Circulating Isovalerylcarnitine and Lung Cancer Risk: Evidence from Mendelian Randomization and Pre-Diagnostic Blood Measurements

Karl Smith-Byrne*[1°]; Agustin Cerani*[2]; Florence Guida[1]; Sirui Zhou[2]; Antonio Agudo[3]; Krasimira Aleksandrova[4,5]; Aurelio Barricarte[7,8]; Miguel Rodríguez Barranco[8,9,10]; Christoph H. Bochers[2,11]; Inger Torhild Gram[12]; Jun Han[11]; Christopher I Amos[6], Rayjean J. Hung[13]; Kjell Grankvist[14]; Therese Haugdhal Nøst[12]; Liher Imaz[15,16]; María Dolores Chirlaque-López [17,8]; Mikael Johansson[18]; Rudolf Kaaks[19,20]; Tilman Kühn[19]; Richard M Martin[21]; James D. McKay[1]; Valeria Pala[22]; Hilary A Robbins[1]; Torkjel M. Sandanger[12]; David Schibli[11]; Matthias B. Schulze[4,5]; Ruth C. Travis[23]; Paolo Vineis[24]; Elisabete Weiderpass[1]; Paul Brennan[1]; Mattias Johansson[1○]; J. Brent Richards[2,25,26○].

[1]Genomic Epidemiology Branch, International Agency for Research on Cancer (IARC-WHO), Lyon, France.

[2]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada / Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada

[3]Unit of Nutrition and Cancer, Cancer Epidemiology Research Program. Institut Català d'Oncologia, Spain

[4]Nutrition, Immunity and Metabolism Senior Scientist Group, Department of Nutrition and Gerontology, German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Nuthetal, Germany

[5]University of Potsdam, Institute of Nutritional Science, Potsdam, Germany

[6] Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Texas

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

[8]CIBER in Epidemiology and Public Health (CIBERESP), Madrid, Spain

[9]Escuela Andaluza de Salud Pública (EASP), Granada, Spain

[10]Instituto de Investigación Biosanitaria ibs.GRANADA, Granada, Spain

[11]University of Victoria–Genome British Columbia Proteomics Centre, Victoria, BC, Canada / Division of Medical Sciences, University of Victoria, Victoria, BC, Canada

[12]Faculty of Health Sciences, Department of Community Medicine, University of Tromsø, The Arctic University of Norway, Norway

[13]Prosserman Centre for Health Research, Mount Sinai Hospital, Toronto

[14]Department of Medical Biosciences, Umeå University

[15]Ministry of Health of the Basque Government, Public Health Division of Gipuzkoa, Donostia-San Sebastian, Spain

[16]Biodonostia Health Research Institute, Donostia-San Sebastian, Spain

[17]Department of Epidemiology, Regional Health Council, IMIB-Arrixaca, Murcia University, Murcia, Spain

[18]Department of Radiation Sciences, Umeå University

[19]German Cancer Research Center (DKFZ), Heidelberg, Department of Cancer Epidemiology.

[20]Translational Lung Research Center (TLRC) Heidelberg, Member of the German Center for Lung Research (DZL), Germany

[21]Clinical Epidemiology & Public Health, University of Bristol, Bristol

[22]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano

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

[24]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK

[25]Division of Endocrinology, Department of Medicine & Department of Human Genetics, McGill University, Montreal, Quebec, Canada

[26]Department of Twin Research and Genetic Epidemiology, King's College London, Strand, London, United Kingdom

*** Shared first authorship**

° Corresponding and senior authors:

- Brent Richards, M.D., M.Sc. - 3755 Côte Ste-Catherine Road, Suite H-413 Montréal, Québec H3T 1E2; Tel: 514-340-8222 ext. 4362; brent.richards@mcgill.ca
- Mattias Johansson, Ph.D.: johanssonm@iarc.fr

Running Title: Isovalerylcarnitine and Lung Cancer

Keywords: Isovalerylcarnitine, Lung Cancer, Metabolite, Prospective, EPIC

Word count: 2,996

The authors declare no potential conflicts of interest.

Disclaimer: 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.

Funding

This study was supported by grants from the US National Cancer Institute under award number U19CA203654 and Cancer Research UK (C18281/A19169), as well as the

Canadian Institutes of Health Research, the Canadian Foundation for Innovation, the Fonds de Recherche Santé Québec (FRSQ), and the FRQS Clinical Research Scholarship. Metabolite GWAS studies were conducted within TwinsUK, which is funded by the Wellcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility, and Biomedical Research Centre based at Guy's and St Thomas's NHS Foundation Trust in partnership with King's College London. The INTEGRAL-ILCCO OncoArray data collection was supported by National Institute of Health under award number U19CA203654.

Abstract

Background: Tobacco exposure causes 8 of 10 lung cancers, and identifying additional risk factors is challenging due to confounding introduced by smoking in traditional observational studies.

Materials and Methods: We used Mendelian randomization (MR) to screen 207 metabolites for their role in lung cancer predisposition using independent genome-wide-association studies (GWAS) of blood metabolite levels (n = 7,824) and lung cancer risk (n = 29,266 cases / 56,450 controls). A nested case control study (656 cases and 1,309 matched controls) was subsequently performed using pre-diagnostic blood samples to validate MR association with lung cancer incidence data from population-based cohorts (EPIC and NSHDS).

Results: An MR-based scan of 207 circulating metabolites for lung cancer risk identified that blood isovalerylcarnitine (IVC) was associated with a decreased odds of lung cancer after accounting for multiple testing (Log10-OR = 0.43, 95% CI: 0.29-0.63). Molar measurement of IVC in pre-diagnostic blood found similar results (Log10-OR = 0.39, 95% CI: 0.21-0.72). Results were consistent across lung cancer sub-types.

Conclusions: Independent lines of evidence support an inverse association of elevated circulating IVC with lung cancer risk through a novel methodological approach that integrates genetic and traditional epidemiology to efficiently identify novel cancer biomarkers.

Impact: Our results find compelling evidence in favor of a protective role for a circulating metabolite, IVC, in lung cancer aetiology. From the treatment of a Mendelian disease, isovaleric acidemia, we know that circulating IVC is modifiable through a restricted protein diet or glycine and L-carnatine supplementation. IVC may represent a modifiable and inversely associated biomarker for lung cancer.

Introduction

Lung cancer causes 1.8 million deaths worldwide and is the leading cause of cancer death globally(1). Tobacco causes 8 of 10 lung cancers,(2) and a number of environmental and occupational exposures, such as radon(3) and radiation(4), are well-described. Nonetheless, little is known about specific biochemical modifiable risk factors for lung cancer. In many Western countries, a large proportion of lung cancer cases now occur in former or never smokers(5). Identifying additional modifiable risk factors for lung cancer beyond smoking is therefore of great interest, may identify individuals at risk, and provide other prevention targets.

Causal inference in humans can be biased by confounding, where the exposure and the outcome share a common cause. Research into lung cancer aetiology is particularly challenging as many putative risk factors, including health conditions, socio-economic factors, and biomarkers(6) strongly associate with smoking behaviours, which induces confounding in traditional epidemiological studies. Mendelian randomization (MR)(7), which uses germline genetic variants as instrumental variables, is less prone to confounding because it relies upon the random segregation of alleles at meiosis and their random allocation at conception, thereby breaking association with nearly all confounding factors(8).

However, the causal interpretation of MR estimates relies on several major assumptions (**Figure 1, top**). First, the genetic proxy must be robustly associated with the exposure. Second, the genetic proxy must not be associated with factors that confound the exposure-outcome association. Third, the genetic proxy must affect the

outcome only via the exposure, i.e. the absence of horizontal pleiotropy[6]. Fourth, genetic proxies cannot increase the exposure in some subjects and decrease it in others: the effect must be consistent in the same direction or null(9). Several novel statistical methods and qualitative analyses have recently emerged to evaluate violations of these assumptions. However, horizontal pleiotropy can be reduced in metabolite studies by using genetic variants that influence the metabolite and are located in, or close to, genes whose roles in determining metabolite levels have been previously well-described. Since hundreds of metabolite enzymatic pathways have been studied over the past century(10), a wealth of information is available to identify such genetic variants and assess potential bias due to horizontal pleiotropy(11,12). Additionally, genetic and biological variability affecting blood and other tissues metabolic profile are well documented to promote oncogenesis and cancer proliferation(13,14). Therefore, metabolomics-based MR can help overcome a main limitation of MR when the genetic determinants of candidate metabolites act upon genes involved in metabolic pathways, while offering a rationale for biomarker discovery in cancer.

Recent large-scale genome-wide association studies (GWAS) have identified the genetic determinants of hundreds of biomarkers, such as metabolites(15). Therefore, two-sample MR, where the exposure and outcome are assessed in different studies(16), could be used to screen for the effect of these metabolites on disease risk if a large GWAS has been conducted for the disease(17). These results could then be assessed using direct measurement of the metabolite in appropriate case-control studies, providing converging evidence from different methods that are subject to different limitations and biases(18,19).

Our objective was to identify metabolic risk factors for lung cancer risk using an approach that integrates Mendelian randomization with direct metabolite analysis in pre-diagnostic sample from large-population cohorts.

Materials and Methods

Overall study design

Our goal was to identify aetiological metabolic markers of lung cancer risk using two independent but complementary designs: an exploratory two-sample MR in large GWAS, with validation for the importance of the most promising metabolites in pre-diagnostic blood from case-control studies nested in large population cohorts (Figure 1). STROBE-MR(20) and STROBE(21) reporting guidelines were followed for MR and case-control studies, respectively.

Mendelian randomization

Study populations and data sources

SNP-metabolite association data were obtained from a metabolite GWAS in 7,824 subjects of European descent from two population-based cohorts using the Metabolon platform(15). SNP-lung cancer risk associations were extracted from a recent large-scale lung cancer GWAS with 29,266 cases and 56,450 controls of European descent(22). All studies received ethical approval from their respective review committees/boards and all participants provided written consent.

Statistical analysis

Of the 400 metabolites assayed in 7,824 individuals using the Metabolon platform(15), 207 circulating metabolites had SNPs associated at genome-wide

significance ($P<5x10^{-8}$). SNPs were clumped at linkage disequilibrium, r^2, > 0.001. After data harmonization 207 metabolites with 555 unique SNPs were included in analyses (Figure 1, eTable 1). Where metabolites had only one available SNP, a Wald estimate was used to estimate the effect on lung cancer risk. Where multiple SNPs were available for a metabolite, odds ratio (OR) were estimated using a likelihood-based MR approach (ML)(16). A False Discovery Rate (FDR) was applied to adjust for multiple-hypothesis testing from these primary MR analyses using all available instruments for the 207 metabolites investigated.

Sensitivity analyses

For metabolites with statistically significant ML-based and FDR-adjusted effects, we ran weighted-median(23), weighted-mode(24), MR-Egger(25) and MR-RAPS(26) sensitivity analyses that can provide pleiotropy-robust estimates in the presence of bias from horizontal pleiotropy and can quantify net directional pleiotropy (using the MR-Egger intercept(25)). Heterogeneity of the SNP estimates—an indication of horizontal pleiotropy—was evaluated using Cochran's and Rücker's Q(27).

Bidirectional MR was used to estimate potential lung cancer effects on metabolite concentrations as a sensitivity analysis to assess the correct orientation of MR estimates.

For metabolites with available *cis*-SNPs (within 1Mb of a gene known to influence metabolite levels), a separate secondary MR analysis was conducted using the Wald ratio. MR analysis using *cis*-SNPs are less likely to be subject to pleiotropy(12) and provide a biological rationale for SNP-metabolite associations(9).

Further, a stringent Bayesian colocalization analysis was performed to assess confounding by linkage disequilibrium(28).

SNP associations with metabolite-pathway components and lung cancer risk factors were searched to qualitatively evaluate biological plausibility and pleiotropy, respectively, in Phenoscanner(29), KEGG(30), OMIM(31), eQTLgen(32) and GTEx(33) database as described in eMethods.

Additional analyses stratified by histological cancer subtypes (adenocarcinoma: 11,273 cases, 55,483 controls; squamous cell carcinoma: 7,426, 55,627; small cell carcinoma: 2,664, 21,444) and by smoking status (never: 2,355, 7,504; ever: 23,223, 16,964) were performed for metabolites that remained significant after sensitivity analyses.

Prospective nested case-control study

Study Populations

Metabolites with robust evidence of an effect on lung cancer risk following all genetic analyses were subsequently measured using pre-diagnostic plasma samples from a prospective case-control study nested within the European Prospective Investigation into Cancer (EPIC)(34) and Nutrition and The Northern Swedish Health and Disease Study (NSHDS)(35).

The EPIC study is a large multi-centre prospective cohort that recruited participants between 1992 and 1998(34). For the case-control study, participants were taken from the 238,816 individuals from the centres in Netherlands, UK, France,

Germany, Spain, and Italy who donated a blood sample at study recruitment. NSHDS is an ongoing prospective cohort of the population in Västerbotten County, Sweden. At the end of follow-up for the current study sample in 2014, 99,404 study participants who donated a blood sample at enrolment had been recruited. Further cohort information is provided in eMethods.

All study participants gave written informed consent to participate in the study and the research was approved by the participating countries' local ethics committees and IARC's Ethical Review Committee.

Outcome and study design

Incident lung cancer was defined based on the International Classification of Diseases for Oncology (ICD-O-2) and included all invasive cancers coded as "C34". Cases were chosen to maximize time to from blood collection to diagnosis (min: 2.5 years; 97% cases over 5 years).

At the time of diagnosis of an index case, two cohort participants that were alive and free of cancer (excluding non-melanoma skin cancer) were randomly selected as controls and matched(36) based on study centre, sex, date of blood collection (±12 months), and age at blood collection (± 3 months, relaxed up to ± 5 years) as shown in eFigure 1. To adjust for smoking and maximize power in smoking-stratified analyses, one control in each matched-set was also matched on the index case's smoking status from 5 categories: never smokers, short and long term quitters among former smokers (<10 years and 10 years since quitting, respectively), and light and heavy smokers among current smokers (<15 cigarettes and 15 cigarettes per day,

respectively). The overall sampling strategy yielded 649 cases and 1,296 matched controls.

Molar metabolite measurement

Molar metabolite concentrations in plasma samples were quantified at The Metabolomics Innovation Centre (TMIC, University of Victoria, Canada) by ultrahigh-performance liquid chromatography–tandem mass spectrometry (UPLC-MS/MS) operated in the multiple-reaction monitoring mode and expressed in nmol/L as described in eMethods.

Statistical analyses

Descriptive statistics were conducted for anthropometric and lifestyle characteristics between cases and controls. Metabolite concentrations were Log_{10}-transformed to allow for direct comparison between case-control and MR estimates, which were measured on this scale. Linear regression was used to test for linear trends among controls by strata of selected characteristics (sex, age, body mass index, and smoking traits).

Primary analysis involved a conditional logistic regression model to examine the statistical association of the prioritized metabolites with lung cancer risk, conditioned on the matching factors and adjusted for age, BMI, and smoking characteristics (cigarettes per day and smoking duration). Secondary analyses were repeated in subgroups according to histology and smoking status (*never/ever*). Additional analyses by quartile of metabolites delimited in controls for lung cancer

overall and by the abovementioned subgroups were also conducted. Statistical analyses were performed using R (*TwoSampleMR(37) Coloc(28)*; The R project(38)).

Results

MR Analyses

After FDR correction (5%), three metabolites were associated with lung cancer risk: arachidonate(20-4n6), 1-arachidonoylglycerophosphocholine, and isovalerylcarnitine (IVC) (eTable1:2). However, only IVC remained associated with lung cancer risk after pleiotropy-robust analyses (weighted-median(23), weighted-mode(24), MR-Egger(25) and MR-RAPS(26)). As determined *a priori*, IVC was therefore the only metabolite further investigated. A genetically determined increment in blood IVC concentration (log_{10}) was associated with a reduced risk of lung cancer (OR_{ML}: 0.43, 95%CI: 0.29-0.63, N_{SNP}=6; eTable2). Similar results were observed for IVC and lung cancer risk when stratified by histological subtypes (small cell carcinoma: OR_{ML}: 0.19, 95%CI: 0.07-0.50; squamous cell carcinoma: OR_{ML}: 0.39, 95%CI: 0.21-0.72; and adenocarcinoma: OR_{ML}: 0.52, 95%CI: 0.31-0.88) and by smoking status (ever: OR_{ML}: 0.43, 95%CI: 0.22-0.64; and never: OR_{ML}: 0.76, 95%CI: 0.49-1.02) (Figure 2, eTable3).

Sensitivity analyses: MR assumption evaluation

A *cis*-SNP (rs9635324) was identified for IVC using the ProGeM package(11). This SNP is located downstream (5.5kb) from the isovaleryl dehydrogenase gene (*IVD*), which was confirmed using KEGG's(30) enzyme codes. IVD's substrate is the metabolite isovaleryl-CoA, whose carnitine circulating form is isovalerylcarnitine, IVC (Figure 3). Mutations at the *IVD* locus render this enzyme inactive and leads to

isovaleric acidemia, an autosomal recessive inborn error of leucine catabolism characterized by an accumulation of IVC in whole blood(39). Moreover, the cis-SNP allele associated with lower IVC was consistently associated with higher *IVD* enzyme gene expression in whole blood (eQTLgen data(32,40)) and in lung tissue from GTEx(33). Thus, impaired IVD enzyme function leads to higher blood IVC, whereas higher functional enzyme expression is associated with low IVC levels. Overall, these findings provide a clear biological rationale for the *cis*-SNP-metabolite association via *IVD*'s enzyme, supporting MR assumptions(9) (Figure 1). MR analyses using only this *cis*-SNP supported IVC's primary MR results overall (Log_{10} OR: 0.27, 95%CI: 0.14-0.54) and by histological subtype (Figure 3, eTable3).

The colocalization analysis (Figure 1) found an 80% posterior probability that a single signal (*cis*-SNP rs9635324) at the genomic locus around *IVD* affects both circulating IVC levels and lung cancer risk. Additional sensitivity analyses revealed no evidence of horizontal pleiotropy or heterogeneity (eTables 4 & 5). Notably, the only association identified for the *cis*-SNP rs9635324 was with IVC, supporting the validity of this instrument (eTable 6). Finally, bidirectional MR analysis showed no association of lung cancer with IVC levels (eTable 7).

Prospective nested case-control study

Data from 656 cases and 1,309 matched controls were included in the analysis. The mean age at blood collection was 56 years for both controls and cases, and for cases the mean time between pre-diagnostic blood collection and diagnosis was 7 years (range: 2-10 years). Among controls, IVC concentrations were higher among men compared to women, participants with higher BMI, and among participants who

smoked (driven by higher proportion of smoking in males who have higher IVC on average than females), smoked a greater number of cigarettes per day, and who smoked for a greater number of years (Table 2).

The primary conditional logistic regression analysis showed that a 10-fold increment in blood IVC was associated with 52% lower risk of lung cancer (log_{10}-OR: 0.52, 95%CI: 0.32-0.86). After adjusting for detailed smoking exposure (smoking duration and cigarettes per day) and BMI, the association between blood IVC was accentuated and resembled that of the MR analysis (log_{10}-OR: 0.39, 95%CI: 0.21-0.72) with no difference in precision ($SE_{minimally\ adjusted}$=0.27 vs. $SE_{Fully\ Adjusted}$=0.24).

Stratified analysis by histological subtypes and smoking yielded similar OR estimates to that of the primary analysis, although confidence intervals included one, indicating that these sub-group analyses may have benefitted from a larger sample size. Risk analyses by quartiles of IVC with lung cancer can be found in eTable 8.

Discussion

In this study, we integrated genetic (MR) and traditional epidemiology study designs as an efficient and novel approach to identify lung cancer biomarkers with plausible aetiological involvement. In the initial MR analyses, we tested 207 metabolites and identified IVC as associated with lung cancer risk. Subsequent direct blood measurement of IVC in pre-diagnostic blood samples from large prospective

case-control studies independently supported an association of IVC with lung cancer risk.

Aetiological research on lung cancer is hampered by the wide-ranging impact of smoking, not only on lung cancer risk, but also on many putative risk factors. MR largely overcomes this confounding by relying upon random assignment of alleles at conception, yet it can yield biased estimates when its assumptions are violated(7). The most problematic assumption is the lack of horizontal pleiotropy. The study design we have followed helps to mitigate this bias since the enzymatic and genetic determinants of IVC have been previously described, allowing us to use only SNPs near enzymes known to influence IVC levels directly. It is possibly, but unlikely, that such *cis*-SNPs act on lung cancer via pathways independent of IVC. Furthermore, in Phenoscanner(29), a database with over 65 billion published SNP associations, we identified no associations between the SNPs used as proxies of IVC and smoking characteristics. We thus conclude that the observed relationship between IVC and lung cancer is independent of smoking.

We next analysed the concentrations of IVC using pre-diagnostic blood samples from a case-control study nested within two large population cohorts. This analysis allowed us to carefully evaluate the epidemiological properties of IVC and its relation to lung cancer risk using direct measurements. This analysis confirmed the inverse association between IVC and lung cancer risk, and careful adjustment for smoking characteristics further accentuated the association. Taken together, these data are consistent with a role for IVC metabolism in lung cancer aetiology in humans. Nonetheless, future work should aim to replicate these findings in larger cohorts and

investigate the IVC-lung cancer association among never smokers in a sample with greater power for stratified analyses.

IVC is a carnitine substrate of the enzyme isovaleryl-CoA dehydrogenase, which is involved in the degradation of leucine and fatty acids. Leucine is, in turn, an essential amino acid that is involved in metabolic regulation via the mTORC1 complex, which may influence cancer development through intracellular signals regulating cellular growth and proliferation(41). Leucine also regulates the cellular availability of glutamine, a major player in cancer proliferation and drug-resistance via metabolic rewiring(42). More proximally, IVC is a selective activator of calpain, an inducer of apoptosis(43,44), thus lower cellular IVC levels may interfere with programmed cell death. Whilst there is limited epidemiological evidence in the literature on the importance of IVC in cancer, circulating IVC has previously been inversely associated with endometrial cancer(45). Despite this evidence, the specific biological pathway from IVC to lung cancer pathophysiology remains to be elucidated.

Since cancer's first portrayal as a metabolic disease over a hundred years ago(10), a deeper understanding of the metabolic heterogeneity and adaptability of cancerous tissue(46) has yielded novel metabolism-targeted therapies(42,47). Similarly, genetic and biological variability affecting several tissues' metabolic profile are well documented to impact cancer risk and proliferation(13,14). The treatment of isovaleric acidemia shows that IVC can be modified by a restricted protein diet, glycine and L-carnitine supplementation(39), yet this remains to be investigated in lung cancer.

Much of the published biomarker research has used MR to test existing hypotheses reported in the observational and clinical literature due to its robustness to classic epidemiology biases(48,49). In contrast, here we demonstrate that a conservative set of MR-based decision criteria, leveraging features unique to metabolites—involved in cancer biology—, can be used as a primary step to generate strong statistical evidence in favor of a metabolite, or a set of these, from a large panel of candidate metabolites. Given the prohibitively high cost of measuring a full panel of metabolites/proteins in an adequately powered sample, our study demonstrates an efficient approach to identifying plausible aetiological biomarkers that can readily be applied to other cancer outcomes.

Limitations

This approach, however, is not without limitations. Not all known metabolites are available on commercial panels, therefore our study did not include all known blood metabolites. Further, while *cis* instruments for a biomarker may generate strong MR evidence[8], their identification is non-trivial. We thus advise caution when using MR to scan for biomarkers where *cis* instruments are not available to corroborate MR signals from *trans* genetic variation. Last, but perhaps most importantly, we provide evidence that IVC is important in the aetiology of lung cancer, but this does not preclude other metabolites in the IVC pathway having biological effects on lung cancer and further studies are required to fully investigate each constituent of the pathway.

Conclusion

We found elevated levels of IVC inversely associated with lung cancer risk in both MR and nested case-control studies, thus providing evidence in favor of a

protective role for IVC in lung cancer aetiology. Further research is required to clarify the mechanisms by which IVC may influence lung cancer development. More generally, we present a methodological approach for biomarker discovery that allows for efficient identification of biomarkers using MR, to be followed up by direct measurements in well-designed epidemiological studies.

Acknowledgements

We thank our collaborators from the International Lung Cancer Consortium (ILCCO) Adonina Tardon, Angela Risch, Angeline Andrew, Chu Chen, David Christiani, Demetrios Albanes, Erich Wichmann, Gadi Rennert, Geoffrey Liu, Hans Brunnström, Heike Bickeböller, Hongbing Shen, Jian-Min Yuan, John K. Field, John R. McLaughlin, Kjell Grankvist, Lambertus A. Kiemeney, Loic Le Marchand, M. Dawn Teare, Maria Teresa Landi, Matthew B. Schabath, Melinda C. Aldrich, Mikael Johansson, Neil Caporaso, Olle Melander, Philip Lazarus, Richard Houlston, Sanjay S. Shete, Shan Zienolddiny, Stephen Lam, Stig E. Bojesen, Susanne Arnold, Thorunn Rafnar, Victoria Stevens, Ying Wang and Yun-Chul Hong.

References

1. Golan S, Eggener S. GLOBOCAN 2012 v 1.0, cancer incidence and mortality worldwide: IARC CancerBase No. 11. Journal of Endourology2015.
2. Forouzanfar MH, Afshin A, Alexander LT, Biryukov S, Brauer M, Cercy K*, et al.* Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet **2016** doi 10.1016/S0140-6736(16)31679-8.
3. Cheng ES, Egger S, Hughes S, Weber M, Steinberg J, Rahman B*, et al.* Systematic review and meta-analysis of residential radon and lung cancer in never-smokers. Eur Respir Rev **2021**;30(159) doi 10.1183/16000617.0230-2020.
4. Gilbert ES. Ionising radiation and cancer risks: what have we learned from epidemiology? Int J Radiat Biol **2009**;85(6):467-82 doi 10.1080/09553000902883836.
5. Barta JA, Powell CA, Wisnivesky JP. Global epidemiology of lung cancer. Annals of Global Health2019.
6. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F*, et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nature Genetics2019.
7. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. BMJ (Online) **2018** doi 10.1136/bmj.k601.
8. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol **2003**;32(1):1-22 doi 10.1093/ije/dyg070.
9. Swanson SA, Hernan MA. The challenging interpretation of instrumental variable estimates under monotonicity. Int J Epidemiol **2018**;47(4):1289-97 doi 10.1093/ije/dyx038.
10. Koppenol WH, Bounds PL, Dang CV. Otto Warburg's contributions to current concepts of cancer metabolism. Nat Rev Cancer **2011**;11(5):325-37 doi 10.1038/nrc3038.
11. Stacey D, Fauman EB, Ziemek D, Sun BB, Harshfield EL, Wood AM*, et al.* ProGeM: A framework for the prioritization of candidate causal genes at molecular quantitative trait loci. Nucleic Acids Research **2019** doi 10.1093/nar/gky837.
12. Swerdlow DI, Kuchenbaecker KB, Shah S, Sofat R, Holmes MV, White J*, et al.* Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. International Journal of Epidemiology **2016** doi 10.1093/ije/dyw088.
13. Pavlova NN, Thompson CB. The Emerging Hallmarks of Cancer Metabolism. Cell Metab **2016**;23(1):27-47 doi 10.1016/j.cmet.2015.12.006.
14. Min HY, Lee HY. Oncogene-Driven Metabolic Alterations in Cancer. Biomol Ther (Seoul) **2018**;26(1):45-56 doi 10.4062/biomolther.2017.211.
15. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J*, et al.* An atlas of genetic influences on human blood metabolites. Nature Genetics **2014** doi 10.1038/ng.2982.
16. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genetic Epidemiology **2013** doi 10.1002/gepi.21758.
17. Dastani Z, Hivert MF, Timpson N, Perry JR, Yuan X, Scott RA*, et al.* Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-

ethnic meta-analysis of 45,891 individuals. PLoS Genet **2012**;8(3):e1002607 doi 10.1371/journal.pgen.1002607.
18. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. Int J Epidemiol **2016**;45(6):1866-86 doi 10.1093/ije/dyw314.
19. Munafò MR, Davey Smith G. Robust research needs many lines of evidence. Nature **2018** doi 10.1038/d41586-018-01023-3.
20. Smith GD, Davies NM, Dimou N, Egger M, Gallo V, Golub R*, et al.* STROBE-MR: Guidelines for strengthening the reporting of Mendelian randomization studies. PeerJ Preprints; 2019. Report nr 2167-9843.
21. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP*, et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med **2007**;4(10):e296 doi 10.1371/journal.pmed.0040296.
22. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC*, et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nature Genetics **2017** doi 10.1038/ng.3892.
23. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genetic Epidemiology **2016** doi 10.1002/gepi.21965.
24. Hartwig FP, Smith GD, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. International Journal of Epidemiology **2017** doi 10.1093/ije/dyx102.
25. Bowden J, Smith GD, Burgess S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. International Journal of Epidemiology **2015** doi 10.1093/ije/dyv080.
26. Zhao Q, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S. Small. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. arXiv preprint **2018**.
27. Bowden J, Spiller W, Del Greco FM, Sheehan N, Thompson J, Minelli C*, et al.* Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. International Journal of Epidemiology **2018** doi 10.1093/ije/dyy101.
28. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C*, et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genetics **2014** doi 10.1371/journal.pgen.1004383.
29. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB*, et al.* PhenoScanner: A database of human genotype-phenotype associations. Bioinformatics **2016** doi 10.1093/bioinformatics/btw373.
30. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research1999.
31. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). Human Mutation **2000** doi 10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G.
32. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B*, et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv **2018** doi 10.1101/447367.

33. Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. Biopreservation and Biobanking2015.
34. Riboli E, Hunt K, Slimani N, Ferrari P, Norat T, Fahey M*, et al.* European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutrition **2002** doi 10.1079/phn2002394.
35. Hallmans G, Agren A, Johansson G, Johansson A, Stegmayr B, Jansson JH*, et al.* Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort - evaluation of risk factors and their interactions. Scand J Public Health Suppl **2003**;61:18-24 doi 10.1080/14034950310001432.
36. Integrative Analysis of Lung Cancer E, Risk Consortium for Early Detection of Lung C, Guida F, Sun N, Bantis LE, Muller DC*, et al.* Assessment of Lung Cancer Risk on the Basis of a Biomarker Panel of Circulating Proteins. JAMA Oncol **2018**;4(10):e182078 doi 10.1001/jamaoncol.2018.2078.
37. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D*, et al.* The MR-base platform supports systematic causal inference across the human phenome. eLife **2018** doi 10.7554/eLife.34408.
38. R Development Core Team. R: A language and environment for statistical computing. Viena, Austria2018.
39. Vockley J, Ensenauer R. Isovaleric acidemia: New aspects of genetic and phenotypic heterogeneity. American Journal of Medical Genetics - Seminars in Medical Genetics **2006** doi 10.1002/ajmg.c.30089.
40. Zhernakova DV, Deelen P, Vermaat M, Van Iterson M, Van Galen M, Arindrarto W*, et al.* Identification of context-dependent expression quantitative trait loci in whole blood. Nature Genetics **2017** doi 10.1038/ng.3737.
41. Jewell JL, Kim YC, Russell RC, Yu FX, Park HW, Plouffe SW*, et al.* Differential regulation of mTORC1 by leucine and glutamine. Science **2015** doi 10.1126/science.1259472.
42. Altman BJ, Stine ZE, Dang CV. From Krebs to clinic: glutamine metabolism to cancer therapy. Nat Rev Cancer **2016**;16(10):619-34 doi 10.1038/nrc.2016.71.
43. Pontremoli S, Melloni E, Viotti PL, Michetti M, Di Lisa F, Siliprandi N. Isovalerylcarnitine is a specific activator of the high calcium requiring calpain forms. Biochem Biophys Res Commun **1990**;167(1):373-80 doi 10.1016/0006-291x(90)91775-n.
44. Ferrara F, Bertelli A, Falchi M. Evaluation of carnitine, acetylcarnitine and isovalerylcarnitine on immune function and apoptosis. Drugs Exp Clin Res **2005**;31(3):109-14.
45. Gaudet MM, Falk RT, Stevens RD, Gunter MJ, Bain JR, Pfeiffer RM*, et al.* Analysis of serum metabolic profiles in women with endometrial cancer and controls in a population-based case-control study. Journal of Clinical Endocrinology and Metabolism **2012** doi 10.1210/jc.2012-1490.
46. Porporato PE, Filigheddu N, Pedro JMB, Kroemer G, Galluzzi L. Mitochondrial metabolism and cancer. Cell Res **2018**;28(3):265-80 doi 10.1038/cr.2017.155.
47. da Veiga Moreira J, Hamraz M, Abolhassani M, Schwartz L, Jolicoeur M, Peres S. Metabolic therapies inhibit tumor growth in vivo and in silico. Sci Rep **2019**;9(1):3153 doi 10.1038/s41598-019-39109-1.
48. Smith Byrne K, Appleby PN, Key TJ, Holmes MV, Fensom GK, Agudo A*, et al.* The role of plasma microseminoprotein-beta in prostate cancer: an observational nested case–control and Mendelian randomization study in the European prospective investigation into cancer and nutrition. Annals of Oncology **2019** doi 10.1093/annonc/mdz121.

49. Fanidi A, Carreras-Torres R, Larose TL, Yuan JM, Stevens VL, Weinstein SJ*, et al.* Is high vitamin B12 status a cause of lung cancer? International Journal of Cancer **2019** doi 10.1002/ijc.32033.

Tables

Table 1. Characteristics of control and lung cancer participants

	Controls (n = 1,296)	**Cases (n = 649)**
Sex, *n* (%)		
Male	723 (55.7)	364 (56.1)
Female	573 (44.3)	285 (43.9)
Age at blood collection, *mean* (95% CI)	56.7 (56.2-57.1)	56.7 (56.1-57.3)
Body Mass Index (BMI), *mean* (95% CI)	26.3 (26.1-26.5)	26.3 (26.0-26.6)
Smoking Status, *n* (%)		
Never	357 (27.3)	74 (11.2)
Previous	401 (30.6)	163 (24.9)

Current	538 (41.1)	412 (62.8)
Cigarettes per day, *mean* (95% CI)	9.1 (8.6-9.6)	14.7 (13.9-15.4)
Smoking duration (Years), *mean* (95% CI)	21.6 (20.7-22.5)	31.1 (29.8-32.3)
Time since quitting (Years), *mean* (95% CI)	7.3 (6.7-7.9)	3.0 (2.3-3.8)

Table 2. Mean blood isovalerylcarnitine concentration in controls by selected characteristics

	Isovalerylcarnitine (nmol/L)			
	Controls (1,296)		Case (649)	
	N	**Mean (95% CI)**	***N***	**Mean (95% CI)**
Sex				
Male	723	73.8 (71.7-75.8)	364	69.5 (66.6-72.5)
Female	573	55.7 (53.5-58.1)	285	54.9 (51.7-58.3)
Age at blood collection (Years)				
< 50 Years	251	62.1 (52.8-71.4)	128	66.9 (53.9-79.8)
≥ 50 and < 55 Years	237	67.6 (62.7-72.5)	113	67.6 (60.5-74.7)
≥ 55 and < 60 Years	340	66.6 (63.3-69.9)	170	64.7 (60.0-69.3)
≥ 60 and < 65 Years	328	66.2 (61.3-71.2)	164	60.5 (53.6-67.4)
≥ 65 and < 70 Years	73	71.5 (61.4-81.6)	39	58.7 (44.9-72.5)
≥ 70 Years	67	61.4 (52.8-71.4)	35	45.5 (26.3-64.7)
Body Mass Index				
≤ 25	521	58.9 (56.5-61.5)	269	56.1 (52.6-59.6)
> 25 and ≤ 30	562	69.7 (68.5-76.3)	277	67.9 (64.6-71.4)
> 30	212	72.4 (68.5-76.3)	102	68.9 (63.3-74.6)
Smoking Status				
Never	357	63.1 (60.0-66.1)	74	57.7 (51.0-64.5)
Previous	401	68.9 (65.9-71.8)	163	65.8 (61.2-70.4)
Current	538	65.4 (62.9-67.9)	412	63.1 (60.2-65.9)
Cigarettes per day				
< 0	357	62.9 (59.9-65.9)	74	57.8 (51.0-64.5)
≥ 1 and < 5	110	61.1 (55.6-66.6)	24	68.3 (56.4-80.2)
≥ 5 and < 10	224	64.5 (60.6-68.3)	97	61.3 (55.4-67.3)
≥ 10 and < 15	167	65.8 (61.3-70.3)	96	61.7 (55.7-67.6)
≥ 15	333	70.7 (67.5-73.8)	335	63.7 (60.5-66.9)
Smoking duration (Years)				
0	357	62.9 (59.9-65.9)	74	57.8 (51.0-64.5)
≥ 1 and < 20	204	63.4 (59.4-67.4)	47	64.3 (55.8-72.8)
≥ 20 and < 30	205	67.3 (63.2-71.4)	112	55.5 (49.5-61.5)

≥ 40	489	67.7 (65.1-70.3)	395	66.0 (63.0-69.1)

Headings from eTables for reference (Online-only)

eTable 1. Genetic association data for blood metabolite levels and lung cancer risk for metabolite-associated single nucleotide polymorphisms (SNPs) from Shin et al. and McKay et al. genome-wide associations studies.

eTable 2. Mendelian randomization estimates (Likelihood Method) for metabolites statistically significantly associated with metabolite levels after false dicovery rate adjustment for multiple-hypothesis testing.

eTable3. Mendelian randomization results using cis and all independent GWAS single nucleotide polymorphisms (SNP) isovalerylcarnitine instruments for multiple methods.

eTable 4. Quantification of directional horizontal pleiotropy from MR-Egger Intercept.

eTable 5.Cochran and Rucker's Q Statistics for Model Heterogeneity for IVW and MR-Egger Methods.

eTable 6. Available information on significant associations for MR instruments from available GWAS.

eTable 7. Bi-directional MR analysis of lung cancer on isovalerylcarnitine levels.

eTable 8. Multi-variable adjusted odds ratios (95% CI) for lung cancer by quartile of isovalerylcarnitine, subdivided by selected factors.

Figures Legends

Figure 1. Description of overall study design including data used, methods, assumptions, and sensitivity analyses for Mendelian randomisation, and nested case-control study design and data. The objective of this study was to efficiently identify aetiological metabolic markers of lung cancer risk using two independent designs: an exploratory two-sample Mendelian randomization (MR) and a nested case-control study. **A.** The causal interpretation of MR(7) estimates relies on 4 assumptions: (1) the genetic proxy (single nucleotide polymorphism, SNP) must be associated with the metabolite; (2) the SNP must not be associated with factors that confound the exposure-outcome association; (3), the SNP must affect the outcome only via the exposure: absence of horizontal pleiotropy; (4) the SNP cannot increase the exposure in some subjects and decrease it in others: the effect must be consistent in the same direction or null(9). *Step 1*: SNPs associated with 207 metabolites ($r2>0.001$, $P<5x10^{-8}$) from a metabolomics genome-wide association study (GWAS; N=7,824 European descent)(15) were harmonized with cancer-associated SNP data from a lung cancer GWAS (29,266 cases and 56,450 controls, European descent)(22). A total of 207 metabolites with 555 associated SNPs were included in MR analyses after data harmonization. *Step 2*: given the absence of genetic, exposure and outcome data in the same study population, two-sample MR(16) enabled effect estimation for a panel of metabolites (N=207) on lung cancer risk using for each metabolite the Wald ratio (if only 1 metabolite-associated SNP) or maximum likelihood (ML if more than 1 metabolite-associated SNP). *Step 3*: a false discovery rate (FDR) was applied to adjust for multiple-hypothesis testing (N=207). *Step 4-7*: Following FDR, only 3 metabolites remained statistically associated with lung cancer risk, thus MR assumptions were qualitatively and quantitatively evaluated following novel STROBE-

MR reporting guidelines(20) for their associated SNPs. Weighted median(23), weighted mode(24), MR-Egger(25) and MR-RAPS(26) sensitivity analyses provide estimates robust to bias from horizontal pleiotropy (assumption 3) and allow to quantify net directional pleiotropy (MR-Egger intercept(25)). Only one metabolite, remained associated after these analyses and further sensitivity analyses were performed for its associated SNPs: Cochran's and Rücker´s Q, *cis*-SNP analyses and Phenoscanner(29)/literature searches for SNP associations with lung cancer risk factors. The direction of effect was tested by bidirectional MR. *Step 6*: colocalization analysis estimates the posterior probability that the genomic locus centred on cis-SNPs affects both circulating metabolite levels and lung cancer risk, supporting an aetiological effect(28) (assumption 2). *Step 7*: biological plausibility for the SNP-metabolite association was further assessed by searching in metabolism-based resource KEGG(30), Mendelian genetics resource OMIM(31), gene expression databases eQTLgen(32) and GTEx(33). **B.** Molar concentrations of metabolite(s) with consistent and strong MR evidence of an effect on lung cancer risk were measured using pre-diagnostic blood samples from a nested case-control study based on EPIC(34) (Europe) and NSHDS(35) (Sweden) population based cohorts to further support and estimate the effect of such metabolites on lung cancer risk. This study followed a matched case-control design (649 cases and 1,296 matched controls)(36) where index cases were matched at diagnosis to two controls based on study centre, sex, date of blood collection (±12 months), and age at blood collection (± 3 months, relaxed up to ± 5 years). Additionally, to maximise power in smoking-stratified analyses one control in each matched-set was also matched on the index case's smoking status from 5 categories: never smokers, short and long term quitters among former smokers (<10 years and 10 years since quitting, respectively), and light and

heavy smokers among current smokers (<15 cigarettes and 15 cigarettes per day, respectively).

Figure 2. Two-sample Mendelian randomization and nested case-control study results of the estimated effect of isovalerylcarnitine on lung cancer risk. Subjects with higher blood isovalerylcarnitine (Log_{10} units) had on average 57% (OR: 0.43, 95% CI: 0.29-0.63; adjusted for multiple-hypothesis testing) and 61% (OR: 0.39, 95% CI: 0.21-0.72) lower risk of lung cancer as independently estimated by MR and nested case-control studies, respectively.

Figure 3. Biological plausibility for the association between *IVD* gene's *cis* genetic variant and isovalerylcarnitine levels. A) A cis-SNP, rs9635324, (i.e. SNPs in or within 1Mb of a gene known to influence metabolite levels) was identified for IVC (ProGeM package(11)), which is located downstream (5.5kb) from the isovaleryl dehydrogenase gene (IVD) and was confirmed using KEGG's(30) enzyme codes. IVD's substrate is the metabolite isovaleryl-CoA, whose carnitine cellular and circulating form is IVC. **B**) The *cis*-SNP allele associated with lower IVC was consistently associated with higher IVD enzyme gene expression in whole blood (eQTLgen data(32,40)) and in lung tissue from GTEx(33). It is known from Mendelian genetics(31) that impaired IVD enzyme function leads to higher blood IVC, as determined by inborn errors of metabolism(39), whereas higher functional enzyme expression is associated with low IVC levels. This evidence provides a clear biological rationale for the *cis*-SNP-metabolite association via IVD's enzyme, supporting Mendelian randomization assumptions(9) and the validity of its results.