



OPEN

Exploring the contribution of lifestyle to the impact of education on the risk of cancer through Mendelian randomization analysis

Loukas Zagkos^{1,2}✉, Alexander Schwinges³, Hasnat A. Amin¹, Terry Dovey¹ & Fotios Drenos¹

Educational attainment (EA) has been linked to the risk of several types of cancer, despite having no expected direct biological connection. In this paper, we investigate the mediating role of alcohol consumption, smoking, vegetable consumption, fruit consumption and body mass index (BMI) in explaining the effect of EA on 7 cancer groupings. Large-scale genome wide association study (GWAS) results were used to construct the genetic instrument for EA and the lifestyle factors. We conducted GWAS in the UK Biobank sample in up to 335,024 individuals to obtain genetic association data for the cancer outcomes. Univariable and multivariable two-sample Mendelian randomization (MR) analyses and mediation analyses were then conducted to explore the causal effect and mediating proportions of these relations. MR mediation analysis revealed that reduced lifetime smoking index accounted for 81.7% (49.1% to 100%) of the protective effect of higher EA on lower respiratory cancer. Moreover, the effect of higher EA on lower respiratory cancer was mediated through vegetable consumption by 10.2% (4.4% to 15.9%). We found genetic evidence that the effect of EA on groups of cancer is due to behavioural changes in avoiding well established risk factors such as smoking and vegetable consuming.

Cancer is a risk to health and the primary cause of death worldwide^{1,2}. An estimated 19.3 million new cancer cases and almost 10.0 million cancer deaths occurred in 2020³. A steady increase in mortality and incidence, particularly in developed countries⁴, calls for more effective cancer prevention⁵. While improvements in survival rates reflect progress in medical technology and healthcare, the rising incidence of cancer has been attributed to generational changes in obesity, lowered physical activity, a difference in diet and other lifestyle factors^{6,7}. The fact that increasing cancer rates offset higher survival rates and lead to higher absolute mortality⁸ signifies the importance of tackling modifiable risk factors that lead to high incidence rates.

Educational attainment (EA) predicts cancer outcomes^{9–11} and is a central driver between socio-economic status (SES) and health¹². For example, 22% of US cancer deaths could be prevented if all Americans had the cancer death rates of college-educated Americans⁵. Although there is no direct biological link between EA and cancer risk, it is believed that EA leads to more effective self-management and habits¹². Various behaviourally related modifiable risk factors in low EA/SES have been studied. The most prominent factors identified include: alcohol consumption, physical inactivity, obesity, cigarette smoking, as well as low fruit and vegetable consumption^{13–17}.

Studies specifically targeting the mediating effect of modifiable risk factors in EA and cancer risk are sparse. Some studies assess all-cause mortality^{18,19} while other studies address general cancer risk in the context of SES^{20–22}. Quantitative assessments of the mediating effect of risk factors from such observational data are based on multivariable analysis, a method aiming to disentangle the effects of multiple variables on the outcome²³.

¹Department of Life Sciences, College of Health, Medicine and Life Sciences, Brunel University London, Kingston Lane, Uxbridge, London UB8 3PH, UK. ²Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, St Mary's Campus, London W2 1PG, UK. ³Department of Infectious Diseases, Faculty of Medicine, National Heart & Lung Institute, Imperial College London, Cale Street, London SW3 6LY, UK. ✉email: l.zagkos@imperial.ac.uk; Fotios.drenos@brunel.ac.uk

While the results of such assessments provide direction for further experimental studies, their validity is limited. Observational studies are also at risk of systemic biases between groups²⁴. Furthermore, clustering of risky behaviours is common²⁵, whereby individuals readily engage with a variety of behaviours, both risky and mitigating, the negative outcome that the authors wish to investigate, potentially introducing further bias. This makes it difficult to measure and correct for all risky behaviours and unmeasured risk factors might constitute confounders²⁶ using traditional experimental investigative techniques. Moreover, data is commonly collected at distinct times, which does not capture a lifetime exposure²⁷.

Mendelian Randomization is an analytical method to assess whether a risk factor has a causal effect on an outcome of interest, using genetic variants as instrumental variables²⁸. The approach treats genetic variants as proxy measures for clinical interventions on risk factors, and thus has been extensively shown to anticipate the results of a randomised control trial²⁹. The aim of this study was to explore the mediating role of lifestyle factors in explaining the effect of EA on the risk of cancer, using individual level data from the UK Biobank (UKB) and summary statistics from reliable published genome wide association studies (GWAS). We considered five lifestyle factors individually and simultaneously: number of alcoholic drinks per week, lifetime smoking index, BMI, fruit consumption and vegetable consumption. These five lifestyle factors were used to interrogate the underlying mechanisms by which EA affects the risk of cancer.

Methods

Data sources

Outcome data source

Individual level data on cancer incidence were obtained from UK Biobank (UKB), a prospective population study with detailed information about approximately 500,000 participants³⁰. The data was collected between 2006 and 2011 and participants volunteered to provide biological samples for the measurement of biochemical markers and subsequent genotyping, anthropomorphic measures through a number of collection centres in the UK, and sociodemographic, lifestyle and health behaviours information through a series of in-person and online questionnaires. The processes for genotyping and data management have been described in depth³¹. Phenotype data were obtained from UKB data-fields 41,270 and 40,006 for 30 site-specific cancers. In this analysis, up to 335,024 UKB participants of European ancestry were considered, after excluding samples with relatedness of first or second degree and samples with discordant genetic and reported sex. Cancer cases were defined according to ICD-10 codes (international classification of diseases, 10th revision), obtained through linkage to national cancer registries. Genetic associations were estimated with 7 cancer groupings: digestive (7695 cases and 327,356 controls), female reproductive (3,612 cases and 177,190 controls), head and neck (1331 cases and 334,378 controls), lower gastrointestinal (GI) tract (6545 cases and 328,601 controls), lower respiratory (2307 cases and 332,457 controls), male reproductive (8988 cases and 149,131 controls) and upper gastrointestinal (GI) tract (1300 cases and 149,131 controls). Cancer groupings were generated to maximise statistical power and were determined based on their location in the body. Those with a cancer diagnosis were considered a case only for their chronologically first reported cancer type. This was done to distinguish between primary cancers originating in the specific tissue and secondary cancers metastasising from another location.

Exposure data source

To assess EA, we used publicly available summary statistics from a Social Science Genetic Association Consortium (SSGAC) meta-analysis of GWAS³². The primary meta-analysis combined 3 quality-controlled cohort-level results from studies in Europe and USA and was conducted on approximately 3 million individuals of European ancestry. Education years were measured for all samples over the age of 30. In this work, we used meta-analysis GWAS results from all discovery cohorts except 23andMe, conducted on approximately 800,000 samples, 442,183 of which were UKB participants.

Lifestyle data sources

Alcohol consumption, body mass index (BMI), smoking, fruit and vegetable consumption were assessed as potentially mediating risk factors in this work. Publicly available summary level data were used for alcoholic drinks consumed per week from Saunders et al.^{33,34}, who conducted a GWAS meta-analysis using data from 60 cohorts on 2,965,643 individuals. To capture smoking behaviour, we used genetic summary statistics on lifetime smoking index, measured in 462,690 UKB participants of European ancestry who had phenotype data and passed genotype inclusion criteria³⁵. Following a method previously reported³⁶, smoking status, age at initiation in years, age at cessation in years and number of cigarettes smoked per day were combined into a lifetime smoking index. Genetic estimates for BMI were obtained from the Genetic Investigation of Anthropometric Traits consortium (GIANT) GWAS meta-analysis of 681,275 samples of European ancestry³⁷, around 450,000 of which were UKB participants. To assess fruit and vegetable consumption, we used publicly available GWAS summary statistics from the MR-base³⁸ for binary traits ‘fruit consumers’ and ‘vegetable consumers’, conducted on 64,949 UKB participants of European ancestry. These variables were generated as consumption over the last 24 h. A number of Yes/No questions relating to eating particular food groups or items were given to the participants following the pattern ‘Did you eat any <food-group> yesterday?’ providing examples of relevant foods and a picture. The participants completed the questionnaire in the assessment centre or online in four separate occasions within a year (<http://biobank.ctsu.ox.ac.uk/crystal/docs/DietWebQ.pdf>).

Genome wide association studies

Genome wide association studies were conducted to obtain associations between 9,420,314 genotyped and imputed single nucleotide polymorphisms (SNPs) and 7 cancer groupings: digestive, female reproductive, head

and neck, lower GI, lower respiratory, male reproductive and upper GI in up to 335,024 unrelated participants of white British ancestry. The SNPs tested were located in the autosomes. Only SNPs with a minor allele frequency greater than 0.01 and a Hardy–Weinberg equilibrium p-value greater than 10^{-6} were considered. The association of each SNP was tested using a linear regression model, adjusting for sex, age and the first 4 genetic principal components to control for population structure.

Statistical analysis

Genetic instruments

Genetic variants were considered as instrumental variables for EA and the lifestyle factors in the MR analysis if they were bi-allelic, had a minor allele frequency (MAF) greater than 0.01 and a Hardy–Weinberg equilibrium P-value greater than 10^{-6} . For EA, lifetime smoking index, drinks per week and BMI, genetic variants below the genome-wide significance threshold ($p < 5 \times 10^{-8}$) were selected as instruments, whereas for fruit and vegetable consumption, we considered a less stringent p-value threshold ($p < 10^{-5}$), as it was the lowest threshold that provided robust signals for the two dietary GWAS results. We identified independent SNPs after clumping summary estimates, using a linkage disequilibrium (LD) threshold of $r^2 < 0.001$ and a clumping window of 10 Mb. For the LD estimates between the genetic variants, we used the 1000 genomes phase 3 European reference panel³⁹. To ensure that the first MR assumption holds, only genetic variants with an F-statistic greater than 10 were included in the analysis⁴⁰. To further test the validity of the MR assumptions, we identified traits that associate with the genetic instruments used in this work, using the SNP nexus platform (<https://www.snp-nexus.org/v4/>). In the two-sample MR setting, genetic variants were excluded from the analysis when the direction of effects between exposure and outcome associations could not be inferred (in the case of palindromic SNPs with MAF greater than 0.42). Genetic instruments comprised 413 independent SNPs for EA, 507 independent variants for BMI, 126 for lifetime smoking index, 10 for drinks per week, 32 for fruit consuming and 21 for vegetable consuming.

Two-sample univariable Mendelian randomization

The current study used two-sample univariable and multivariable MR analysis in a multistep process to assess the relationships between EA, each of the 5 lifestyle factors and 7 cancer outcomes of interest. First, we tested the associations between EA and 7 cancer categories. We used categories over site-specific cancer types to maximise statistical power. Following this, we assessed the associations of EA with the 5 possible mediating lifestyle factors. Lifestyle factors were then tested against the 7 cancer categories simultaneously through MVMR. The random-effects inverse variance weighted (IVW) method was used as the main method for univariable MR analysis, which provides precise causal estimates, under the assumption that all genetic variants are valid instrumental variables⁴¹. The three instrumental variable assumptions dictate that the genetic instrument is associated with the exposure, is independent of any confounders of the exposure and outcome association and is associated with the outcome only via the exposure. In sensitivity analysis, we conducted the MR-Egger method⁴² to detect possible violations due to pleiotropic effects within genetic variants in the analysis and MR-weighted median method, which reports an accurate effect estimate, given that at least 50% of the weight in the analysis comes from valid instruments⁴³. In addition, we performed MR-PRESSO, a method which detects and removes outlier SNPs based on their contribution to heterogeneity⁴⁴. The I^2 statistic was calculated to detect heterogeneity among the MR estimates obtained from multiple genetic variants. MR-IVW effect estimates were deemed statistically significant if association p-values were smaller than the Bonferroni corrected threshold $0.05/n$, where n represents the total number of independent tests in each part of the analysis (EA with 7 cancer groups: $p < 0.05/7 = 7.14 \times 10^{-3}$, EA with 5 lifestyle factors: $p < 0.01$ and 5 lifestyle factors and EA with 7 cancer groups: $p < 1.19 \times 10^{-3}$). To quantify the amount of bias due to sample overlap in the MR effect estimates, we used the MRLap method⁴⁵, which estimates corrected MR estimates, accounting for potential bias. We reported the p-value corresponding to the test statistic used to test for differences between the observed and corrected MR estimates.

Two-sample multivariable Mendelian randomization

To estimate the direct effect of each of the lifestyle factors on the risk of cancer, we performed multivariable Mendelian randomization (MVMR) analysis⁴⁶. This method allows the use of multiple genetic variants associated with more than one risk factor as instruments to identify the causal effect of each risk factor on the outcome, independent of the rest of the risk factors. To obtain the list of genetic instruments for MVMR, we first merged, before clumping, the SNPs associated with each lifestyle factor or the exposure below their respective p-value thresholds as determined in the univariable MR analysis, and then clumped these genetic variants using for each variant the smallest p-value of their association with each lifestyle factor. Following this process, we generated a list of independent genetic variants that are associated with at least one lifestyle factor, as the MVMR paradigm dictates. All estimates were reported as odds ratio (OR) per unit increase in exposure, together with their 95% confidence interval (95% CI).

Proportion of lifestyle factor mediation

Network Mendelian randomization (network-MR) was conducted⁴⁷ using the MR effect estimates and standard errors obtained previously to calculate the direct and indirect effect of EA on cancer risk. This was done per cancer outcome and per possible mediator, using their MVMR estimates, to obtain fractions of mediated and non-mediated effects. The indirect effect of EA on cancer through a lifestyle factor was estimated by multiplying the effect of EA on that factor times the effect of the lifestyle factor on the cancer outcome. The total effect was the estimated MR effect of EA on cancer. The direct effect was estimated by subtracting the indirect effect from the total effect, estimated from the first step MR. Last, the mediation proportion for each lifestyle factor

was calculated by dividing the indirect effect over the total effect. To derive standard errors of the mediation proportion estimates, we used the delta method⁴⁸.

Statistical software

Analysis was conducted in R version 4.0.2⁴⁹, two-sample analyses and sensitivity analyses were performed using the “TwoSampleMR” v.0.5.6⁵⁰ and “MRPRESSO” v1.0⁴⁴ R packages. Figures were produced using the R package “forestplot” v3.1.1⁵¹. GWAS was conducted using PLINK 1.90 command line tool (www.cog-genomics.org/plink/1.9/). This study is reported based on the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines Supplementary Table 1.

Ethics approval

This study is based on publicly available data and the informed consent and ethical review were acquired in all the original studies. The study is reported following the STROBE-MR statement.

Results

Individual SNP estimates of the per-allele effects on EA, BMI, lifetime smoking index, drinks per week, fruit consumption and vegetable consumption are reported in Supplementary Table 2. Sample size, number of cases and included ICD-10 codes per cancer group are shown in Supplementary Table 3. Graphical representation of the model can be found in the directed acyclic graph (DAG) in Fig. 1. Univariable and multivariable MR estimates, sensitivity analysis results, MR-Egger intercepts and genetic heterogeneity statistics are provided in Supplementary Tables 4–6. Associations were considered statistically significant in this work if MR-IVW and MR-PRESSO estimates were significant after multiple testing correction, MR-Egger and MR-weighted median estimates had the same direction of effect and MR-Egger intercept was not significant ($p > 0.05$).

The associations of EA with the lifestyle factors and the various groups of cancer were obtained per one standard deviation (sd) increase, which corresponds to 3.6 years of additional education in the UKB. MR-IVW results revealed two associations between EA and the odds of cancer, which were below the Bonferroni adjusted p-value threshold of $p = 0.05/7 = 7.14 \times 10^{-3}$. One sd increase in EA was associated with lower odds of lower respiratory tract cancer (OR: 0.40, 95% CI: 0.30 to 0.54), lower odds of upper GI cancer (OR: 0.59, 0.43 to 0.82) and lower odds of digestive cancer (OR: 0.81, 0.69 to 0.94) (Fig. 2). At a $p = 0.05/5 = 0.01$ Bonferroni threshold, MR-IVW results indicated that increasing EA was associated with lower BMI (beta: -0.24, -0.28 to -0.20) and lower lifetime smoking index (beta: -0.22, -0.24 to -0.20). One sd increase in EA was also associated with increased odds of fruit (OR: 1.12, 1.10 to 1.14) and vegetable consumption (OR: 1.08, 1.07 to 1.11). Weighted median and MR-Egger effects had similar effect estimates to the IVW method. The effect of EA on BMI was quite heterogeneous with an I^2 statistic of 89% but a consistent effect between the MR-IVW, MR-Weighted median and MR-PRESSO methods. MR results of the effect of higher genetically predicted EA on lifestyle factors are summarised in Fig. 3. Moreover, using SNP nexus, the identified traits are unlikely to be potential confounders of the associations tested (Supplementary Table 7). MRlap method results suggested that there was no significant effect of sample overlap in the calculated MR estimates (Supplementary Table 8).

MVMR analysis revealed seven significant effects of the lifestyle factors on the risk of cancer groups, below $p = 0.05/42 = 1.19 \times 10^{-3}$. Increasing lifetime smoking index was associated with increasing odds of lower respiratory cancer (OR: 31.4, 17.8 to 55.6), head and neck cancer (OR: 5.96, 2.95 to 12.03), upper gastrointestinal cancer (OR: 4.12, 2.20 to 7.70) and digestive cancer (OR: 1.67, 1.25 to 2.24). Higher genetically predicted BMI was associated with increased odds of upper GI cancer (OR: 1.64, 1.30 to 2.07) and lower respiratory cancer (OR: 1.65,

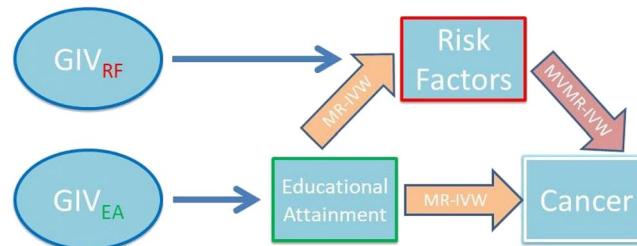


Figure 1. Study design. Network Mendelian randomization (MR) was conducted to identify the mediation proportion through lifestyle factors of the effect of education on the risk of cancer. Genetic instrumental variables were selected for each exposure based on their association below the genome-wide significance threshold, $p < 5 \times 10^{-8}$. To obtain robust signals, we used a less stringent threshold for fruit and vegetable consuming in the UKB ($p < 10^{-5}$). Potentially causal estimates were produced using Mendelian randomization inverse variance weighted (MR-IVW) method as our main approach. MR sensitivity analyses were also conducted (MR-Egger, MR-weighted median, MR-PRESSO) to assess the robustness of the results. The simultaneous effects of each lifestyle factor on cancer were estimated using multivariable Mendelian randomization (MVMR-IVW). Mediation percentage through a lifestyle factor was obtained by dividing the indirect effect over the total effect.

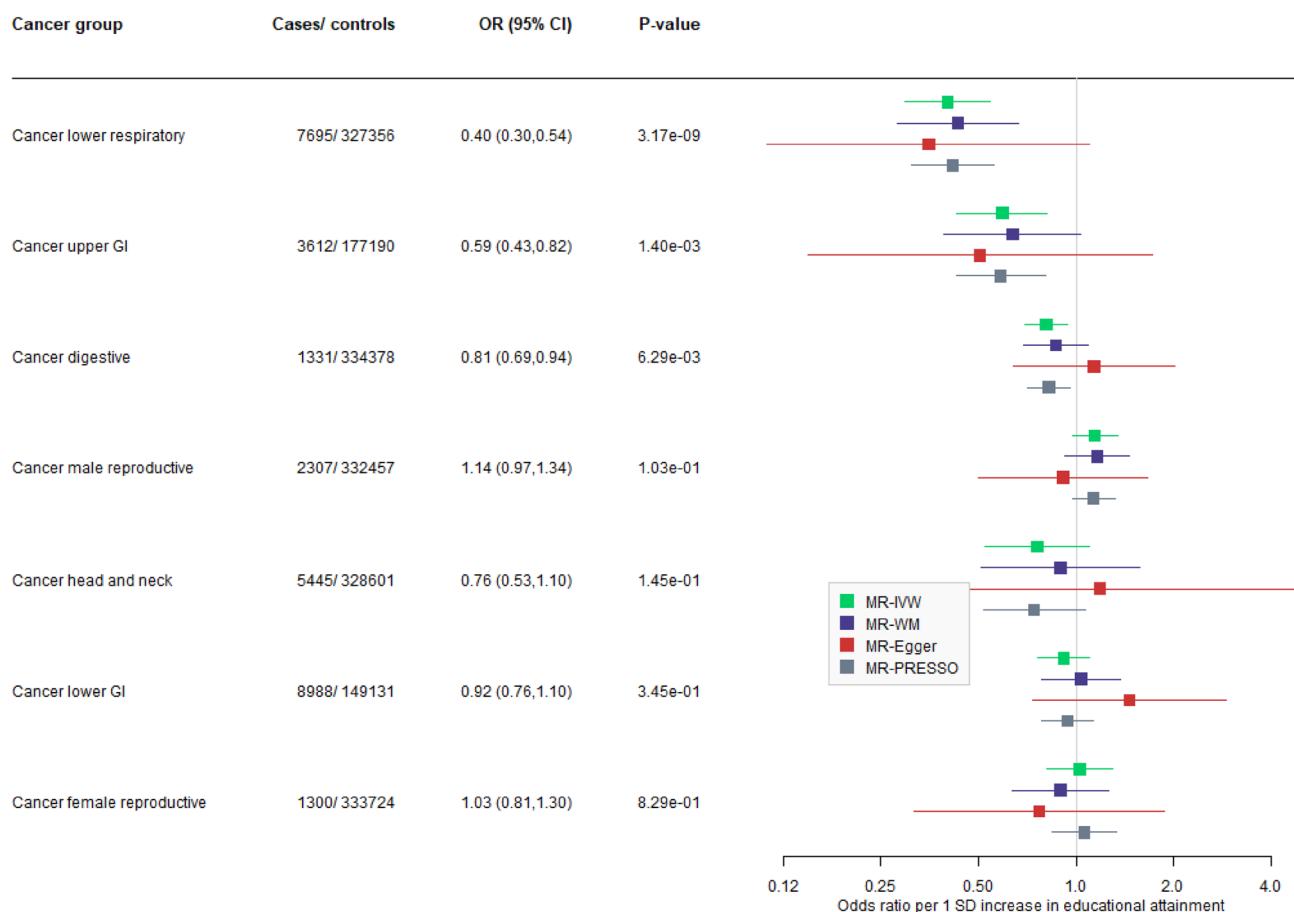


Figure 2. Two-sample Mendelian randomization (MR) estimates per 1 standard deviation increase in educational attainment for 7 cancer groups: digestive, female reproductive, head and neck, lower respiratory, lower gastrointestinal (GI) tract, male reproductive and upper GI tract. Associations were considered statistically significant if MR-IVW and MR-PRESSO p-values were smaller than $0.05/7 = 7.14 \times 10^{-3}$, MR-Egger and MR-weighted median effect estimates were in the same direction and the MR-Egger intercept was not significant ($p > 0.05$).

1.31 to 2.07). Last, vegetable consuming was associated with lower odds of lower respiratory cancer (OR: 0.32, 0.21 to 0.52) (Fig. 4).

MR mediation analysis revealed two significant mediation ratios, after correcting for multiple testing (Supplementary Table 9). The largest part of the protective effect of increased EA on lower odds of lower respiratory cancer was mediated through smoking by 81.7% (50.5 to 100%). Interestingly, vegetable consumption was also mediating factor for the link between EA and lower respiratory cancer with 10.2% (4.4 to 15.9%).

Discussion

The objective of this work was to identify the mediating lifestyle factors linking educational attainment to risk of cancer. We investigated 5 well supported lifestyle factors, including alcohol consumption, body mass index, fruit consumption, lifestyle smoking and vegetable consumption, using summary statistics from large cohorts. In our genetic analysis we found evidence of associations between EA and lower respiratory, upper GI and digestive cancers. MR mediation analysis results revealed that on average, 81.7% of the protective effect of EA on lower respiratory cancer was mediated by lifetime smoking and 10.2% by vegetable consumption.

The findings for the effect of EA on cancer risk are largely in agreement with literature. All potential causal associations found agree with estimates based on observational data found in literature^{5,52–54}. MR methods therefore provide high-level evidence for a causal relationship between EA and cancer. Regarding the effect of EA on the lifestyle factors, the current study also supports existing findings. Previous comparisons of high-school and college EA estimated a 56%⁵⁵ and 64%⁵⁶ lower smoking status. Observational studies indicate that there is a negative association between EA and BMI in higher-income countries⁵⁷. Existing literature provides no association estimates of EA with fruit and vegetable consumption but suggests an effect of socio-economic status on fruit and slightly lower on vegetable consumption^{58,59}. Moreover, increasing EA is associated with increased alcohol intake frequency in MR studies⁶⁰. The MR estimates of smoking on lower respiratory cancer are in agreement with observational studies^{61,62,63} for current smokers, but also with existing MR studies⁶⁴. Moreover, genetically predicted BMI has been found to be positively associated with the risk of oesophageal cancer^{65,66}. Previous observational study on mediation analysis for EA on lung cancer found that, adjustment for smoking

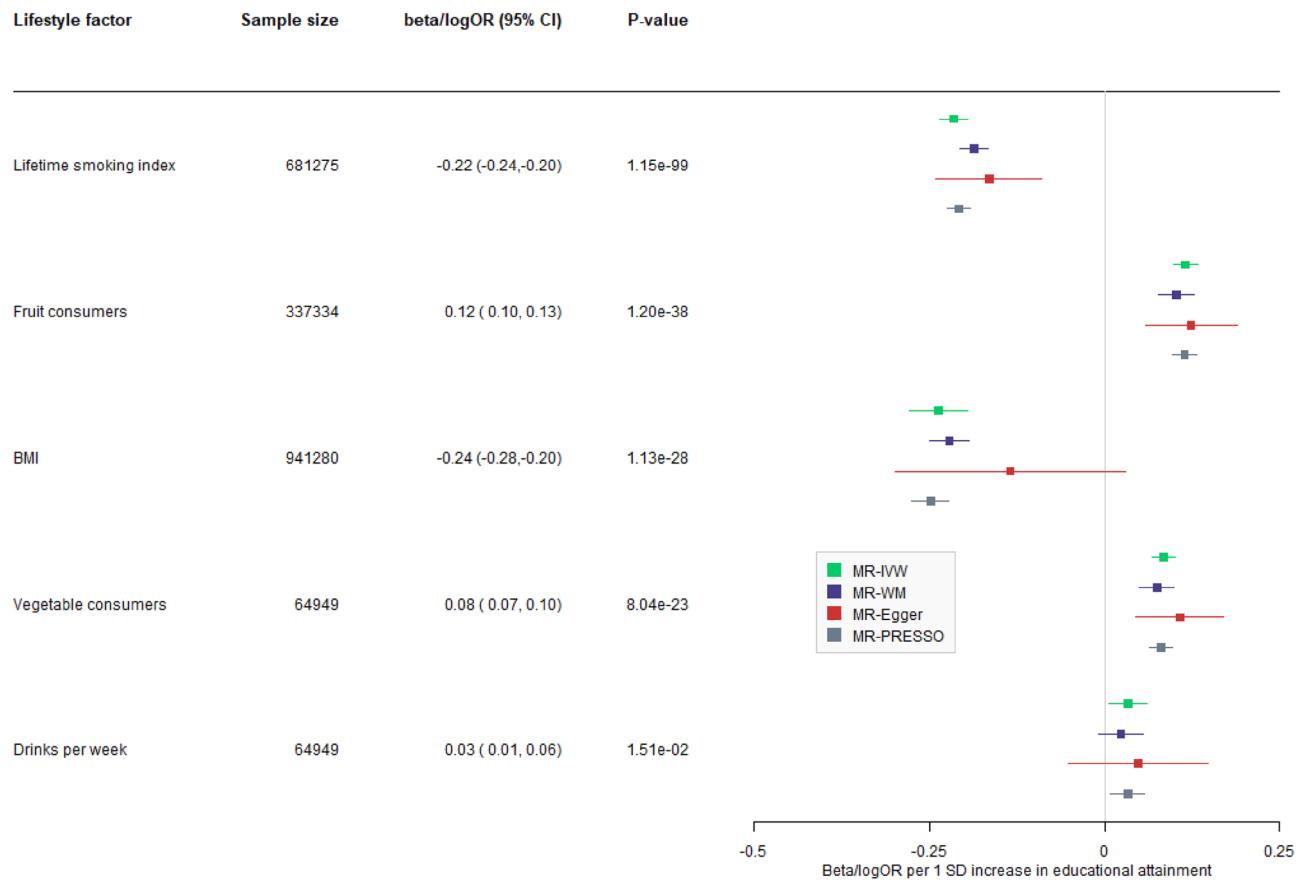


Figure 3. Two-sample Mendelian randomization (MR) estimates per 1 standard deviation increase in educational attainment for 5 lifestyle factors: body mass index (BMI), lifetime smoking index, drinks consumed per week, fruit consumed and vegetable consumed in the past 24 h. Associations were considered statistically significant if MR-IVW and MR-PRESSO p-values were smaller than $0.05/5 = 0.01$, MR-Egger and MR-weighted median effect estimates were in the same direction and the MR-Egger intercept was not significant ($p > 0.05$).

decreased relative educational differences of lung cancer incidence by 50 to 70%⁶⁷. Our MR network study gives a comparable mediation proportion of the protective effect of EA on lower respiratory cancer through lifetime exposure to smoking, around 80% of the total effect, since MR corrects for unknown confounding factors. Last, we identified a significant mediation proportion of the protective effect of EA on lung cancer through vegetable consumption, which is consistent with numerous observational studies suggesting a protective role of fruit and vegetables in lung cancer aetiology^{68–70}.

Prior to making a number of key conclusions based on the data presented in this study, it is prudent to consider some of the limitations. The UKB sample has been shown to not be fully representative of the UK population. Individuals in the UKB have a higher likelihood, compared to the UK population, of being lean, non-smokers, non-drinkers and being older and female. This “healthy volunteer” bias is also affecting the total cancer incidence which may have introduced bias in our results. However, the assessment of effects of exposures on health outcomes in non-representative samples is still generalisable⁷¹. Fruit and vegetable consumption GWAS did not yield many strong genetic instruments compared to other factors possibly due to self-reported information and the limited time this accounts for. In addition, we excluded UKB participants from the lifestyle factors, where possible. However, partial sample overlap of EA, BMI, fruit and vegetable consumption with cancer incidence may have biased some of the MR effect estimates away from the null. Moreover, MR makes the assumption that all associations are linear, however, existing studies have shown a J-shape association between alcohol consumption and cancer risks^{72,73}. In addition, the diagnosis of one cancer type may affect the surveillance, screening, or diagnostic practices for other cancer types. This could introduce bias if the ascertainment of the second cancer is influenced by the awareness or diagnosis of the first cancer. Last, due to the lack of availability of individual level data, we couldn't test if there were any interactions between the exposure and the mediators.

The statistical power of a MR study depends on how much variation in the exposure is explained by the chosen genetic instrumental variables, the sample size and the true causal association between exposure and outcome⁷⁴. Low cancer prevalence limits statistical power in this study, which may in turn explain the lack of statistically significant mediators identified beyond smoking and BMI. In addition, low vegetable and fruit consumption heritability could be limiting the usefulness of their genetic instruments. MVMR corrects for overlap in pathways, however, it is not able to deal with unmeasured confounders possibly interacting with associations⁴⁶. Furthermore, MVMR could be subject to weak instrument bias⁷⁵. A violation of the assumed linearity of causal effects could further bias the estimate⁷⁶.

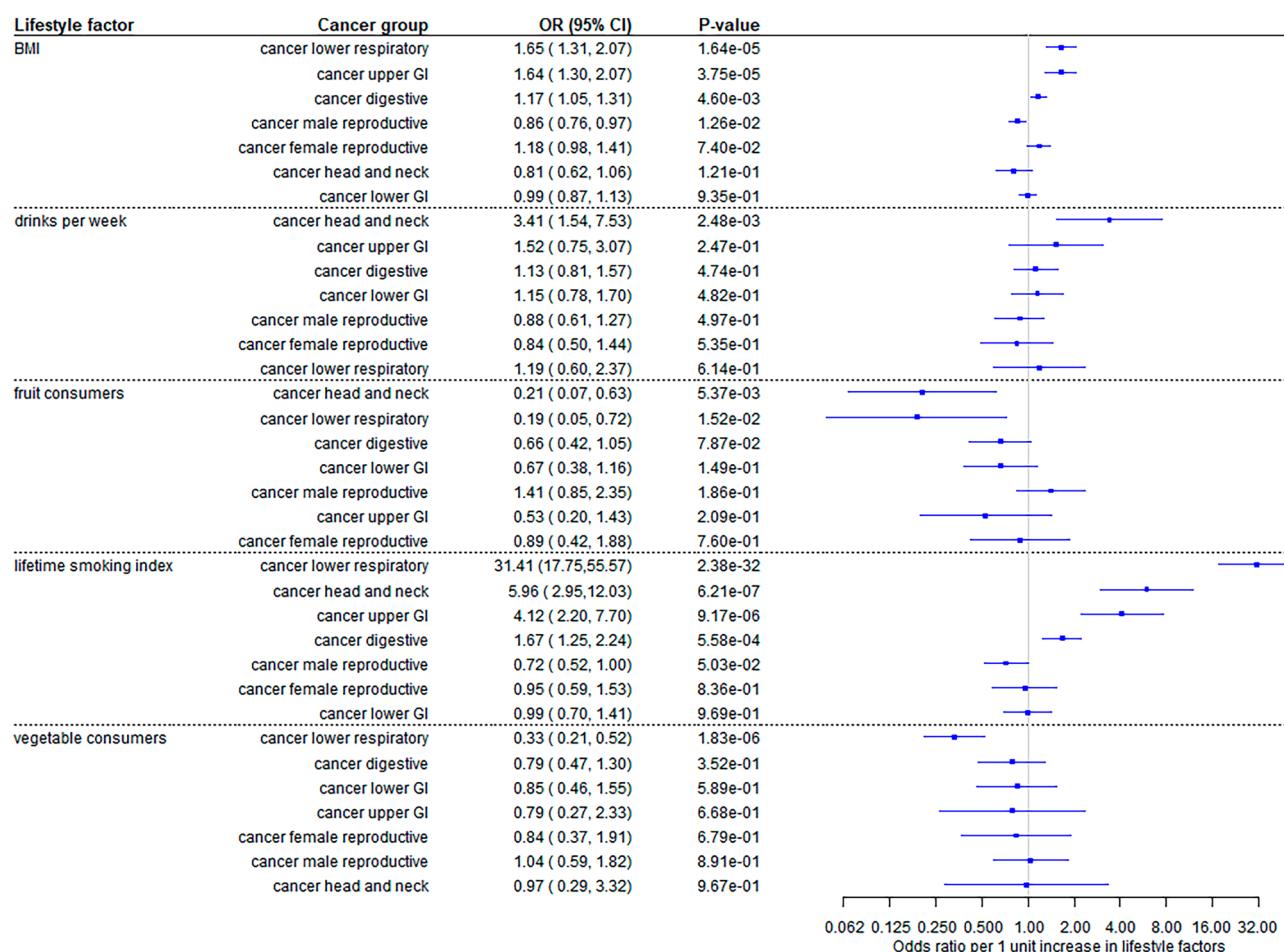


Figure 4. Two-sample multivariable Mendelian randomization (MVMR) estimates per 1 unit increase in 5 lifestyle factors for 7 cancer groups. Associations were considered statistically significant if MVMR-IVW p-values were smaller than $0.05/42 = 1.19 \times 10^{-3}$.

The results presented in this work provide several avenues of future research. The central limitation of statistical power has to be overcome with larger sample sizes or more accurate measurements. Further research on risk factors mediating cancer risk and EA as exposure also hinges on that requirement. Another potential, more methodological avenue of research is concerned with the addition of functional genetic variants to the developed pipeline. This would equip us with more tools to produce accurate estimates and validation thereof.

Medical science has rightly focused on how to treat people with cancer, however, any attempts to prevent the development of cancers are of utmost importance. Our finding, as well as work from others, indicate that the number of years in education has a proportional impact on cancer. Although it is not possible for everyone to reach the same level of EA, this work identifies our priorities in achieving similar benefits through targeted interventions. Currently, the detrimental effects of smoking and obesity are part of primary and secondary school curriculum in several countries. Given their relatively recent inclusion though, it is still early to quantify their effectiveness, as cancer is more common in older individuals that left education before the focus in healthier lifestyles. Nevertheless, our results suggest that we may see the gains of this strategy in the future and that a focus to educate children in primary and secondary education on the dangers of smoking and how to better maintain a healthy weight should be adopted more widely. In addition, investigations should tailor interventions to accommodate people that have already left education at secondary school level. Focusing our efforts of “good-health” education on the identified factors is likely to have the biggest impact on cancer rates.

Data availability

UK Biobank individual level data used in this work can be accessed after applying for access at <https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access>. Genetic association data are publicly available in the original studies.

Received: 8 March 2023; Accepted: 10 February 2024

Published online: 13 March 2024

References

1. Heron, M. & Anderson, R. N. Changes in the leading cause of death: Recent patterns in heart disease and cancer mortality. *NCHS Data Brief* **254**, 1–8 (2016).
2. Gjertsen, F., Bruzzone, S. & Griffiths, C. E. Burden of suicide presented as one of the leading causes of death: Uncover facts or misrepresent statistics?. *J. Glob. Health* <https://doi.org/10.7189/jogh.09.010401> (2019).
3. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**(3), 209–249 (2021).
4. Leuven, E., Plug, E. & Ronning, M. Education and cancer risk. *Labour Econ.* **43**, 106–121 (2016).
5. Siegel, R. L. *et al.* An assessment of progress in cancer control. *CA Cancer J. Clin.* **68**(5), 329–339 (2018).
6. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**(6), 394–424 (2018).
7. Bingham, S. & Riboli, E. Diet and cancer - The European prospective investigation into cancer and nutrition. *Nat. Rev. Cancer* **4**(3), 206–215 (2004).
8. Wilson, L., Bhatnagar, P. & Townsend, N. Comparing trends in mortality from cardiovascular disease and cancer in the United Kingdom, 1983–2013: Joinpoint regression analysis. *Popul. Health Metr.* <https://doi.org/10.1186/s12963-017-0141-5> (2017).
9. Williams, D. R., Priest, N. & Anderson, N. B. Understanding associations among race, socioeconomic status, and health: Patterns and prospects. *Health Psychol.* **35**(4), 407–411 (2016).
10. Shohamim, S. *et al.* Residential area deprivation predicts fruit and vegetable consumption independently of individual educational level and occupational social class: A cross sectional population study in the Norfolk cohort of the European Prospective Investigation into Cancer (EPIC-Norfolk). *J. Epidemiol. Commun. H* **58**(8), 686–691 (2004).
11. Teng, A. M., Atkinson, J., Disney, G., Wilson, N. & Blakely, T. Changing socioeconomic inequalities in cancer incidence and mortality: Cohort study with 54 million person-years follow-up 1981–2011. *Int. J. Cancer* **140**(6), 1306–1316 (2017).
12. Mirowsky, J. & Ross, C. E. *Education, Social Status, and Health* 1st edn. (Routledge, 2003).
13. Boylan, S. *et al.* Socio-economic circumstances and food habits in Eastern, Central and Western European populations. *Public Health Nutr.* **14**(4), 678–687 (2011).
14. Adler, N. E. *et al.* Socioeconomic status and health. The challenge of the gradient. *Am. Psychol.* **49**(1), 15–24 (1994).
15. Warnakulasuriya, S. Significant oral cancer risk associated with low socioeconomic status. *Evid. Based Dent.* **10**(1), 4–5 (2009).
16. Geyer, S., Hemstrom, O., Peter, R. & Vagero, D. Education, income, and occupational class cannot be used interchangeably in social epidemiology. Empirical evidence against a common practice. *J. Epidemiol. Community Health* **60**(9), 804–10 (2006).
17. Ovrum, A. Socioeconomic status and lifestyle choices: Evidence from latent class analysis. *Health Econ.* **20**(8), 971–984 (2011).
18. Nordahl, H. *et al.* Education and cause-specific mortality: The mediating role of differential exposure and vulnerability to behavioral risk factors. *Epidemiology* **25**(3), 389–396 (2014).
19. Gallo, V. *et al.* Social inequalities and mortality in Europe—results from a large multi-national cohort. *PLoS One* **7**(7), e39013 (2012).
20. Hastert, T. A., Ruterbusch, J. J., Beresford, S. A., Sheppard, L. & White, E. Contribution of health behaviors to the association between area-level socioeconomic status and cancer mortality. *Soc. Sci. Med.* **148**, 52–58 (2016).
21. Lund Nilsen, T. I., Johnsen, R. & Vatten, L. J. Socio-economic and lifestyle factors associated with the risk of prostate cancer. *Br. J. Cancer* **82**(7), 1358–1363 (2000).
22. Conway, D. I. *et al.* Socioeconomic inequalities and oral cancer risk: A systematic review and meta-analysis of case-control studies. *Int. J. Cancer* **122**(12), 2811–2819 (2008).
23. Jepsen, P., Johnsen, S. P., Gillman, M. W. & Sorensen, H. T. Interpretation of observational studies. *Heart* **90**(8), 956–960 (2004).
24. Hernan, M. A. The hazards of hazard ratios. *Epidemiology* **21**(1), 13–15 (2010).
25. Schuit, A. J., van Loon, A. J., Tijhuis, M. & Ocke, M. Clustering of lifestyle risk factors in a general adult population. *Prev. Med.* **35**(3), 219–224 (2002).
26. Smith, G. D. & Ebrahim, S. “Mendelian randomization”: Can genetic epidemiology contribute to understanding environmental determinants of disease?. *Int. J. Epidemiol.* **32**(1), 1–22 (2003).
27. Blakely, T., McKenzie, S. & Carter, K. Misclassification of the mediator matters when estimating indirect effects. *J. Epidemiol. Community Health* **67**(5), 458–466 (2013).
28. Davies, N. M., Holmes, M. V. & Davey, S. G. Reading Mendelian randomization studies: A guide, glossary, and checklist for clinicians. *BMJ* **362**, k601 (2018).
29. Thanassoulis, G. & O'Donnell, C. J. Mendelian Randomization nature's randomized trial in the post-genome era. *JAMA J. Am. Med. Assoc.* **301**(22), 2386–2388 (2009).
30. Sudlow, C. *et al.* UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PloS Med.* **12**(3), e1001779 (2015).
31. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**(7726), 203 (2018).
32. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat. Genet.* **54**(4), 437 (2022).
33. Saunders, G. R. B. *et al.* Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature* **612**(7941), 720 (2022).
34. Liu, M. Z. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**(2), 237 (2019).
35. Wootton, R. E. *et al.* Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: A Mendelian randomization study. *Psychol. Med.* **50**(14), 2435–2443 (2020).
36. Leffondré, K., Abrahamowicz, M., Xiao, Y. L. & Siemiatycki, J. Modelling smoking history using a comprehensive smoking index: Application to lung cancer. *Stat. Med.* **25**(24), 4132–4146 (2006).
37. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in similar to 700,000 individuals of European ancestry. *Hum. Mol. Genet.* **27**(20), 3641–3649 (2018).
38. Elsworth B L, M., Alexander, T., Liu, Y., Matthews, P., Hallett, J., Bates, P., Palmer, T., Haberland, V., Smith, G D, Zheng, J., Haycock, P., Gaunt, T R, Hemani, G. The MRC IEU OpenGWAS data infrastructure. *BioRxiv* (2020).
39. Genomes Project C *et al.* An integrated map of genetic variation from 1092 human genomes. *Nature* **491**(7422), 56–65 (2012).
40. Palmer, T. M. *et al.* Using multiple genetic variants as instrumental variables for modifiable risk factors. *Stat. Methods Med. Res.* **21**(3), 223–242 (2012).
41. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**(7), 658–665 (2013).
42. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**(2), 512–525 (2015).
43. Bowden, J., Smith, G. D., Haycock, P. C. & Burgess, S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**(4), 304–314 (2016).
44. Verbanck, M., Chen, C. Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**(5), 693 (2018).
45. Mounier, N. & Kutalik, Z. Bias correction for inverse variance weighting Mendelian randomization. *Genet. Epidemiol.* **47**(4), 314–331 (2023).

46. Burgess, S., Dudbridge, F. & Thompson, S. G. Re: Multivariable Mendelian randomization: The use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181**(4), 290–291 (2015).
47. Burgess, S., Daniel, R. M., Butterworth, A. S. & Thompson, S. G. Network Mendelian randomization: Using genetic variants as instrumental variables to investigate mediation in causal pathways. *Int. J. Epidemiol.* **44**(2), 484–95 (2015).
48. Sanderson, E. Multivariable Mendelian randomization and mediation. *CSH Perspect. Med.* <https://doi.org/10.1101/cshperspect.a038984> (2021).
49. RC T.R: A language and environment for statistical computing. R Foundation for Statistical Computing (2020).
50. Hemani, G. Z. J. et al. The MR-base platform supports systematic causal inference across the human genome. *eLife* <https://doi.org/10.7554/eLife.34408> (2018).
51. Gordon M L, T, Gordon, M M. R Package ‘forestplot’. Advanced Forest Plot Using ‘grid’ Graphics (2016).
52. Siegel, R., Ward, E., Brawley, O. & Jemal, A. Cancer statistics, 2011: The impact of eliminating socioeconomic and racial disparities on premature cancer deaths. *CA Cancer J. Clin.* **61**(4), 212–236 (2011).
53. Gornick, M. E., Eggers, P. W. & Riley, G. F. Associations of race, education, and patterns of preventive service use with stage of cancer at time of diagnosis. *Health Serv. Res.* **39**(5), 1403–1427 (2004).
54. Mouw, T. et al. Education and risk of cancer in a large cohort of men and women in the United States. *PLoS One* **3**(11), e3639 (2008).
55. Cavelaars, A. E. et al. Educational differences in smoking: International comparison. *BMJ* **320**(7242), 1102–1107 (2000).
56. de Walque, D. Does education affect smoking behaviors? Evidence using the Vietnam draft as an instrument for college education. *J. Health Econ.* **26**(5), 877–895 (2007).
57. He, J. B., Chen, X. J., Fan, X. T., Cai, Z. H. & Huang, F. Is there a relationship between body mass index and academic achievement? A meta-analysis. *Public Health* **167**, 111–124 (2019).
58. De Irala-Estevez, J. et al. A systematic review of socio-economic differences in food habits in Europe: Consumption of fruit and vegetables. *Eur. J. Clin. Nutr.* **54**(9), 706–714 (2000).
59. Cooke, L. J. et al. Demographic, familial and trait predictors of fruit and vegetable consumption by pre-school children. *Public Health Nutr.* **7**(2), 295–302 (2004).
60. Rosoff, D. B. et al. Educational attainment impacts drinking behaviors and risk for alcohol dependence: Results from a two-sample Mendelian randomization study with ~780,000 participants. *Mol. Psychiatr.* **26**(4), 1119–1132 (2021).
61. O’Keeffe, L. M. et al. Smoking as a risk factor for lung cancer in women and men: A systematic review and meta-analysis. *BMJ Open* **8**(10), e021611 (2018).
62. Minami, Y. & Tateno, H. Associations between cigarette smoking and the risk of four leading cancers in Miyagi Prefecture, Japan: A multi-site case-control study. *Cancer Sci.* **94**(6), 540–547 (2003).
63. Pesch, B. et al. Cigarette smoking and lung cancer—relative risk estimates for the major histological types from a pooled analysis of case-control studies. *Int. J. Cancer* **131**(5), 1210–1219 (2012).
64. Larsson, S. C. et al. Smoking, alcohol consumption, and cancer: A mendelian randomization study in UK Biobank and international genetic consortia participants. *PloS Med.* **17**(7), e1003178 (2020).
65. Scherubl, H. Excess body weight and gastrointestinal cancer risk. *Visc. Med.* **37**(4), 261–266 (2021).
66. Vithayathil, M. et al. Body size and composition and risk of site-specific cancers in the UK Biobank and large international consortia: A mendelian randomization study. *PloS Med.* **18**(7), e1003706 (2021).
67. Menvielle, G. et al. The role of smoking and diet in explaining educational inequalities in lung cancer incidence. *J. Natl. Cancer Inst.* **101**(5), 321–330 (2009).
68. Wang, C., Yang, T., Guo, X. F. & Li, D. The associations of fruit and vegetable intake with lung cancer risk in participants with different smoking status: A meta-analysis of prospective cohort studies. *Nutrients* **11**(8), 1791 (2019).
69. Vieira, A. R. et al. Fruits, vegetables and lung cancer risk: A systematic review and meta-analysis. *Ann. Oncol.* **27**(1), 81–96 (2016).
70. Wang, J. et al. Citrus fruit intake and lung cancer risk: A meta-analysis of observational studies. *Pharmacol. Res.* **166**, 105430 (2021).
71. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *Am. J. Epidemiol.* **186**(9), 1026–34 (2017).
72. Park, J. H., Han, K., Hong, J. Y., Park, Y. S. & Park, J. O. Association between alcohol consumption and pancreatic cancer risk differs by glycaemic status: A nationwide cohort study. *Eur. J. Cancer* **163**, 119–27 (2022).
73. Zhang, X. Y. et al. Alcohol consumption and risk of cardiovascular disease, cancer and mortality: a prospective cohort study. *Nutr. J.* **20**(1), (2021).
74. Brion, M. J., Shakhsbazov, K. & Visscher, P. M. Calculating statistical power in Mendelian randomization studies. *Int. J. Epidemiol.* **42**(5), 1497–1501 (2013).
75. Burgess, S. & Thompson, S. G. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Stat. Med.* **30**(11), 1312–1323 (2011).
76. Rees, J. M. B., Wood, A. M. & Burgess, S. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Stat. Med.* **36**(29), 4705–4718 (2017).

Author contributions

Conceptualization: F.D. and A.S., Funding acquisition: F.D., Formal Analysis: L.Z. and A.S., Data curation: L.Z., A.S. and H.A., Supervision: F.D. and T.D., Methodology: L.Z., A.S. and F.D., Writing – original draft: L.Z., A.S., F.D. and T.D., Writing – review & editing: L.Z., A.S., H.A., F.D. and T.D.

Funding

The work was supported by a Brunel Research Initiative and Enterprise Fund to FD. The funders were not involved in the analysis and interpretation of the data, in the writing of the report, or in the decision to submit the paper for publication. This research has been conducted using the UK Biobank Resource under project 44566 (<https://www.ukbiobank.ac.uk/2018/12/genetic-and-non-genetic-factors-able-to-predict-and-modify-the-risk-of-different-types-of-cancer/>).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-54259-7>.

Correspondence and requests for materials should be addressed to L.Z. or F.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024