of the study country?	
Is this submission one of a set of linked or companion papers, which should be considered for publication together?	No
Please indicate which of the following standard reporting guidelines were used for this submission:	STROBE - For cohort and case-control studies
If you selected "Other" in the previous question, which reporting guideline did you use for this submission?	
Is this a solicited manuscript?	No
I confirm that authors of this paper agree to be accountable for all aspects of the work, such that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.	I confirm this statement.
Does your manuscript contain figures which you would like published in color? You will be charged US \$600 for each figure reproduced in color in print. Charges are accrued per figure, not per panel.	Yes, I understand that I will be charged for color reproduction and would like at least one figure printed in color
How many figures would you like printed in color? as follow-up to "Does your manuscript contain figures which you would like published in color? You will be charged US \$600 for each figure reproduced in color in print. Charges are accrued per figure, not per panel."	1
Which figures would you like printed in color? as follow-up to "Does your manuscript contain figures which you would like published in color? You will be charged US \$600 for each figure reproduced in color in print. Charges are accrued per figure, not per panel."	Figure number: 2
Please provide the name, address, email address, and telephone number of the person to whom the invoice for color printing should be sent. as follow-up to "Does your manuscript contain figures which you would like published in color? You will be charged US \$600 for each figure reproduced in	Karina Zaluski c/o IARC 150, cours Albert Thomas zaluskik@iarc.fr +33472738485

color in print. Charges are accrued per figure, not per panel."	
Does this manuscript submission contain supplemental materials to be published online?	Yes
<p>Please note the following limits on the total number of supplemental tables plus supplemental figures:</p> <ul style="list-style-type: none"> • Articles, Reviews, Systematic Reviews, and Meta-Analyses: 8 • Commentaries and Mini-Reviews: 4 • Brief Communications: 2 • Correspondence, Responses, and Editorials: 1 <p>Does this submission exceed these limits? If so, select <i>Yes</i> and provide a justification in the resulting box.</p> <p>as follow-up to "Does this manuscript submission contain supplemental materials to be published online?"</p>	No
<p>Please check the box below to confirm the following statement:</p> <p>I hereby grant to Oxford University Press a non-exclusive license for the duration of the copyright period to publish the data supplement/supplementary materials in all languages and media. I also grant to Oxford University Press the right to grant third party permissions to republish the data supplement/supplementary materials in whole or parts thereof in any medium without limitation. As the author, I retain the copyright and all other rights in the data supplement/supplementary materials.</p> <p>as follow-up to "Does this manuscript submission contain supplemental materials to be published online?"</p>	I confirm this statement.
Excluding Supplementary Figures, how many figures are included in your manuscript?	2
Excluding Supplementary Tables, how many tables are included in your manuscript?	3
Choose one statement regarding	There are no directly related manuscripts or abstracts, published or unpublished, by

manuscripts directly related to this submission:	any author(s) of this paper.
Have any authors' names been removed since the original manuscript submission? For any authors that have been removed, please provide the editorial office with written permission from the author to be removed. An email is acceptable.	No
Have any authors been added since the original manuscript submission?	No authors have been added
Does your manuscript contain reference to work cited as personal communication, unpublished data, or a manuscript in preparation, submitted for publication, or in press?	No
Does this submission include previously published tables, figures, or text (either reprinted or adapted) for which the copyright is held by another publisher?	No
Question 1: The Work under Consideration for Publication Read Instructions associated with this question. Did you or your institution at any time receive payments or services from a third party (government, commercial, private foundation, etc.) for any aspect of the submitted work (including but not limited to grants, data monitoring board, study design, manuscript preparation, statistical analysis, etc.)?	No
Question 2: Relevant Financial Activities Outside the Written Work Read the Instructions. Do you have financial relationships (regardless of amount of compensation) with entities as described? You should report relationships that were present during the 36 months prior to publication . Err on the side of full disclosure.	No
Question 3: Intellectual Property -- Patents and Copyrights Do you have any patents, whether planned, pending or issued, broadly relevant to the work?	No
Question 4: Relationships Not Covered Above Are there other relationships or activities	No

that readers could perceive to have influenced, or that give the appearance of potentially influencing, what you wrote in the submitted work?	
Order of Authors:	<p>Erikka Lofffield, PhD</p> <p>Magdalena Stepien, PhD</p> <p>Vivian Viallon, PhD</p> <p>Laura Trijsburg, PhD</p> <p>Joseph Rothwell, PhD</p> <p>Nivonirina Robinot, BSc</p> <p>Carine Biessy, BSc</p> <p>Ingvar A Bergdahl, PhD</p> <p>Stina Bodén, MSc</p> <p>Mattias B Schulze, PhD</p> <p>Manuela Bergman, PhD</p> <p>Elisabete Weiderpass, MD, MSc, PhD</p> <p>Julie A Schmidt, PhD</p> <p>Raul Zamora-Ros, PhD</p> <p>Therese Haugdahl Nøst, PhD</p> <p>Torkjel M Sandanger, PhD</p> <p>Emily Sonestedt, PhD</p> <p>Bodil Ohlsson, PhD</p> <p>Verena Katzke, PhD</p> <p>Rudolf Kaaks, PhD</p> <p>Fulvio Ricceri, PhD</p> <p>Anne Tjonneland, PhD</p> <p>Christina C Dahm, PhD</p> <p>Maria-Jose Sanchez, PhD</p> <p>Antonia Trichopoulou, PhD</p> <p>Rosario Tumino, MD, MSc, DLSHTM</p> <p>Maria-Dolores Chirlaque, PhD</p> <p>Giovanna Masala, PhD</p> <p>Eva Ardanaz, PhD</p> <p>Roel Vermeulen, PhD</p> <p>Paul Brennan, PhD</p> <p>Demetrius Albanes, MD</p> <p>Stephanie J Weinstein, PhD</p> <p>Augustin Scalbert, PhD</p> <p>Neal D Freedman, PhD MPH</p> <p>Marc J Gunter, PhD</p>

	Mazda Jenab, PhD
	Rashmi Sinha, PhD
	Pekka Keski-Rahkonen, PhD
	Pietro Ferrari, PhD

International Agency for Research on Cancer



150 cours Albert Thomas
69372 Lyon cedex 08, France

Nutrition and Metabolism Section
Nutritional Methodology and Biostatistics Group
Tel.: +33 4 72 73 80 31
Fax: + 33 4 72 73 83 61
E-mail: FerrariP@iarc.fr
<http://www.iarc.fr>

Ref.: NMB/PF/kz/21/02/JNCI

24 February 2021

Dear Dr Ganz and the JNCI Editorial Board,


We thank you and the Reviewers for the feedback on our manuscript "Novel biomarkers of habitual alcohol intake and associations with risk of pancreatic and liver cancers and liver disease mortality" (JNCI-20-1776), a research initiative led by scientists at the International Agency for Research on Cancer and at the US National Cancer Institute. We carefully revised the manuscript according to the Reviewers' comments that contributed to further improve the text. You will find it enclosed, together with our point-by-point response letter.

Based on the comments received, we now refer to our study as a multi-stage study design rather than a replication, and we have revised the manuscript accordingly. We carefully explained the reasons for not conducting a proper discovery-replication study with features acquired by the same laboratory platforms at different times for each of the three components of our study. We also added information on the stability of 2-hydroxy-3-methylbutyric acid based on 1-year intraclass correlation coefficient and emphasized the need to examine 2-hydroxy-3-methylbutyric acid (and other candidate biomarkers) in an alcohol feeding trial. Several other details related to the design, analysis and interpretation of the results of our study were further clarified in the response letter and in the text, as requested by the Reviewers. In an attempt to comprehensively amend the text in line with the Reviewers' suggestions, the text now slightly exceeds the 3300-word limit.

Once again we confirm that the authors of this research paper have directly participated in the planning, execution, or analysis of the study, and have read and approved the final version submitted. The contents of this manuscript have not been copyrighted or published previously. The contents of this manuscript are not under consideration for publication elsewhere. The contents of this manuscript will not be copyrighted, submitted, or published elsewhere while acceptance by the Journal is under consideration.

Please do not hesitate to let us know if you require any further information. Thank you in advance for your kind consideration.

Yours sincerely,



Pietro Ferrari, PhD

Head, Nutritional Methodology and Biostatistics Group

Comments from the Editors

Because the accepted versions of manuscripts will now be published on the Journal site before they undergo copyediting and typesetting by the Journal, we are asking authors to carefully proofread their manuscripts for correct spelling and grammar as part of the revision process.

Reviewers raise important concerns regarding the design of discovery and verification of novel biomarkers including need for better use of dimension reduction techniques. As noted by the reviewers, the current design is not a true replication study but a multi-stage study. Please address these and the other concerns they raise as noted below:

We thank the editors for giving us the opportunity to respond to the Reviewer's comments and improve the manuscript. We also recognize that our study is better characterized as a multi-stage study design rather than a replication. We have revised the manuscript accordingly. Please refer to our responses to the Reviewers' concerns below for a detailed explanation of our rationale.

Fundamental details of the biomarker stability /decay and time-frame in relation to alcohol intake (acute, chronic, or some combination) that it actually represents?

The editor highlights an important consideration; however, studies on the kinetics of candidate biomarkers that we identified are currently lacking. We did, however, find data on 1-year intraclass correlation coefficients for 2-hydroxy-3-methylbutyric acid (i.e., alpha-hydroxyisovalerate) ranging from 0.76 to 0.49 in independent samples of 60 women with blood collections at baseline and 1-year in the Shanghai Women's Health Study and 30 adults (14 women and 16 men) with blood collections at baseline and 1-year in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial, respectively. These ICCs indicate low to moderate within-subject variability (i.e., good to moderate reliability) over 1-year. We have added this information to the discussion (page 13) and have also highlighted the need for studying 2-hydroxy-3-methylbutyric acid and other candidate biomarkers of alcohol intake in an alcohol feeding trial, which is better suited to establishing the dose-response relationship between alcohol and these candidate biomarkers, as well as the timeframe of biomarker decay/stability in relation to alcohol intake (page 14).

As noted in the manuscript, alcohol intake is well reported and validated in the EPIC study – correlation of 0.79 between FFQ and 12 and 24 hour recalls over a year was clearly higher than the validation correlation for any other nutrient or food group. This is consistent with a large body of literature and raises the question of potential value added over the questionnaire intake for this marker.

Although validation studies have shown larger correlations between dietary questionnaire and 24-hour dietary recall (24-hdr) measurements of alcohol intake than for most other dietary constituents, this information may not reflect the level of accuracy. Alcohol drinking is a sensitive exposure to recall, making it prone to systematic underreporting across types of assessments. As a result, estimates of validity coefficients for alcohol intake could be inflated (biased upward) by correlation between errors in the two types of self-reported assessments, questionnaires and 24-hdrs, making the validity better than its nominal true level. We have expanded the introduction (page 5) to address this issue.

For these reasons, there is a need for objective assessments of long-term alcohol use that reliably distinguish light from moderate and heavy drinkers. Moreover, the features identified in this study may help clarify associations of alcohol intake with disease risk by providing insights into mechanisms underlying these associations. We have added text to the discussion (page 14) to highlight the potential added value of these markers in etiologic studies.

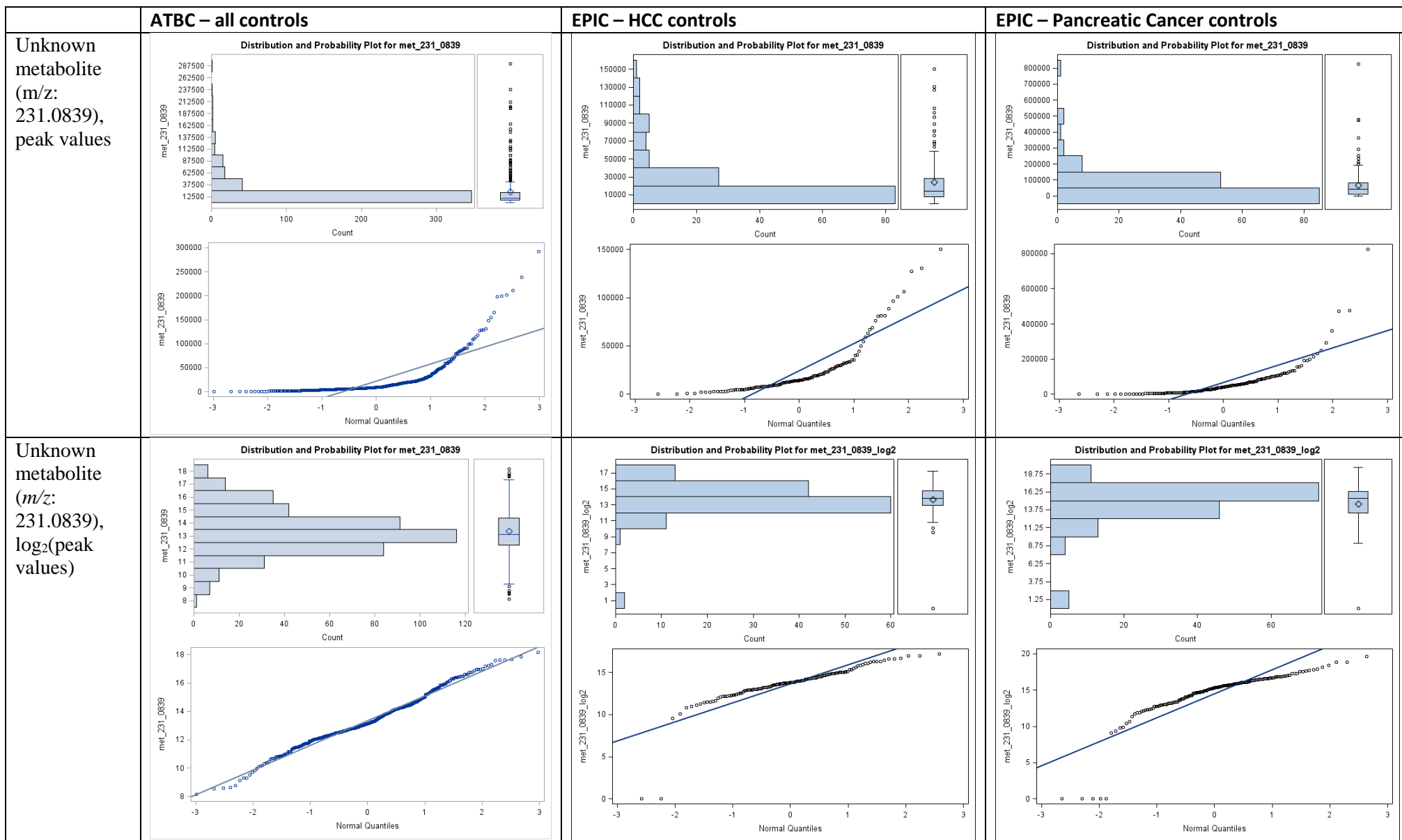
Comments from the Reviewers

Please note: All the comments to authors we have received are included below, regardless of the numbering of the reviewers.

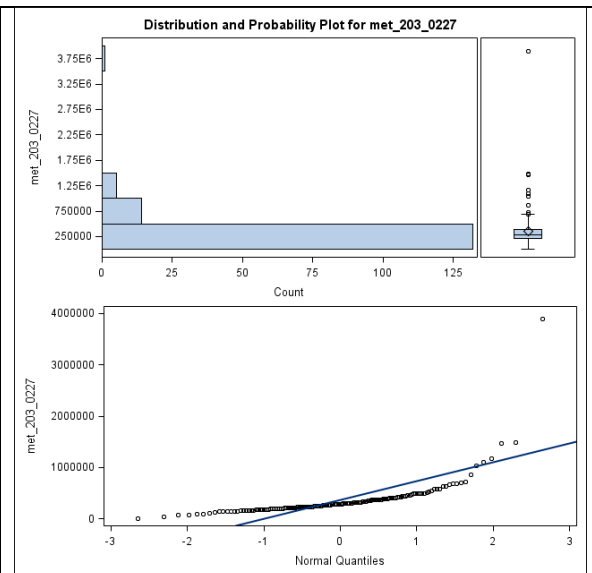
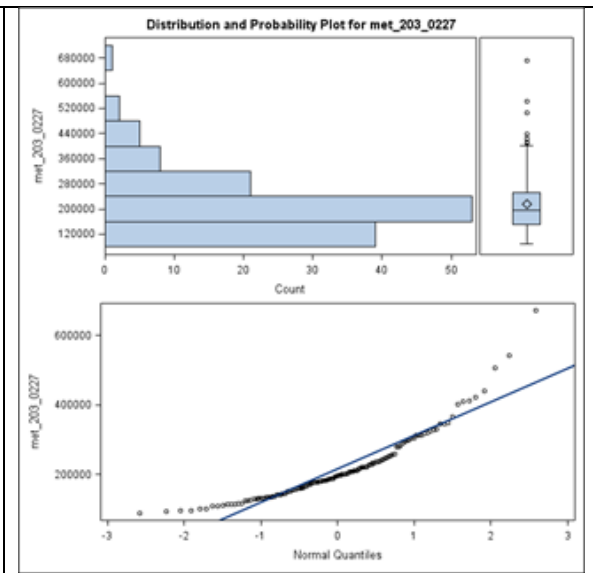
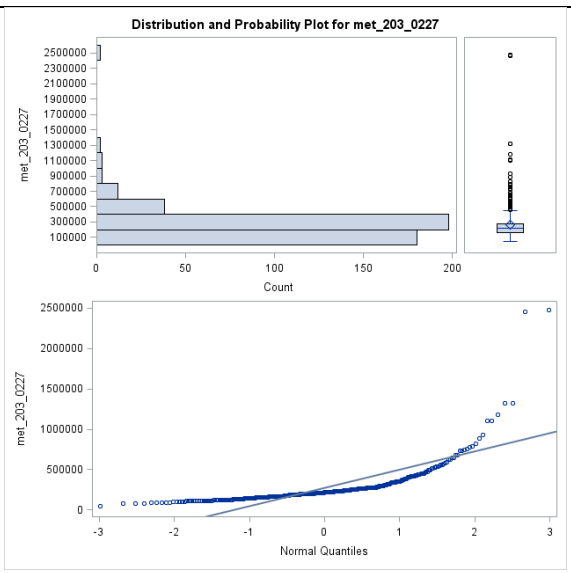
Reviewer 1: Major concerns with the statistical analyses:

The distribution of the feature intensities should be checked to ensure that the log₂ transformation is appropriate.

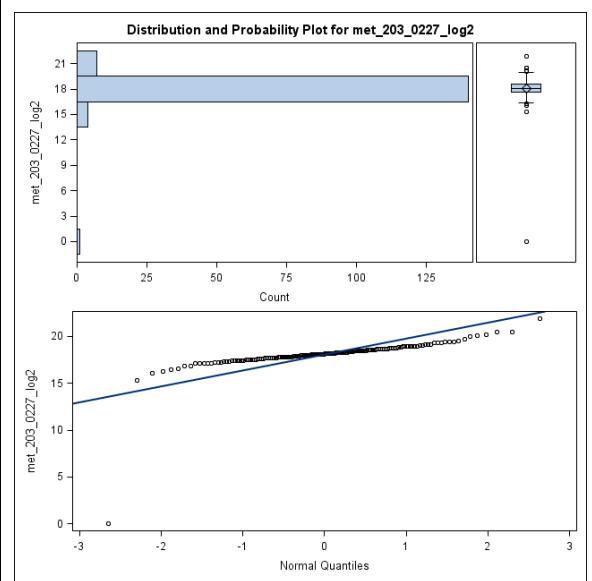
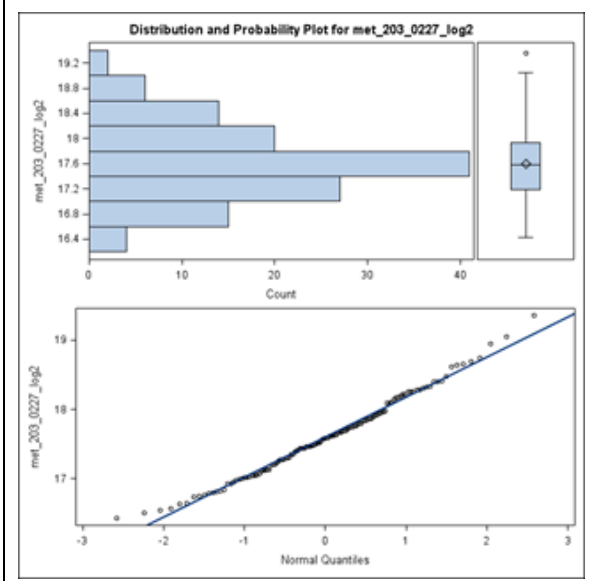
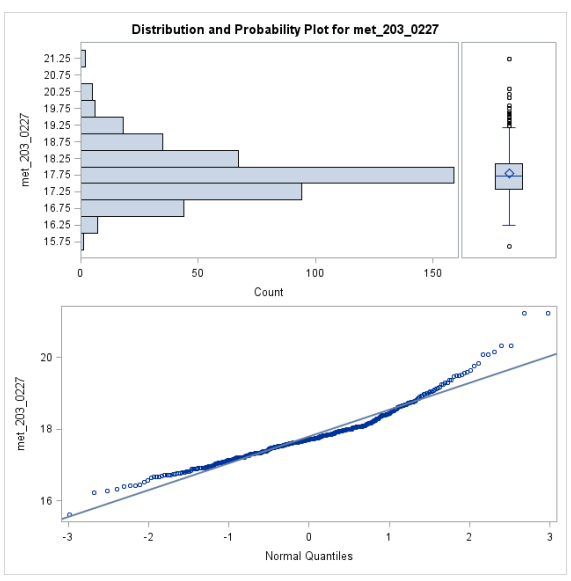
Metabolite distributions tend to be non-normal and are often skewed right, but identifying an optimal, universal transformation for 100s to 1000s of metabolites is unrealistic. Therefore, in metabolomics studies, it has become standard practice to log-transform features (i.e., relative peak levels) to improve symmetry and approximate normality. This is particularly useful in correlation analyses to make proper statistical inference for correlation values. Using a logarithm of base 2 is also common in metabolomics studies because it lends itself to the simple interpretation of regression parameters in logistic regression models, where feature concentration values are related to the risk of a given disease. The exponential of regression parameter indicates the increase in disease OR associated with a 1-unit increase corresponding to a doubling of the exposure variable, i.e. the feature level in this case. We further standardized the continuous variables such that a 1-unit increase corresponded to a 1-standard deviation increase on the log-scale. In this way, disease-specific OR estimates were computed for comparable increases of the exposure variables, in turn, alcohol intake and feature levels, as reported in Table 3. We used histograms and QQ-plots to show the distribution of each metabolite feature that was correlated with alcohol, and we confirmed that log-transformation substantially improved normality and symmetry for both metabolites in EPIC and ATBC (see plots below).



2-hydroxy-3-methylbutyric acid, peak values



2-hydroxy-3-methylbutyric acid, log₂(peak values)



It's not clear why features selected in the discovery dataset were carried to the second dataset, then a subset of these features was then, again, carried to the third dataset. All features should be considered separately in all three data sets via dimension reduction techniques, and the final set should be chosen by the union of these selected features.

Multiple approaches for selecting metabolite features of interest have been employed in recent years across hundreds of metabolomics studies, but, to the best of our knowledge, there is no consensus on an optimal approach as it likely varies by study question and design. In the current study, we worked with metabolomics data from four independent sets of data that were generated over a period of four years at four distinct times: 1. EPIC cross-sectional study, 2. EPIC hepatocellular carcinoma (HCC) and 3. pancreatic cancer nested case-control studies, and 4. ATBC nested liver cancer and liver disease mortality case-control study. Although the metabolomics data were generated by the same metabolomics lab at IARC, there are unavoidable differences in retention times and overall performance of the LC columns used over the years.

Consequently, we did not attempt to align all the features from the different datasets. Rather, we manually matched features from a single discovery dataset to those from the other datasets. For this reason, we adopted a multi-stage design to optimize the information extracted from each set of data and to limit the number of multiple comparisons across successive replication phases. We began our discovery process with the EPIC cross-sectional sample, which was the largest of 3 EPIC datasets and did not rely on controls from nested cancer studies. Out of the 6,597 features in the EPIC cross-sectional sample, 133 (98 RP+ and 35 RP-) were associated to alcohol intake, after FDR correction. Based on spectral data, these features were manually matched to features acquired in controls (n=280) from the two EPIC nested case-control studies on pancreatic cancer and hepatocellular carcinoma, resulted in 49 matching features (38 RP+ and 11 RP-), of which 10 features (7 RP+ and 3 RP-) were correlated with alcohol intake, after Bonferroni correction. In the third and final stage, the 7 RP+ features were successfully matched to features measured in ATBC controls. Note that in ATBC study only RP+ mode was available, and all 7 features were correlated with alcohol intake.

In summary, our multi-stage design was motivated by the challenge of individually matching features that were generated at different time points, and it was agreed by our team of biostatisticians, chemists, and epidemiologists as the best suited design to the data and question at hand. In agreement with the Reviewer's suggestion, we have updated the definition of our design from a true replication design to a multi-stage design.

It's also not clear why false-discovery rate was used in the first dataset while Bonferroni was used in the second. Suggest adopt supervised dimension reduction conditional on the same selection criteria instead of stepwise selection.

The first stage of analysis was exploratory in nature and included 6,597 features, many of which were highly correlated with each other, indicating that accounting for multiple comparisons via FDR-corrected p-values was reasonable to account for multiple comparisons. As the number of tests in the second and third sets were restricted to those features that were matched between stage 1 and 2 and then between stages 2 and 3, respectively, we deemed it prudent to adopt the more conservative Bonferroni threshold for statistical significance.

In line with the Reviewer's suggestion, we conducted complementary analysis using a supervised dimension reduction technique. Features were selected via the LASSO regression, which was applied

to the EPIC cross-sectional sample. LASSO selected 14 features, including one with m/z and retention time similar to features that we identified in univariate analyses as part of the metabolite 2-hydroxy-3-methylbutyric acid. We also implemented elastic-net regression. For $\alpha = 0.1$ ($\alpha=1$ corresponds to the pure lasso; $\alpha=0$ to pure ridge regression), 54 features were identified, 3 of which had retention time close to 2.78 that originated from the 2-hydroxy-3-methylbutyric acid.

We acknowledge that we could have used various study designs and analytical strategies. As shown by our additional analyses, LASSO, elastic net and other multivariate machine learning techniques would have selected a limited number of correlated features; thus, an advantage of the methodology we chose, based on univariate tests and correction for multiple testing, is that univariate tests tend to repeatedly select features originating from the same metabolite; since some of the features were not retained after matching across different sets of data, techniques that tend to select one or only a few features out of several candidates would offer an efficient yet suboptimal analytical strategy in our setting.

Leave-one-out cross validation does not seem appropriate in this particular case, because the samples are matched case-control.

The analysis in question only used data from controls; however, in line with the Reviewer's comment, we have removed this analysis from the manuscript owing to concerns about its limited value.

Reviewer 2: `Review of Lofffield et al JNCI-20-1776

This paper describes untargeted metabolomics analysis that lead to a biomarker for alcohol (2-hydroxy-3-methylbutyric acid) which was correlated with FFQ-based alcohol intake and was associated with risk of liver cancer, pancreatic cancer, and fatal liver disease in EPIC and or ATBC.

If these results hold up in further replication, they could be quite important, but there are many practical issues to consider. If it was possible to fit one (or both) of the biomarkers into a targeted metabolomics array that could be run at much lower cost than untargeted chips and if batch effects could be reduced or eliminated, then these results may be quite significant. I think that the paper needs to touch on these issues (replication, cost, and data quality) much more than it currently does.

Thank you for your comments and feedback. We agree that these issues are important and have added the following text to our discussion (page 12): "Additionally, targeted metabolomics panels that can simultaneously measure multiple alcohol-related metabolites , including 2-hydroxy-3-methylbutyric acid and related compounds, should be developed to measure absolute concentrations, which will enable comparisons and pooling of data across studies, supporting replication and improving risk estimation; this is especially important for diseases such as pancreatic cancer, for which the literature is suggestive [40] yet inconsistent [41]."

We also think it is important to note that in the current analysis we developed and applied a pre-processing pipeline to correct unwanted laboratory features in the data, including variability introduced by well plate and batch. Additionally, we considered the correlations between residuals of metabolite features and alcohol intake so that we could adjust features for batch variables in addition to other potential confounders, like age and smoking, that were also relevant to alcohol intake.

Other comments:

Abstract - not important to indicate base of logarithmic transformation if expressed per SD (log2). This applies elsewhere.

We have made this change throughout the text.

Abstract states that this is a potential biomarker for habitual alcohol intake but is anything known about how quickly it itself is metabolized? Could this be a biomarker of recent alcohol intake instead?

As noted in our response to the editor (above), the rate at which a metabolite is broken down is an important consideration; however, studies on the kinetics of the candidate biomarkers that we identified are lacking. We did, however, find data on the 1-year reliability of 2-hydroxy-3-methylbutyric acid in previously published metabolomics methods studies, and we have added this information to the discussion (page 13). In the discussion, we have also highlighted the need for studying 2-hydroxy-3-methylbutyric acid and other candidate biomarkers of alcohol intake in an alcohol feeding trial (page 14), which is better suited to establishing the dose-response relationship between alcohol and these candidate biomarkers.

Methods: I am annoyed by the way in which the word "replication" is used throughout the paper. There are two results that need replication

1. The observed relationship between the two metabolites and alcohol use from the EPIC FFQ found in the untargeted analysis. (Although technically what is described is not really a replication study but is rather a multi-stage study, with associations being winnowed down at each stage).

This concern was also raised by the first Reviewer (see detailed response above), and we agree that a multi-stage design better characterizes our approach. Thus, we have revised the text to reflect this change in description and to better explain our rationale for this approach.

2. The association between the two metabolites and liver cancer (HCC specifically) and pancreatic cancer in the EPIC nested case control study. The ATBC study case-control study serves as only a partial replication of what is found in the EPIC, much more replication work is needed, in particular these studies are all quite small, and the number of total associations considered is quite large, lessening confidence (and widening confidence intervals).

The Reviewer is correct that the metabolite-endpoint associations are only partly replicated and that more work is needed to corroborate our findings; this includes replicating and extending these findings in other existing nested case-control sets and metabolomics consortia. We have highlighted this in the discussion (page 12).

Page 6 second paragraph is especially confusing. Attention shifts in the third sentence to the cases from the case control studies case control studies, this threw me a bit since I was expecting to hear about the two "replication" studies for the relationship between alcohol use and the metabolites first, before getting into the etiological studies (case-control studies). I suggest reorganizing this paragraph and the next to discuss first the discovery and then the two replication studies of the association between self-reported alcohol and metabolites, before describing the case control studies. This should be much clearer and would better correspond to Figure 1.

Thank you for this suggestion. We have made this change (pages 6 & 7), which improved clarity and readability.

Page 8. Make sure that alcohol intake in the EPIC and ATBC is consistently referred to as self-reported.

We have added “self-reported” as a qualifier of alcohol intake throughout the manuscript.

Give a citation for the "residual method" in linear regression models. Actually, the residual method is biased and suffers loss of power compared to just including the variable of interest into the model. Here the variable of interest could be the residuals for the feature intensities. Suggest including these in the model for alcohol intake (rather than calculating the correlation of the residuals of the feature intensities with the residuals of alcohol intake). It probably doesn't matter much here, but in general it is better.

We have added the following citation to the methods (page 8) for the residual method: Kleinbaum, D. G., Kupper, L. K. and Muller, K. E. (1987) Applied regression analysis and other multivariable methods. Duxbury Press, Belmont, CA. We agree with the limitations mentioned by the Reviewer. It is noteworthy that the residual method is prone to conservative results, which is a desirable feature in an exploratory setting. Also, over a standard adjustment in regression models, the residual method has the advantage of allowing the role of specific covariates to be controlled for, separately for each variable analyzed in a correlation study. In this study, well plate and batch indicators were used to adjust metabolomics feature concentrations only, but not self-reported alcohol intake.

In agreement with the Reviewer's request, we ran additional analyses using linear models with raw values of self-reported alcohol as the independent variable and either raw values of feature intensities or residuals of feature intensities, (to correct for laboratory factors) as the dependent variable. In either case, models were adjusted for the other covariates (smoking intensity, etc.). Results were overall very similar to the ones presented in the manuscript, and included the same features associated with 2-hydroxy-3-methylbutyric acid and the unknown compound.

Page 8. As a coffee-drinker I would prefer that you say that coffee drinking is associated with "reduced risk" of liver cancer and liver disease, rather than just "risk".

Yes, we agree and have clarified this in the text on page 8: “Coffee drinking and coffee-associated metabolites have been strongly associated with lower risk of liver cancer and liver disease mortality in ATBC”.

Page 10: What was the partial correlation between the two metabolites found in the discovery stage and self-reported alcohol intake in the three cohorts? This is a better way of expressing the results of linear regression analysis than dichotomizing the data and then doing AUROC analysis in my opinion.

As clarified earlier, we estimated the correlations between the residuals of self-reported alcohol intake, adjusted for age, sex, country (in EPIC only), body mass index, smoking status and intensity, and coffee consumption, and the residuals of each feature adjusted for the same potential confounders as well as plate number, position within the plate (row and column indexes), and the study (EPIC stage 2) or batch indicator (ATBC stage 3).

In agreement with the Reviewer’s comment, we have removed the AUROC analysis and highlighted the correlations between residuals of the two metabolites and between residuals of each of the two metabolites and self-reported alcohol intake (page 10). We have summarized the correlations below for your convenience:

	Correlation coefficient between residuals*		
	Alcohol	Unknown metabolite (m/z: 231.0839)	2-hydroxy-3-methylbutyric acid (m/z: 203.0227)
EPIC stage 1 (n = 454 participants)			
Alcohol	1.00		
Unknown metabolite (m/z: 231.0839)	0.41	1.00	
2-hydroxy-3-methylbutyric acid (m/z: 203.0227)	0.26	0.23	1.00
EPIC stage 2 (n = 280 controls)			
Alcohol	1.00		
Unknown metabolite (m/z: 231.0839)	0.38	1.00	
2-hydroxy-3-methylbutyric acid (m/z: 203.0227)	0.24	0.25	1.00
ATBC stage 3 (n = 438 controls)			
Alcohol	1.00		
Unknown metabolite (m/z: 231.0839)	0.40	1.00	
2-hydroxy-3-methylbutyric acid (m/z: 203.0227)	0.40	0.54	1.00

Page 11: It seems odd that self-reported alcohol was not a predictor of HCC or PC but the metabolite was. After all the metabolite was discovered by determining it was strongly associated with self-reported alcohol (not of true alcohol consumption). If self-reported alcohol is a very poor predictor of true alcohol consumption, then it seems kind of serendipitous that something highly correlated with self-reports turned out to be such a good predictor of HCC and PC risk.

The Reviewer’s skepticism is warranted, yet we think that there is a logical explanation for the seemingly incongruent findings. First, it is important to note that the correlations between self-reported alcohol intake and each metabolite, although highly statistically significant, are modest, ranging from 0.38 to 0.41 and from 0.24 to 0.40 across the three data sets for the unknown metabolite (m/z: 231.0839) and 2-hydroxy-3-methylbutyric acid, respectively. This observation is compatible with the fact that feature intensities show varying levels of association with the endpoints, as compared to self-reported alcohol. Despite wider confidence intervals the odds ratios (ORs) for self-reported alcohol are consistent with ORs for the unknown compound, which is likely linked to ethanol metabolism. The larger ORs for 2-hydroxy-3-methylbutyric, consistently observed across the four endpoints, may reflect that 2-hydroxy-3-methylbutyric acid is not a constituent or a byproduct of alcohol intake; rather, its level may reflect a relevant biological response to alcohol intake that potentially plays a role in the etiology of multiple chronic diseases. The fact that associations of alcohol with 2-hydroxy-3-methylbutyric acid were robust across all 3 studies is also likely not a coincidence given that previous studies have reported fairly large values of 1-year ICCs,

indicating that within person or biologic variability over time is lower than between person variability. Finally, the larger correlations between self-reported alcohol with the unknown metabolite than with 2-hydroxy-3-methylbutyric acid could be in line with this reasoning.

More generally, was self-reported alcohol consumption considered as a covariate (adjustment) variable in the cancer risk analyses? I.e. do the two metabolites significantly improve the model over just using self-reported alcohol? (Presumably they do in the EPIC case-control study). Clearly this is a result that will need further confirmation/replication in the future.

To follow-up on the Reviewer's question, we added self-reported alcohol intake to the model and added this to the methods (page 9). The OR estimates for the metabolites did not change substantially. Moreover, in these models each metabolite was more strongly associated with the disease endpoint than self-reported alcohol was. We present these results the revised version of Table 3 and in the results (page 11) and discussion (pages 11 and 14).

Figure 1. The red box indicates two metabolites identified but only lists one (the unknown compound).

Thank you for catching this. The figure has been revised accordingly.

Reviewer 3: The authors identified biomarkers related to alcohol consumption and investigated the associations with risk of pancreatic and liver cancers as well as liver disease mortality, using two European cohort studies. The use of untargeted metabolomics in the discovery and replication datasets is a major strength of this study that may help address potential limitations of self-reported alcohol intake. In general, the paper is nicely written with well-rounded metabolomics approaches. However, some revisions in text and data presentation seem to be required. Please see the following comments and suggestions:

Major comments:

* The authors observed much weaker or null associations for hepatocellular carcinoma (HCC) and pancreatic and liver cancers when using questionnaire-derived alcohol intake as opposed to alcohol-associated metabolites. Could it be due to the use of a single assessment of alcohol?

It is certainly possible that participants' alcohol intake changed over time, and we have noted on page 14 that having a single assessment of alcohol intake in both EPIC and ATBC is a limitation of the study: "Additionally, self-reported alcohol intake and blood measures were assessed in each study at baseline only; therefore, we are unable to account for changes in alcohol intake or metabolites over time."

* Alcohol metabolism and susceptibility to its metabolites can be different between men and women as authors applied different categories to classify high vs low consumers in the EPIC replication dataset. In Table 3, have authors explored sex-specific associations for disease risk in EPIC? Sex was not taken into account in the model.

The results summarized in table 3 are from conditional logistic regression models, and in the EPIC nested case-control studies sex was a matching factor. ATBC included men only.

We did not originally conduct sex-stratified analyses due to limited sample size. However, as suggested by the Reviewer, we undertook additional analyses to examine interactions by sex in EPIC (see below). For both HCC and pancreatic cancer, the interaction term between sex and each metabolite or self-reported alcohol intake was not statistically significant. Although the OR for 2-hydroxy-3-methylbutyric acid and HCC is statically significant in men but not women, the confidence intervals are overlap, indicating that OR estimates are homogeneous. Overall, sex stratified analyses and tests for interaction lack statistical power owing to small sample size. Therefore, we have included the results for your review but do not think they warrant inclusion in the manuscript.

Associations of self-reported alcohol and alcohol-related metabolites with HCC and pancreatic cancer in EPIC stratified by sex							
	Men (n=87 case-control sets)			Women (n=41 case-control sets)			Interaction with Sex
	OR	(95% CI)	p-value	OR	(95% CI)	p-value	p-value
HCC, EPIC (128 case-control sets)							
Alcohol intake (12g/day)	1.03	(0.87-1.23)	0.72	1.97	(0.93-4.17)	0.08	0.45
Alcohol intake (1-SD (log ₂))	0.67	(0.42-1.07)	0.09	1.23	(0.53-2.87)	0.63	0.40
Unknown compound (1-SD (log ₂)) ²	0.99	(0.59-1.67)	0.97	0.84	(0.23-2.99)	0.78	0.40
2-hydroxy-3-methylbutyric acid (1-SD (log ₂)) ³	3.76	(1.69-8.35)	0.001	2.63	(0.66-10.4)	0.17	0.50
Pancreatic cancer, EPIC (152 case-control sets)							
Men (n=60 case-control sets)			Women (n=92 case-control sets)				
Alcohol intake (12g/day)	1.15	(0.89-1.50)	0.29	0.86	(0.59-1.24)	0.41	0.27
Alcohol intake (1-SD (log ₂))	1.27	(0.73-2.21)	0.40	0.87	(0.58-1.31)	0.50	0.30
Unknown compound (1-SD (log ₂)) ²	2.23	(1.02-4.87)	0.05	0.96	(0.70-1.30)	0.77	0.06
2-hydroxy-3-methylbutyric acid (1-SD (log ₂)) ³	2.02	(0.95-4.28)	0.07	1.52	(0.98-2.37)	0.06	0.88

¹ Models for hepatocellular carcinoma (HCC) were adjusted for body mass index (BMI, kg/m²), waist circumference (cm), recreational and household physical activity (Met-hours/week), a composite variable comprising smoking status and intensity (Never, Current: 1-15 cig/day, Current: 16-25 cig/day, Current: 26+ cig/day, Former: quit ≤ 10 years, Former: quit 11-20 years, Former: quit 20+ years, Current, occasional pipe/cigar use, Current/Former: missing, Unknown), level of educational attainment, and coffee intake (grams/day, log₂-transformed); models for pancreatic cancer were adjusted for BMI (kg/m²), sex-specific physical activity categories and smoking;

² Unknown compound (*m/z*=231.0839);

³ 2-hydroxy-3-methylbutyric acid (*m/z*=203.0227).

* Does EPIC have any data on smoking intensity that can be included as a covariate?

Thank you for highlighting this. There was a mistake in our table 3 footnote regarding covariate adjustment for smoking in the EPIC nested case-control studies. We did in fact adjust for a comprehensive composite variable comprising smoking status and intensity (Never, Current: 1-15

cig/day, Current: 16-25 cig/day, Current: 26+ cig/day, Former: quit \leq 10 years, Former: quit 11-20 years, Former: quit 20+ years, Current, occasional pipe/cigar use, Current/Formers: missing, Unknown). We have revised the table footnote accordingly. At the same time, we updated the models for pancreatic cancer and HCC since these mistakenly used the feature residuals rather than the log₂ value of feature signals. ORs did not meaningfully change, and we have updated all ORs in the tables and text.

Minor comments:

* The authors may want to define what m/z, f1 and f2 stand for when mentioned first in the text.

We have made these changes to the text on pages 7, 8, and 9.

* Page 8, line 2: abd ackground - typo?

Thank you for catching this. It now reads "and background".

Novel biomarkers of habitual alcohol intake and associations with risk of pancreatic and liver cancers and liver disease mortality

Erikka Loftfield, PhD ^{1*}; Magdalena Stepien, PhD ^{2*}; Vivian Viallon, PhD ³; Laura Trijsburg, PhD ³; Joseph Rothwell, PhD ^{2,5,6}; Nivonirina Robinot, MSc ⁴; Carine Biessy, BSc ³; Ingvar A. Bergdahl, PhD ⁷; Stina Bodén, MSc ⁸; Matthias B. Schulze, PhD ^{9,10}; Manuela Bergman, PhD ^{9,10}; Elisabete Weiderpass, MD, MSc, PhD ¹¹; Julie A. Schmidt, PhD ¹²; Raul Zamora-Ros, PhD ¹³; Therese H. Nøst, PhD ¹⁴; Torkjel M Sandanger, PhD ¹⁴; Emily Sonestedt, PhD ¹⁵; Bodil Ohlsson, PhD ¹⁵; Verena Katzke, PhD ¹⁶; Rudolf Kaaks, PhD ¹⁶; Fulvio Ricceri, PhD ¹⁷; Anne Tjønneland, PhD ¹⁸; Christina C. Dahm, PhD ¹⁹; Maria-Jose Sánchez, PhD ^{20,21,22}; Antonia Trichopoulou, PhD ²³; Rosario Tumino, MD, MSc, DLSHTM²⁴; María-Dolores Chirlaque, PhD ^{25,26}; Giovanna Masala, PhD ²⁷; Eva Ardanaz, PhD ^{28,29,30}; Roel Vermeulen, PhD ³¹; Paul Brennan, PhD ³²; Demetrius Albanes, MD¹; Stephanie J. Weinstein, PhD¹; Augustin Scalbert, PhD ⁴; Neal D. Freedman, PhD ¹; Marc J. Gunter, PhD ²; Mazda Jenab, PhD ²; Rashmi Sinha, PhD ¹; Pekka Keski-Rahkonen, PhD ⁴ ‡; Pietro Ferrari, PhD ³ ‡†

* these first authors contributed equally

‡ these senior authors contributed equally

† corresponding author

¹ Metabolic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA.

² Nutritional Epidemiology Group, International Agency for Research on Cancer (IARC-WHO), Lyon, France.

³ Nutritional Epidemiology and Biostatistics Group, International Agency for Research on Cancer (IARC-WHO), Lyon, France.

⁴ Biomarkers Group, International Agency for Research on Cancer (IARC-WHO), Lyon, France.

⁵ Centre for Epidemiology and Population Health (U1018), Generations and Health team, Faculté de Médecine, Université Paris-Saclay, UVSQ, INSERM, Villejuif, France.

⁶ Gustave Roussy, F-94805, Villejuif, France.

⁷ Biobank Research Unit, Umeå University, Sweden.

⁸ Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden.

⁹ Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany.

¹⁰ Institute of Nutritional Science, University of Potsdam, Nuthetal, Germany.

- ¹¹ International Agency for Research on Cancer, World Health Organization.
- ¹² Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom.
- ¹³ Unit of Nutrition and Cancer, Epidemiology Research Program, Catalan Institute of Oncology, Bellvitge Biomedical Research Institute (IDIBELL), Hospitalet de Llobregat (Barcelona), Spain.
- ¹⁴ Department of Community Medicine, UiT- The Arctic University of Norway, Tromsø, Norway.
- ¹⁵ Department of Clinical Sciences in Malmö, Lund University, Malmö, Sweden.
- ¹⁶ Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
- ¹⁷ Department of Clinical and Biological Sciences, University of Turin, Italy; Unit of Epidemiology, Regional Health Service ASL TO3, Grugliasco (TO), Italy.
- ¹⁸ Danish Cancer Society Research Center; University of Copenhagen, Department of Public Health
- ¹⁹ Department of Public Health, Aarhus University, Denmark.
- ²⁰ Escuela Andaluza de Salud Pública (EASP), Granada, Spain; Instituto de Investigación Biosanitaria ibs.GRANADA, Granada, Spain.
- ²¹ Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain.
- ²² Department of Preventive Medicine and Public Health, University of Granada, Granada, Spain.
- ²³ Hellenic Health Foundation, Athens, Greece.
- ²⁴ Cancer Registry and Histopathology Department, Provincial Health Authority (ASP 7) Ragusa , Italy.
- ²⁵ Department of Epidemiology, Regional Health Council, IMIB-Arrixaca, Murcia University, Murcia, Spain.
- ²⁶ CIBER in Epidemiology and Public Health (CIBERESP), Madrid, Spain.
- ²⁷ Cancer Risk Factors and Life-Style Epidemiology Unit, Institute for Cancer Research, Prevention and Clinical Network - ISPRO, Florence, Italy.
- ²⁸ Navarra Public Health Institute, Pamplona, Spain.
- ²⁹ IdiSNA, Navarra Institute for Health Research, Pamplona, Spain.
- ³⁰ CIBER Epidemiology and Public Health CIBERESP, Madrid, Spain.
- ³¹ Institute for Risk Assessment Sciences, Division of Environmental Epidemiology, Utrecht University, Utrecht, The Netherlands.
- ³² Genetic Epidemiology Group, International Agency for Research on Cancer (IARC-WHO), Lyon, France.

Notes: The authors have no potential conflicts of interest to disclose. Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors

alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

Acknowledgements: EPIC Umeå investigators thank the Västerbotten Intervention Programme and the County Council of Västerbotten for providing data and samples and acknowledge the contribution from Biobank Sweden, supported by the Swedish Research Council (VR 2017-00650). We thank the National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands, for their contribution and ongoing support to the EPIC Study.

Availability of data and materials: 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>

Funding: This work was supported by the Intramural Research Program of the National Cancer Institute at the National Institutes of Health. For EPIC-Oxford, it is: Cancer Research UK C8221/A29017 and C8221/A19170, and Medical Research Council MR/M012190/1. RZ-R was supported by the “Miguel Servet” program (CP15/00100) from the Institute of Health Carlos III (Co-funded by the European Social Fund (ESF) - ESF investing in your future). This work was supported in part by the French National Cancer Institute (L’Institut National du Cancer; INCA; grant numbers 2009-139 and 2014-1-RT-02-CIRC-1; PI: M. Jenab). For pancreatic cancer in EPIC the work was supported by internal IARC funds.

Abstract

Background: Alcohol is an established risk factor for several cancers, but modest alcohol-cancer associations may be missed due to measurement error in self-reported assessments. Biomarkers of habitual alcohol intake may provide novel insight into the relationship between alcohol and cancer risk.

Methods: Untargeted metabolomics was used to identify metabolites correlated with self-reported habitual alcohol intake in a discovery dataset from the European Prospective Investigation into Cancer and Nutrition (EPIC; n=454). Significant correlations were tested in independent datasets of controls from case-control studies nested within EPIC (n=280) and the Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC; n=438) study. Conditional logistic regression was used to estimate odds ratios (OR) and 95% confidence intervals for associations of alcohol-associated metabolites and self-reported alcohol intake with risk of pancreatic cancer, hepatocellular carcinoma (HCC), liver cancer, and liver disease mortality in the contributing studies.

Results: Two metabolites displayed a dose-response association with self-reported alcohol intake: 2-hydroxy-3-methylbutyric acid and an unidentified compound ($m/z(+)$:231.0839). A 1-SD (\log_2) increase in levels of 2-hydroxy-3-methylbutyric acid was associated with risk of HCC (OR=2.54; 1.51-4.27) and pancreatic cancer (OR=1.43; 1.03-1.99) in EPIC and liver cancer (OR=2.00; 1.44-2.77) and liver disease mortality (OR=1.98; 1.51-2.60) in ATBC. Conversely, a 1-SD (\log_2) increase in questionnaire-derived alcohol intake was not associated with HCC or pancreatic cancer in EPIC or liver cancer in ATBC but was associated with liver disease mortality (OR=2.19; 1.60-2.86) in ATBC.

Conclusions: 2-Hydroxy-3-methylbutyric acid is a candidate biomarker of habitual alcohol intake that may advance the study of alcohol and cancer risk in population-based studies.

Keyword: alcohol intake, untargeted metabolomics, 2-hydroxy-3-methylbutyric acid, biomarkers, EPIC, ATBC.

In 2016, an estimated 2.8 million deaths, corresponding to 6.8% and 2.2% of age-standardized deaths in men and women, respectively, were attributed to alcohol use worldwide [1]. Excessive alcohol consumption is an established risk factor for many acute and chronic health conditions [2], including cancers of the upper aerodigestive tract, female breast, liver, colon, and rectum [3]. However, the relationship of alcohol, particularly light-to-moderate alcohol consumption, with other cancer sites remains controversial [4].

Self-reported alcohol intake is, like other dietary factors, prone to underreporting [5]. Validation studies have shown larger correlations for alcohol intake measured via dietary questionnaire and 24-hour dietary recall than those many other dietary constituents; however, this information may not reflect the level of accuracy since alcohol is a sensitive exposure, making it susceptible to under-reporting across self-reported assessments. Consequently, the extent and distribution of exposure misclassification is unknown [6], and it is likely that observed associations between alcohol use and disease risk in prospective studies are attenuated and that estimates of alcohol-attributable death and disease are underestimated. Biomarkers of liver function and oxidative stress are used to study alcohol-related liver injury and alcoholic liver disease (ALD) [7, 8], but most alcohol consumers, particularly light-to-moderate consumers, will never manifest ALD. There are also biomarkers of recent (e.g., ethyl glucuronide) and heavy alcohol use (e.g., carbohydrate deficient transferrin and phosphatidylethanol (PEth)) [9-11]. However, biomarkers of habitual alcohol use, including light-to-moderate drinking, are needed to better assess alcohol exposure in epidemiological studies and to improve risk estimates for diseases including cancer where modest associations may exist.

Metabolomics is a powerful tool for discovering dietary biomarkers. When used in an untargeted mode, it can detect a wide range of compounds in biological samples including metabolites formed during digestion, metabolism, and microbial fermentation [12, 13], making it well-suited for discovering novel biomarkers of exposure or response to habitual alcohol consumption. Herein we applied a multi-stage design, using untargeted metabolomics and independent discovery and test datasets, to identify serum metabolites associated with habitual alcohol consumption among free-living individuals with a wide range of intake. We then estimated the associations of these candidate

alcohol biomarkers with risk of pancreatic cancer, liver cancers, and liver disease mortality in the European Prospective Investigation into Cancer and Nutrition (EPIC) study and the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC).

METHODS

Study design

EPIC recruitment and study procedures, including dietary assessment methods and blood collection are described extensively elsewhere [14]. Briefly, EPIC is a large cohort study of over half a million men and women recruited between 1992 and 2000 in 23 European centers. Diet, including average daily alcohol intake, over the 12 months before enrolment was assessed by validated country-specific food frequency questionnaires (FFQ) designed to capture local dietary habits with high compliance. Country-specific self-reported alcohol intake was calculated based on the estimated average glass volume and ethanol content for wine, beer, cider, sweet liquor, distilled spirits, or fortified wines, using information collected in standardized 24-hr dietary recalls from a subset of the cohort [15]. The correlation between alcohol intake estimated by FFQ and 24-hour dietary recall was 0.79 [16]. Blood samples were collected and stored at -196°C under liquid nitrogen at the International Agency for Research on Cancer (IARC) for all countries except Sweden (-80°C freezers), and Denmark (-150°C, nitrogen vapor).

Our study included a discovery and two independent test datasets (**Figure 1**). The discovery set (n=454) was nested in the EPIC cross-sectional study [17, 18]. The first test set included control subjects from two EPIC nested case-control studies of hepatocellular carcinoma (HCC; n=128) and pancreatic cancer (n=152) with untargeted metabolomics data [19-21]. The second test set included two nested case-control studies in the ATBC cohort of male Finnish smokers [22]. In ATBC, participants reported on demographics, lifestyle, and medical history via questionnaires and donated a fasting serum sample at baseline, which was stored at -70°C. For this study, we excluded controls (as well as cases) with missing self-reported alcohol intake (n=72) and those with samples that failed laboratory analysis (n=18); of the remaining 864 observations, n=438 were controls .

In EPIC, non-metastatic incident HCC (n=128) and pancreatic cancer (n=152) cases, were matched 1:1 with cancer-free controls on study center, sex, age at blood collection (± 1 year), date (± 6 months) and time of the day (± 2 h) of blood collection, fasting status, and, for women, exogenous hormone use. Follow-up was based on a combination of methods, including health insurance records, registries, and active follow-up [14]. Approval for the EPIC study was obtained from the IARC ethics review board (Lyon, France) and local review bodies of participating institutions. In ATBC, participants were passively followed during the post-intervention period via linkage with the Finnish Cancer Registry and death registry. Liver cancer (n=229) and liver disease mortality (n=248) cases were individually matched 1:1 with controls, selected by incidence density sampling, on baseline age (± 5 years) and serum draw date (± 30 days) [23]. After excluding ATBC cases and controls with missing data, 192 and 199 complete liver cancer and liver disease mortality case-control set remained. Approval for the ATBC study was obtained from the Institutional Review Boards of National Cancer Institute (Bethesda, Maryland), and the National Public Health Institute of Finland. EPIC and ATBC studies were conducted according to the guidelines of the Declaration of Helsinki; all participants provided written informed consent.

Metabolomics analyses

Sample analysis, data pre-processing, matching of features across datasets, and compound identification are described in detail in the **Supplementary Methods**. Briefly, all samples were analyzed by the same laboratory at IARC with a UHPLC-QTOF-MS system (1290 Binary LC system, 6550 QTOF mass spectrometer; Agilent Technologies, Santa Clara, CA) using reversed phase chromatography and electrospray ionization. Raw data were processed using Agilent MassHunter Qualitative analysis B.06.00, ProFinder B.08.00, and Mass Profiler Professional B.12.1 software with Agilent's recursive feature finding procedure. The m/z (mass to charge ratio) values of the features of interest were searched against the Human Metabolome Database (HMDB) [24] and METLIN [25]. Compound identity was confirmed by comparison of chemical standards and representative samples.

Statistical analyses

We used an integrated workflow for metabolomics data analysis [26]. Features detected in <50% of the discovery set samples and background features, (i.e., feature intensities present in all blanks with ratio of geometric mean intensities of non-blank:blank samples <5) were excluded. Feature intensities were \log_2 -transformed. Study participants with >50% missing features and those identified as outliers by a PCA-based approach were excluded [27]. Missing values were imputed within each plate by a K-nearest neighbours method, with K=10 [28]. Last, feature intensities measured across plates within any single batch were normalised by applying a random forest-based approach to correct for unwanted variation [29]. In the EPIC discovery set and test sets, these steps were applied on feature matrices acquired in positive and negative modes separately. In ATBC, these steps were applied on each batch.

In the discovery and test sets, self-reported alcohol intake (g/day) was adjusted for age, sex, country (in EPIC only), body mass index (BMI, kg/m^2), smoking status and intensity, coffee consumption (g/day, log-transformed) via the residual method in linear regression models [30]. Coffee drinking and coffee-associated metabolites have been strongly associated with lower risk of liver cancer and liver disease mortality in ATBC [23, 31]; for consistency, coffee drinking was considered a potential confounder across discovery and test sets. Residuals for feature intensities were also adjusted for well plate number within the analytical batch, position within the plate (row and column indexes), and the study (EPIC HCC or pancreatic cancer) or batch indicator (ATBC) as random effects. We used the principal component partial- R^2 (PC-PR2) method [32] to quantify the contribution of alcohol and potential confounders to the variability of the 67 features intensities that were statistically significantly associated with self-reported alcohol intake in the discovery set [33].

We calculated Pearson correlation coefficients using the residuals for self-reported alcohol intake and for feature intensities; correlations with a false discovery rate (FDR)-corrected p-value <0.05 were considered statistically significant, and each feature in this set (f_i) was carried forward for testing in our multistage design. After the discovery stage, f_1 residual-adjusted correlation coefficients were computed and corrected by the more conservative Bonferroni method. The correlations between f_i features and self-reported alcohol with a p-value <0.05/ f_1 were considered

statistically significant comprised a second set of features (f_2) that were carried forward to the next stage in ATBC. Again, correlations between the residuals of self-reported alcohol intake and of feature intensities were calculated. The linearity of the association between standardized residuals of 2-hydroxy-3-methylbutyric acid and self-reported alcohol intake was evaluated with cubic regression splines with 5 knots [34], by comparing the log-likelihood of models with and without the non-linear terms to a chi-distribution with 2 degrees of freedom.

We estimated odds ratios (OR) and 95% confidence intervals (95% CI) for candidate features and HCC and pancreatic cancer in EPIC and liver cancer and fatal liver disease in ATBC using conditional logistic regression models. In crude models (conditioned on the matching criteria only), multivariable models, and multivariable models additionally adjusting for self-reported alcohol intake, \log_2 -transformed feature intensities were centered and scaled (i.e., mean=0, standard deviation=1) to ensure comparability of OR across different endpoints.

All statistical analyses were performed using the Statistical Analysis Software, release 9.4 (SAS Institute Inc., Cary, NC, USA) and R version 3.6.0 [35].

RESULTS

Population characteristics

Baseline participant characteristics are presented in **Table 1**. In the EPIC discovery set, most participants were women (57.5%) and never (52.2%) or former (26.4%) smokers. In the set of EPIC HCC and pancreatic cancer controls, there was a higher percentage of men (52.7%) and a lower percentage of never smokers (46.2%) than in the discovery set. In the set of ATBC liver cancer and liver disease death controls, all participants were Finnish men and current smokers. Median self-reported alcohol intake was 10.0 g/day, 6.6 g/day, and 11.5 g/day in the EPIC discovery, EPIC and ATBC test sets, respectively.

Biomarker discovery analysis

After excluding participant samples identified as outliers or as having too many missing values, the final discovery set (stage 1) comprised 451 and 452 study participants in positive and

negative ionization mode datasets, respectively. The final EPIC test set (stage 2) comprised 271 and 277 study participants in positive and negative ionization datasets, respectively. Residuals of 205 features in the discovery set were significantly correlated with residuals of self-reported alcohol intake (163 features in positive and 42 features in negative ionization mode; **Figure 1**), with correlation coefficients ranging from -0.29 to 0.50 in log-log plots (**Table S1**).

Of the 205 features in the discovery set, 51 features in positive and 16 features in negative ionization mode ($f_1=67$) matched by mass and retention time with equivalent features in the EPIC test set, and PC-PR2 analyses showed that self-reported alcohol intake explained >7% of variability in the feature intensities ($f_1=67$; **Figure 2**). Residuals of $f_2=10$ features were statistically significantly correlated with residuals of self-reported alcohol intake (**Table 2**). The first two features corresponded to a compound that could not be unequivocally identified, but had an identical mass, isotope pattern, ion formation (mostly $[M+Na]^+$ and $[M+HCOOH-H]^-$) and retention time to ethyl glucoside (HMDB0029968) [37]. However, chromatograms (**Supplementary Methods**) indicated a lack of specificity, and although fragmentation of the $[M+Na]^+$ ion could not be induced, our results suggest the unknown is a combination of ethyl- α -D-glucoside, ethyl- β -D-glucoside, and an additional structural isomer. The remaining eight features corresponded to a single compound, which was confirmed by comparison with an authentic standard as 2-hydroxy-3-methylbutyric acid (HMDB0000407). Residuals of all seven positive ionization mode features selected in the EPIC test set were positively correlated with residuals of self-reported alcohol in the ATBC test set (stage 3; **Table 2**).

For subsequent analyses, the feature with the greatest chromatographic intensity (i.e., main feature) for each metabolite was used (**Table 2**). In each of the three datasets, the residuals of the main features for the two candidate metabolites were significantly correlated, with correlation coefficients ranging from 0.23 in the EPIC discovery set to 0.54 in the ATBC test set. The test for non-linearity with cubic regression splines using restricted regression spline was borderline significant for residuals of 2-hydroxy-3-methylbutyric acid and self-reported alcohol intake ($p=0.06$; **Figure S1**).

Disease risk associations

In multivariable models (**Table 3**), 2-hydroxy-3-methylbutyric acid was associated with increased odds of HCC ($OR_{1-SD}=2.54$; 1.51, 4.27) and pancreatic cancer ($OR_{1-SD}=1.43$; 1.03, 1.99) in EPIC, as well as liver cancer ($OR_{1-SD}=2.00$; 1.44, 2.77) and fatal liver disease ($OR_{1-SD}=2.16$; 1.63, 2.86) in ATBC; associations remained following adjustment for self-reported alcohol intake. The unknown candidate biomarker was associated with increased odds of liver cancer ($OR_{1-SD}=1.70$; 95% CI: 1.29, 2.25) and liver disease mortality ($OR=1.98$; 95% CI: 1.51-2.60) in ATBC, and these associations were also independent of self-reported alcohol intake. However, the unknown was not associated with HCC or pancreatic cancer in EPIC. Self-reported alcohol intake was not associated with HCC ($OR_{1-SD}=0.78$; 95% CI: 0.56, 1.09) or pancreatic cancer risk ($OR_{1-SD}=1.03$; 0.77, 1.39) in EPIC, but was strongly associated with liver disease mortality ($OR_{1-SD}=2.19$; 95% CI, 1.60, 2.98) in ATBC. The alcohol findings are in line with previously published EPIC and ATBC analyses [36-38].

DISCUSSION

Using untargeted metabolomics data from a discovery and two independent sets of cancer-free controls to validate correlations between candidate metabolite feature and self-reported alcohol, we found two serum metabolites that were highly correlated with self-reported habitual alcohol intake. One compound was identified as 2-hydroxy-3-methylbutyric acid; the other remains unknown but is likely a combination of isomers of ethyl glucoside. Of note, ethyl- α -D-glucoside is a known constituent of some alcoholic beverages [39]. Notably, 2-hydroxy-3-methylbutyric acid was strongly associated with HCC and pancreatic cancer risks in EPIC, and with liver cancer and fatal liver disease in ATBC, and these associations remained after adjustment for self-reported alcohol intake. This suggests that 2-hydroxy-3-methylbutyric acid, which is not a constituent or a by-product of alcohol intake, may reflect a relevant biological response to alcohol intake that potentially plays a role in the aetiology of multiple chronic diseases. In contrast, self-reported alcohol intake was only consistently associated with liver disease mortality risk in ATBC. Further research is needed to elucidate the potential metabolic cascade from alcohol drinking to 2-hydroxy-3-methylbutyric acid to disease and to replicate and extend the observed associations. Additionally, targeted metabolomics panels that can simultaneously measure multiple alcohol-related metabolites using authentic standards, including 2-

hydroxy-3-methylbutyric acid and related compounds, should be developed to measure absolute concentrations, which will enable comparisons and pooling of data across studies, supporting replication and improving risk estimation; this is especially important for diseases such as pancreatic cancer, for which the literature is suggestive [40] yet inconsistent [41].

Prior population-based studies have used a targeted or semi-targeted metabolomics approach to identify alcohol-specific metabolomic profiles of self-reported alcohol intake. Three studies, including one in EPIC, used targeted metabolomics, measuring 123 to 163 metabolites, to gain insight into metabolic pathways linking alcohol drinking to human health [42-44]; ten alcohol-metabolite associations were common to all three studies and included phosphatidylcholines (PCs), LysoPCs, acylcarnitines and sphingomyelins. Of note, PCs contribute to the formation of PEth in human tissues [45], which is a known biomarker of recent and heavy alcohol consumption used to diagnose alcohol abuse [46, 47]. A fourth targeted study used nuclear magnetic resonance to evaluate cross-sectional associations of 76 lipids, fatty acids, amino acids, ketone bodies and gluconeogenesis-related metabolites with alcohol consumption [48]. The endogenous metabolites identified by these targeted platforms did not overlap with the compounds most highly correlated with self-reported alcohol intake in our untargeted study, underscoring the breadth of the metabolome and discovery potential of untargeted metabolomics methods.

Metabolomics analyses that limit biomarker discovery to previously annotated compounds have also identified several alcohol-related biomarkers. For example, using prediagnostic serum samples from a nested breast cancer case-control study within a U.S. cohort, self-reported alcohol intake was associated with 16 of the 617 annotated metabolites, including 2-hydroxy-3-methylbutyric acid, 2,3-dihydroxyisovaleric acid (i.e., 2,3-hydroxy-3-methylbutyric acid), ethyl glucuronide and several endogenous metabolites related to androgen metabolism [49]. Other cross-sectional analyses, measuring hundreds of metabolites, also found associations of 2-hydroxy-3-methylbutyric acid, 2,3-dihydroxyisovaleric acid (i.e., 2,3-hydroxy-2-methylbutyric acid) and ethyl glucuronide with self-reported alcohol intake using prediagnostic serum [50, 51]. However, these studies did not test associations in multiple, independent datasets, and estimated correlations in cases and controls combined. One study, which reported using discovery and replication sets, evaluated associations

between self-reported alcohol intake and 356 known metabolites among 1500 African Americans and carried significant metabolites forward for testing in a smaller set of 477 African Americans [52]. This study found that alcohol was associated with five 2-hydroxybutyrate-related metabolites including 2-hydroxy-3-methylbutyric acid [52]. Also using a multi-stage design, a Japanese study of 107 metabolites identified positive associations between 2-hydroxybutyric acid and self-reported alcohol intake in a discovery set and independent test set [53].

The production of 2-hydroxy-3-methylbutyric acid and other hydroxybutyric acid-related metabolites is linked to the rate of hepatic glutathione synthesis, which can increase considerably in response to oxidative stress or detoxification of xenobiotics in the liver [54]. A targeted metabolomics investigation in EPIC found evidence suggesting that glutathione metabolism is involved in the development of HCC [20]. Additionally, 2-hydroxy-3-methylbutyric acid is a product of branched-chain amino acid metabolism, which has been linked to alcohol drinking [53, 55]. Finally, prior research on metabolite variability reported 1-year intraclass correlation coefficients for 2-hydroxy-3-methylbutyric acid (i.e., alpha-hydroxyisovalerate) ranging from 0.76 to 0.49 in independent samples of 60 Chinese women and 30 US men and women, respectively [56], suggesting low to moderate within-subject variability (i.e., good to moderate reliability) over one year.

To our knowledge, this study is unique in its untargeted metabolomics approach without preselected metabolites and its use of a multi-stage design to test the associations of thousands of metabolite features with self-reported alcohol intake in a large discovery dataset and then retest candidate metabolite features in two independent sets of cancer-free controls. By considering nearly 7,000 features, many of which are correlated, we greatly increased the number of potential candidates, but we also incurred stronger penalisation for multiple testing. Consequently, our approach may have missed features that did not meet stringent statistical significance thresholds. A strength of our approach was the use of three large independent datasets although matching features across sets may have resulted in the loss of relevant information. Other potential limitations relate to generalizability, measurement error, and changes in alcohol use over time. Circulating metabolite levels reflect environmental exposures as well as host and microbial metabolism [57-59], and identification of candidate biomarkers that are sufficiently specific to ethanol and generalizable to diverse populations

is challenging. Measurement error, both systematic and random, is inherent to self-reported assessments [60-62] and likely biases association estimates in aetiological studies as well as biomarker discovery studies. Additionally, self-reported alcohol intake and blood measures were assessed in each study at baseline only; therefore, we are unable to account for changes in alcohol intake or metabolites over time. Despite our use of cutting-edge untargeted metabolomics methods, a robust study design, and an aetiological component to evaluate the associations of our candidate biomarkers with disease outcomes, we cannot dismiss the possibility that our findings were impacted by measurement error in self-reported alcohol intake.

In summary, we observed robust correlations between self-reported habitual alcohol intake and 2-hydroxy-3-methylbutyric acid and an unidentified compound in a discovery set and two independent test sets of cancer-free participants. Associations for 2-hydroxy-3-methylbutyric acid with risk of HCC and pancreatic cancer in the EPIC study and with liver cancer in ATBC were stronger than those for either self-reported alcohol intake or the unidentified compound. Both candidate biomarkers were associated with liver endpoints independent of self-reported alcohol intake, indicating value beyond being correlates of intake. In conclusion, 2-hydroxy-3-methylbutyric acid is a promising candidate biomarker for studying the relationship between habitual alcohol intake and health [49-52], but further research, preferably in the context a randomized-controlled trial, is needed to better characterize the relationship between 2-hydroxy-3-methylbutyric acid and alcohol at varying levels of intake.

References

1. Global Burden of Disease Collaborators. Alcohol use and burden for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 2018;392(10152):1015-1035.
2. World Health Organization. *Alcohol fact sheet*. <http://www.who.int/mediacentre/factsheets/fs349/en/>.
3. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Personal habits and indoor combustions. Volume 100 E. A review of human carcinogens. *IARC Monogr Eval Carcinog Risks Hum* 2012;100(Pt E):1-538.
4. World Cancer Research Fund. Continuous Update Project Expert Report 2018. In. *Alcoholic drinks and the risk of cancer*.
5. Klatsky AL, Udaltsova N, Li Y, *et al*. Moderate alcohol intake and cancer: the role of underreporting. *Cancer Causes Control* 2014;25(6):693-9.
6. Kroke A, Klipstein-Grobusch K, Hoffmann K, *et al*. Comparison of self-reported alcohol intake with the urinary excretion of 5-hydroxytryptophol:5-hydroxyindole-3-acetic acid, a biomarker of recent alcohol intake. *Br J Nutr* 2001;85(5):621-7.
7. Das SK, Nayak P, Vasudevan DM. Biochemical markers for alcohol consumption. *Indian J Clin Biochem* 2003;18(2):111-8.
8. Das SK, Vasudevan DM. Biochemical diagnosis of alcoholism. *Indian J Clin Biochem* 2005;20(1):35-42.
9. Peterson K. Biomarkers for alcohol use and abuse - A summary. *Alcohol Research & Health* 2004;28(1):30-37.
10. Torrente MP, Freeman WM, Vrana KE. Protein biomarkers of alcohol abuse. *Expert Rev Proteomics* 2012;9(4):425-36.
11. Helander A, Bottcher M, Dahmen N, *et al*. Elimination Characteristics of the Alcohol Biomarker Phosphatidylethanol (PEth) in Blood during Alcohol Detoxification. *Alcohol and Alcoholism* 2019;54(3):251-257.
12. Scalbert A, Brennan L, Manach C, *et al*. The food metabolome: A window over dietary exposure. *Am J Clin Nutr* 2014;99(6):1286-1308.
13. Edmands WMB, Ferrari P, Rothwell JA, *et al*. Polyphenol metabolome in human urine and its association with intake of polyphenol-rich foods across European countries. *Am J Clin Nutr* 2015;102(4):905-913.
14. 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.
15. Slimani N, Ferrari P, Ocke M, *et al*. 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* 2000;54(12):900-17.
16. Kaaks R, Slimani N, Riboli E. 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* 1997;26 Suppl 1:S26-36.
17. Slimani N, Bingham S, Runswick S, *et al*. Group level validation of protein intakes estimated by 24-hour diet recall and dietary questionnaires against 24-hour urinary nitrogen in the European Prospective Investigation into Cancer and Nutrition (EPIC) calibration study. *Cancer Epidemiol Biomarkers Prev* 2003;12(8):784-95.
18. Rothwell JA, Keski-Rahkonen P, Robinot N, *et al*. A Metabolomic Study of Biomarkers of Habitual Coffee Intake in Four European Countries. *Mol Nutr Food Res* 2019;63(22):e1900659.
19. Stepien M, Keski-Rahkonen P, Kiss A, *et al*. Metabolic perturbations prior to hepatocellular carcinoma diagnosis - Findings from a prospective observational cohort study. *Int J Cancer*. 2021 Feb 1;148(3):609-625.
20. 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. *Int J Cancer* 2016;138(2):348-60.

21. Gasull M, Pumarega J, Kiviranta H, *et al.* Methodological issues in a prospective study on plasma concentrations of persistent organic pollutants and pancreatic cancer risk within the EPIC cohort. *Environ Res* 2019;169:417-433.
22. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. The ATBC Cancer Prevention Study Group. *Ann Epidemiol* 1994;4(1):1-10.
23. Lofftfield E, Rothwell JA, Sinha R, *et al.* Prospective Investigation of Serum Metabolites, Coffee Drinking, Liver Cancer Incidence, and Liver Disease Mortality. *J Natl Cancer Inst* 2020;112(3):286-294.
24. Wishart DS, Feunang YD, Marcu A, *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* 2018;46(D1):D608-D617.
25. Smith CA, O'Maille G, Want EJ, *et al.* METLIN: a metabolite mass spectral database. *Ther Drug Monit* 2005;27(6):747-51.
26. Kirpich AS, Ibarra M, Moskalenko O, *et al.* SECIMTools: a suite of metabolomics data analysis tools. *BMC Bioinformatics* 2018;19(1):151.
27. Edmands WM, Barupal DK, Scalbert A. MetMSLine: an automated and fully integrated pipeline for rapid processing of high-resolution LC-MS metabolomic datasets. *Bioinformatics* 2015;31(5):788-90.
28. Do KT, Wahl S, Raffler J, *et al.* Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* 2018;14(10):128.
29. Fan S, Kind T, Cajka T, *et al.* Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data. *Anal Chem* 2019;91(5):3590-3596.
30. Kleinbaum DG, Kupper LK, Muller KE. *Applied regression analysis and other multivariable methods*. Belmont, CA: Duxbury Press; 1987.
31. Lai GY, Weinstein SJ, Albanes D, *et al.* The association of coffee intake with liver cancer incidence and chronic liver disease mortality in male smokers. *Br J Cancer* 2013;109(5):1344-51.
32. Fages A, Ferrari P, Monni S, *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):1074-1083.
33. Perrier F, Novoloaca A, Ambatipudi S, *et al.* Identifying and correcting epigenetics measurements for systematic sources of variation. *Clin Epigenetics* 2018;10:38.
34. Chambers J, Hastie T, Pregibon D. Statistical Models in S: Chapter 7. Generalized additive models. *Heidelberg, 1990*, p. 317-321. Physica-Verlag HD.
35. R Core Team. R: A language and environment for statistical computing. In: R Foundation for Statistical Computing; 2013.
36. 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):1686-95.
37. Rohrmann S, Linseisen J, Vrieling A, *et al.* Ethanol intake and the risk of pancreatic cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Cancer Causes Control* 2009;20(5):785-94.
38. Schwartz LM, Persson EC, Weinstein SJ, *et al.* Alcohol consumption, one-carbon metabolites, liver cancer and liver disease mortality. *PLoS One* 2013;8(10):e78156.
39. Mishima T, Harino S, Sugita J, *et al.* Plasma kinetics and urine profile of ethyl glucosides after oral administration in the rat. *Biosci Biotechnol Biochem* 2008;72(2):393-7.
40. Naudin S, Li K, Jaouen T, *et al.* Lifetime and baseline alcohol intakes and risk of pancreatic cancer in the European Prospective Investigation into Cancer and Nutrition study. *Int J Cancer* 2018;143(4):801-812.
41. World Cancer Research Fund/American Institute for Cancer Research. Continuous Update Project Expert Report 2018. Diet, nutrition, physical activity and pancreatic cancer. In; 2018.
42. Jaremek M, Yu Z, Mangino M, *et al.* Alcohol-induced metabolomic differences in humans. *Transl Psychiatry* 2013;3:e276.
43. van Roekel EH, Trijsburg L, Assi N, *et al.* Circulating Metabolites Associated with Alcohol Intake in the European Prospective Investigation into Cancer and Nutrition Cohort. *Nutrients* 2018;10(5).

44. Lacruz ME, Kluttig A, Tiller D, *et al.* Cardiovascular Risk Factors Associated With Blood Metabolite Concentrations and Their Alterations During a 4-Year Period in a Population-Based Cohort. *Circ Cardiovasc Genet* 2016;9(6):487-494.
45. Brühl A, Faldum A, Löffelholz K. Degradation of phosphatidylethanol counteracts the apparent phospholipase D-mediated formation in heart and other organs. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 2003;1633(2):84-89.
46. Walther L, de Bejczy A, Lof E, *et al.* Phosphatidylethanol is superior to carbohydrate-deficient transferrin and gamma-glutamyltransferase as an alcohol marker and is a reliable estimate of alcohol consumption level. *Alcohol Clin Exp Res* 2015;39(11):2200-8.
47. Zheng Y, Beck O, Helander A. Method development for routine liquid chromatography-mass spectrometry measurement of the alcohol biomarker phosphatidylethanol (PEth) in blood. *Clin Chim Acta* 2011;412(15-16):1428-35.
48. Wurtz P, Cook S, Wang Q, *et al.* Metabolic profiling of alcohol consumption in 9778 young adults. *Int J Epidemiol* 2016;45(5):1493-1506.
49. Playdon MC, Ziegler RG, Sampson JN, *et al.* Nutritional metabolomics and breast cancer risk in a prospective study. *Am J Clin Nutr* 2017;106(2):637-649.
50. Guertin KA, Moore SC, Sampson JN, *et al.* Metabolomics in nutritional epidemiology: identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations. *Am J Clin Nutr* 2014;100(1):208-17.
51. Playdon MC, Sampson JN, Cross AJ, *et al.* Comparing metabolite profiles of habitual diet in serum and urine. *Am J Clin Nutr* 2016;104(3):776-89.
52. Zheng Y, Yu B, Alexander D, *et al.* Metabolomic patterns and alcohol consumption in African Americans in the Atherosclerosis Risk in Communities Study. *Am J Clin Nutr* 2014;99(6):1470-8.
53. Harada S, Takebayashi T, Kurihara A, *et al.* Metabolomic profiling reveals novel biomarkers of alcohol intake and alcohol-induced liver injury in community-dwelling men. *Environ Health Prev Med* 2016;21(1):18-26.
54. Lord RS, Bralley JA. Clinical applications of urinary organic acids. Part I: Detoxification markers. *Altern Med Rev* 2008;13(3):205-15.
55. Pallister T, Jennings A, Mohny RP, *et al.* Characterizing Blood Metabolomics Profiles Associated with Self-Reported Food Intakes in Female Twins. *PLoS One* 2016;11(6):e0158568.
56. Sampson JN, Boca SM, Shu XO, *et al.* Metabolomics in epidemiology: sources of variability in metabolite measurements and implications. *Cancer Epidemiol Biomarkers Prev* 2013;22(4):631-40.
57. Vippera K, O'Keefe SJ. Intestinal microbes, diet, and colorectal cancer. *Current Colorectal Cancer Reports* 2013;9(1):95-105.
58. Putignani L, Dallapiccola B. Foodomics as part of the host-microbiota-exposome interplay. *J Proteomics* 2016;147:3-20.
59. Shin SY, Fauman EB, Petersen AK, *et al.* An atlas of genetic influences on human blood metabolites. *Nature Genetics* 2014;46(6):543-550.
60. Kipnis V, Subar AF, Midthune D, *et al.* Structure of dietary measurement error: Results of the OPEN biomarker study. *Am J Epidemiol* 2003;158(1):14-21.
61. Prentice RL, Mossavar-Rahmani Y, Huang Y, *et al.* Evaluation and comparison of food records, recalls, and frequencies for energy and protein assessment by using recovery biomarkers. *Am J Epidemiol* 2011;174(5):591-603.
62. Willett W. *Nutritional epidemiology*. Oxford: Oxford University Press; 2013.
63. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 1995;57(1):289-300.

Figure legends

Figure 1. Flowchart of the multi-stage study, displaying features and samples size of the EPIC cross-sectional study that was used as a discovery set (stage 1 and the independent sets of cancer-free controls from EPIC (stage 2) and ATBC (stage 3) (blue box), as well as of the aetiological analyses in nested-case-control studies (red box).

Figure 2. PC-PR2 (Principal Component Partial R^2) analysis to quantify the contribution of potential confounder variables to the variability of the set of $f_1=67$ feature intensities that were statistically significantly associated to alcohol intake in the discovery set.

Table 1. Descriptive statistics of the EPIC and ATBC samples used to identify and confirm associations of metabolite features with self-reported alcohol intake.

	EPIC Discovery (stage 1) ¹ (n=454)	EPIC controls (stage 2) ² (n=280)	ATBC controls (stage 3) ³ (n=438)
Men (%)	42.5	52.7	100
BMI (median kg/m ² ; 10-90 th %)	25.8 (20.9-31.6)	26.6 (20.7-34.1)	26.2 (22.5-31.3)
Age (median years; 10-90 th %)	55.2 (42.5-63.9)	59.4 (49.0-68.6)	56.0 (51.0-63.0)
Smoking status (%)			
Current	18.5	19.2	100
Former	26.4	33.5	
Never	52.2	46.2	
Unknown	2.9	1.1	
Smoking intensity (median cig/day; 10-90 th %)	11.5 (2-26)	15 (4-30)	20 (10-30)
Country (%)			
France	14.5	0.4	
Italy	34.8	18.5	
Spain	-	10.0	
United Kingdom	-	17.1	
The Netherlands	-	10.3	
Greece	12.3	10.7	
Germany	38.3	24.9	
Denmark	-	8.2	
Finland	-	-	100
Alcohol non-drinkers (%) ⁴	8	14	9
Alcohol intake (median g/day; 10 th -90 th %)			
Men	21.4 (1.3-50.4)	14.9 (1.0-51.7)	11.5 (0.2-42.1)
Women	5.2 (0.02-24.9)	2.0 (0.01-23.3)	--
Coffee intake (median g/day; 10 th -90 th %)	146.3 (21.4, 580.2)	190 (3, 857)	550 (220-1,100)

¹EPIC cross-sectional sample;

²Controls from both liver and pancreatic cancer EPIC nested case-control studies;

³Controls from liver cancer and liver disease mortality ATBC nested case-control studies excluding those with missing data on alcohol intake;

⁴Alcohol non-drinkers are considered as those with alcohol intake ≤ 0.1 g/day.

Table 2. Feature-specific intensity and reproducibility (coefficient of variation=CV) in quality control (QC) samples, and adjusted Pearson correlation coefficients (r) with alcohol intake in the discovery and independent test sets.

m/z ⁴	RT ⁵ (min)	Method	Associated metabolite	QC samples ¹ (n=38)		EPIC Discovery (stage 1; n=454) ²			EPIC controls (state 2; n=280) ³		ATBC controls (stage 3; n=438)	
				Mean intensity	CV (%)	r	p-value	q-value ⁶	r	p-value ⁷	r	p-value ⁸
231.0839 ⁹	0.89	RP+	Unknown	58378	18.5	0.41	1.2 x 10 ⁻¹⁹	4.4 x 10 ⁻¹⁶	0.38	7.0 x 10 ⁻¹¹	0.40	6.3 x 10 ⁻¹⁸
253.0925	0.93	RP-	Unknown	11140	13.2	0.39	2.6 x 10 ⁻¹⁸	4.6 x 10 ⁻¹⁵	0.32	3.2 x 10 ⁻⁸	- ¹⁰	-
203.0227 ⁹	2.78	RP+	2-hydroxy-3-methylbutyric acid	204079	14.8	0.26	1.9 x 10 ⁻⁸	2.0 x 10 ⁻⁶	0.24	5.3 x 10 ⁻⁵	0.40	1.1 x 10 ⁻¹⁸
217.9895	2.78	RP+	2-hydroxy-3-methylbutyric acid	36539	11.7	0.30	9.0 x 10 ⁻¹¹	2.1 x 10 ⁻⁸	0.25	2.3 x 10 ⁻⁵	0.38	2.4 x 10 ⁻¹⁶
250.0134	2.78	RP+	2-hydroxy-3-methylbutyric acid	122838	12.5	0.28	9.0 x 10 ⁻¹⁰	1.6 x 10 ⁻⁷	0.27	8.2 x 10 ⁻⁶	0.40	3.5 x 10 ⁻¹⁸
221.0605	2.78	RP+	2-hydroxy-3-methylbutyric acid	56192	11.2	0.28	2.6 x 10 ⁻⁹	3.2 x 10 ⁻⁷	0.25	2.1 x 10 ⁻⁵	0.39	1.9 x 10 ⁻¹⁷
218.9958	2.78	RP+	2-hydroxy-3-methylbutyric acid	115590	11.7	0.28	1.3 x 10 ⁻⁹	2.1 x 10 ⁻⁷	0.26	1.8 x 10 ⁻⁵	0.40	1.7 x 10 ⁻¹⁸
235.0479	2.78	RP+	2-hydroxy-3-methylbutyric acid	34447	15.5	0.20	2.3 x 10 ⁻⁵	1.0 x 10 ⁻³	0.26	2.1 x 10 ⁻⁵	0.38	4.2 x 10 ⁻¹⁶
117.0559	2.78	RP-	2-hydroxy-3-methylbutyric acid	211842	12.1	0.28	1.3 x 10 ⁻⁹	2.2 x 10 ⁻⁷	0.28	2.0 x 10 ⁻⁶	- ¹⁰	-
261.9788	2.78	RP-	2-hydroxy-3-methylbutyric acid	15985	11.9	0.27	7.2 x 10 ⁻⁹	8.3 x 10 ⁻⁷	0.28	2.7 x 10 ⁻⁶	- ¹⁰	-

¹ Quality control samples within the discovery set;

² The analyses of features acquired in positive and negative modes used data from 451 and 452 participants, respectively, after the exclusion of outliers and samples with too many missing values;

³ The analyses of features acquired in positive and negative modes used data from 271 and 277 participants, respectively, after the exclusion of outliers and samples with too many missing values;

⁴ m/z= monoisotopic mass divided by the charge state values, as observed in the discovery set;

⁵ Retention time;

⁶ Q-values associated to False Discovery Rate (FDR) procedure to correct for multiple testing [63], alpha=0.05;

⁷ Threshold for statistical significance corrected with Bonferroni method for multiple testing, equal to 0.0007463 (0.05/f₁, with f₁=67).

⁸ Threshold for statistical significance corrected with Bonferroni method for multiple testing, equal to 0.007 (0.05/f₃, with f₃=7);

⁹ Feature chosen for analysis of disease see Table 3;

¹⁰ Feature not available in ATBC.

Table 3. Crude and adjusted odds ratios (OR, 95% CI) of self-reported alcohol intake (12 g/day) and the main features of the unknown compound and 2-hydroxy-3-methylbutyric acid (per 1-SD) with hepatocellular carcinoma (HCC; 129 case-control sets) and pancreatic cancer (152 case-control sets) in EPIC, and with liver cancer (194 case-control sets) and liver disease mortality (201 case-control sets) in ATBC

	Crude models			Adjusted models ¹			Alcohol-adjusted models ²		
	OR	(95% CI)	p-value	OR	(95% CI)	p-value	OR	(95% CI)	p-value
HCC, EPIC (128 case-control sets)									
Alcohol intake (12g/day)	1.13	(1.00, 1.27)	0.05	1.04	(0.89, 1.20)	0.65			
Alcohol intake (1-SD (log ₂))	0.93	(0.73, 1.20)	0.59	0.78	(0.56, 1.09)	0.14			
Unknown compound (1-SD (log ₂)) ³	1.27	(0.92, 1.76)	0.15	1.01	(0.66, 1.52)	0.98	1.23	(0.75, 2.01)	0.40
2-hydroxy-3-methylbutyric acid (1-SD (log ₂)) ⁴	2.28	(1.52, 3.43)	7.0 x 10 ⁻⁵	2.54	(1.51, 4.27)	4.2 x 10 ⁻⁴	3.12	(1.74, 5.56)	4.2 x 10 ⁻⁴
Pancreatic cancer, EPIC (152 case-control sets)									
Alcohol intake (12g/day)	1.07	(0.92, 1.25)	0.36	1.04	(0.88, 1.24)	0.65			
Alcohol intake (1-SD (log ₂))	1.08	(0.83, 1.40)	0.58	1.03	(0.77, 1.39)	0.83			
Unknown compound (1-SD (log ₂)) ³	1.15	(0.92, 1.46)	0.22	1.10	(0.91, 1.41)	0.48	1.10	(0.83, 1.46)	0.50
2-hydroxy-3-methylbutyric acid (1-SD (log ₂)) ⁴	1.43	(1.07, 1.92)	0.02	1.43	(1.03, 1.99)	0.03	1.46	(1.03, 2.06)	0.03
Liver cancer, ATBC (192 case-control sets)									
Alcohol intake (12g/day)	1.25	(1.09, 1.43)	1.2 x 10 ⁻³	1.17	(1.01, 1.36)	0.03			
Alcohol intake (1-SD (log ₂))	1.33	(1.05, 1.67)	0.016	1.23	(0.94, 1.60)	0.13			
Unknown compound (1-SD (log ₂)) ³	1.34	(1.07, 1.68)	0.01	1.70	(1.29, 2.25)	2.0 x 10 ⁻⁴	1.76	(1.28, 2.41)	5.0 x 10 ⁻⁴
2-hydroxy-3-methylbutyric acid (1-SD (log ₂)) ⁴	2.08	(1.53, 2.82)	2.7 x 10 ⁻⁶	2.00	(1.44, 2.77)	3.4 x 10 ⁻⁵	2.07	(1.43, 2.98)	9.9 x 10 ⁻³
Liver disease mortality, ATBC (199 case-control sets)									
Alcohol intake (12g/day)	1.38	(1.22, 1.55)	1.1 x 10 ⁻⁷	1.32	(1.16, 1.50)	1.6 x 10 ⁻⁵			
Alcohol intake (1-SD (log ₂))	2.37	(1.78, 3.14)	2.8 x 10 ⁻⁸	2.19	(1.60, 2.98)	8.4 x 10 ⁻⁷			
Unknown compound (1-SD (log ₂)) ³	2.11	(1.63, 2.72)	1.0 x 10 ⁻⁸	1.98	(1.51, 2.60)	8.6 x 10 ⁻⁷	1.65	(1.24, 2.20)	7.0 x 10 ⁻⁴
2-hydroxy-3-methylbutyric acid (1-SD (log ₂)) ⁴	2.26	(1.73, 2.95)	2.1 x 10 ⁻⁹	2.16	(1.63, 2.86)	9.6 x 10 ⁻⁸	1.85	(1.38, 2.48)	3.9 x 10 ⁻⁵

¹ Models for hepatocellular carcinoma (HCC) were adjusted for body mass index (BMI, kg/m²), waist circumference (cm), recreational and household physical activity (Met-hours/week), a composite variable for smoking status and intensity (Never, Current: 1-15 cig/day, Current: 16-25 cig/day, Current: 26+ cig/day, Former: quit <=

10 years, Former: quit 11-20 years, Former: quit 20+ years, Current, occasional pipe/cigar/ use, Current/Former: missing, Unknown), level of educational attainment, and coffee intake ((log₂)grams/day); models for pancreatic cancer were adjusted for BMI (kg/m²), sex-specific physical activity categories and the composite variable for smoking status and intensity; ATBC liver cancer and fatal liver disease models were adjusted for age (years), BMI (kg/m²), leisure time physical activity, smoking intensity (cigarettes/day), level of educational attainment, and coffee intake ((log₂)grams/day);

² Models were further adjusted for self-reported alcohol intake ((log₂)grams/day);

³ Unknown compound (m/z=231.0839);

⁴ 2-hydroxy-3-methylbutyric acid (m/z=203.0227).

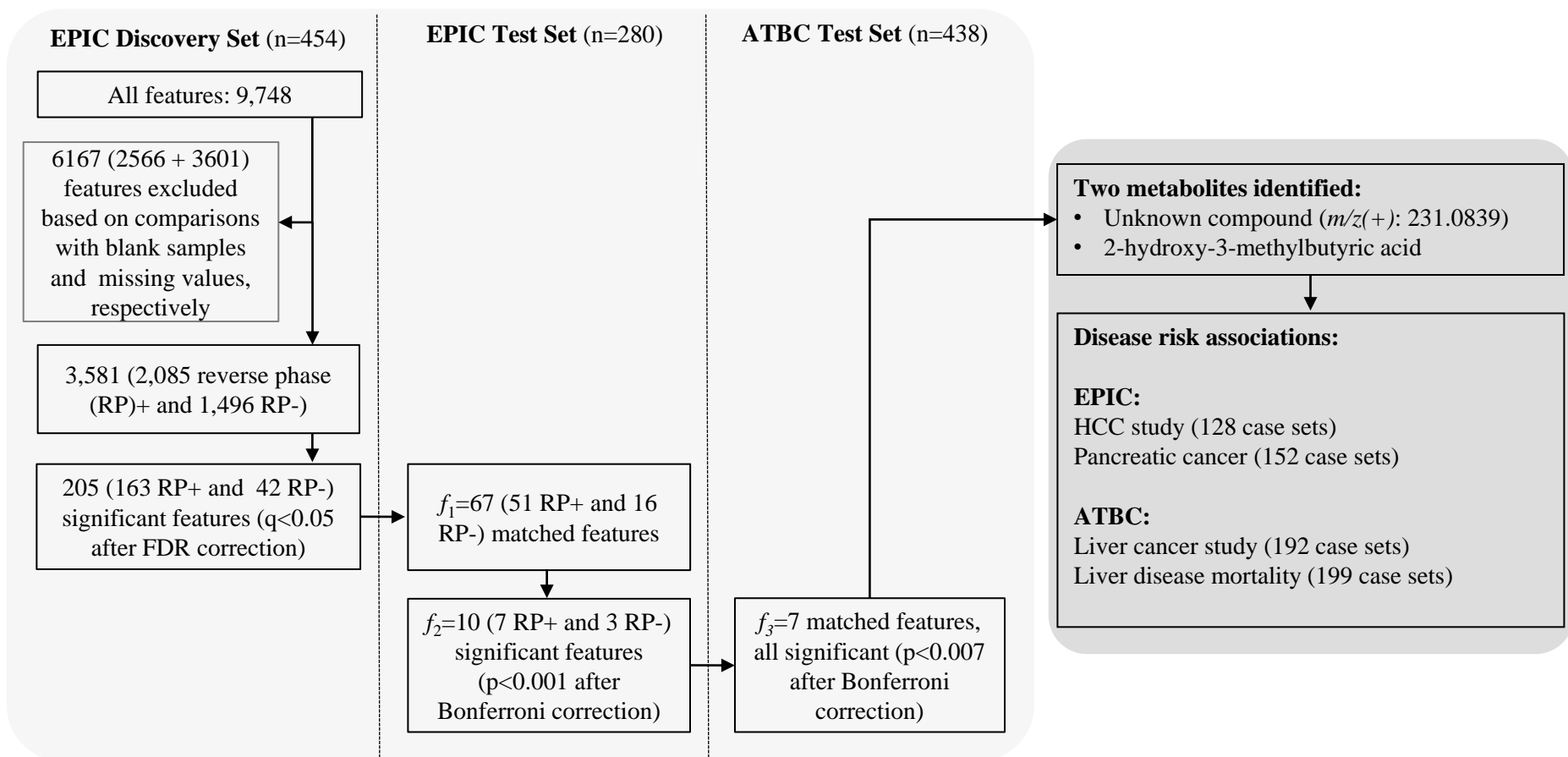
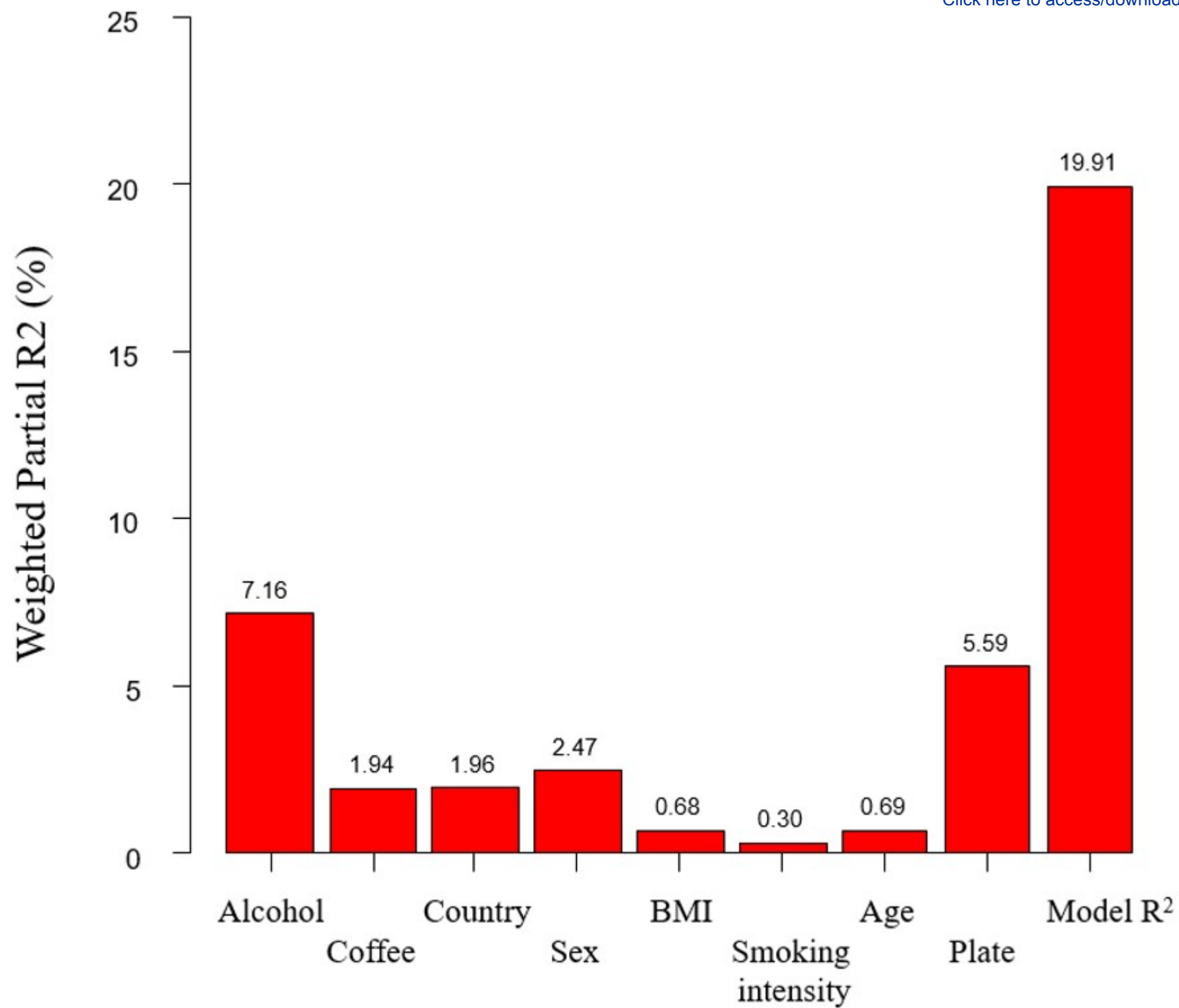


Figure 2





Click here to access/download

Supplemental Material, Other
pub_form-coversheet-peer_reviewed - JNCI
submission.pdf



Click here to access/download

Supplemental Materials

Table S1 + new table S2 + Fig S1 + Supp methods
resubmission.pdf



Novel biomarkers of habitual alcohol intake and associations with risk of pancreatic and liver cancers and liver disease mortality

Erikka Loftfield, PhD ^{1*}; Magdalena Stepien, PhD ^{2*}; Vivian Viallon, PhD ³; Laura Trijbsburg, PhD ³; Joseph Rothwell, PhD ^{2,5,6}; Nivonirina Robinot, MSc ⁴; Carine Biessy, BSc ³; Ingvar A. Bergdahl, PhD ⁷; Stina Bodén, MSc ⁸; Matthias B. Schulze, PhD ^{9,10}; Manuela Bergman, PhD ^{9,10}; Elisabete Weiderpass, MD, MSc, PhD ¹¹; Julie A. Schmidt, PhD ¹²; Raul Zamora-Ros, PhD ¹³; Therese H. Nøst, PhD ¹⁴; Torkjel M Sandanger, PhD ¹⁴; Emily Sonestedt, PhD ¹⁵; Bodil Ohlsson, PhD ¹⁵; Verena Katzke, PhD ¹⁶; Rudolf Kaaks, PhD ¹⁶; Fulvio Ricceri, PhD ¹⁷; Anne Tjønneland, PhD ¹⁸; Christina C. Dahm, PhD ¹⁹; Maria-Jose Sánchez, PhD ^{20,21,22}; Antonia Trichopoulou, PhD ²³; Rosario Tumino, MD, MSc, DLSHTM²⁴; María-Dolores Chirlaque, PhD ^{25,26}; Giovanna Masala, PhD ²⁷; Eva Ardanaz, PhD ^{28,29,30}; Roel Vermeulen, PhD ³¹; Paul Brennan, PhD ³²; Demetrius Albanes, MD¹; Stephanie J. Weinstein, PhD¹; Augustin Scalbert, PhD⁴; Neal D. Freedman, PhD ¹; Marc J. Gunter, PhD ²; Mazda Jenab, PhD ²; Rashmi Sinha, PhD ¹; Pekka Keski-Rahkonen, PhD ^{4 ‡}; Pietro Ferrari, PhD ^{3 ‡†}

* these first authors contributed equally

‡ these senior authors contributed equally

† corresponding author

¹ Metabolic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA.

² Nutritional Epidemiology Group, International Agency for Research on Cancer (IARC-WHO), Lyon, France.

³ Nutritional Epidemiology and Biostatistics Group, International Agency for Research on Cancer (IARC-WHO), Lyon, France.

⁴ Biomarkers Group, International Agency for Research on Cancer (IARC-WHO), Lyon, France.

⁵ Centre for Epidemiology and Population Health (U1018), Generations and Health team, Faculté de Médecine, Université Paris-Saclay, UVSQ, INSERM, Villejuif, France.

⁶ Gustave Roussy, F-94805, Villejuif, France.

⁷ Biobank Research Unit, Umeå University, Sweden.

⁸ Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden.

⁹ Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany.

¹⁰ Institute of Nutritional Science, University of Potsdam, Nuthetal, Germany.

- ¹¹ International Agency for Research on Cancer, World Health Organization.
- ¹² Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom.
- ¹³ Unit of Nutrition and Cancer, Epidemiology Research Program, Catalan Institute of Oncology, Bellvitge Biomedical Research Institute (IDIBELL), Hospitalet de Llobregat (Barcelona), Spain.
- ¹⁴ Department of Community Medicine, UiT- The Arctic University of Norway, Tromsø, Norway.
- ¹⁵ Department of Clinical Sciences in Malmö, Lund University, Malmö, Sweden.
- ¹⁶ Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
- ¹⁷ Department of Clinical and Biological Sciences, University of Turin, Italy; Unit of Epidemiology, Regional Health Service ASL TO3, Grugliasco (TO), Italy.
- ¹⁸ Danish Cancer Society Research Center; University of Copenhagen, Department of Public Health
- ¹⁹ Department of Public Health, Aarhus University, Denmark.
- ²⁰ Escuela Andaluza de Salud Pública (EASP), Granada, Spain; Instituto de Investigación Biosanitaria IBS GRANADA, Granada, Spain.
- ²¹ Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain.
- ²² Department of Preventive Medicine and Public Health, University of Granada, Granada, Spain.
- ²³ Hellenic Health Foundation, Athens, Greece.
- ²⁴ Cancer Registry and Histopathology Department, Provincial Health Authority (ASP 7) Ragusa, Italy.
- ²⁵ Department of Epidemiology, Regional Health Council, IMIB-Arrixaca, Murcia University, Murcia, Spain.
- ²⁶ CIBER in Epidemiology and Public Health (CIBERESP), Madrid, Spain.
- ²⁷ Cancer Risk Factors and Life-Style Epidemiology Unit, Institute for Cancer Research, Prevention and Clinical Network - ISPRO, Florence, Italy.
- ²⁸ Navarra Public Health Institute, Pamplona, Spain.
- ²⁹ IdiSNA, Navarra Institute for Health Research, Pamplona, Spain.
- ³⁰ CIBER Epidemiology and Public Health CIBERESP, Madrid, Spain.
- ³¹ Institute for Risk Assessment Sciences, Division of Environmental Epidemiology, Utrecht University, Utrecht, The Netherlands.
- ³² Genetic Epidemiology Group, International Agency for Research on Cancer (IARC-WHO), Lyon, France.

Notes: The authors have no potential conflicts of interest to disclose. Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors

alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

Acknowledgements: EPIC Umeå investigators thank the Västerbotten Intervention Programme and the County Council of Västerbotten for providing data and samples and acknowledge the contribution from Biobank Sweden, supported by the Swedish Research Council (VR 2017-00650). We thank the National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands, for their contribution and ongoing support to the EPIC Study.

Availability of data and materials: 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>

Funding: This work was supported by the Intramural Research Program of the National Cancer Institute at the National Institutes of Health. For EPIC-Oxford, it is: Cancer Research UK C8221/A29017 and C8221/A19170, and Medical Research Council MR/M012190/1. RZ-R was supported by the “Miguel Servet” program (CP15/00100) from the Institute of Health Carlos III (Co-funded by the European Social Fund (ESF) - ESF investing in your future). This work was supported in part by the French National Cancer Institute (L’Institut National du Cancer; INCA; grant numbers 2009-139 and 2014-1-RT-02-CIRC-1; PI: M. Jenab). For pancreatic cancer in EPIC the work was supported by internal IARC funds.

Abstract

Background: Alcohol is an established risk factor for several cancers, but modest alcohol-cancer associations may be missed due to measurement error in self-reported assessments. ~~The identification of biomarkers of habitual alcohol intake may enhance evidence on the role of alcohol in cancer onset~~ **provide novel insight into the relationship between alcohol and cancer risk.**

Formatted: Highlight

Methods: Untargeted metabolomics was used to identify metabolites correlated with self-reported habitual alcohol intake in a discovery dataset from the European Prospective Investigation into Cancer and Nutrition (EPIC; n=454). Significant correlations were tested in independent datasets of controls from case-control studies nested within EPIC (n=280) and the Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC; n=438) study. Conditional logistic regression was used to estimate odds ratios (OR) and 95% confidence intervals for associations of alcohol-associated metabolites and self-reported alcohol intake with risk of pancreatic cancer, hepatocellular carcinoma (HCC), liver cancer, and liver disease mortality in the contributing studies.

Results: Two metabolites displayed a dose-response association with self-reported alcohol intake: 2-hydroxy-3-methylbutyric acid and an unidentified compound ($m/z(+):231.0839$). A 1-SD (\log_2) increase in levels of 2-hydroxy-3-methylbutyric acid was associated with risk of HCC (OR=2.54; 1.51-4.27) and pancreatic cancer (OR=1.43; 1.03-1.99) in EPIC and liver cancer (OR=2.00; 1.44-2.77) and liver disease mortality (OR=1.98; 1.51-2.60) in ATBC. Conversely, a 1-SD (\log_2) increase in questionnaire-derived alcohol intake was not associated with HCC or pancreatic cancer in EPIC or liver cancer in ATBC but was associated with liver disease mortality (OR=2.19; 1.60-2.86) in ATBC.

Conclusions: 2-Hydroxy-3-methylbutyric acid is a candidate biomarker of habitual alcohol intake that may advance the study of alcohol and cancer risk in population-based studies.

Keyword: alcohol intake, untargeted metabolomics, 2-hydroxy-3-methylbutyric acid, biomarkers, EPIC, ATBC

In 2016, an estimated 2.8 million deaths, corresponding to 6.8% and 2.2% of age-standardized deaths in men and women, respectively, were attributed to alcohol use worldwide [1]. Excessive alcohol consumption is an established risk factor for many acute and chronic health conditions [2], including cancers of the upper aerodigestive tract, female breast, liver, colon, and rectum [3]. However, the relationship of alcohol, particularly light-to-moderate alcohol consumption, with other cancer sites remains controversial [4].

Self-reported alcohol intake is, like other dietary factors, prone to underreporting [5].

Validation studies have shown larger correlations for alcohol intake measured via dietary questionnaire and 24-hour dietary recall than those for most many other dietary constituents; however, this information may not reflect the level of accuracy since alcohol is a sensitive exposure, making it susceptible to under-reporting across self-reported assessments. Consequently, the extent and distribution of exposure misclassification is unknown [6], and it is likely that observed associations between alcohol use and disease risk in prospective studies are attenuated and that estimates of alcohol-attributable death and disease are underestimated. Biomarkers of liver function and oxidative stress are used to study alcohol-related liver injury and alcoholic liver disease (ALD) [7, 8], but most alcohol consumers, particularly light-to-moderate consumers, will never manifest ALD. There are also biomarkers of recent (e.g., ethyl glucuronide) and heavy alcohol use (e.g., carbohydrate deficient transferrin and phosphatidylethanol (PEth)) [9-11]. However, biomarkers of habitual alcohol use, including light-to-moderate drinking, are needed to better assess alcohol exposure in epidemiological studies and to improve risk estimates for diseases including cancer where modest associations may exist.

Metabolomics is a powerful tool for discovering dietary biomarkers. When used in an untargeted mode, it can detect a wide range of compounds in biological samples including metabolites formed during digestion, metabolism, and microbial fermentation [12, 13], making it well-suited for discovering novel biomarkers of exposure or response to habitual alcohol consumption. Herein we applied a multi-stage design, using untargeted metabolomics and independent discovery and test datasets, to identify serum metabolites associated with habitual alcohol consumption among free-living individuals with a wide range of intake. We then estimated the associations of these candidate

alcohol biomarkers with risk of pancreatic cancer, liver cancers, and liver disease mortality in the European Prospective Investigation into Cancer and Nutrition (EPIC) study and the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC).

METHODS

Study design

EPIC recruitment and study procedures, including dietary assessment methods and blood collection are described extensively elsewhere [14]. Briefly, EPIC is a large cohort study of over half a million men and women recruited between 1992 and 2000 in 23 European centers. Diet, including average daily alcohol intake, over the 12 months before enrolment was assessed by validated country-specific food frequency questionnaires (FFQ) designed to capture local dietary habits with high compliance. Country-specific self-reported alcohol intake was calculated based on the estimated average glass volume and ethanol content for wine, beer, cider, sweet liquor, distilled spirits, or fortified wines, using information collected in standardized 24-hr dietary recalls from a subset of the cohort [15]. The correlation between alcohol intake estimated by FFQ and 24-hour dietary recall was 0.79 [16]. Blood samples were collected and stored at -196°C under liquid nitrogen at the International Agency for Research on Cancer (IARC) for all countries except Sweden (-80°C freezers), and Denmark (-150°C, nitrogen vapor).

Our study included a discovery and two independent test datasets (Figure 1). The discovery set (n=454) was nested in the EPIC cross-sectional study [17, 18]. The first test set included control subjects from two EPIC nested case-control studies of hepatocellular carcinoma (HCC; n=128) and pancreatic cancer (n=152) with untargeted metabolomics data [19-21]. The second test set included two nested case-control studies in the ATBC cohort of male Finnish smokers [22]. In ATBC, participants reported on demographics, lifestyle, and medical history via questionnaires and donated a fasting serum sample at baseline, which was stored at -70°C. For this study, we excluded controls (as well as cases) with missing self-reported alcohol intake (n=72) and those with samples that failed

laboratory analysis (n=18); of the remaining 864 observations, n=438 were controls ~~and were included in the second test set.~~

~~Complete case control sets were used to calculate risk estimates.~~ In EPIC, non-metastatic incident HCC (n=128) and pancreatic cancer (n=152) cases, were matched 1:1 with cancer-free controls on study center, sex, age at blood collection (± 1 year), date (± 6 months) and time of the day (± 2 h) of blood collection, fasting status, and, for women, exogenous hormone use. Follow-up was based on a combination of methods, including health insurance records, registries, and active follow-up [14]. Approval for the EPIC study was obtained from the IARC ethics review board (Lyon, France) and local review bodies of participating institutions. In ATBC, participants were passively followed during the post-intervention period via linkage with the Finnish Cancer Registry and death registry. Liver cancer (n=229) and liver disease mortality (n=248) cases were individually matched 1:1 with controls, selected by incidence density sampling, on baseline age (± 5 years) and serum draw date (± 30 days) [23]. ~~After excluding ATBC cases and controls with missing data, 192 and 199 complete case control sets remained in our~~ liver cancer and liver disease mortality ~~case-control set remained~~analytic samples, respectively. Approval for the ATBC study was obtained from the Institutional Review Boards of National Cancer Institute (Bethesda, Maryland), and the National Public Health Institute of Finland. EPIC and ATBC studies were conducted according to the guidelines of the Declaration of Helsinki; all participants provided written informed consent.

Metabolomics analyses

Sample analysis, data pre-processing, matching of features across datasets, and compound identification are described in detail in the **Supplementary Methods**. Briefly, all samples were analyzed by the same laboratory at IARC with a UHPLC-QTOF-MS system (1290 Binary LC system, 6550 QTOF mass spectrometer; Agilent Technologies, Santa Clara, CA) using reversed phase chromatography and electrospray ionization. Raw data were processed using Agilent MassHunter Qualitative analysis B.06.00, ProFinder B.08.00, and Mass Profiler Professional B.12.1 software with Agilent's recursive feature finding procedure. The m/z (mass to charge ratio) values of the features of interest were searched against the Human Metabolome Database (HMDB) [24] and METLIN

[25]. Compound identity was confirmed by comparison of chemical standards and representative samples.

Statistical analyses

We used an integrated workflow for metabolomics data analysis [26]. Features detected in <50% of the discovery set samples and background features, (i.e., feature intensities present in all blanks with ratio of geometric mean intensities of non-blank:blank samples <5) were excluded. Feature intensities were \log_2 -transformed. Study participants with >50% missing features and those identified as outliers by a PCA-based approach were excluded [27]. Missing values were imputed within each plate by a K-nearest neighbours method, with K=10 [28]. Last, feature intensities measured across plates within any single batch were normalised by applying a random forest-based approach to correct for unwanted variation [29]. In the EPIC discovery set and test sets, these steps were applied on feature matrices acquired in positive and negative modes separately. In ATBC, these steps were applied on each batch.

In the discovery and test sets, self-reported alcohol intake (g/day) was adjusted for age, sex, country (in EPIC only), body mass index (BMI, kg/m²), smoking status and intensity, coffee consumption (g/day, log-transformed) via the residual method in linear regression models [30]. Coffee drinking and coffee-associated metabolites have been strongly associated with lower risk of liver cancer and liver disease mortality in ATBC [23, 31]; for consistency, coffee drinking was considered a potential confounder across discovery and test sets. Residuals for feature intensities were also adjusted for well plate number within the analytical batch, position within the plate (row and column indexes), and the study (EPIC HCC or pancreatic cancer control set) or batch indicator (ATBC control set) as random effects. We used the principal component partial-R² (PC-PR2) method [32] to quantify the contribution of alcohol and potential confounders to the variability of the 67 features intensities that were statistically significantly associated with self-reported alcohol intake in the discovery set [33].

We calculated Pearson correlation coefficients using the residuals for self-reported alcohol intake and for feature intensities; correlations with a false discovery rate (FDR)-corrected p-value < 0.05 were considered statistically significant, and each feature in this set (f_i) was carried

forward for testing in our multistage design. After the discovery stage, f_1 residual-adjusted correlation coefficients were computed and corrected by the more conservative Bonferroni method in the set independent set of EPIC controls. The correlations between f_1 features and self-reported alcohol with a p-value $<0.05/f_1$ were considered statistically significant comprised a second set of features (f_2) that were carried forward to the next stage in ATBC. Again, correlations between the residuals of self-reported alcohol intake and of feature intensities were calculated. The linearity of the association between standardized residuals of 2-hydroxy-3-methylbutyric acid and self-reported alcohol intake was evaluated with cubic regression splines with 5 knots [34], by comparing the log-likelihood of models with and without the non-linear terms to a chi-distribution with 2 degrees of freedom.

We estimated odds ratios (OR) and 95% confidence intervals (95% CI) for candidate features and HCC and pancreatic cancer in EPIC and liver cancer and fatal liver disease in ATBC using conditional logistic regression models. In crude models (conditioned on the matching criteria only), multivariable models, adjusting for potential confounders, and multivariable models additionally adjusting for self-reported alcohol intake, \log_2 -transformed feature intensities were centered and scaled (i.e., mean=0, standard deviation=1) to ensure comparability of OR across different endpoints.

All statistical analyses were performed using the Statistical Analysis Software, release 9.4 (SAS Institute Inc., Cary, NC, USA) and R version 3.6.0 [35].

RESULTS

Population characteristics

Baseline participant characteristics are presented in Table 1. In the EPIC discovery set, most participants were women (57.5%) and never (52.2%) or former (26.4%) smokers. In the set of EPIC HCC and pancreatic cancer controls that was used first to test correlations between candidate features from the discovery set and self-reported alcohol intake, there was a higher percentage of men (52.7%) and a lower percentage of never smokers (46.2%) than in the discovery set. In the set of ATBC liver cancer and liver disease death controls that was used second to test correlations between remaining candidate features and self-reported alcohol intake, all participants were Finnish men and current

Formatted: Font: Not Italic, Highlight

Formatted: Font: Not Italic

smokers. Median self-reported alcohol intake was 10.0 g/day, 6.6 g/day, and 11.5 g/day in the EPIC discovery, EPIC and ATBC test sets, respectively.

Biomarker discovery analysis

After excluding participant samples identified as outliers or as having too many missing values, the final discovery set (stage 1) comprised 451 and 452 study participants in positive and negative ionization mode datasets, respectively. The final EPIC test set (stage 2) comprised 271 and 277 study participants in positive and negative ionization datasets, respectively. Residuals of 205 features in the discovery set were significantly correlated with residuals of self-reported alcohol intake (163 features in positive and 42 features in negative ionization mode; **Figure 1**), with correlation coefficients ranging from -0.29 to 0.50 in log-log plots (**Table S1**).

Of the 205 features in the discovery set, 51 features in positive and 16 features in negative ionization mode ($f_1=67$) matched by mass and retention time with equivalent features in the EPIC test set, and PC-PR2 analyses showed that self-reported alcohol intake explained >7% of variability in the feature intensities ($f_1=67$; **Figure 2**). Residuals of $f_2=10$ features were statistically significantly correlated with residuals of self-reported alcohol intake (**Table 2**). The first two features corresponded to a compound that could not be unequivocally identified, but had an identical mass, isotope pattern, ion formation (mostly $[M+Na]^+$ and $[M+HCOOH-H]^+$) and retention time to ethyl glucoside (HMDB0029968) [37]. However, chromatograms (**Supplementary Methods**) indicated a lack of specificity, and although fragmentation of the $[M+Na]^+$ ion could not be induced, our results suggest the unknown is a combination of ethyl- α -D-glucoside, ethyl- β -D-glucoside, and an additional structural isomer. The remaining eight features corresponded to a single compound, which was confirmed by comparison with an authentic standard as 2-hydroxy-3-methylbutyric acid (HMDB0000407). Residuals of all seven positive ionization mode features selected in the EPIC test set were positively correlated with residuals of self-reported alcohol in the ATBC test set (stage 3; **Table 2**).

For subsequent analyses, the feature with the greatest chromatographic intensity (i.e., main feature) for each metabolite was used (**Table 2**). In each of the three datasets, the residuals of the main

features for the two candidate metabolites were significantly correlated, with correlation coefficients ranging from 0.23 in the EPIC discovery set to 0.54 in the ATBC test set. The test for non-linearity with cubic regression splines using restricted regression spline was borderline significant for residuals of 2-hydroxy-3-methylbutyric acid and self-reported alcohol intake (p=0.06; **Figure S1**).

Disease risk associations

In multivariable models (**Table 3**), 2-hydroxy-3-methylbutyric acid was associated with increased odds of HCC (OR_{1-SD}=2.54; 1.51, 4.27) and pancreatic cancer (OR_{1-SD}=1.43; 1.03, 1.99) in EPIC, as well as liver cancer (OR_{1-SD}=2.00; 1.44, 2.77) and fatal liver disease (OR_{1-SD}=2.16; 1.63, 2.86) in ATBC, and these associations remained following adjustment for self-reported alcohol intake. The unknown candidate biomarker was associated with increased odds of liver cancer (OR_{1-SD}=1.70; 95% CI: 1.29, 2.25) and liver disease mortality (OR=1.98; 95% CI: 1.51-2.60) in ATBC, and these associations were also independent of self-reported alcohol intake. However, the unknown was not associated with HCC or pancreatic cancer in EPIC. Self-reported alcohol intake was not associated with HCC (OR_{1-SD}=0.78; 95% CI: 0.56, 1.09) or pancreatic cancer risk (OR_{1-SD}=1.03; 0.77, 1.39) in EPIC, but was strongly associated with liver disease mortality (OR_{1-SD}=2.19; 95% CI, 1.60, 2.98) in ATBC. The alcohol findings are in line with previously published EPIC and ATBC analyses [36-38].

DISCUSSION

Using untargeted metabolomics data from a discovery and two independent sets of cancer-free controls to validate correlations between candidate metabolite feature and self-reported alcohol, we found two serum metabolites that were highly correlated with self-reported habitual alcohol intake. One compound was identified as 2-hydroxy-3-methylbutyric acid; the other remains unknown but is likely a combination of isomers of ethyl glucoside. Of note, ethyl- α -D-glucoside is a known constituent of some alcoholic beverages [39]. Notably, 2-hydroxy-3-methylbutyric acid was strongly associated with HCC and pancreatic cancer risks in EPIC, and with liver cancer and fatal liver disease in ATBC, and these associations remained even after adjustment for self-reported alcohol intake. This suggests that 2-hydroxy-3-methylbutyric acid, which is not a constituent or a by-product of alcohol

intake, may reflect a relevant biological response to alcohol intake that potentially plays a role in the aetiology of multiple chronic diseases. In contrast, self-reported alcohol intake was only consistently associated with ~~risk of~~ liver disease mortality ~~risk~~ in ATBC. Further research is needed to elucidate the potential metabolic cascade from alcohol drinking to 2-hydroxy-3-methylbutyric acid to disease and to replicate and extend the observed associations ~~between higher levels of 2-hydroxy-3-methylbutyric acid and greater risk of pancreatic cancer and liver endpoints~~. Additionally, targeted metabolomics panels that can simultaneously measure multiple alcohol-related metabolites using authentic standards, including 2-hydroxy-3-methylbutyric acid and related compounds, should be developed to measure absolute concentrations, which will enable comparisons and pooling of data across studies, supporting replication and improving risk estimation; this is especially important for diseases such as pancreatic cancer, for which the literature is suggestive [40] yet inconsistent [41].

Formatted: Highlight

Prior population-based studies have used a targeted or semi-targeted metabolomics approach to identify alcohol-specific metabolomic profiles of self-reported alcohol intake. Three studies, including one in EPIC, used targeted metabolomics, measuring 123 to 163 metabolites, to gain insight into metabolic pathways linking alcohol drinking to human health [42-44]; ten alcohol-metabolite associations were common to all three studies and included phosphatidylcholines (PCs), LysoPCs, acylcarnitines and sphingomyelins. Of note, PCs contribute to the formation of PEth in human tissues [45], which is a known biomarker of recent and heavy alcohol consumption used to diagnose alcohol abuse [46, 47]. A fourth targeted study used nuclear magnetic resonance to evaluate cross-sectional associations of 76 lipids, fatty acids, amino acids, ketone bodies and gluconeogenesis-related metabolites with alcohol consumption [48]. The endogenous metabolites identified by these targeted platforms did not overlap with the compounds most highly correlated with self-reported alcohol intake in our untargeted study, underscoring the breadth of the metabolome and discovery potential of untargeted metabolomics methods.

Metabolomics analyses that limit biomarker discovery to previously annotated compounds have also identified several alcohol-related biomarkers. For example, using prediagnostic serum samples from a nested breast cancer case-control study within a U.S. cohort, self-reported alcohol intake was associated with 16 of the 617 annotated metabolites, including 2-hydroxy-3-methylbutyric

acid, 2,3-dihydroxyisovaleric acid (i.e., 2,3-hydroxy-3-methylbutyric acid), ethyl glucuronide and several endogenous metabolites related to androgen metabolism [49]. Other cross-sectional analyses, measuring hundreds of metabolites, also found associations of 2-hydroxy-3-methylbutyric acid, 2,3-dihydroxyisovaleric acid (i.e., 2,3-hydroxy-2-methylbutyric acid) and ethyl glucuronide with self-reported alcohol intake using prediagnostic serum [50, 51]. However, these studies did not test associations in multiple, independent datasets, and estimated correlations in cases and controls combined rather than in controls only. One study, which reported using discovery and replication sets, evaluated associations between self-reported alcohol intake and 356 known metabolites among 1500 African Americans and carried significant metabolites forward for testing in a smaller set of 477 African Americans [52]. This study found that alcohol was associated with five 2-hydroxybutyrate-related metabolites including 2-hydroxy-3-methylbutyric acid [52]. Also using a multi-stage design, a Japanese study of 107 metabolites identified positive associations between 2-hydroxybutyric acid and self-reported alcohol intake in a discovery set and independent test set of Japanese men [53].

The production of 2-hydroxy-3-methylbutyric acid and other hydroxybutyric acid-related metabolites is linked to the rate of hepatic glutathione synthesis, which can increase considerably in response to oxidative stress or detoxification of xenobiotics in the liver [54]. A targeted metabolomics investigation in EPIC found evidence suggesting that glutathione metabolism is involved in the development of HCC [20]. Additionally, 2-hydroxy-3-methylbutyric acid is a product of branched-chain amino acid metabolism, which has been linked to alcohol drinking [53, 55]. Finally, prior research on metabolite variability in free living populations over time reported 1-year intraclass correlation coefficients for 2-hydroxy-3-methylbutyric acid (i.e., alpha-hydroxyisovalerate) ranging from 0.76 to 0.49 in independent samples of 60 Chinese women and 30 US men and women, respectively [56], suggesting low to moderate within-subject variability (i.e., good to moderate reliability) over one year.

To our knowledge, this study is unique in its untargeted metabolomics approach without preselected metabolites and its use of a multi-stage design to test the associations of thousands of metabolite features with self-reported alcohol intake in a large discovery dataset and then retest candidate metabolite features in two independent sets of cancer-free controls. By considering nearly

7,000 features, many of which are correlated, we greatly increased the number of potential candidates, but we also incurred stronger penalisation for multiple testing. Consequently, our approach may have missed features that did not meet stringent ~~statistical significance threshold~~~~thresholds for statistical significance~~. A strength of our approach was the use of **three large independent datasets** although matching features ~~by mass and retention time~~ across sets may have resulted in the loss of relevant information. Other potential limitations relate to generalizability, measurement error, **and changes in alcohol use over time**. Circulating metabolite levels reflect environmental exposures as well as host and microbial metabolism [57-59], and identification of candidate biomarkers that are sufficiently specific to ethanol and generalizable to diverse populations is challenging. Measurement error, both systematic and random, is inherent to self-reported assessments [60-62], ~~including alcohol intake~~, and likely biases association estimates in ~~not only~~ aetiological studies, ~~but also in as well as~~ biomarker discovery studies. **Additionally, self-reported alcohol intake and blood measures were assessed in each study at baseline only; therefore, we are unable to account for changes in alcohol intake or metabolites over time**. Despite our use of cutting-edge untargeted metabolomics methods, a robust study design, and an aetiological component to evaluate the associations of our candidate biomarkers with disease outcomes, we cannot dismiss the possibility that our findings were impacted by measurement error in self-reported alcohol intake.

In summary, we observed robust correlations between self-reported habitual alcohol intake and 2-hydroxy-3-methylbutyric acid and an unidentified compound in a discovery set and two independent **test** sets of cancer-free participants. Associations for 2-hydroxy-3-methylbutyric acid with risk of HCC and pancreatic cancer in the EPIC study and with liver cancer in ATBC were stronger than those for either self-reported alcohol intake or the unidentified compound, **and b. Both candidate biomarkers were associated with liver endpoints independent of self-reported alcohol intake, indicating value beyond being correlates of self-reported alcohol intake**. In conclusion, 2-hydroxy-3-methylbutyric acid is a promising candidate biomarker for studying the relationship between habitual alcohol intake and health [49-52], but further research, preferably in the context a randomized-controlled trial, is needed to better characterize the relationship between 2-hydroxy-3-methylbutyric acid and alcohol at varying levels of intake.

References

1. Global Burden of Disease Collaborators. Alcohol use and burden for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 2018;392(10152):1015-1035.
2. World Health Organization. *Alcohol fact sheet*. <http://www.who.int/mediacentre/factsheets/fs349/en/>.
3. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Personal habits and indoor combustions. Volume 100 E. A review of human carcinogens. *IARC Monogr Eval Carcinog Risks Hum* 2012;100(Pt E):1-538.
4. World Cancer Research Fund. Continuous Update Project Expert Report 2018. In. *Alcoholic drinks and the risk of cancer*.
5. Klatsky AL, Udaltsova N, Li Y, *et al*. Moderate alcohol intake and cancer: the role of underreporting. *Cancer Causes Control* 2014;25(6):693-9.
6. Kroke A, Klipstein-Grobusch K, Hoffmann K, *et al*. Comparison of self-reported alcohol intake with the urinary excretion of 5-hydroxytryptophol:5-hydroxyindole-3-acetic acid, a biomarker of recent alcohol intake. *Br J Nutr* 2001;85(5):621-7.
7. Das SK, Nayak P, Vasudevan DM. Biochemical markers for alcohol consumption. *Indian J Clin Biochem* 2003;18(2):111-8.
8. Das SK, Vasudevan DM. Biochemical diagnosis of alcoholism. *Indian J Clin Biochem* 2005;20(1):35-42.
9. Peterson K. Biomarkers for alcohol use and abuse - A summary. *Alcohol Research & Health* 2004;28(1):30-37.
10. Torrente MP, Freeman WM, Vrana KE. Protein biomarkers of alcohol abuse. *Expert Rev Proteomics* 2012;9(4):425-36.
11. Helander A, Bottcher M, Dahmen N, *et al*. Elimination Characteristics of the Alcohol Biomarker Phosphatidylethanol (PEth) in Blood during Alcohol Detoxification. *Alcohol and Alcoholism* 2019;54(3):251-257.
12. Scalbert A, Brennan L, Manach C, *et al*. The food metabolome: A window over dietary exposure. *Am J Clin Nutr* 2014;99(6):1286-1308.
13. Edmands WMB, Ferrari P, Rothwell JA, *et al*. Polyphenol metabolome in human urine and its association with intake of polyphenol-rich foods across European countries. *Am J Clin Nutr* 2015;102(4):905-913.
14. 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.
15. Slimani N, Ferrari P, Ocke M, *et al*. 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* 2000;54(12):900-17.
16. Kaaks R, Slimani N, Riboli E. 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* 1997;26 Suppl 1:S26-36.
17. Slimani N, Bingham S, Runswick S, *et al*. Group level validation of protein intakes estimated by 24-hour diet recall and dietary questionnaires against 24-hour urinary nitrogen in the European Prospective Investigation into Cancer and Nutrition (EPIC) calibration study. *Cancer Epidemiol Biomarkers Prev* 2003;12(8):784-95.
18. Rothwell JA, Keski-Rahkonen P, Robinot N, *et al*. A Metabolomic Study of Biomarkers of Habitual Coffee Intake in Four European Countries. *Mol Nutr Food Res* 2019;63(22):e1900659.
19. Stepien M, Keski-Rahkonen P, Kiss A, *et al*. Metabolic perturbations prior to hepatocellular carcinoma diagnosis - Findings from a prospective observational cohort study. *Int J Cancer*. 2021 Feb 1;148(3):609-625.
20. 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. *Int J Cancer* 2016;138(2):348-60.

21. Gasull M, Pumarega J, Kiviranta H, *et al.* Methodological issues in a prospective study on plasma concentrations of persistent organic pollutants and pancreatic cancer risk within the EPIC cohort. *Environ Res* 2019;169:417-433.
22. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. The ATBC Cancer Prevention Study Group. *Ann Epidemiol* 1994;4(1):1-10.
23. Loftfield E, Rothwell JA, Sinha R, *et al.* Prospective Investigation of Serum Metabolites, Coffee Drinking, Liver Cancer Incidence, and Liver Disease Mortality. *J Natl Cancer Inst* 2020;112(3):286-294.
24. Wishart DS, Feunang YD, Marcu A, *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* 2018;46(D1):D608-D617.
25. Smith CA, O'Maille G, Want EJ, *et al.* METLIN: a metabolite mass spectral database. *Ther Drug Monit* 2005;27(6):747-51.
26. Kirpich AS, Ibarra M, Moskalenko O, *et al.* SECIMTools: a suite of metabolomics data analysis tools. *BMC Bioinformatics* 2018;19(1):151.
27. Edmands WM, Barupal DK, Scalbert A. MetMSLine: an automated and fully integrated pipeline for rapid processing of high-resolution LC-MS metabolomic datasets. *Bioinformatics* 2015;31(5):788-90.
28. Do KT, Wahl S, Raffler J, *et al.* Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* 2018;14(10):128.
29. Fan S, Kind T, Cajka T, *et al.* Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data. *Anal Chem* 2019;91(5):3590-3596.
30. Kleinbaum DG, Kupper LK, Muller KE. *Applied regression analysis and other multivariable methods*. Belmont, CA: Duxbury Press; 1987.
31. Lai GY, Weinstein SJ, Albanes D, *et al.* The association of coffee intake with liver cancer incidence and chronic liver disease mortality in male smokers. *Br J Cancer* 2013;109(5):1344-51.
32. Fages A, Ferrari P, Monni S, *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):1074-1083.
33. Perrier F, Novoloaca A, Ambatipudi S, *et al.* Identifying and correcting epigenetics measurements for systematic sources of variation. *Clin Epigenetics* 2018;10:38.
34. Chambers J, Hastie T, Pregibon D. *Statistical Models in S: Chapter 7. Generalized additive models*. Heidelberg, 1990, p. 317-321. Physica-Verlag HD.
35. R Core Team. *R: A language and environment for statistical computing*. In: R Foundation for Statistical Computing; 2013.
36. 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):1686-95.
37. Rohrmann S, Linseisen J, Vrieling A, *et al.* Ethanol intake and the risk of pancreatic cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Cancer Causes Control* 2009;20(5):785-94.
38. Schwartz LM, Persson EC, Weinstein SJ, *et al.* Alcohol consumption, one-carbon metabolites, liver cancer and liver disease mortality. *PLoS One* 2013;8(10):e78156.
39. Mishima T, Harino S, Sugita J, *et al.* Plasma kinetics and urine profile of ethyl glucosides after oral administration in the rat. *Biosci Biotechnol Biochem* 2008;72(2):393-7.
40. Naudin S, Li K, Jaouen T, *et al.* Lifetime and baseline alcohol intakes and risk of pancreatic cancer in the European Prospective Investigation into Cancer and Nutrition study. *Int J Cancer* 2018;143(4):801-812.
41. World Cancer Research Fund/American Institute for Cancer Research. *Continuous Update Project Expert Report 2018. Diet, nutrition, physical activity and pancreatic cancer*. In; 2018.
42. Jaremek M, Yu Z, Mangino M, *et al.* Alcohol-induced metabolomic differences in humans. *Transl Psychiatry* 2013;3:e276.
43. van Roekel EH, Trijsburg L, Assi N, *et al.* Circulating Metabolites Associated with Alcohol Intake in the European Prospective Investigation into Cancer and Nutrition Cohort. *Nutrients* 2018;10(5).

44. Lacruz ME, Kluttig A, Tiller D, *et al.* Cardiovascular Risk Factors Associated With Blood Metabolite Concentrations and Their Alterations During a 4-Year Period in a Population-Based Cohort. *Circ Cardiovasc Genet* 2016;9(6):487-494.
45. Brühl A, Faldum A, Löffelholz K. Degradation of phosphatidylethanol counteracts the apparent phospholipase D-mediated formation in heart and other organs. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 2003;1633(2):84-89.
46. Walther L, de Bejczy A, Lof E, *et al.* Phosphatidylethanol is superior to carbohydrate-deficient transferrin and gamma-glutamyltransferase as an alcohol marker and is a reliable estimate of alcohol consumption level. *Alcohol Clin Exp Res* 2015;39(11):2200-8.
47. Zheng Y, Beck O, Helander A. Method development for routine liquid chromatography-mass spectrometry measurement of the alcohol biomarker phosphatidylethanol (PEth) in blood. *Clin Chim Acta* 2011;412(15-16):1428-35.
48. Wurtz P, Cook S, Wang Q, *et al.* Metabolic profiling of alcohol consumption in 9778 young adults. *Int J Epidemiol* 2016;45(5):1493-1506.
49. Playdon MC, Ziegler RG, Sampson JN, *et al.* Nutritional metabolomics and breast cancer risk in a prospective study. *Am J Clin Nutr* 2017;106(2):637-649.
50. Guertin KA, Moore SC, Sampson JN, *et al.* Metabolomics in nutritional epidemiology: identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations. *Am J Clin Nutr* 2014;100(1):208-17.
51. Playdon MC, Sampson JN, Cross AJ, *et al.* Comparing metabolite profiles of habitual diet in serum and urine. *Am J Clin Nutr* 2016;104(3):776-89.
52. Zheng Y, Yu B, Alexander D, *et al.* Metabolomic patterns and alcohol consumption in African Americans in the Atherosclerosis Risk in Communities Study. *Am J Clin Nutr* 2014;99(6):1470-8.
53. Harada S, Takebayashi T, Kurihara A, *et al.* Metabolomic profiling reveals novel biomarkers of alcohol intake and alcohol-induced liver injury in community-dwelling men. *Environ Health Prev Med* 2016;21(1):18-26.
54. Lord RS, Bralley JA. Clinical applications of urinary organic acids. Part I: Detoxification markers. *Altern Med Rev* 2008;13(3):205-15.
55. Pallister T, Jennings A, Mohny RP, *et al.* Characterizing Blood Metabolomics Profiles Associated with Self-Reported Food Intakes in Female Twins. *PLoS One* 2016;11(6):e0158568.
56. Sampson JN, Boca SM, Shu XO, *et al.* Metabolomics in epidemiology: sources of variability in metabolite measurements and implications. *Cancer Epidemiol Biomarkers Prev* 2013;22(4):631-40.
57. Vipperla K, O'Keefe SJ. Intestinal microbes, diet, and colorectal cancer. *Current Colorectal Cancer Reports* 2013;9(1):95-105.
58. Putignani L, Dallapiccola B. Foodomics as part of the host-microbiota-exposome interplay. *J Proteomics* 2016;147:3-20.
59. Shin SY, Fauman EB, Petersen AK, *et al.* An atlas of genetic influences on human blood metabolites. *Nature Genetics* 2014;46(6):543-550.
60. Kipnis V, Subar AF, Midthune D, *et al.* Structure of dietary measurement error: Results of the OPEN biomarker study. *Am J Epidemiol* 2003;158(1):14-21.
61. Prentice RL, Mossavar-Rahmani Y, Huang Y, *et al.* Evaluation and comparison of food records, recalls, and frequencies for energy and protein assessment by using recovery biomarkers. *Am J Epidemiol* 2011;174(5):591-603.
62. Willett W. *Nutritional epidemiology*. Oxford: Oxford University Press; 2013.
63. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 1995;57(1):289-300.

Figure legends

Figure 1. Flowchart of the multi-stage study, displaying features and samples size of the EPIC cross-sectional study that was used as a discovery set (stage 1) and the independent sets of cancer-free controls from EPIC (stage 2) and ATBC (stage 3) (blue box), as well as of the aetiological analyses in nested-case-control studies (red box).

Figure 2. PC-PR2 (Principal Component Partial R^2) analysis to quantify the contribution of potential confounder variables to the variability of the set of $f_1=67$ feature intensities that were statistically significantly associated to alcohol intake in the discovery set.

Table 1. Descriptive statistics of the EPIC and ATBC samples used to identify and confirm associations of metabolite features with self-reported alcohol intake.

	EPIC Discovery (stage 1) ¹ (n=454)	EPIC controls (stage 2) ² (n=280)	ATBC controls (stage 3) ³ (n=438)
Men (%)	42.5	52.7	100
BMI (median kg/m ² ; 10-90 th %)	25.8 (20.9-31.6)	26.6 (20.7-34.1)	26.2 (22.5-31.3)
Age (median years; 10-90 th %)	55.2 (42.5-63.9)	59.4 (49.0-68.6)	56.0 (51.0-63.0)
Smoking status (%)			
Current	18.5	19.2	100
Former	26.4	33.5	
Never	52.2	46.2	
Unknown	2.9	1.1	
Smoking intensity (median cig/day; 10-90 th %)	11.5 (2-26)	15 (4-30)	20 (10-30)
Country (%)			
France	14.5	0.4	
Italy	34.8	18.5	
Spain	-	10.0	
United Kingdom	-	17.1	
The Netherlands	-	10.3	
Greece	12.3	10.7	
Germany	38.3	24.9	
Denmark	-	8.2	
Finland	-	-	100
Alcohol non-drinkers (%) ⁴	8	14	9
Alcohol intake (median g/day; 10 th -90 th %)			
Men	21.4 (1.3-50.4)	14.9 (1.0-51.7)	11.5 (0.2-42.1)
Women	5.2 (0.02-24.9)	2.0 (0.01-23.3)	--
Coffee intake (median g/day; 10 th -90 th %)	146.3 (21.4, 580.2)	190 (3, 857)	550 (220-1,100)

¹EPIC cross-sectional sample;

²Controls from both liver and pancreatic cancer EPIC nested case-control studies;

³Controls from liver cancer and liver disease mortality ATBC nested case-control studies excluding those with missing data on alcohol intake;

⁴Alcohol non-drinkers are considered as those with alcohol intake ≤ 0.1 g/day.

Table 2. Feature-specific intensity and reproducibility (coefficient of variation=CV) in quality control (QC) samples, and adjusted Pearson correlation coefficients (r) with alcohol intake in the discovery and **independent test** sets.

m/z ⁴	RT ⁵ (min)	Method	Associated metabolite	QC samples ¹ (n=38)		EPIC Discovery (stage 1; n=454) ²			EPIC controls (state 2; n=280) ³		ATBC controls (stage 3; n=438)	
				Mean intensity	CV (%)	r	p-value	q-value ⁶	r	p-value ⁷	r	p-value ⁸
231.0839 ⁹	0.89	RP+	Unknown	58378	18.5	0.41	1.2 x 10 ⁻¹⁹	4.4 x 10 ⁻¹⁶	0.38	7.0 x 10 ⁻¹¹	0.40	6.3 x 10 ⁻¹⁸
253.0925	0.93	RP-	Unknown	11140	13.2	0.39	2.6 x 10 ⁻¹⁸	4.6 x 10 ⁻¹⁵	0.32	3.2 x 10 ⁻⁸	- ¹⁰	-
203.0227 ⁹	2.78	RP+	2-hydroxy-3-methylbutyric acid	204079	14.8	0.26	1.9 x 10 ⁻⁸	2.0 x 10 ⁻⁶	0.24	5.3 x 10 ⁻⁵	0.40	1.1 x 10 ⁻¹⁸
217.9895	2.78	RP+	2-hydroxy-3-methylbutyric acid	36539	11.7	0.30	9.0 x 10 ⁻¹¹	2.1 x 10 ⁻⁸	0.25	2.3 x 10 ⁻⁵	0.38	2.4 x 10 ⁻¹⁶
250.0134	2.78	RP+	2-hydroxy-3-methylbutyric acid	122838	12.5	0.28	9.0 x 10 ⁻¹⁰	1.6 x 10 ⁻⁷	0.27	8.2 x 10 ⁻⁶	0.40	3.5 x 10 ⁻¹⁸
221.0605	2.78	RP+	2-hydroxy-3-methylbutyric acid	56192	11.2	0.28	2.6 x 10 ⁻⁹	3.2 x 10 ⁻⁷	0.25	2.1 x 10 ⁻⁵	0.39	1.9 x 10 ⁻¹⁷
218.9958	2.78	RP+	2-hydroxy-3-methylbutyric acid	115590	11.7	0.28	1.3 x 10 ⁻⁹	2.1 x 10 ⁻⁷	0.26	1.8 x 10 ⁻⁵	0.40	1.7 x 10 ⁻¹⁸
235.0479	2.78	RP+	2-hydroxy-3-methylbutyric acid	34447	15.5	0.20	2.3 x 10 ⁻⁵	1.0 x 10 ⁻³	0.26	2.1 x 10 ⁻⁵	0.38	4.2 x 10 ⁻¹⁶
117.0559	2.78	RP-	2-hydroxy-3-methylbutyric acid	211842	12.1	0.28	1.3 x 10 ⁻⁹	2.2 x 10 ⁻⁷	0.28	2.0 x 10 ⁻⁶	- ¹⁰	-
261.9788	2.78	RP-	2-hydroxy-3-methylbutyric acid	15985	11.9	0.27	7.2 x 10 ⁻⁹	8.3 x 10 ⁻⁷	0.28	2.7 x 10 ⁻⁶	- ¹⁰	-

¹ Quality control samples within the discovery set;

² The analyses of features acquired in positive and negative modes used data from 451 and 452 participants, respectively, after the exclusion of outliers and samples with too many missing values;

³ The analyses of features acquired in positive and negative modes used data from 271 and 277 participants, respectively, after the exclusion of outliers and samples with too many missing values;

⁴ m/z= monoisotopic mass divided by the charge state values, as observed in the discovery set;

⁵ Retention time;

⁶ Q-values associated to False Discovery Rate (FDR) procedure to correct for multiple testing [63], alpha=0.05;

⁷ Threshold for statistical significance corrected with Bonferroni method for multiple testing, equal to 0.0007463 (0.05/f₁, with f₁=67).

⁸ Threshold for statistical significance corrected with Bonferroni method for multiple testing, equal to 0.007 (0.05/f₃, with f₃=7);

⁹ Feature chosen for analysis of disease see Table 3;

¹⁰ Feature not available in ATBC.

Table 3. Crude and adjusted odds ratios (OR, 95% CI) of self-reported alcohol intake (12 g/day) and the main features of the unknown compound and 2-hydroxy-3-methylbutyric acid (per 1-SD) with hepatocellular carcinoma (HCC; 129 case-control sets) and pancreatic cancer (152 case-control sets) in EPIC, and with liver cancer (194 case-control sets) and liver disease mortality (201 case-control sets) in ATBC

	Crude models			Adjusted models ¹			Alcohol-adjusted models ²		
	OR	(95% CI)	p-value	OR	(95% CI)	p-value	OR	(95% CI)	p-value
HCC, EPIC (128 case-control sets)									
Alcohol intake (12g/day)	1.13	(1.00, 1.27)	0.05	1.04	(0.89, 1.20)	0.65			
Alcohol intake (1-SD (log ₂))	0.93	(0.73, 1.20)	0.59	0.78	(0.56, 1.09)	0.14			
Unknown compound (1-SD (log ₂)) [‡]	1.27	(0.92, 1.76)	0.15	1.01	(0.66, 1.52)	0.98	1.23	(0.75, 2.01)	0.40
2-hydroxy-3-methylbutyric acid (1-SD (log ₂)) [‡]	2.28	(1.52, 3.43)	7.0 x 10 ⁻⁵	2.54	(1.51, 4.27)	4.2 x 10 ⁻⁴	3.12	(1.74, 5.56)	4.2 x 10 ⁻⁴
Pancreatic cancer, EPIC (152 case-control sets)									
Alcohol intake (12g/day)	1.07	(0.92, 1.25)	0.36	1.04	(0.88, 1.24)	0.65			
Alcohol intake (1-SD (log ₂))	1.08	(0.83, 1.40)	0.58	1.03	(0.77, 1.39)	0.83			
Unknown compound (1-SD (log ₂)) [‡]	1.15	(0.92, 1.46)	0.22	1.10	(0.91, 1.41)	0.48	1.10	(0.83, 1.46)	0.50
2-hydroxy-3-methylbutyric acid (1-SD (log ₂)) [‡]	1.43	(1.07, 1.92)	0.02	1.43	(1.03, 1.99)	0.03	1.46	(1.03, 2.06)	0.03
Liver cancer, ATBC (192 case-control sets)									
Alcohol intake (12g/day)	1.25	(1.09, 1.43)	1.2 x 10 ⁻³	1.17	(1.01, 1.36)	0.03			
Alcohol intake (1-SD (log ₂))	1.33	(1.05, 1.67)	0.016	1.23	(0.94, 1.60)	0.13			
Unknown compound (1-SD (log ₂)) [‡]	1.34	(1.07, 1.68)	0.01	1.70	(1.29, 2.25)	2.0 x 10 ⁻⁴	1.76	(1.28, 2.41)	5.0 x 10 ⁻⁴
2-hydroxy-3-methylbutyric acid (1-SD (log ₂)) [‡]	2.08	(1.53, 2.82)	2.7 x 10 ⁻⁶	2.00	(1.44, 2.77)	3.4 x 10 ⁻⁵	2.07	(1.43, 2.98)	9.9 x 10 ⁻³
Liver disease mortality, ATBC (199 case-control sets)									
Alcohol intake (12g/day)	1.38	(1.22, 1.55)	1.1 x 10 ⁻⁷	1.32	(1.16, 1.50)	1.6 x 10 ⁻⁵			
Alcohol intake (1-SD (log ₂))	2.37	(1.78, 3.14)	2.8 x 10 ⁻⁸	2.19	(1.60, 2.98)	8.4 x 10 ⁻⁷			
Unknown compound (1-SD (log ₂)) [‡]	2.11	(1.63, 2.72)	1.0 x 10 ⁻⁸	1.98	(1.51, 2.60)	8.6 x 10 ⁻⁷	1.65	(1.24, 2.20)	7.0 x 10 ⁻⁴
2-hydroxy-3-methylbutyric acid (1-SD (log ₂)) [‡]	2.26	(1.73, 2.95)	2.1 x 10 ⁻⁹	2.16	(1.63, 2.86)	9.6 x 10 ⁻⁸	1.85	(1.38, 2.48)	3.9 x 10 ⁻³

¹ Models for hepatocellular carcinoma (HCC) were adjusted for body mass index (BMI, kg/m²), waist circumference (cm), recreational and household physical activity (Met-hours/week), a composite variable for smoking status and intensity (Never, Current: 1-15 cig/day, Current: 16-25 cig/day, Current: 26+ cig/day, Former: quit <=

10 years, Former: quit 11-20 years, Former: quit 20+ years, Current, occasional pipe/cigar/ use, Current/Former: missing, Unknown), level of educational attainment, and coffee intake ((log₂)grams/day); models for pancreatic cancer were adjusted for BMI (kg/m²), sex-specific physical activity categories and the composite variable for smoking status and intensity; ATBC liver cancer and fatal liver disease models were adjusted for age (years), BMI (kg/m²), leisure time physical activity, smoking intensity (cigarettes/day), level of educational attainment, and coffee intake ((log₂)grams/day);

² Models were further adjusted for self-reported alcohol intake ((log₂)grams/day);

³ Unknown compound (m/z=231.0839);

⁴ 2-hydroxy-3-methylbutyric acid (m/z=203.0227).



Click here to access/download

CONSORT Checklist

Lofffield STROBE checklist.pdf





[Click here to access/download](#)

Revised Manuscript Checklist
Loftfield STROBE checklist resubmission.pdf

