Phenome-wide Mendelian randomisation analysis of 378,142 cases reveals risk factors for eight common cancers

Molly Went[1]*, Amit Sud[1,2,3,4,5]*, Charlie Mills[1]*, Abi Hyde[1*#], Richard Culliford[1], Philip Law[1,] Jayaram Vijayakrishnan[1], Ines Gockel[6], Carlo Maj[7], Johannes Schumacher[7], Claire Palles[8], Martin Kaiser[1,9], Richard Houlston[1]

1. Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK.
2. Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA.
3. Broad Institute of MIT and Harvard, Cambridge, MA, USA.
4. Harvard Medical School, Boston, MA, USA.
5. Department of Immuno-Oncology, Nuffield Department of Medicine, University of Oxford, Oxford, UK.
6. Department of Visceral, Transplant, Thoracic and Vascular Surgery, University Hospital of Leipzig, Leipzig, Germany.
7. Center for Human Genetics, University Hospital of Marburg, Marburg, Germany.
8. Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK.
9. The Royal Marsden Hospital NHS Foundation Trust, London, UK.

*contributed equally

[#]present address: Department of Engineering, University of Cambridge, Cambridge, UK.

Correspondence: Molly Went, Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey SM2 5NG, United Kingdom; e-mail: molly.went@icr.ac.uk

Key words: Cancer, aetiology, risk, Mendelian randomisation, genome wide association study.

Number of words: 151 abstract, 3975 main text, 1494 methods.

Number of references: 70

ABSTRACT

For many cancers there are only a few well-established risk factors. Here, we use summary data from genome-wide association studies (GWAS) in a Mendelian randomisation (MR) phenome-wide association study (PheWAS) to identify potentially causal relationships for over 3,000 traits. Our outcome datasets comprise 378,142 cases across breast, prostate, colorectal, lung, endometrial, oesophageal, renal, and ovarian cancers, as well as 485,715 controls. We complement this analysis by systematically mining the literature space for supporting evidence. In addition to providing supporting evidence for well-established risk factors (smoking, alcohol, obesity, lack of physical activity), we also find sex steroid hormones, plasma lipids, and telomere length as determinants of cancer risk. A number of the molecular factors we identify may prove to be potential biomarkers. Our analysis, which highlights aetiological similarities and differences in common cancers, should aid public health prevention strategies to reduce cancer burden. We provide a R/Shiny app (https://software.icr.ac.uk/app/mrcan) to visualise findings.

INTRODUCTION

Cancer is currently the third major cause of death with an estimated 18.1 million new cases and nearly 10 million cancer deaths in 2020[1]. By 2030 it is predicted there are likely to be 26 million new cancer cases and 17 million cancer-related deaths annually[2]. Such projections have renewed efforts to identify risk factors to inform cancer prevention programmes.

For many cancers, despite significant epidemiological research, there are few well-established risk factors. Although randomised-controlled trials (RCTs) are the gold standard for establishing causal relationships, they are often impractical or unfeasible because of cost, time, and ethical issues. Conversely, case-control studies can be complicated by biases such as reverse causation and confounding. Mendelian randomisation (MR) is an analytical strategy that uses germline genetic variants as instrumental variables (IVs) to infer potentially causal relationships (**Fig. 1A**)[3]. The random assortment of these genetic variants at conception mitigates against reverse causation bias. Moreover, in the absence of pleiotropy (*i.e*. the presence of an association between variants and disease through additional pathways), MR can provide unconfounded disease risk estimates. Elucidating disease causality using MR is gaining popularity especially given the availability of data from large genome-wide association studies (GWAS) and well-developed analytical frameworks[3].

Most MR studies of cancer have been predicated on assumptions about disease aetiology or have sought to evaluate purported associations from conventional observational epidemiology[3,4]. A recently proposed agnostic strategy, termed MR-PheWAS, integrates the phenome-wide association study (PheWAS) with MR methodology to identify potential causal relationships considering hitherto previously unexamined traits[5].

To identify potentially causal relationships for eight common cancers: breast, prostate, colorectal (CRC), lung, endometrial, oesophageal, renal cell carcinoma (RCC), ovarian, and reveal intermediates of risk, we conducted a MR-PheWAS study utilising 378,142 cases and 485,715 controls. We integrated findings with a systematic mining of the literature space to provide supporting evidence and derive a more comprehensive description of disease aetiology (**Fig. 1B**)[6].

RESULTS

Phenotypes and genetic instruments

After filtering we analysed 3,661 traits, proxied by 336,191 genetic variants in conjunction with summary genetic data from published GWAS of breast, prostate, colorectal, lung, endometrial, oesophageal, renal, and ovarian cancers (**Table 1**; **Supplementary Table 1**). The number of single nucleotide polymorphisms (SNPs) used as genetic instruments for each trait ranged from one to 1,335. **Fig. 2** shows the power of our MR study to identify potentially causal relationships between each of the genetically defined traits and each cancer type. The median proportion of variance explained (PVE) by SNPs used as IVs for each of the 3,661 traits evaluated as risk factors was 3.4% (0.01–84%). Our power to demonstrate relationships *a priori* for each cancer type reflects in part inevitably the size of respective GWAS datasets (**Supplementary Table 2**).

Causal associations predicted by MR

To aid interpretation, we grouped traits related to established cancer risk factors (*i.e.* smoking, obesity and alcohol) and those for which current evidence is inconclusive into the following categories, using a similar approach to Markozannes *et al*[4]: cardiometabolic; dietary intake; anthropometrics; immune and inflammatory factors; fatty acid (FA) and lipoprotein metabolism; lifestyle, reproduction, education and behaviour; metabolomics and proteomics; miscellaneous.

Given the large number of traits being evaluated, we categorised the support for potentially causal relationships between non-binary traits and cancers into four hierarchical levels of statistical significance *a priori*: robust, probable, suggestive, and non-significant (**Fig. 3**; **Methods**). Out of the 27,066 graded associations, MR analyses provided robust evidence for a potentially causal relationship with 123 phenotypes (0.5% of total MR analyses), 174 with probable evidence (0.6% of total), 1,652 with suggestive evidence (6% of total). Across the eight cancer types, the largest number of robust associations were observed for endometrial cancer with 37 robust associations, followed by RCC (n = 32), CRC (n = 21), lung (n = 20), breast (n = 10), oesophageal (n = 3) and prostate cancer (n = 1). No robust MR associations were observed for ovarian cancer (**Supplementary Table 3**).

Across all the cancer types, anthropometric traits showed the highest number of robust relationships (n = 32; 0.1%), followed by lifestyle, reproduction, education, and behaviour (n = 17; 0.06%). No

robust associations were observed for dietary intake or cardiometabolic categories (**Supplementary Table 3**).

To visualise the strength and direction of effect of the relationship between each of the traits examined and risk of each cancer type and, where appropriate, their respective subtypes we provide a R/Shiny app (https://software.icr.ac.uk/app/mrcan). **Fig. 4** shows a screenshot of the app for selected traits across the eight different types of cancer.

Many of the identified potentially causal relationships, especially those that were statistically robust or probable, have been reported in previous MR studies and are related to established risk factor categories[4,7,8]. Notably: (i) the relationship between metrics of increased body mass index (BMI) with an increased risk of colorectal (*Robust*, OR_{SD} = 1.19, 95% CI: 1.11 - 1.27, $P = 2.01 \times 10^{-7}$), lung (*Suggestive*, OR_{SD} = 1.22, 95% CI: 1.11 - 1.34, $P = 3.25 \times 10^{-5}$), renal (*Robust*, OR_{SD} = 1.63, 95% CI: 1.44 - 1.85, $P = 2.19 \times 10^{-14}$), endometrial (*Robust*, OR_{SD} = 1.90, 95% CI: 1.67 - 2.15, $P = 3.92 \times 10^{-23}$) and ovarian (*Suggestive*, OR_{SD} = 1.11, 95% CI: 1.01 - 1.22, $P = 2.98 \times 10^{-2}$) cancers[9]; (ii) cigarette smoking with an increased risk of lung cancer[10]; (iii) traits related to higher alcohol consumption and increased risk of oesophageal (*Suggestive*, OR_{SD} = 2.69, 95% CI: 1.58 - 4.49, $P = 2.76 \times 10^{-4}$), CRC (*Suggestive*, OR_{SD} = 1.39, 95% CI: 1.01 - 1.91, $P = 4.53 \times 10^{-2}$), lung (*Probable*, OR_{SD} = 1.55, 95% CI: 1.18 - 2.04, $P = 1.49 \times 10^{-3}$), RCC (*Suggestive*, OR_{SD} = 1.25, 95% CI: 1.03 - 1.53, $P = 2.42 \times 10^{-2}$), endometrial (*Suggestive*, OR_{SD} = 1.23, 95% CI: 1.01 - 1.8515, $P = 4.41 \times 10^{-2}$) and ovarian (*Suggestive*, OR_{SD} = 1.22, 95% CI: 1.05 - 1.40, $P = 7.32 \times 10^{-3}$) cancers[11]; (iv) traits indicative of reduced physical activity and sedentary behaviour with an increased risk of multiple cancers, including breast, lung, colorectal and endometrial[12]. As anticipated, exposure traits pertaining to cigarette smoking were not causally related to lung cancer in never smokers. Paradoxically, but as reported in previous MR analyses, increased BMI was associated with reduced risk of prostate (*Suggestive*, OR_{SD} = 0.82, 95% CI: 0.70 - 0.95, $P = 1.03 \times 10^{-2}$) and breast (*Probable*, OR_{SD} = 0.84, 95% CI: 0.76 - 0.93, $P = 8.40 \times 10^{-4}$) cancer, and an inverse relationship between smoking and prostate cancer risk was shown[9,13]. Our analysis also supports the reported relationship between higher levels of sex hormone-binding globulin with reduced endometrial cancer risk (*Robust*, OR_{SD} = 0.81, 95% CI: 0.74 - 0.89, $P = 9.00 \times 10^{-6}$) and a relationship between testosterone with risk of endometrial (*Probable*, OR_{SD} = 1.48, 95% CI: 1.12 - 1.96, $P = 5.32 \times 10^{-3}$) and breast (*Probable*, OR_{SD} = 1.24, 95% CI: 1.09 - 1.42, $P = 1.43 \times 10^{-3}$) cancer[14,15]. Notably, exposure traits related to testosterone levels were only predicted to be causally associated with luminal-A and luminal-B breast cancer subtypes.

We found associations between genetically predicted high serum vitamin B12 with increased risks of CRC (*Suggestive*, OR_{SD} = 1.09, 95% CI: 1.01 - 1.18, $P = 2.53 \times 10^{-2}$) and prostate (*Suggestive*, OR_{SD} = 1.08, 95% CI: 1.02 - 1.14, $P = 8.87 \times 10^{-3}$) cancer, higher serum calcium (*Suggestive*, OR_{SD} = 1.19, 95% CI: 1.05 - 1.35, $P = 5.92 \times 10^{-3}$) and 25-hydroxyvitamin-D (*Suggestive*, OR_{SD} = 1.18, 95% CI: 1.00 - 1.38, $P = 4.63 \times 10^{-2}$) with an increased risk of RCC, higher blood selenium with decreased risks of CRC (*Suggestive*, OR_{SD} = 0.91, 95% CI: 0.85 - 0.98, $P = 9.49 \times 10^{-3}$) and oesophageal (*Suggestive*, OR_{SD} = 0.84, 95% CI: 0.72 - 0.99, $P = 3.42 \times 10^{-2}$) cancer and higher methionine (*Suggestive*, OR_{SD} = 0.09, 95% CI: 0.01 - 0.99, $P = 4.90 \times 10^{-2}$) and zinc (*Suggestive*, OR_{SD} = 0.94, 95% CI: 0.89 - 0.99, $P = 1.77 \times 10^{-2}$) with reduced CRC risk. We observed no association between genetically predicted blood levels of circulating carotenoids or vitamins B6 and E for any of the cancers. With respect to dietary intake our analysis demonstrated associations between genetically predicted higher levels of coffee intake (*Probable*, OR_{SD} = 0.67, 95% CI: 0.55 - 0.82, $P = 1.03 \times 10^{-4}$), oily fish (*Probable*, OR_{SD} = 0.66, 95% CI: 0.52 - 0.84, $P = 5.41 \times 10^{-4}$), and cheese intake (*Probable*, OR_{SD} = 0.75, 95% CI: 0.64 - 0.89, $P = 1.08 \times 10^{-3}$) with reduced CRC risk and associations between genetically predicted beef (*Suggestive*, OR_{SD} = 1.65, 95% CI: 1.05 - 2.60, $P = 3.07 \times 10^{-2}$) and poultry (*Suggestive*, OR_{SD} = 2.10, 95% CI: 1.06 - 4.16, $P = 3.24 \times 10^{-2}$) intake and elevated CRC risk.

In terms of glucose homeostasis, no relationship between genetically predicted blood glucose or glycated haemoglobin was shown for any of the eight cancers. However, higher levels of genetically predicted levels of fasting insulin (*Probable*, OR_{SD} = 1.78, 95% CI: 1.25 - 2.52, $P = 1.33 \times 10^{-3}$) and insulin growth factor 1 (IGF-1) (*Suggestive*, OR_{SD} = 1.06, 95% CI: 1.01 - 1.12, $P = 3.26 \times 10^{-2}$) and lower proinsulin (*Probable*, OR_{SD} = 0.89, 95% CI: 0.82 - 0.96, $P = 3.09 \times 10^{-3}$) showed associations with CRC. Additionally, an association between proinsulin and RCC (*Suggestive*, OR_{SD} = 0.80, 95% CI: 0.67 - 0.96, $P = 1.50 \times 10^{-2}$), fasting insulin and lung (*Suggestive*, OR_{SD} = 1.40, 95% CI: 1.03 - 1.90, $P = 3.29 \times 10^{-2}$) and endometrial (*Suggestive*, OR_{SD} = 1.76, 95% CI: 1.02 - 3.03, $P = 4.24 \times 10^{-2}$) cancers, and IGF-1 levels and breast cancer (*Probable*, OR_{SD} = 1.07, 95% CI: 1.02 - 1.13, $P = 6.21 \times 10^{-3}$) was observed.

Amongst genetically predicted higher levels of lipoproteins, the only associations were between high density lipoprotein cholesterol (HDL-C) and breast cancer risk (*Probable*, OR_{SD} = 1.08, 95% CI: 1.03 - 1.12, $P = 6.28 \times 10^{-4}$) , low density lipoprotein cholesterol (LDL-C) an elevated risk of CRC (*Suggestive*, OR_{SD} = 1.10, 95% CI: 1.01 - 1.20, $P = 2.18 \times 10^{-2}$), and total cholesterol and increasing ovarian cancer risk (*Suggestive*, OR_{SD} = 1.05, 95% CI: 1.01 - 1.09, $P = 2.67 \times 10^{-2}$). Genetically predicted levels of

plasma FAs showed an association with reduced cancer risk. Specifically, for the omega-6 polyunsaturated FAs, increased levels of arachidonic acid (20:4n6) (*Suggestive*, OR_{SD} = 1.04, 95% CI: 1.02 - 1.05, $P = 6.11 \times 10^{-5}$) and gamma-linoleic acid (18:3n6) (*Suggestive*, OR_{SD} = 35.29, 95% CI: 13.65 - 91.24, $P = 1.94 \times 10^{-13}$) and lower levels of linoleic acid (18:2n6) (*Suggestive*, OR_{SD} = 0.96, 95% CI: 0.95 - 0.97, $P = 3.11 \times 10^{-13}$) and adrenic acid (22:4n6) (*Suggestive*, OR_{SD} = 3.28, 95% CI: 2.34 - 4.59, $P = 5.88 \times 10^{-12}$) with increased risk of CRC; for the omega-3 polyunsaturated FAs, linoleic acid (*Suggestive*, OR_{SD} = 1.02, 95% CI: 1.00 - 1.04, $P = 3.05 \times 10^{-2}$) and eicosapentaenoic acid (*Suggestive*, OR_{SD} = 0.42, 95% CI: 0.19 - 0.94, $P = 3.44 \times 10^{-2}$) showed an association with ovarian cancer risk while arachidonic acid was associated with endometrial cancer (*Suggestive*, OR_{SD} = 0.98, 95% CI: 0.97 - 0.99, $P = 2.83 \times 10^{-3}$). Performing a leave-one-out and single SNP analysis (**Supplementary Table 4** and **5,** respectively) we found, similar to previously published work, that the majority of associations with respect to omega-3 and omega-6 fatty acids are driven by correlated associations within the *FADS* locus [16,17].

A relationship between longer lymphocyte telomere length (LTL) and an increased risk of six of the eight cancer types was identified - RCC (*Robust,* OR_{SD} = 2.01, 95% CI: 1.65 - 2.45, $P = 3.27 \times 10^{-12}$), lung (*Robust,* OR_{SD} = 1.61, 95% CI: 1.41 - 1.84, $P = 2.48 \times 10^{-12}$), breast (*Probable,* OR_{SD} = 1.12, 95% CI: 1.04 - 1.20, $P = 2.07 \times 10^{-3}$), prostate (*Probable,* OR_{SD} = 1.25, 95% CI: 1.10 - 1.43, $P = 9.77 \times 10^{-4}$), colorectal (*Suggestive*, OR_{SD} = 1.13, 95% CI: 1.00 - 1.28, $P = 4.24 \times 10^{-2}$) and ovarian cancer (*Suggestive*, OR_{SD} = 1.18, 95% CI: 1.05 - 1.33, $P = 4.88 \times 10^{-3}$).

In addition to a robust association between higher HLA-DR dendritic plasmacytoid levels and risk of prostate cancer (OR_{SD} = 1.05, 95% CI: 1.03 - 1.06, $P = 5.22 \times 10^{-10}$), 26 probable associations between genetically predicted levels of other circulating immune and inflammatory factors were shown across the cancers studied. These included higher levels of IL-18 with reduced risk of lung cancer (*Probable*, OR_{SD} = 0.89, 95% CI: 0.83 - 0.96, $P = 2.00 \times 10^{-3}$), with specificity for lung cancer in never smokers. For proteomic traits, we conducted a Bayesian colocalisation analysis to determine whether genetic variants influencing protein levels and cancer risk are shared by considering the strongest proteomic associations with a clear gene target and a cis-IV (*i.e.* within 1Mb; **Methods**) with *P*-value $<1\times10^{-6}$ in the outcome cancer. We identified KDEL motif-containing protein 2 (*KDELC2)* and RCC, as well as Copine-1 (*CPNE1)* and Immunoglobulin superfamily containing leucine-rich repeat protein 2 (*ISLR2)* and breast cancer as having a high posterior probability of a shared variant (*i.e.* $PP_{H4} > 0.8$). In contrast, Kunitz-type protease inhibitor 2 (*SPINT2)* and prostate cancer, as well as Semaphorin-3G

(*SEMA3G)* and CRC, were shown to have distinct variants at the gene target (*i.e*. PP_{H3} >0.8; **Supplementary Table 6**). Results for the IV at Histo-blood group ABO system transferase (*ABO*) with ovarian cancer were indeterminate (PP_{H4} = 0.67 and PP_{H3} = 0.33).

Our MR analysis provides support for a relationship between rectal polyps and CRC (β = 95.59, Standard Error (SE) = 4.99, *P* = 6.88 x 10^{-82})[18], benign breast disease and breast cancer[19], and oesophageal reflux with risk of oesophageal cancer (β = 0.27, SE = 0.08, *P* = 1.30 x 10^{-3}) (**Supplementary Table 7**)[20]. Other associations included possible relationships between pulmonary fibrosis and lung cancer[21], as well as the relationship between a diagnosis of schizophrenia and lung cancer (β = 0.10, SE = 0.04, *P* = 2.89 x 10^{-2}), which has been previously reported in conventional epidemiological studies[22]. It was noteworthy, however, that we did not find evidence to support the purported relationship between hypertension and risk of developing RCC[23]. Similarly, our analysis did not provide evidence to support a causal relationship between either type 1 or type 2 diabetes and an increased cancer risk.

Multivariable MR of biologically related traits

Selected traits within our analysis may show pleiotropic effects with other traits and work by Burgess *et al*[24] has shown that MR can only assess the causal effect of a risk factor on an outcome by using genetic variants that are solely associated with the risk factor of interest. To address pleiotropy we performed multivariable MR (MVMR) as a form of mediation analysis focusing on known biologically related traits. Specifically, we examined the role of IGF-1 and height on breast and colorectal cancer risk[25]; lipid traits on breast and colorectal cancer risk[26,27]; and fasting insulin, sex hormone-binding globulin levels (SHBG), BMI and testosterone on endometrial cancer risk[28](**Supplementary Table 8**). In the MVMR analysis of HDL-C, LDL-C and triglyceride levels, we found the relationship of increasing HDL cholesterol with breast cancer risk and increasing LDL-C with colorectal cancer risk remained significant in a model accounting for these biologically related traits (OR_{MVMR}= 1.06, P_{MVMR}= 0.03 and OR_{MVMR}= 1.09, P_{MVMR}= 0.04, respectively). Considering height and IGF-1 and their association with CRC risk and breast cancer risk, IGF-1 remained significantly associated with breast cancer risk (OR_{MVMR}= 1.06, P_{MVMR}= 0.049), while height remained significantly associated with colorectal cancer risk (OR_{MVMR}= 1.06, P_{MVMR}= 0.045). In contrast IGF-1 became non-significant (*P*= 0.16), which may suggest that the relationship between IGF-1 levels and CRC is mediated through the relationship with height. Finally, MVMR of fasting insulin, SHBG, BMI and testosterone and their effect on endometrial cancer, attenuated the significance of association (*P* > 0.5) of fasting insulin and bioavailable

testosterone with the outcome, while SHBG and BMI remained significant, but with a modest decrease in effect size (OR_{MVMR}= 0.61, P_{MVMR}= 0.02 and OR_{MVMR}= 1.65, P_{MVMR}= $6.37x10^{-5)}$. Hence this suggests that bioavailable testosterone and fasting insulin do not have an independent effect on endometrial cancer risk and the associations are likely to be mediated, at least in part, through SHBG and BMI.

Literature-mined support for MR defined relationships

To provide support for the associations and to gain molecular insights into the underlying biological basis of relationships we performed triangulation through systematic literature mining. We identified 55,105 literature triples across the eight different cancer types and 680,375 literature triples across the MR defined putative risk factors (**Supplementary Table 9**). Overlapping risk factor-cancer pairings from our MR analysis yielded on average 49 potential causal relationships. **Supplementary Table 10** stratifies the literature space size by trait category while recognising that identified relationships with a small literature space could be reflective of deficiencies in semantic mapping relationships with large literature spaces supporting triangulation. **Supplementary Table 11** provides the complete list of potential mediators for each trait. Illustrating the use of triangulation using a large literature space (defined herein as >50 triples) to support potentially causal relationships, **Fig. 5** highlights four notable examples (IGF-1, LAG-3, IL-18, and PRDX1).

IGF-1, which is reported to play a role in multiple cancers, appears to mediate its effect in part through beta-catenin and BRAF signalling, modulating CRC and breast cancer risk[29]. Whilst LAG-3 inhibition is an attractive therapeutic target in restoring T-cell function, we demonstrate genetically elevated LAG-3 levels as being associated with reduced CRC, endometrial and lung cancer. In all three of these cancers, the association appears to be at least partly mediated through IL-10. The seemingly paradoxical relationship between LAG-3 levels and tumourgenesis may reflect potentiation of T-cell function by serum LAG-3 rather than cell membrane expressed LAG-3[30]. We identify genetically predicted IL-18 levels as being associated with an increased risk of lung cancer. Our literature mining also supports a role for the decoy inhibitory protein, IL-18BP as being a mediator of lung cancer risk as well as IL-10, IL-12, IL-4 and TNF[31]. Finally, PRDX1, a member of the peroxiredoxin family of antioxidant enzymes, interacts with the androgen receptor to enhance its transactivation resulting in increased EGFR-mediated signalling and an increased prostate cancer risk[32].

DISCUSSION

By performing a MR-PheWAS we have been able to agnostically examine the relationship between multiple traits and the risk of eight different cancer types, restricted only by the availability of suitable genetic instruments. Importantly, many of the traits we examined have not previously been the subject of conventional epidemiological studies or been assessed by MR. Comparing our work with a recent systematic review of the previously published MR studies of cancer, less than 10% of the MR exposures in this study had been the subject of previous investigations[4]. In addition, 85% of those traits which we found were significant had not previously been examined. Even for risk factors that were examined in many previous analyses, the number of cases and controls in our study has afforded greater power to identify potential causal associations. This has allowed us to exclude large effects on cancer risk for most exposure traits examined.

In addition to predicting causal relationships for the well-established lifestyle traits, which validates our approach, we implicate other lifestyle factors that have been putatively associated by observational epidemiology contributing to cancer risk. For example, the protective effects of physical activity (*Suggestive*) with lung cancer risk, oily fish (*Probable*) for CRC risk and fresh/dried fruit intake (*Probable*) for breast cancer risk. Several of the potentially causal relationships we identify have been the subject of studies of individual traits and include the association between longer LTL with increased risk of RCC and lung cancers (*Robust*); sex steroid hormones and risk of breast and endometrial cancer and circulating lipids with CRC and breast cancer. Clustering of MR predicted causal effect sizes for each trait cancer relationship highlights the importance of risk factors common to many cancers but also reveal differences in their impact in part likely to be reflective of underlying biology (**Fig. 6**).

Using genetic instruments for plasma proteome constituents has allowed us to identify hitherto unexplored potential risk factors for a number of the cancers, including: the cytokine like molecule, FAM3D, which plays a role in host defence against inflammation associated carcinogenesis with lung cancer[33]; the autophagy associated cytokine cardiotrophin-1 with lung (*Probable*), endometrial (*Suggestive*), prostate (*Suggestive*) and breast (*Suggestive*) cancer and the tumour progression associated antigen CD63 with endometrial cancer[34,35]. Levels of these and other plasma proteins potentially represent biomarkers worthy of future prospective studies. Furthermore, for proteomic traits with *cis*-IVs previous work has found that an MR association with colocalization evidence is

associated with a higher likelihood of a particular target-indication pair being successful in drug discovery[36].

A principal assumption in MR is that variants used as IVs are associated with the exposure trait under investigation. We therefore used SNPs associated with exposure traits at genome-wide significance. Furthermore, only IVs from European populations were used to limit bias from population stratification. Our MR analysis does, however, have limitations. Firstly, we were limited to studying phenotypes with genetic instruments available, moreover traits such as food intake or television watching can be highly correlated with other exposures making deconvolution of the causal risk factor problematic[37–39]. While MVMR can be used to account for the correlation between traits, calculation of conditional F-statistics for dietary traits yielded weak instruments ($F < 3$), which precludes their inclusion in an MVMR model due to weak instrument bias. Secondly, correcting for multiple testing guards against false positives especially when based on a single exposure outcome. However, the potential for false negatives is not unsubstantial. Since we have not adjusted for between trait correlations, our associations are inevitably conservative. Thirdly, for several traits, we had limited power to demonstrate associations of small effect. Fourthly, not unique to our MR analysis, is the inability of our study to deconvolute time-varying effects of genetic variants as evidenced by the relationship between obesity and breast cancer risk[40]. Finally, as with all MR studies, excluding pleiotropic IVs is challenging. To address this, we incorporated information from weighted median and mode-based estimate methods, to classify the strength of potentially causal associations. For groups of traits susceptible to pleiotropy (*e.g.*, lipids) we also demonstrated how their incorporation into a MVMR model can affect the relationship between these traits and outcome. There are inevitably limitations to such modelling as exemplified by the strong relationship between plasma FA and risk of CRC which has been shown to be driven by the pleiotropic *FADS* locus which has a profound effect on the metabolism of multiple FA through its gene expression[41].

A major concern articulated regarding any MR-PheWAS is the need to provide supporting evidence from alternative sources. Herein we have sought to address this by conducting a systematic interrogation of the literature space and potentially identify intermediates to explain relationships. Furthermore, we performed MVMR to deconvolute relationships where multiple traits appear to influence cancer risk. Although literature mined data can be noisy and driven by publication bias, we have been able to provide a narrative of the potentially causal relationships for several risk factors, which are attractive candidates for molecular validation.

342

343 While complementary studies are required to delineate the exact biological mechanisms
344 underpinning associations, our analysis does however highlight important targets for primary
345 prevention of cancer in the population. The limited power to robustly characterise relationships
346 between some exposure traits and cancer in this study, provides an impetus for larger MR studies.
347 Finally, we recognise that MR is not infallible and replication and triangulation of findings using
348 different data sources, and if possible, benchmarking against RCTs is highly desirable. Such efforts
349 could identify additional factors as targets to reduce the overall burden of cancer.

METHODS

Ethics approval

The analysis was undertaken using published GWAS data, hence ethical approval was not required.

Study design

Our study had four elements. Firstly, the identification of genetic variants serving as instruments for exposure traits under investigation; secondly, the acquisition of GWAS data for the eight cancers; thirdly, MR analysis; fourthly, triangulation through literature mining to provide supporting evidence for potential causal relationships (**Fig. 1B**).

Genetic variants serving as instruments

SNPs considered genetic instruments, were identified from published studies or MR-Base (**Supplementary Table 2**). For each SNP, the corresponding effect estimate on a trait expressed in *per* standard deviation (SD) units (assuming a *per* allele effect) and standard error (SE) was obtained. Only SNPs with a minor allele frequency >0.01 and a trait association of *P*-values $<5 \times 10^{-8}$ in a European population GWAS were considered as instruments. We excluded correlated SNPs at a linkage disequilibrium threshold of $r^2 > 0.01$, retaining SNPs with the strongest effect. For binary traits we restricted our analyses to traits with a medical diagnosis, excluding cancer. We removed duplicate exposure traits based on manual curation.

Cancer GWAS summary statistics

To examine the association of each genetic instrument with cancer risk, we used summary GWAS effect estimates from: (1) Online consortia resources, for breast (BCAC; https://bcac.ccge.medschl.cam.ac.uk/, accessed July 2022) and prostate cancer (PRACTICAL; http://practical.icr.ac.uk/; accessed July 2022)[42,43]; (2) GWAS Catalog (https://www.ebi.ac.uk/gwas/), for ovarian, CRC, endometrial and lung cancers (accessed September 2022)[44–46]; (3) Investigators of published work, for RCC and oesophageal cancer[47–49]. Cancer subtype summary statistics were available for lung, breast, and ovarian cancers. As the UK Biobank was used to obtain genetic instruments for many traits investigated, the CRC and oesophageal GWAS association statistics were recalculated from primary data excluding UK Biobank samples to avoid sample overlap bias (**Table 1**). Single nucleotide polymorphisms were harmonised

to ensure that the effect estimates of SNPs on exposure traits and cancer risk referenced the same allele (**Supplementary Table 12)**[50].

Statistical analysis

For each SNP, effects were estimated for cancer as an odds ratio (OR) per SD unit increase in the putative risk factor (ORSD), with 95% confidence intervals (CIs), using the Wald ratio[51]. For traits with multiple SNPs as IVs, causal effects were estimated under an inverse variance weighted random-effects (IVW-RE) model as the primary measurement as it is robust in the presence of pleiotropic effects, provided any heterogeneity is balanced at mean zero (**Supplementary Tables 3, 13-15**)[52]. Weighted median estimate (WME) and mode-based estimates (MBE) were obtained to assess the robustness of findings (**Supplementary Table 16)**[53,54]. Directional pleiotropy was assessed using MR-Egger regression (**Supplementary Table 17**)[55]. The MR Steiger test was used to infer the direction of potentially causal effect for continuous exposure traits (**Supplementary Table 18**)[56]. For this we estimated the PVE using Cancer Research UK lifetime risk estimates for each tumour type (**Supplementary Table 19**). A leave-one-out strategy under the IVW-RE model was employed to assess the potential impact of outlying and pleiotropic SNPs (**Supplementary Table 4**)[57]. This sensitivity analysis tests the effect of performing MR on the IVs leaving one SNP out in turn. It can be used to identify when one SNP is driving the association as, when this SNP is removed, we can expect to see an attenuation of the MR association significance. Because two-sample MR of a binary risk factor and a binary outcome can be biased, we primarily considered whether there exists a significant non-zero effect, and only report ORs for consistency[58]. For proteomic traits which had an IV located *cis* (+/- 1Mb) of the gene target we performed colocalisation using coloc[59]. This enumerates the four possible configurations of causal variants for two traits, calculating support for each model based on a Bayes factor. Adopting prior probabilities of p_1, $p_2 = 1\times10^{-4}$ and $p_{12} = 1\times10^{-5}$, a posterior probability ≥0.80 was considered as supporting a specific model. For analyses of selected traits using MVMR we used the *mv_multiple* function in the TwoSampleMR package. MVMR was applied to investigate which of these traits within the same category had independent pleiotropic effects on a specific cancer. We restricted our MVMR analyses to traits which had ≥2 IVs and for which we had access to full summary statistics required for the analysis. Statistical analyses were performed using the TwoSampleMR package v0.5.6 (https://github.com/MRCIEU/TwoSampleMR) and MendelianRandomization package in R (v3.4.0)[50].

Estimation of study power

The power of MR to predict a causal relationship depends on the PVE by the instrument[60]. We excluded instruments with a F-statistic <10 since these are considered indicative of evidence for weak instrument bias[61]. We calculated conditional F-statistics for the traits using the *condFstat* function in the MendelianRandomzation package[62] (**Supplementary Table 20**). We estimated the genetic correlation between traits using Linkage-Disequilibrium Adjusted Kinships (LDAK) software (**Supplementary Table 21**). We derived LD matrices for the genetic variants using the *ld_matrix* function in TwoSampleMR. We estimated study power, stipulating a *P*-value of 0.05 for each target *a priori* across a range of effect sizes as *per* Brion *et al.* (**Supplementary Table 2**)[63]. Since power estimates for binary exposure traits and binary outcomes in a two-sample setting are unreliable, we did not estimate study power for binary traits[58].

Assignment of statistical significance

The support for a causal relationship with non-binary traits was categorised into four hierarchical levels of statistical significance *a priori*: robust ($P_{IVW\text{-}RE} < 1.4\times10^{-5}$; corresponding to a *P*-value of 0.05 after Bonferroni correction for multiple testing (0.05/3,500), P_{WME} or $P_{MBE} < 0.05$, predicted true causal direction and >1 IVs), probable ($P_{IVW\text{-}RE} < 0.05$, P_{WME} or $P_{MBE} < 0.05$, predicted true causal direction and >1 IVs), suggestive ($P_{IVW\text{-}RE} < 0.05$ or $P_{WALD} < 0.05$), and non-significant ($P_{IVW\text{-}RE} \geq 0.05$ or $P_{WALD} \geq 0.05$) (**Supplementary Table 22)**. Robust associations are those that remain significant after correcting for multiple testing, the predicted direction of the effect is predicted to be from the exposure to the cancer risk and multiple MR methods report a significant association. We consider these associations to have the strongest statistical evidence, by virtue of the concordance between various MR methods and statistical validation tests. Probable associations are those that do not remain significant after correcting for multiple testing, but the remaining conditions are the same as for robust traits. We include this classification to account for the large number of traits tested in this analysis, noting that when taken in isolation these traits may be reported as having potentially causal associations with cancer. Suggestive traits are those in which show significance *P* <0.05, but where one of the following conditions are flouted: the direction of effect may not be predicted to be from exposure to cancer outcome, or there is no significant consensus between the multiple MR methods. Additionally, significant associations for which only one SNP could be used as an IV are classified as suggestive. This was chosen to reflect the potential uncertainties that arise when performing MR using a Wald ratio test with a single IV. Finally, all other traits are classified as non-significant, indicating that it is unlikely that there is any potentially causal association. While non-significant

associations can be due to low statistical power, they also indicate that a moderate causal effect is unlikely. For binary traits we classified associations as being supported ($P < 0.05$) or not supported ($P > 0.05$; **Supplementary Tables 6, 23-25**).

Support for causality

To strengthen evidence for causal relationships predicted from the MR analysis we exploited the semantic predications in Semantic MEDLINE Database (SemMedDB), which is based on all PubMed citations[64]. Within SemMedDB pairs of terms connected by a predicate which are collectively known as 'literature triples' (*i.e.* 'subject term 1' – predicates – 'object term 2'). These literature triples represent semantic relationships between biological entities derived from published literature. To interrogate SemMedDB we queried MELODI Presto and EpiGraphDB to facilitate data mining of epidemiological relationships for molecular and lifestyle traits[65–67]. For each putative risk factor-cancer pair the set of triples were overlapped, and common terms identified to reveal potentially causal pathways and inform aetiology. Based on the information profile of all literature mined triples, we considered literature spaces with >50 literature triples as being viable, corresponding to 90% of the information content[68]. We complemented this systematic text mining by referencing reports from the World Cancer Research Fund/American Institute for Cancer Research, and the International Agency for Cancer Research Global Cancer Observatory, as well as querying specific putative relationships in PubMed[69,70].

DATA AVAILABILITY

Genetic instruments can be obtained through MR-Base or from published work (**Supplementary Table 2**). Summary GWAS cancer data are available from: https://bcac.ccge.medschl.cam.ac.uk/bcacdata/ (breast cancer); http://practical.icr.ac.uk/blog/?page_id=8088 (prostate cancer); GWAS Catalogue ID: GCST004481 (ovarian cancer); GWAS Catalogue ID: GCST006464 (endometrial cancer); GWAS Catalogue ID: GCST004748 (lung cancer); direct communication with consortia (renal and esophageal cancers); - phs001415.v1.p1, phs001315.v1.p1, phs001078.v1.p1, phs001903.v1.p1, phs001856.v1.p1 and phs001045.v1.p1 (US based studies) and GWAS Catalog ID: GCST90129505 (European based studies) colorectal cancer. Source data are provided within the supplementary tables of this paper.

CODE AVAILABILITY

We provide custom code used to generate the results presented in this study at https://github.com/houlstonlab/MR-PheWAS

REFERENCES

1. 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**, 209–249 (2021).
2. International Agency for Research on Cancer. *World Cancer Report 2008*. (International Agency for Research on Cancer, 2008).
3. Thompson, J. STEPHEN BURGESS, SIMON G. THOMPSON. Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation. Boca Raton: CRC Press. *Biometrics* vol. 73 356–356 Preprint at https://doi.org/10.1111/biom.12674 (2017).
4. Markozannes, G. *et al.* Systematic review of Mendelian randomization studies on risk of cancer. *BMC Med.* **20**, 41 (2022).
5. Millard, L. A. C. *et al.* MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization. *Scientific Reports* vol. 5 Preprint at https://doi.org/10.1038/srep16645 (2015).
6. Mathison, S. Why Triangulate? *Educational Researcher* vol. 17 13–17 Preprint at https://doi.org/10.3102/0013189x017002013 (1988).
7. Cancer risk factors. *WCRF International* https://www.wcrf.org/diet-activity-and-cancer/risk-factors/ (2022).
8. Website. https://gco.iarc.fr/causes/.
9. Bhaskaran, K. *et al.* Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults. *Lancet* **384**, 755–765 (2014).
10. Doll, R. & Hill, A. B. The mortality of doctors in relation to their smoking habits; a preliminary report. *Br. Med. J.* **1**, 1451–1455 (1954).
11. Bagnardi, V., Blangiardo, M., La Vecchia, C. & Corrao, G. A meta-analysis of alcohol drinking and cancer risk. *British Journal of Cancer* vol. 85 1700–1705 Preprint at https://doi.org/10.1054/bjoc.2001.2140 (2001).
12. Schmid, D. & Leitzmann, M. F. Television Viewing and Time Spent Sedentary in Relation to Cancer Risk: A Meta-Analysis. *JNCI: Journal of the National Cancer Institute* vol. 106 Preprint at

https://doi.org/10.1093/jnci/dju098 (2014).

13. Islami, F., Moreira, D. M., Boffetta, P. & Freedland, S. J. A systematic review and meta-analysis of tobacco use and prostate cancer mortality and incidence in prospective cohort studies. *Eur. Urol.* **66**, 1054–1064 (2014).
14. Allen, N. E. *et al.* Endogenous sex hormones and endometrial cancer risk in women in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Endocr. Relat. Cancer* **15**, 485–497 (2008).
15. Key, T., Appleby, P., Barnes, I., Reeves, G. & Endogenous Hormones and Breast Cancer Collaborative Group. Endogenous sex hormones and breast cancer in postmenopausal women: reanalysis of nine prospective studies. *J. Natl. Cancer Inst.* **94**, 606–616 (2002).
16. Borges, M. C. *et al.* Role of circulating polyunsaturated fatty acids on cardiovascular diseases risk: analysis using Mendelian randomization and fatty acid genetic association data from over 114,000 UK Biobank participants. *BMC Med.* **20**, 210 (2022).
17. May-Wilson, S. *et al.* Pro-inflammatory fatty acid profile and colorectal cancer risk: A Mendelian randomisation analysis. *Eur. J. Cancer* **84**, 228–238 (2017).
18. Stryker, S. J. *et al.* Natural history of untreated colonic polyps. *Gastroenterology* **93**, 1009–1013 (1987).
19. Hartmann, L. C. *et al.* Benign breast disease and the risk of breast cancer. *N. Engl. J. Med.* **353**, 229–237 (2005).
20. Lagergren, J., Bergström, R., Lindgren, A. & Nyrén, O. Symptomatic Gastroesophageal Reflux as a Risk Factor for Esophageal Adenocarcinoma. *New England Journal of Medicine* vol. 340 825–831 Preprint at https://doi.org/10.1056/nejm199903183401101 (1999).
21. Turner-Warwick, M., Lebowitz, M., Burrows, B. & Johnson, A. Cryptogenic fibrosing alveolitis and lung cancer. *Thorax* vol. 35 496–499 Preprint at https://doi.org/10.1136/thx.35.7.496 (1980).
22. Catts, V. S., Catts, S. V., O'Toole, B. I. & Frost, A. D. J. Cancer incidence in patients with schizophrenia and their first-degree relatives - a meta-analysis. *Acta Psychiatr. Scand.* **117**, 323–336 (2008).
23. Alcala, K. *et al.* The relationship between blood pressure and risk of renal cell carcinoma. *Int. J. Epidemiol.* **51**, 1317–1327 (2022).
24. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic

variants to estimate causal effects. *Am. J. Epidemiol.* **181**, 251–260 (2015).

25. Larsson, S. C. *et al.* Insulin-like growth factor-1 and site-specific cancers: A Mendelian randomization study. *Cancer Med.* **9**, 6836–6842 (2020).
26. Beeghly-Fadiel, A. *et al.* A Mendelian randomization analysis of circulating lipid traits and breast cancer risk. *Int. J. Epidemiol.* **49**, 1117–1131 (2020).
27. Johnson, K. E. *et al.* The relationship between circulating lipids and breast cancer risk: A Mendelian randomization study. *PLoS Med.* **17**, e1003302 (2020).
28. Hazelwood, E. *et al.* Identifying molecular mediators of the relationship between body mass index and endometrial cancer risk: a Mendelian randomization analysis. *BMC Med.* **20**, 125 (2022).
29. Desbois-Mouthon, C. *et al.* Insulin and IGF-1 stimulate the β-catenin pathway through two signalling cascades involving GSK-3β inhibition and Ras activation. *Oncogene* vol. 20 252–259 Preprint at https://doi.org/10.1038/sj.onc.1204064 (2001).
30. Burnell, S. E. A. *et al.* Seven mysteries of LAG-3: a multi-faceted immune receptor of increasing complexity. *Immunother Adv* **2**, ltab025 (2022).
31. Zhou, T. *et al.* IL-18BP is a secreted immune checkpoint and barrier to IL-18 immunotherapy. *Nature* **583**, 609–614 (2020).
32. Park, S.-Y. *et al.* Peroxiredoxin 1 interacts with androgen receptor and enhances its transactivation. *Cancer Res.* **67**, 9294–9303 (2007).
33. Liang, W. *et al.* FAM3D is essential for colon homeostasis and host defense against inflammation associated carcinogenesis. *Nat. Commun.* **11**, 5912 (2020).
34. Akkoc, Y. *et al.* Tumor-derived CTF1 (cardiotrophin 1) is a critical mediator of stroma-assisted and autophagy-dependent breast cancer cell migration, invasion and metastasis. *Autophagy* **19**, 306–323 (2023).
35. Koh, H. M., Jang, B. G. & Kim, D. C. Prognostic Value of CD63 Expression in Solid Tumors: A Meta-analysis of the Literature. *In Vivo* **34**, 2209–2215 (2020).
36. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).

37. May-Wilson, S. *et al.* Large-scale GWAS of food liking reveals genetic determinants and genetic correlations with distinct neurophysiological traits. *Nat. Commun.* **13**, 2743 (2022).
38. Abdellaoui, A., Dolan, C. V., Verweij, K. J. H. & Nivard, M. G. Gene–environment correlations across geographic regions affect genome-wide association studies. *Nature Genetics* vol. 54 1345–1354 Preprint at https://doi.org/10.1038/s41588-022-01158-0 (2022).
39. Wade, K. H. *et al.* Applying Mendelian randomization to appraise causality in relationships between nutrition and cancer. *Cancer Causes Control* **33**, 631–652 (2022).
40. Swanson, S. A., Tiemeier, H., Ikram, M. A. & Hernán, M. A. Nature as a Trialist?: Deconstructing the Analogy Between Mendelian Randomization and Randomized Trials. *Epidemiology* **28**, 653–659 (2017).
41. May-Wilson, S. *et al.* Pro-inflammatory fatty acid profile and colorectal cancer risk: A Mendelian randomisation analysis. *European Journal of Cancer* vol. 84 228–238 Preprint at https://doi.org/10.1016/j.ejca.2017.07.034 (2017).
42. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* **52**, 572–581 (2020).
43. Schumacher, F. R. *et al.* Author Correction: Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **51**, 363 (2019).
44. Phelan, C. M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet.* **49**, 680–691 (2017).
45. O'Mara, T. A. *et al.* Identification of nine new susceptibility loci for endometrial cancer. *Nat. Commun.* **9**, 3166 (2018).
46. McKay, J. D. *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* **49**, 1126–1132 (2017).
47. Fernandez-Rozadilla, C. *et al.* Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nat. Genet.* (2022) doi:10.1038/s41588-022-01222-9.
48. Scelo, G. *et al.* Genome-wide association study identifies multiple risk loci for renal cell carcinoma. *Nat. Commun.* **8**, 15724 (2017).

589 49. Schröder, J. *et al.* GWAS meta-analysis of 16 790 patients with Barrett's oesophagus and oesophageal
590 adenocarcinoma identifies 16 novel genetic risk loci and provides insights into disease aetiology beyond
591 the single marker level. *Gut* (2022) doi:10.1136/gutjnl-2021-326698.
592 50. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human
593 phenome. *Elife* **7**, (2018).
594 51. Burgess, S., Small, D. S. & Thompson, S. G. A review of instrumental variable estimators for Mendelian
595 randomization. *Statistical Methods in Medical Research* vol. 26 2333–2355 Preprint at
596 https://doi.org/10.1177/0962280215597579 (2017).
597 52. Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data
598 Mendelian randomization. *Stat. Med.* **36**, 1783–1802 (2017).
599 53. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian
600 Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.*
601 **40**, 304–314 (2016).
602 54. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian
603 randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* **46**, 1985–1998 (2017).
604 55. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect
605 estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
606 56. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely
607 measured traits using GWAS summary data. *PLoS Genet.* **13**, e1007081 (2017).
608 57. Corbin, L. J. *et al.* BMI as a Modifiable Risk Factor for Type 2 Diabetes: Refining and Understanding
609 Causal Estimates Using Mendelian Randomization. *Diabetes* **65**, 3002–3007 (2016).
610 58. Burgess, S. & Labrecque, J. A. Mendelian randomization with a binary exposure variable: interpretation
611 and presentation of causal estimates. *Eur. J. Epidemiol.* **33**, 947–952 (2018).
612 59. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies
613 using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
614 60. Stock, J. H., Wright, J. H. & Yogo, M. A Survey of Weak Instruments and Weak Identification in
615 Generalized Method of Moments. *Journal of Business & Economic Statistics* vol. 20 518–529 Preprint at

https://doi.org/10.1198/073500102288618658 (2002).

61. Staiger, D. & Stock, J. Instrumental Variables Regression with Weak Instruments. Preprint at https://doi.org/10.3386/t0151 (1994).

62. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).

63. Brion, M.-J. A., Shakhbazov, K. & Visscher, P. M. Calculating statistical power in Mendelian randomization studies. *Int. J. Epidemiol.* **42**, 1497–1501 (2013).

64. Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G. & Rindflesch, T. C. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics* **28**, 3158–3160 (2012).

65. Elsworth, B. & Gaunt, T. R. MELODI Presto: a fast and agile tool to explore semantic triples derived from biomedical literature. *Bioinformatics* vol. 37 583–585 Preprint at https://doi.org/10.1093/bioinformatics/btaa726 (2021).

66. Liu, Y. *et al.* EpiGraphDB: a database and data mining platform for health data science. *Bioinformatics* **37**, 1304–1311 (2021).

67. Vabistsevits, M., Robinson, T., Elsworth, B., Liu, Y. & Gaunt, T. Integrating Mendelian randomization and literature-mined evidence for breast cancer risk factors. *bioRxiv* (2022) doi:10.1101/2022.07.19.22277795.

68. Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* vol. 27 623–656 Preprint at https://doi.org/10.1002/j.1538-7305.1948.tb00917.x (1948).

69. Cancer risk factors. *WCRF International* https://www.wcrf.org/diet-activity-and-cancer/risk-factors/ (2022).

70. Website. https://gco.iarc.fr/causes/.

ACKNOWLEDGMENTS

R.S.H. acknowledges grant support from Cancer Research UK (C1298/A8362), the Wellcome Trust (214388) and Myeloma UK. A.S. is in receipt of a National Institute for Health Research (NIHR) Academic Clinical Lectureship, funding from the Royal Marsden Biomedical Research Centre, a Starter Grant from the Academy of Medical Sciences and is the recipient of a Wellcome Trust Early Career Award (227000/Z/23/Z). M.K. is supported by a fellowship from the David Forbes-Nixon Foundation. We acknowledge pump-priming funding from the Royal Marsden Biomedical Research Centre Early Diagnosis, Detection and Stratified Prevention Theme. This is a summary of independent research supported by the NIHR Biomedical Research Centre at the Royal Marsden NHS Foundation Trust and the Institute of Cancer Research. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. Support from the DJ Fielding Medical Research Trust is also acknowledged. A.H. was in receipt of a summer studentship from the Genetics Society. We thank Alex Cornish for providing code and critically appraising the manuscript.

The breast cancer genome-wide association analyses for BCAC and CIMBA were supported by Cancer Research UK (PPRPGM-Nov20\100002, C1287/A10118, C1287/A16563, C1287/A10710, C12292/A20861, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565) and the Gray Foundation, The National Institutes of Health (CA128978, X01HG007492- the DRIVE consortium), the PERSPECTIVE project supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344) and the Ministère de l'Économie, Science et Innovation du Québec through Genome Québec and the PSRSIIRI-701 grant, the Quebec Breast Cancer Foundation, the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), the European Union's Horizon 2020 Research and Innovation Programme (634935 and 633784), the Post-Cancer GWAS initiative (U19 CA148537, CA148065 and CA148112 - the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer (CRN-87521), the Komen Foundation for the Cure, the Breast Cancer Research Foundation and the Ovarian Cancer Research Fund. All studies and funders are listed in Zhang H et al (Nat Genet, 2020).

The colorectal cancer genome-wide association analysis was supported by Ulrike Peters (GECCO), Stephanie Schmit (CCFR), Stephen Gruber (CORECT), Ian Tomlinson (CORGI, SCOT), and Malcolm Dunlop (SOCCS). Full study details and funders are listed in Fernandez-Rozadilla C et al (Nat Genet,

2023). The Prostate cancer genome-wide association analyses are supported by the Canadian Institutes of Health Research, European Commission's Seventh Framework Programme grant agreement n° 223175 (HEALTH-F2-2009-223175), Cancer Research UK Grants C5047/A7357, C1287/A10118, C1287/A16563, C5047/A3354, C5047/A10692, C16913/A6135, and The National Institute of Health (NIH) Cancer Post-Cancer GWAS initiative grant: No. 1 U19 CA 148537-01 (the GAME-ON initiative). We would also like to thank the following for funding support: The Institute of Cancer Research and The Everyman Campaign, The Prostate Cancer Research Foundation, Prostate Research Campaign UK (now PCUK), The Orchid Cancer Appeal, Rosetrees Trust, The National Cancer Research Network UK, The National Cancer Research Institute (NCRI) UK. We are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust. The Prostate Cancer Program of Cancer Council Victoria also acknowledge grant support from The National Health and Medical Research Council, Australia (126402, 209057, 251533, , 396414, 450104, 504700, 504702, 504715, 623204, 940394, 614296,), VicHealth, Cancer Council Victoria, The Prostate Cancer Foundation of Australia, The Whitten Foundation, PricewaterhouseCoopers, and Tattersall's. EAO, DMK, and EMK acknowledge the Intramural Program of the National Human Genome Research Institute for their support. Genotyping of the OncoArray was funded by the US National Institutes of Health (NIH) [U19 CA 148537 for ELucidating Loci Involved in Prostate cancer SuscEptibility (ELLIPSE) project and X01HG007492 to the Center for Inherited Disease Research (CIDR) under contract number HHSN268201200008I] and by Cancer Research UK grant A8197/A16565. Additional analytic support was provided by NIH NCI U01 CA188392 (PI: Schumacher). Funding for the iCOGS infrastructure came from: the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118, C1287/A 10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 – the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund. The BPC3 was supported by the U.S. National Institutes of Health, National Cancer Institute (cooperative agreements U01-CA98233 to D.J.H., U01-CA98710 to S.M.G., U01-CA98216 toE.R., and U01-CA98758 to B.E.H., and Intramural Research Program of NIH/National Cancer Institute, Division of Cancer Epidemiology and Genetics). CAPS GWAS study was supported by the Swedish Cancer Foundation (grant no 09-0677, 11-484, 12-823), the Cancer Risk Prediction Center

(CRisP; www.crispcenter.org), a Linneus Centre (Contract ID 70867902) financed by the Swedish Research Council, Swedish Research Council (grant no K2010-70X-20430-04-3, 2014-2269). PEGASUS was supported by the Intramural Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health.

AUTHOR CONTRIBUTIONS

Contribution: M.W, A.S, C.M. and R.S.H designed the study. M.W, A.S., C.M., A.H., R.C., and P.L. performed statistical analyses; M.W, A.S., C.M., and R.S.H. drafted the manuscript; all authors reviewed, read, and approved the final manuscript.

COMPETING INTERESTS

The authors declare no competing financial interests.

TABLES AND FIGURES LEGENDS

Figure 1. Principles of Mendelian randomisation (MR) and study overview: (a) Assumptions in MR that need to be satisfied to derive unbiased causal effect estimates. Dashed lines represent direct causal and potential pleiotropic effects that would violate MR assumptions. A, indicates genetic variants used as IVs are strongly associated with the trait; B, indicates genetic variants only influence cancer risk through the trait; C, indicates genetic variants are not associated with any measured or unmeasured confounders of the trait-cancer relationship. SNP, single-nucleotide polymorphism; **(b) Study overview.** Genetic variants serving as instruments for exposure traits under investigation were identified from MRBase or PubMed. GWAS data for the eight cancers was acquired and MR analysis was performed. Results were triangulated through literature mining to provide supporting evidence for potentially causal relationships. Created with BioRender.com. GWAS, genome-wide association study.

Figure 2. Power to predict causal relationships in the Mendelian randomisation analysis across the eight different cancers. Each line represents an individual trait with the line colour indicating the F-statistic, a measure of instrument strength. The analysis of most traits is well powered across a modest range of odds ratios. Generally, better powered traits are those with a higher F-statistic. F-stat: F-statistic.

Figure 3. Hierarchical classification of associations. Potentially causal relationships between non-binary traits and cancers were categorised into four hierarchical levels of statistical significance *a priori*; robust ($P_{IVW\text{-}RE} < 1.4\times10^{-5}$; corresponding to a *P*-value of 0.05 after Bonferroni correction for multiple testing (0.05/3,500), P_{WME} or $P_{MBE} < 0.05$, predicted true causal direction and >1 IVs), probable ($P_{IVW\text{-}RE} < 0.05$, P_{WME} or $P_{MBE} < 0.05$, predicted true causal direction and >1 IVs), suggestive ($P_{IVW\text{-}RE} < 0.05$ or $P_{WALD} < 0.05$), and non-significant ($P_{IVW\text{-}RE} \geq 0.05$ or $P_{WALD} \geq 0.05$). Weighted median estimates (WME)[53] and mode-based estimates (MBE)[54] were used in addition to an inverse weighted random effects (IVW-RE) model, to assess the robustness of our findings, while MR-Egger regression assessed the extent to which directional pleiotropy could affect causal estimates[55]. MR-Steiger was used to ascertain that the exposure trait influenced the outcome and not *vice versa* [56]. Binary traits were classified associations as being supported ($P < 0.05$) or not supported ($P > 0.05$). MR, Mendelian randomisation; IV, instrumental variable.

Figure 4. **Bubble plot of the potentially causal relationship between selected traits and risk of different cancers.** The columns correspond to different cancer types. The colours on the heatmap correspond to the strength of associations (odds ratio) and their direction (red positively correlated, blue negatively correlated). *P*-values represent the results from two-sided tests and are unadjusted. The size of each node corresponds to the -$\log_{10}$ *P*-value, with increasing size indicating a smaller *P*-value. In the available R/Shiny app (https://software.icr.ac.uk/app/mrcan), moving the cursor on top of each bubble will reveal the underlying MR statistics.

Figure 5. Sankey diagram of literature spaces for exemplar cancer risk factors. These diagrams illustrate the relationship between exposure traits and cancers via their linked literature triples. The thickness of the line connecting two mediating traits indicates the frequency with which that triple is mentioned in the literature. Relationships for: (a) *IGF-1* and colorectal cancer; (b) *IL-18* and lung cancer; (c) *LAG-3* and endometrial cancer; (d) *PRDX1* and prostate cancer. AR: androgen receptor; EGF: epidermal growth factor; EGFR: epidermal growth factor receptor; ESRK: extracellular signal regulated kinases; GMCSF: granulocyte-macrophage colony-stimulating factor; HACII: histocompatibility antigens class II; IFNG: interferon gamma; MM: matrix metalloproteinases; MM9: matrix metalloproteinase 9; PHRP: parathyroid hormone-related protein; PMH: phosphoric monoester hydrolases; PPT: phenylpyruvate tautomerase; PR: progesterone receptor; RIG: recombinant interferon-gamma; TF: transcription factor; TNF: tumour necrosis factor; TSG: tumour suppressor genes; VEGFA: vascular endothelial growth factor A.

Figure 6. Heatmap and dendrogram showing clustering of potentially causal associations between traits and cancer risk. Heatmap based on *Z*-statistics using the clustering method implemented in the pheatmap function within R. Colours correspond to the strength of associations and their direction (red positive association with risk, blue inverse association with risk). Trait classes are annotated on the left. Only traits showing an association for at least one cancer type are shown. Further heatmaps for individual classes of traits are shown in **Supplementary Figures**.

Table 1. Details of cancer genome-wide association studies used in the Mendelian randomisation analysis. The number of cases and controls, the number of studies contributing to the meta-analyses and the associated publication and GWAS catalogue IDs are provided for each cancer GWAS. Where applicable, the number of cases and controls in given histological subtypes are also provided.