STUDY PROTOCOL

A protocol for using human genetic data to identify circulating protein level changes that are the causal consequence of cancer processes

Lisa M. Hobson[1,2*], **Richard M. Martin**[2,3], **Karl Smith-Byrne**[3,4], **George Davey Smith**[1,2,3], **Gibran Hemani**[1,3], **Joseph H. Gilbody**[1,2], **James Yarmolinsky**[1,2,5], **Sarah E.R. Bailey**[6], **Lucy J. Goudswaard**[1,2*], **Philip C. Haycock**[1*]

1 MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, United Kingdom, **2** Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom, **3** Bristol NIHR Biomedical Research Centre, University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol, Bristol, United Kingdom, **4** Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom, **5** Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom, **6** Department of Health and Community Science, University of Exeter, Exeter, United Kingdom

* Lisa.hobson@bristol.ac.uk (LMH); lucy.goudswaard@bristol.ac.uk (LJG); philip.haycock@bristol.ac.uk (PCH)

OPEN ACCESS

Citation: Hobson LM, Martin RM, Smith-Byrne K, Davey Smith G, Hemani G, Gilbody JH, et al. (2025) A protocol for using human genetic data to identify circulating protein level changes that are the causal consequence of cancer processes. PLoS One 20(12): e0312970. https://doi.org/10.1371/journal.pone.0312970

Editor: Yazhou He, The University of Edinburgh, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: October 25, 2024

Accepted: November 11, 2025

Published: December 5, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0312970

Copyright: © 2025 Hobson et al. This is an open access article distributed under the terms of the Creative Commons Attribution License,

Abstract

Introduction

Cancer is a leading cause of death worldwide. Early detection of cancer improves treatment options and patient survival but detecting cancer at the earliest stage presents challenges. Identification of circulating protein biomarkers for cancer risk stratification and early detection is an attractive avenue for potentially minimally invasive screening and early detection methods. This research aims to identify protein level changes that are downstream of genetic liability to lung cancer and colorectal cancer.

Methods and analysis

PRS will be calculated using the PRS continuous shrinkage approach (PRS-CS and PRS-CSx) for colorectal and lung cancer risk. This methodology utilises effect sizes from summary statistics from genome-wide association studies (GWAS) available for the cancers of interest to generate weights via the continuous shrinkage approach which incorporates the strengths of the GWAS associations into the shrinkage applied. This methodology both improves upon previous PRS methods in accuracy as well as improving cross-ancestry application in the PRS-CSx approach. GWAS summary statistics will be from the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) and the International Lung Cancer Consortium (ILCCO). The association between the polygenic risk scores and 2923 proteins measured by the Olink platform in UK Biobank (UKB) participants with protein measures available will

which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: No datasets were generated or analysed during the current study. All relevant data from this study will be made available upon study completion.

Funding: LMH is supported in part by grant MR/N0137941/1 for the GW4 BIOMED MRC DTP, awarded to the Universities of Bath, Bristol, Cardiff and Exeter from the Medical Research Council (MRC)/UKRI RMM is a National Institute for Health Research Senior Investigator (NIHR202411). RMM, LJG and PCH are supported by a Cancer Research UK 25 (C18281/A29019) programme grant (the Integrative Cancer Epidemiology Programme). RMM is also supported by the NIHR Bristol Biomedical Research Centre which is funded by the NIHR (BRC-1215-20011) and is a partnership between University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol. Department of Health and Social Care disclaimer: The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. SERB was supported by an NIHR Advanced Fellowship (NIHR 301666) whilst undertaking this work. Additional support was provided by the Higgins family. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: RMM, LJG and PCH have received funding from Cancer Research UK, a commercial funder. LMH receives funding from the GW4 BioMed2 MRC DTP, a commercial funder. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

be assessed using linear regression under the assumption of linearity in the proteomic data. The proteins identified could represent several different scenarios of association such as forward causation (protein causes cancer), reverse causation (cancer genetic liability causes protein level change), or horizontal pleiotropy bias (no causal relationship exists between the protein and cancer). Forward and reverse Mendelian randomization sensitivity analyses, as well as colocalization analysis, will be performed in efforts to distinguish between these three scenarios. Protein changes identified as causally downstream of genetic liability to cancer could reflect processes occurring prior to, or after, cancer onset. Due to individuals in the UKB having proteins measures at only one timepoint, and because UKB contains a mix of incident and prevalent cases, some protein measures will have been made prior to a cancer diagnosis while others will have been made after a cancer diagnosis. We will explore the strength of association in relation to the time between protein measurement and prevalent or incident cancer diagnosis.

Introduction

Detecting cancer at an early stage is important because patients diagnosed early have a greater chance of being treated with curative intent and so experience increased long-term survival. Cancer is a leading cause of death worldwide [1] with survival rates considerably lower when diagnosis is made at a later stage. For colorectal cancer and lung cancer the 5-year survival is reduced from more than 9 in 10 and 6 in 10, respectively, when diagnosed at stage 1, to 1 in 10 for colorectal cancer and less than 1 in 10 for lung cancer when diagnosed at stage 4 [2]. The NHS Long Term Plan aims to increase early detection of cancers from half to three quarters by the year 2028 to improve cancer survival [3] as currently in England only 54% of cancers are detected early [4]. To meet this goal a number of research challenges need to be addressed, including development of methods for determining cancer risk (i.e., risk stratification) and identifying biomarkers which are effective for detecting cancer at an early stage [5].

One of the challenges to be overcome in improving cancer early detection is the identification of specific biomarkers for the cancers of interest that can be measured by minimally invasive, low-cost methods and are able to be implemented in a clinical setting. One way to address this challenge is the measurement of circulating protein levels in blood serum or plasma, potentially feasible because of the widespread use of blood tests in healthcare. Circulating protein biomarkers are a potentially useful tool for several clinical areas including identifying groups at high risk of the future development of cancer (risk stratification), early detection, disease diagnosis and monitoring biological processes [6,7]. They may provide a minimally invasive means of screening asymptomatic individuals for undiagnosed disease or for diagnosis of symptomatic patients [8,9]. An advantage of measuring protein within the blood is the reduced volume of sample required for analysis and affordability, as opposed to other minimally invasive methods, such as circulating-tumour DNA (ctDNA) sequencing.

For the Olink explore 3072 panel, only 6 μL of plasma or serum is required vs. 4-5mL of plasma to obtain 5–10ng/mL of ctDNA and with the possibility of implementation of protein testing via ELISA, costing around ~£4 per test [10–15].

One approach to biomarker discovery is via prospective cohort studies to identify proteins associated with the incidence of a disease of interest, by measuring protein levels in individuals before diagnosis. These methods require large sample sizes over long periods of time to capture these events, at great financial and time cost [16]. A comparatively inexpensive technique for biomarker discovery has been formalised by Holmes and Davey Smith [17], and involves application of Mendelian randomization (MR) of disease liability as the exposure on protein levels as the outcome (sometimes described as reverse MR or reverse gear MR). Building on this idea, we propose that protein level changes resulting from cancer onset can be identified via an individual's PRS for specific cancers, representing their genetic liability to developing that cancer. Defining the point of "cancer onset" remains difficult, with many possible mechanisms of initiation; for the purpose of this study we will use date of diagnosis to determine prevalent vs. incident cases within the cohort [18].

Proteins associated with genetic liability to cancer could reflect different mechanisms of association. Associations could reflect 'forward causation' where the protein is upstream of and causal for cancer, e.g., P1-6 (Fig 1, panel A) or 'reverse causation' where carcinogenesis is causing the change in downstream protein level, e.g., P7-9 (Fig 1, panel A). Proteins that are associated via forward causation are upstream of the cancer pathway and therefore do not always denote the presence of cancer but could identify potential therapeutic targets for cancer prevention and cancer prediction, these protein levels will likely remain stable over long periods of time. Proteins downstream of cancer development will likely show more variation in levels resulting from the progression of cancer; we will refer to these proteins as "reverse causal". For proteins that cause cancer, most of the variants in the cancer PRS will have no causal relationship to those proteins, e.g., G9 and P6 (Fig 1, panel B), whereas for proteins that are causally associated downstream of cancer liability pathways, all variants in the PRS could contribute to the association signal (Fig 1, panel C). We thus expect a cancer PRS to be better powered for discovery of proteins downstream of cancer development. However, the relative balance in findings reflecting scenarios one ('forward causation') and two ('reverse causation') is likely to depend on the prevalence of cancer, including early or pre-clinical stages, in the sample used to measure the proteins. In general, the higher the prevalence, the greater the number of associations we expect to see reflecting effects of cancer liability pathways on protein concentration. Association of proteins with a genetic liability to cancer can also be due to factors other than genetic liability such as horizontal pleiotropy bias, as illustrated in Fig 1, panel A by G4 and G6; which may negate its use in risk stratification or early detection.

Aims and objective

Cancer early detection remains a challenge, with a lack of specific biomarkers for lung and colorectal cancer. This research aims to identify protein level changes that are causally downstream of genetic liability to lung cancer and colorectal cancer for use as potential biomarkers for these cancers. To achieve this, we will use a combination of PRS, observational analyses, bidirectional Mendelian Randomization and colocalization approaches.

Methods and analysis

Polygenic risk score analyses

PRS can be developed using GWAS summary statistics on the associations of many SNPs across the genome with cancer. In this way millions of SNPs can be combined to develop an individual's PRS for a cancer. A PRS is the sum of the number of copies of risk alleles individuals have for SNPs across the genome, weighted by the effect size of these SNPs in relation to the disease of interest, in this case, cancer [19]. While the initiation of cancer and the factors that contribute to the onset and progression of cancer are still not fully understood [20]; by calculating a PRS using data from all SNPs across the genome, SNPs involved in initiation, promotion and progression of cancer will be captured by this score, reflecting the complex process of cancer development [18].

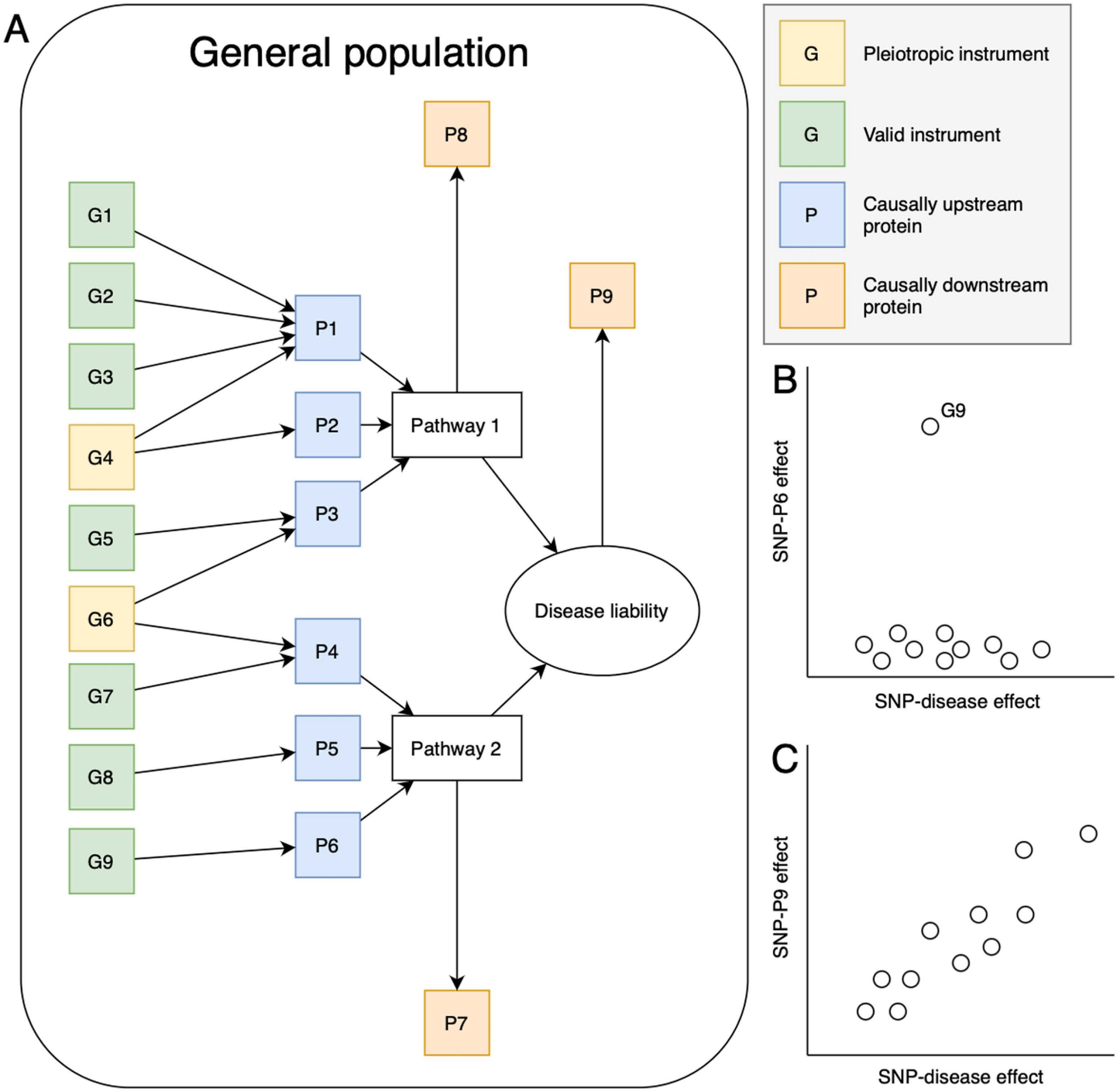

Fig 1. Illustration of the difference pathways which may be reflected by protein associations with genetic variables and disease liability. Panel A shows the relationship between genetic variants (G1-9), upstream proteins (P1-6) and downstream proteins (P7-9) via the disease liability pathways. Panel B shows the relationship between SNPs and upstream protein P6. Panel C shows the relationship between SNPs and downstream protein P9.

https://doi.org/10.1371/journal.pone.0312970.g001

In this study, PRS will be calculated for UKB participants (Application ID: 15825/81499) with proteomic measurements (N = 49,542), individuals with sex-mismatch (derived by comparing genetic sex and reported sex) or individuals with sex-chromosome aneuploidy will be excluded from the analysis (N = 814) as well as highly related individuals related to a 3rd degree to >200 individuals (N = 2) [21]. For colorectal cancer, we will use effect weights derived from GWAS summary statistics of the: i) Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) (GWAS Catalog Accession: GCST90255675) for Europeans; and ii) the Asia Colorectal Cancer Consortium (ACCC)/Korean-National Cancer Centre CRC Study 2 (Korea-NCC2) for GWAS summary statistics for East Asians. For lung cancer, we will use effect weights derived from GWAS summary statistics from the International Lung Cancer Consortium (ILCCO) (GWAS Catalog Accession: GCST004748). Sample selection and quality control within these studies has been previously described [22,23].

PRS will be derived from the PRS-CS approach, using summary statistics from GWAS for the cancer of interest along with an external linkage disequilibrium (LD) reference panel corresponding to the ancestry of the GWAS. The continuous shrinkage approach incorporates the strengths of the GWAS associations into the shrinkage applied to shrink small SNP effects towards zero, while large effects are unaffected [24], generating a posterior effect size for each SNP [25]. These weights will be used to calculate the PRS of UK Biobank participants for colorectal and lung cancer, calculating the sum of risk increasing alleles across all genetic variants weighted by the effect sizes generated by PRS-CS [25,26]. PRS-CSx applies the same methodology as PRS-CS to multi-ancestry GWAS summary statistics, improving generalisability of results to more ancestry groups within the global majority by allowing the use of different ancestry GWAS summary statistics and LD panels for those populations [27]. In an effort to reduce the Eurocentric bias and to increase power in addition to developing a European PRS, we will be utilising GWAS summary statistics for colorectal cancer from European and East Asian ancestries to develop polygenic risk scores [22,28,29].

Cancer subgroup analyses

In addition to a PRS for overall colorectal and lung cancers, we will calculate a PRS for colon cancer and rectal cancer specifically and for lung cancer subgroups (adenocarcinoma, squamous cell carcinoma, small cell carcinoma). Additionally, PRS scores will be calculated for never smokers and ever smokers using weights generated from summary statistics of GWAS for lung cancer in never smokers and lung cancer in ever smokers.

Olink proteins

Olink protein measurements were performed as part of the Pharma Proteomic Project (UKB-PPP) on blood plasma samples using the antibody-based protein Olink Explore 3072 Proximity Extension Assay. Proteomics were generated for 54,219 participants who were considered to be highly representative of the UK Biobank population on baseline characteristics and showed enrichment for selected diseases [30]. The number of participants with colorectal and lung cancers can be seen in Table 1. Quality control, sample selection and data processing has been described previously [30]. Associations between the participants' PRS and 2923 Olink protein measures from the UK Biobank will be tested via linear regression, adjusting for age, sex, principal components and sample storage time where this has an impact on protein level variation [31]. Protein measures will undergo inverse rank normal transformation (INT) for each protein [32]. The number of independent proteins will be calculated using the metaboprep R package [33]. False discovery rate correction will be applied to p-values, proteins with p-value less than the calculated alpha will be prioritised for further analyses.

Sensitivity analyses

Proteins identified from association analyses may reflect different scenarios, including causation or confounding from population stratification or dynastic effects. Some possible scenarios include: (1) a protein may be a cause of cancer risk, which we define as "forward causation"; (2) an alternative scenario is that the protein identified is causally downstream

Table 1. Cancer frequency within the UKB cohort and within the UKB-PPP study participants.

ICD10/ICD9 code	Cancer	Number of cases (UKB)	Number of cases (UKB-PPP)
C18/153	Malignant neoplasm of colon	6015 (1.2%)	630 (0.13%)
C19/1540	Malignant neoplasm of rectosigmoid junction	642 (0.13%)	67 (0.013%)
C20/1541	Malignant neoplasm of rectum	2670 (0.53%)	268 (0.053%)
C34/162	Malignant neoplasm of bronchus and lung	5305 (1.1%)	597 (0.12%)
ICD-O-3 Code	**Histological Subset**	**Number of cases (UKB)**	**Number of cases (UKB-PPP)**
8140, 8211, 8250–8260, 8310, 8323, 8480–8490, 8550	Lung adenocarcinoma	2390 (0.48%)	261 (0.052%)
8070-8072	Lung squamous cell carcinoma	1244 (0.25%)	151 (0.03%)
8041–8042	Lung small cell carcinoma	502 (0.1%)	64 (0.013%)

Number of cases for each cancer type derived from UK Biobank phenotypic data with percentage of cases out of overall individuals is represented in brackets. UKB overall n = 501939, UKB with protein measures (UKB-PPP) n = 52996.

https://doi.org/10.1371/journal.pone.0312970.t001

of cancer liability, which we refer to as "reverse causation"; (3) there is no causal relationship between the protein and cancer and the identified association reflects horizontal pleiotropy, (4) due to population stratification where spurious associations are due to differences in the GWAS population and those that the PRS is calculated on. We will perform various sensitivity analyses to distinguish amongst these scenarios, described below.

Bidirectional Mendelian randomisation sensitivity analyses

MR uses genetic variants, associated with the phenotype of interest as the instrumental variable to assess the effect of the phenotype on an outcome. Due to the random nature of inheritance of genetic variants there is an advantage over observational epidemiology whereby confounders may influence both the exposure and outcome of interest [34–36]. Genetic associations used in MR analyses often come from GWAS summary data, whereby association is conventionally defined by a p-value threshold of 5×10^{-8}.

Assumptions. The three core assumptions of MR, known as the instrumental variable (IV) assumptions (Fig 2), are relevance (IV1) – is the instrumental variable (G) associated with the exposure (E), independence (IV2) – there is no confounding of the association between the instrument (G) and outcome (O) (this can arise through population stratification, dynastic effects and assortative mating) and exclusion restriction (IV3) – the instrumental variable (G) does not act on the outcome (O) except via the exposure (E) no horizontal pleiotropy (red dashed line, IV3) [37–39].

Study design

MR will be performed in the forward and reverse direction: forward MR, where the protein is the exposure and cancer is the outcome, will be used to estimate the effect of selected proteins on the cancer of interest and reverse MR, where cancer is the exposure and protein levels are the outcomes, will be used to estimate the effect of cancer liability on circulating protein concentration [17]. Forward MR will be performed using cis-pQTLs to instrument proteins identified as being associated with the cancer PRS, the threshold for these will be $p < 3.4 \times 10^{-11}$ [40]. Cis-pQTLs will be defined as within < 1Mbp of the protein coding gene and trans-pQTLs will be defined as > 1Mbp away from the protein coding gene [41]. Reverse MR will be performed using SNPs associated with the cancer PRS at a threshold of $p < 5 \times 10^{-8}$ excluding cis-pQTLs for the protein. If association is found in the forward direction this may suggest that the protein is causal for the cancer but if association is found in the reverse direction this may suggest that genetic liability to cancer is causing the protein level change [17]; to elucidate this causality, different MR estimation methods will be employed, the application and conditions of these are described below.

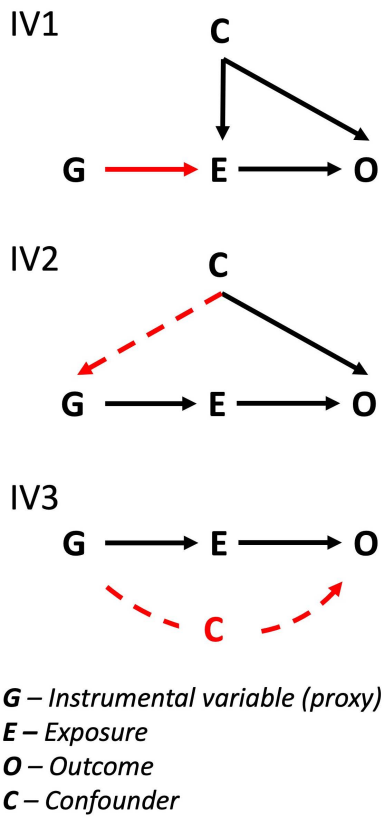

Fig 2. Directed Acyclic Graphs (DAGs) showing the three IV assumptions of Mendelian Randomization. IV1 represents that the assumption that the genetic variant (G) is strongly associated (red arrow) with the exposure (E). IV2 represents the assumption that confounders don't also act on the genetic variant (red dashed arrow) as they do the outcome. IV3 represents the assumption that the genetic variants do not affect the outcome through other routes separate from the exposure (red dashed arrow) (C).

https://doi.org/10.1371/journal.pone.0312970.g002

Instrument & method selection

The strength of instrument will be determined by calculating the F-statistic, a measure of potential weak instrument bias that could arise from the use of IVs as a proxy for the effect of exposure on outcome [39]. The F-statistic takes into account the genetic variance (R^2), the sample size and how many instruments are present. An F-statistic greater than 10 indicates that the bias from weak instruments is small, where this F-stat is less than 10 this indicates a possibility of bias and will be noted [42].

Dependent on the number of SNPs available, the appropriate method of effect-estimation will be selected for MR analyses. For proteins with a single pQTL SNP the Wald ratio will be calculated as the ratio of SNP-outcome/SNP-exposure association [43]. For proteins with two or three independent SNPs, a fixed-effects inverse variance weighted (IVW) model will be used. For four or more independent SNPs, a random effects inverse variance (IVW) model, combining multiple SNP outcome/exposure Wald ratio, will be used [44]. In the event of multiple independent SNPs, pleiotropy will be considered by calculating Cochran's Q statistic, a method for assessing global and individual pleiotropy across instruments [45]. Weighted mode and weighted median methods will also be used when > 10 SNPs are available [46,47].

The MR-PRESSO, weighted mode and weighted median methods will be used to assess IVs for horizontal pleiotropy, violation of IV3 where the IV acts on the outcome not via the exposure, by comparing estimates with and without suspected pleiotropic variants, this will be repeated for both forward and reverse MR [48]. In addition being robust to

pleiotropy methods such as MR robust adjusted profile score (RAPS) also accounts for other potential sources of bias such as weak instruments and measurement error in the exposure [49]. MR-CAUSE (Causal Analysis Using Summary Effect estimates) is another method that can be used when IV3 is violated due to pleiotropic effects of correlated pleiotropy, where the pleiotropic factor is a confounder of the exposure-outcome association versus when the IV has effects on pleiotropic factor independently of the effect of the IV on the exposure this is uncorrelated pleiotropy [50].

When performing reverse MR using a larger numbers of SNPs, clustered heterogeneity can occur when different genetic variants are causally associated via distinct pathways. To assess this, clustering based methods can be used to divide groups based on these estimates of causality [51]. MR-Clust will be used to investigate clustered heterogeneity across IVs and identify potential distinct pathways that make up the effect estimate; clustering works by separating the variants into clusters with additional null and junk clusters, representing no causal effect or those that do not fit within the distinct clusters [51,52]. Another clustering method that will be used is the Noise-Augmented von Mises-Fisher Mixture model (NAvMix), this method allows for variants to belong to multiple clusters based on their probability of membership to that cluster [53,54]. The contamination mixture model method can also be used to cluster into distinct groups based on the IVs causal effect estimate even when invalid IVs are present [55]. PheWAS-based clustering will also be used to cluster SNP associations based on different pathways and thus help identify other causal pathways of the PRS – protein associations found [56]. The methods of MR described make different assumptions and aim to address different violation of the IV assumptions, testing of these different methods has illustrated the variation in accuracy and the need for appropriate method selection based on the datasets used [57].

Data harmonisation

Harmonisation across the GWAS summary statistics datasets will be performed using the TwoSampleMR R package to ensure that the effect on exposure and outcome are in the same direction, using effect allele frequency (EAF) to infer the strand, palindromic SNPs are kept where they are inferable from EAF. [58].

Colocalization

Associations may be due to genomic confounding, where genetic variants in linkage disequilibrium (LD) at the same locus act on the cancer and protein via separate pathways, a form of horizontal pleiotropy bias. Colocalization analyses will be used to assess if genetic associations with cancer and proteins are due to shared causal variants at the same locus through genomic confounding [59,60].

A full overview of the sample selection and filtering of available data is shown in Fig 3, including the number of participants after filtering and results to be taken forward.

Ethics

The colorectal cancer GWAS conducted by Fernandez-Rozadilla et al. (2022) was approved by the South Central Ethics Committee (UK) under the reference number 17/SC/0079 [22].

All studies used in the lung cancer GWAS conducted by McKay et al. (2017) obtained local ethics committee approval and all participants gave informed consent [23].

Application for colorectal cancer site specific GWAS summary statistics from GECCO has been approved.

Application for summary statistics from the Asian Colorectal Cancer Consortium (ACCC) and the Korean-National Cancer Center CRC Study 2 (Korea-NCC2) will be submitted.

UK Biobank was approved by the North West Multi-centre Research Ethics Committee (MREC) as a Research Tissue Bank (RTB) approval renewed in 2021, all participants in the study have given informed consent [61]. Genotype, phenotype and Olink protein measure data access has been obtained under Application ID: 15825/81499.

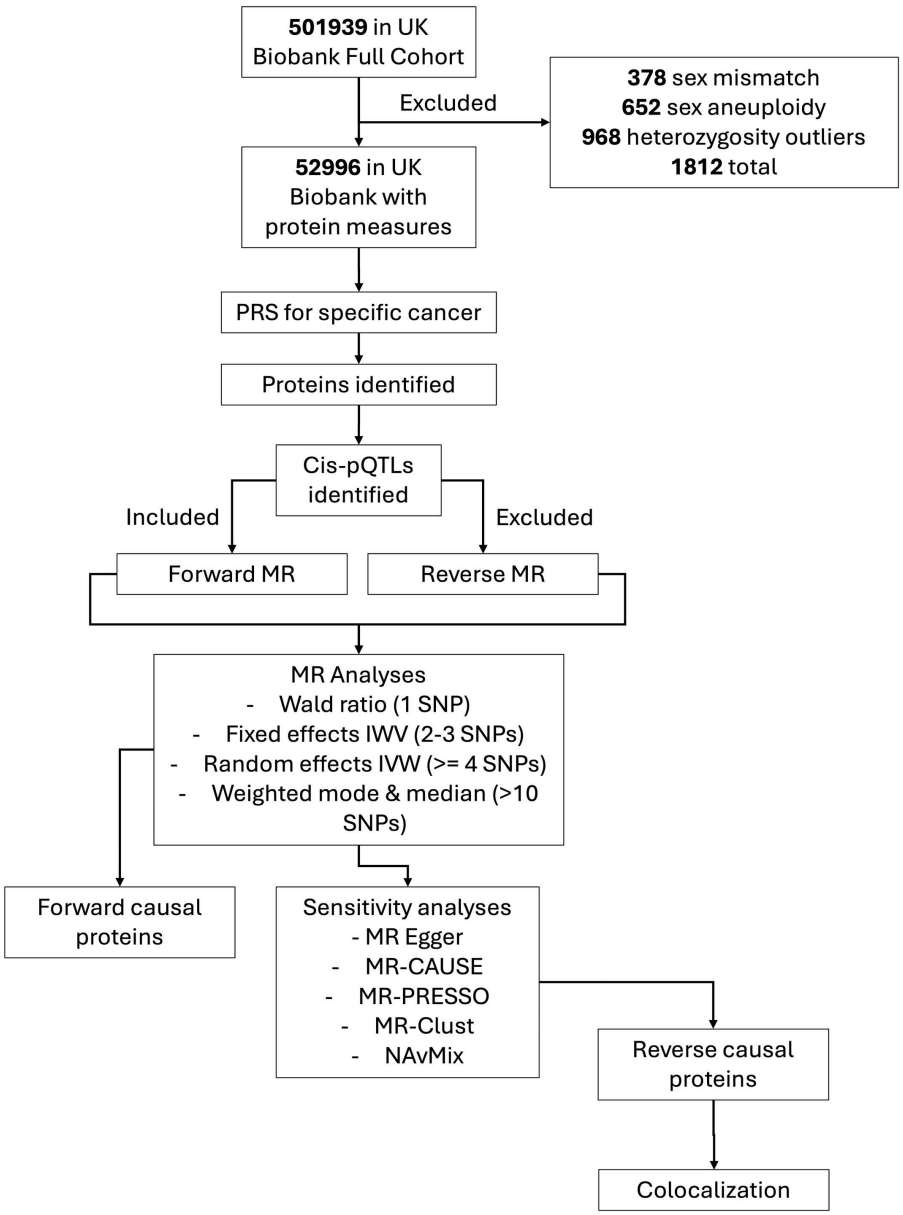

Fig 3. Flowchart of methodology and data selection. UK Biobank data will be filtered based on exclusion criteria and availability of protein data, this will filter into proteins identified. Cis-pQTLs for the proteins will be identified which will either by used as instruments for forward MR or excluded from reverse MR. Various MR analyses will be performed based on number of SNPs available and sensitivity analyses will be used to filter proteins to perform colocalization.

https://doi.org/10.1371/journal.pone.0312970.g003

Further analyses

"Time-to" and "Time-from" diagnosis. In observational analyses, we will evaluate the magnitude of the relationship between proteins, taken either pre or post-diagnosis, and cancer risk. This will involve an analysis of prevalent and incident cancer cases in UKB (Table 2) and a time variable derived from date of cancer diagnosis and time of blood collection [62,63]. To adjust for any variation in protein concentration as a result of sample storage and protein

Table 2. Prevalent and incident cases of cancer from UKB cohort and within the UKB-PPP study participants.

ICD10/ICD9 Code	Cancer	UKB Overall		UKB-PPP	
		Prevalent Cases	Incident Cases	Prevalent Cases	Incident Cases
C18/153	Malignant neoplasm of colon	1029	4986	97	533
C19/1540	Malignant neoplasm of rectosigmoid junction	186	456	22	45
C20/1541	Malignant neoplasm of rectum	574	2096	60	208
C34/162	Malignant neoplasm of bronchus and lung	245	5060	31	566
ICD-O-3 Code	**Histological Subset**	**Prevalent Cases**	**Incident Cases**	**Prevalent Cases**	**Incident Cases**
8140, 8211, 8250–8260, 8310, 8323, 8480–8490, 8550	Lung adenocarcinoma	77	2313	8	253
8070-8072	Lung squamous cell carcinoma	74	1170	14	137
8041–8042	Lung small cell carcinoma	23	479	3	61

Number of incident and prevalent cases for each cancer type derived from UK biobank phenotypic data.

https://doi.org/10.1371/journal.pone.0312970.t002

degradation over time [31], the relationship between storage time and protein level for all protein measures available will be assessed. Proteins are more likely to be causally downstream of cancer onset if the association with cancer is sensitive to time between protein measure and cancer diagnosis, a potential route for differentiating between normal baseline levels and levels that suggest the presence of cancer. If protein levels are detectable prior to patient reported symptoms proteins may be more suited for screening and early detection.

Replication of findings. Replication of protein association and MR will be carried out in the DECODE cohort [64] and EPIC study [65] where proteins are available.

Software

This work will be carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol - http://www.bristol.ac.uk/acrc/.

PRS-CS (https://github.com/getian107/PRScs) and PRS-CSx (https://github.com/getian107/PRScsx) will be used to calculate polygenic risk scores, using R, Python and PLINK.

Metaboprep (https://github.com/MRCIEU/metaboprep) R package will be used to calculate independent proteins [33]. Mendelian Randomization analyses and data harmonisation will be performed using the R packages TwoSampleMR (https://github.com/MRCIEU/TwoSampleMR) and MendelianRandomization (https://cran.r-project.org/web/packages/MendelianRandomization/index.html) [66].

Proteins will be inverse rank normal transformed using the “RankNorm” function in R package “RNOmni” (https://cran.r-project.org/web/packages/RNOmni/index.html) [32].

Patient and public involvement

A summary of the proposed research was presented to members of a patient and public involvement group, with either personal experience with cancer or experience via a family member. The feedback received was that this was very important research and that they believe it would be useful for early detection for cancers that do not yet have specific screening via a blood test. Updates about this study will also be disseminated to the group.

Results of these analyses will be disseminated via the University of Bristol MRC Integrative Epidemiology Unit IEU Portal and submitted as a manuscript to a peer-reviewed journal for publication. All statistical code will be made available via GitHub.

Polygenic risk scores and PRS-protein associations will be returned to the UK Biobank in line with the UK Biobank obligation for researchers outlined [67].

Strengths and limitations of this study

- Strengths of the study:
  - A strength of using PRS in the discovery step of identifying proteins to carry through into MR, is that lifetime genetic liability to cancer is captured with the use of a continuous measure of genetic risk. In contrast, using cancer registry data only captures those who already have a diagnosis and not those who may go on to be diagnosed imminently.
  - This study will use a novel approach to construct PRS using weights generated from SNPs across the genome, we expect this to be more powerful for discovery in comparison with a PRS constructed using only genome-wide significant SNPs.
- Limitations of the study
  - Lack of protein data for diverse population groups within available datasets; therefore, results may not be generalisable to ancestries outside of the European population for whom sufficient protein data was available for this study.
  - UKB participants reflect a subset of the population that volunteered for the study, participants are from a higher socioeconomic position than average which could introduce collider bias.
  - Prevalent cancer cases will reflect a specific subset of the general population with cancer, individuals who have survived cancer and were able to volunteer for the study; potentially introducing survivorship bias.
  - It cannot be ruled out that proteins may reflect effects of processes beyond cancer liability to protein pathways.
  - Lack of staging information for cancer cases within the UKB limiting our ability to distinguish early versus more advanced cancers.
  - The proteomic technology currently used measures protein binding as opposed to protein levels
  - Subgroup analyses are more prone to false positive results which will require replication analysis to ensure these reflect actual associations.
  - Power calculations have not been performed to assess the power of the subgroup analyses.

Author contributions

Conceptualization: Richard M. Martin, Karl Smith-Byrne, Sarah E.R. Bailey, Lucy J. Goudswaard, Philip C. Haycock.

Methodology: Richard M. Martin, Karl Smith-Byrne, Sarah E.R. Bailey, Lucy J. Goudswaard, Philip C. Haycock.

Supervision: Richard M. Martin, Karl Smith-Byrne, Sarah E.R. Bailey, Lucy J. Goudswaard, Philip C. Haycock.

Visualization: Gibran Hemani.

Writing – original draft: Lisa M. Hobson.

Writing – review & editing: Richard M. Martin, Karl Smith-Byrne, George Davey Smith, Gibran Hemani, Joseph H Gilbody, James Yarmolinsky, Sarah E.R. Bailey, Lucy J. Goudswaard, Philip C. Haycock.

References

1. WHO. Cancer. 3 Feb 2022 [cited 1 Nov 2023]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer
2. Why is early cancer diagnosis important? In: Cancer Research UK [Internet]. 2 Apr 2015 [cited 9 Dec 2023]. Available: https://www.cancerresearchuk.org/https%3A//www.cancerresearchuk.org/about-cancer/spot-cancer-early/why-is-early-diagnosis-important
3. NHS England. NHS Long Term Plan. NHS England; 2019 Jan. Report No.: 1.2. Available: https://www.longtermplan.nhs.uk/publication/nhs-long-term-plan/

4. Health Education England. Improving cancer diagnosis and earlier detection. In: Health Education England [Internet]. 15 Mar 2023 [cited 9 Dec 2023]. Available: https://www.hee.nhs.uk/our-work/primary-care/improving-cancer-diagnosis-earlier-detection
5. Crosby D, Bhatia S, Brindle KM, Coussens LM, Dive C, Emberton M, et al. Early detection of cancer. Science. 2022;375(6586):eaay9040. https://doi.org/10.1126/science.aay9040 PMID: 35298272
6. Integrative Analysis of Lung Cancer Etiology and Risk (INTEGRAL) Consortium for Early Detection of Lung Cancer, Guida F, Sun N, Bantis LE, Muller DC, Li P, et al. Assessment of Lung Cancer Risk on the Basis of a Biomarker Panel of Circulating Proteins. JAMA Oncol. 2018;4(10):e182078. https://doi.org/10.1001/jamaoncol.2018.2078 PMID: 30003238
7. Carrasco-Zanini J, Pietzner M, Davitte J, Surendran P, Croteau-Chonka DC, Robins C, et al. Proteomic signatures improve risk prediction for common and rare diseases. Nat Med. 2024;30(9):2489–98. https://doi.org/10.1038/s41591-024-03142-z PMID: 39039249
8. Pavlou MP, Diamandis EP. The cancer cell secretome: a good source for discovering biomarkers? J Proteomics. 2010;73(10):1896–906. https://doi.org/10.1016/j.jprot.2010.04.003 PMID: 20394844
9. Califf RM. Biomarker definitions and their applications. Exp Biol Med (Maywood). 2018;243(3):213–21. https://doi.org/10.1177/1535370217750088 PMID: 29405771
10. Gao Q, Zeng Q, Wang Z, Li C, Xu Y, Cui P, et al. Circulating cell-free DNA for cancer early detection. Innovation (Camb). 2022;3(4):100259. https://doi.org/10.1016/j.xinn.2022.100259 PMID: 35647572
11. Yan Y-Y, Guo Q-R, Wang F-H, Adhikari R, Zhu Z-Y, Zhang H-Y, et al. Cell-Free DNA: Hope and Potential Application in Cancer. Front Cell Dev Biol. 2021;9:639233. https://doi.org/10.3389/fcell.2021.639233 PMID: 33693004
12. Song P, Wu LR, Yan YH, Zhang JX, Chu T, Kwong LN, et al. Limitations and opportunities of technologies for the analysis of cell-free DNA in cancer diagnostics. Nat Biomed Eng. 2022;6(3):232–45. https://doi.org/10.1038/s41551-021-00837-3 PMID: 35102279
13. Kim H, Park KU. Clinical Circulating Tumor DNA Testing for Precision Oncology. Cancer Res Treat. 2023;55(2):351–66. https://doi.org/10.4143/crt.2022.1026 PMID: 36915242
14. Olink® Explore 3072 high-throughput proteomics platform now available: Significantly expands Olink's protein library for biomarker discovery | Olink Holding AB. [cited 5 Apr 2024]. Available: https://investors.olink.com/news-releases/news-release-details/olinkr-explore-3072-high-throughput-proteomics-platform-now/
15. Prostate Specific Antigen (PSA) ELISA for serum or plasma 2-25ng/ml Dialab. [cited 2 May 2024]. Available: https://www.alphalabs.co.uk/z00338
16. Mosley JD, Feng Q, Wells QS, Van Driest SL, Shaffer CM, Edwards TL, et al. A study paradigm integrating prospective epidemiologic cohorts and electronic health records to identify disease biomarkers. Nat Commun. 2018;9(1):3522. https://doi.org/10.1038/s41467-018-05624-4 PMID: 30166544
17. Holmes MV, Davey Smith G. Can Mendelian Randomization Shift into Reverse Gear? Clin Chem. 2019;65(3):363–6. https://doi.org/10.1373/clinchem.2018.296806 PMID: 30692117
18. Balmain A. Peto's paradox revisited: black box vs mechanistic approaches to understanding the roles of mutations and promoting factors in cancer. Eur J Epidemiol. 2022;38(12):1251–8. https://doi.org/10.1007/s10654-022-00933-x
19. Choi SW, Mak TS-H, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15(9):2759–72. https://doi.org/10.1038/s41596-020-0353-1
20. Davey Smith G, Hofman A, Brennan P. Chance, ignorance, and the paradoxes of cancer: Richard Peto on developing preventative strategies under uncertainty. Eur J Epidemiol. 2023;38(12):1227–37. https://doi.org/10.1007/s10654-023-01090-5 PMID: 38147198
21. Mitchell R, Gibran H, Dudding T, Corbin L, Harrison S, Paternoster L. UK Biobank Genetic Data: MRC-IEU Quality Control, version 2. University of Bristol; 2019. https://doi.org/10.5523/BRIS.1OVAAU5SXUNP2CV8RCY88688V
22. Fernandez-Rozadilla C, Timofeeva M, Chen Z, Law P, Thomas M, Schmit S, 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. 2023;55(1):89–99. https://doi.org/10.1038/s41588-022-01222-9 PMID: 36539618
23. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet. 2017;49(7):1126–32. https://doi.org/10.1038/ng.3892 PMID: 28604730
24. van Erp S, Oberski DL, Mulder J. Shrinkage priors for Bayesian penalized regression. Journal of Mathematical Psychology. 2019;89:31–50. https://doi.org/10.1016/j.jmp.2018.12.004
25. Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019;10(1):1776. https://doi.org/10.1038/s41467-019-09718-5 PMID: 30992449
26. Ge T. PRS-CS. 2018. Available: https://github.com/getian107/PRScs
27. Ruan Y, Lin Y-F, Feng Y-CA, Chen C-Y, Lam M, Guo Z, et al. Improving polygenic prediction in ancestrally diverse populations. Nat Genet. 2022;54(5):573–80. https://doi.org/10.1038/s41588-022-01054-7 PMID: 35513724
28. Ge T, Irvin MR, Patki A, Srinivasasainagendra V, Lin Y-F, Tiwari HK, et al. Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. Genome Med. 2022;14(1):70. https://doi.org/10.1186/s13073-022-01074-2 PMID: 35765100

29. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584–91. https://doi.org/10.1038/s41588-019-0379-x PMID: 30926966
30. Sun BB, Chiou J, Traylor M, Benner C, Hsu Y-H, Richardson TG, et al. Plasma proteomic associations with genetics and health in the UK Biobank. Nature. 2023;622(7982):329–38. https://doi.org/10.1038/s41586-023-06592-6 PMID: 37794186
31. Enroth S, Hallmans G, Grankvist K, Gyllensten U. Effects of Long-Term Storage Time and Original Sampling Month on Biobank Plasma Protein Concentrations. EBioMedicine. 2016;12:309–14. https://doi.org/10.1016/j.ebiom.2016.08.038 PMID: 27596149
32. McCaw Z. RNOmni: Rank Normal Transformation Omnibus Test. 2023. Available: https://cran.r-project.org/web/packages/RNOmni/index.html
33. MRCIEU/metaboprep: a pipeline of metabolomics data processing and quality control. [cited 9 Feb 2024]. Available: https://github.com/MRCIEU/metaboprep
34. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey Smith G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. Am J Clin Nutr. 2016;103(4):965–78. https://doi.org/10.3945/ajcn.115.118216 PMID: 26961927
35. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ. 2018;362:k601. https://doi.org/10.1136/bmj.k601 PMID: 30002074
36. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. Int J Epidemiol. 2013;42(4):1134–44. https://doi.org/10.1093/ije/dyt093 PMID: 24062299
37. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ. 2018;362:k601. https://doi.org/10.1136/bmj.k601 PMID: 30002074
38. Zheng J, Baird D, Borges M-C, Bowden J, Hemani G, Haycock P, et al. Recent Developments in Mendelian Randomization Studies. Curr Epidemiol Rep. 2017;4(4):330–45. https://doi.org/10.1007/s40471-017-0128-6 PMID: 29226067
39. Sanderson E, Glymour MM, Holmes MV, Kang H, Morrison J, Munafò MR, et al. Mendelian randomization. Nat Rev Methods Primers. 2022;2:6. https://doi.org/10.1038/s43586-021-00092-5 PMID: 37325194
40. Sun BB, Chiou J, Traylor M, Benner C, Hsu Y-H, Richardson TG, et al. Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. bioRxiv. 2022. p. 2022.06.17.496443. https://doi.org/10.1101/2022.06.17.496443
41. Fauman EB, Hyde C. An optimal variant to gene distance window derived from an empirical definition of cis and trans protein QTLs. BMC Bioinformatics. 2022;23(1). https://doi.org/10.1186/s12859-022-04706-x
42. Burgess S, Thompson SG. Avoiding bias from weak instruments in Mendelian randomization studies. International Journal of Epidemiology. 2011;40(3):755–64. https://doi.org/10.1093/ije/dyr036
43. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. Stat Methods Med Res. 2017;26(5):2333–55. https://doi.org/10.1177/0962280215597579 PMID: 26282889
44. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol. 2013;37(7):658–65. https://doi.org/10.1002/gepi.21758 PMID: 24114802
45. Bowden J, Hemani G, Davey Smith G. Invited Commentary: Detecting Individual and Global Horizontal Pleiotropy in Mendelian Randomization-A Job for the Humble Heterogeneity Statistic? Am J Epidemiol. 2018;187(12):2681–5. https://doi.org/10.1093/aje/kwy185 PMID: 30188969
46. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet Epidemiol. 2016;40(4):304–14. https://doi.org/10.1002/gepi.21965 PMID: 27061298
47. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. Int J Epidemiol. 2017;46(6):1985–98. https://doi.org/10.1093/ije/dyx102 PMID: 29040600
48. 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. 2018;50(5):693–8. https://doi.org/10.1038/s41588-018-0099-7 PMID: 29686387
49. Zhao Q, Wang J, Hemani G, Bowden J, Small DS. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. Ann Statist. 2020;48(3). https://doi.org/10.1214/19-aos1866
50. Morrison J, Knoblauch N, Marcus JH, Stephens M, He X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nat Genet. 2020;52(7):740–7. https://doi.org/10.1038/s41588-020-0631-4 PMID: 32451458
51. Foley CN, Mason AM, Kirk PDW, Burgess S. MR-Clust: clustering of genetic variants in Mendelian randomization with similar causal estimates. Bioinformatics. 2021;37(4):531–41. https://doi.org/10.1093/bioinformatics/btaa778 PMID: 32915962
52. Foley CN. cnfoley/mrclust. 2024. Available: https://github.com/cnfoley/mrclust
53. Grant AJ, Gill D, Kirk PDW, Burgess S. Noise-augmented directional clustering of genetic association data identifies distinct mechanisms underlying obesity. PLoS Genet. 2022;18(1):e1009975. https://doi.org/10.1371/journal.pgen.1009975 PMID: 35085229
54. Grant A. aj-grant/navmix. 2024. Available: https://github.com/aj-grant/navmix
55. Burgess S, Foley CN, Allara E, Staley JR, Howson JMM. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. Nat Commun. 2020;11(1):376. https://doi.org/10.1038/s41467-019-14156-4 PMID: 31953392

56. Darrous L, Hemani G, Davey Smith G, Kutalik Z. PheWAS-based clustering of Mendelian Randomisation instruments reveals distinct mechanism-specific causal effects between obesity and educational attainment. Nat Commun. 2024;15(1):1420. https://doi.org/10.1038/s41467-024-45655-8 PMID: 38360877

57. Hu X, Cai M, Xiao J, Wan X, Wang Z, Zhao H, et al. Benchmarking Mendelian Randomization methods for causal inference using genome-wide association study summary statistics. medRxiv. 2024. p. 2024.01.03.24300765. doi:10.1101/2024.01.03.24300765

58. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7:e34408. https://doi.org/10.7554/eLife.34408 PMID: 29846171

59. Foley CN, Staley JR, Breen PG, Sun BB, Kirk PDW, Burgess S, et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. Nat Commun. 2021;12(1):764. https://doi.org/10.1038/s41467-020-20885-8 PMID: 33536417

60. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10(5):e1004383. https://doi.org/10.1371/journal.pgen.1004383 PMID: 24830394

61. Ethics. [cited 4 Jan 2024]. Available: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics

62. : Data-Field 40005. [cited 15 Feb 2024]. Available: https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=40005

63. : Data-Field 3166. [cited 15 Feb 2024]. Available: https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=3166

64. SCIENCE. In: deCODE genetics [Internet]. 6 Sept 2012 [cited 1 Dec 2023]. Available: https://www.decode.com/research/

65. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. 2002;5(6B):1113–24. https://doi.org/10.1079/PHN2002394 PMID: 12639222

66. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7:e34408. https://doi.org/10.7554/eLife.34408 PMID: 29846171

67. External Info: returning_results. [cited 5 Jan 2024]. Available: https://biobank.ndph.ox.ac.uk/showcase/exinfo.cgi?src=returning_results