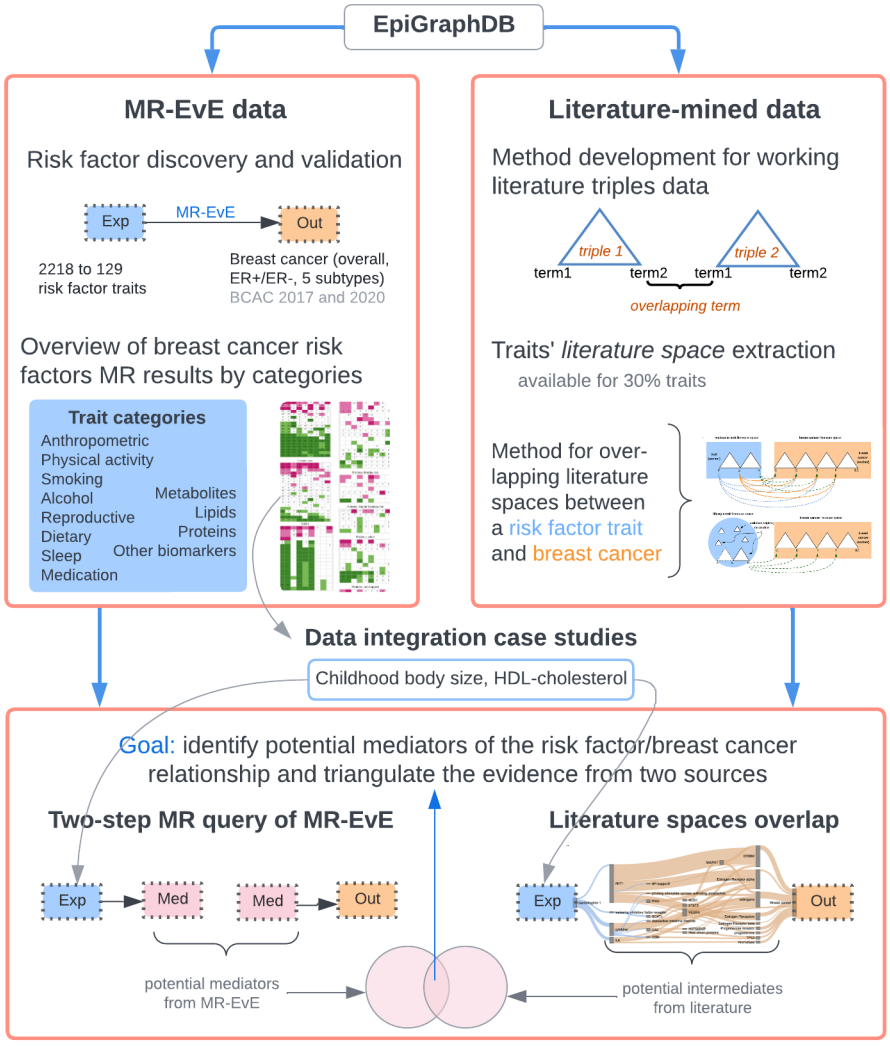
Graphical Abstract

Integrating Mendelian randomization and literature-mined evidence for breast cancer risk factors

Marina Vabistsevits, Tim Robinson, Ben Elsworth, Yi Liu, Tom R Gaunt



Integrating Mendelian randomization and literature-mined evidence for breast cancer risk factors

Marina Vabistsevits[a,b,d], Tim Robinson[a,b], Ben Elsworth[a,b,c], Yi Liu[a,b], Tom R Gaunt[a,b]

[a]*Medical Research Council Integrative Epidemiology Unit at the University of Bristol, University of Bristol, Oakfield 7 House, Oakfield Grove, Bristol, BS8 2BN, UK*
[b]*Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield 7 House, Oakfield Grove, Bristol, BS8 2BN, UK*
[c]*Our Future Health, 2 New Bailey, 6 Stanley Street, Manchester, M3 5GS, UK*
[d]*University of Exeter Medical School, University of Exeter St Luke's Campus, 79 Heavitree Rd, Exeter, EX2 4TH, UK*
[e]*corresponding author: Marina Vabistsevits (marina.vabistsevits@bristol.ac.uk)*

Abstract

Objective: An increasing challenge in population health research is efficiently utilising the wealth of data available from multiple sources to investigate disease mechanisms and identify potential intervention targets. The use of biomedical data integration platforms can facilitate evidence triangulation from these different sources, improving confidence in causal relationships of interest. In this work, we aimed to integrate Mendelian randomization (MR) and literature-mined evidence from the EpiGraphDB biomedical knowledge graph to build a comprehensive overview of risk factors for developing breast cancer.

Methods: We utilised MR-EvE ("Everything-vs-Everything") data to identify candidate risk factors for breast cancer and generate hypotheses for potential mediators of their effect. We also integrated this data with literature-mined relationships, which were extracted by overlapping literature spaces of risk factors and breast cancer. The literature-based discovery (LBD) results were followed up by validation with two-step MR to triangulate the findings from two data sources.

Results: We identified 129 novel and established lifestyle risk factors and molecular traits with evidence of an effect on breast cancer, and made the MR results available in an R/Shiny app (`https://mvab.shinyapps.io/MR_heatmaps/`). We developed an LBD approach for identifying potential mechanistic intermediates of identified risk factors. We present the results of MR and literature evidence integration for two case studies (childhood body size and HDL-cholesterol), demonstrating their complementary functionalities.

Conclusion: We demonstrate that MR-EvE data offers an efficient hypothesis-

generating approach for identifying disease risk factors. Moreover, we show that integrating MR evidence with literature-mined data may be used to identify causal intermediates and uncover the mechanisms behind the disease.

Keywords: mendelian randomization, genome-wide association study, causal inference, breast cancer, literature-based discovery, knowledge graph, literature mining

---

1. Introduction

Triangulation, a process of integrating evidence from multiple methodologies, has become increasingly important in population health research [1, 2]. Different evidence sources are likely to have unique and unrelated sources of bias, so consistent evidence from such sources improves confidence in the causal relationship between a risk factor and outcomes of interest [3]. A potential obstacle to the widespread adaptation of triangulation in epidemiological research is the challenge of integrating and harmonising the large volume of data from different datasets. Multiple platforms [4, 5, 6, 7] have been developed to facilitate access to combined and curated biomedical datasets. Using these data integration platforms will facilitate the study of disease aetiology.

EpiGraphDB (`https://epigraphdb.org`) [4] is a biomedical knowledge graph that was designed to facilitate data mining of epidemiological relationships. The unique advantages of this platform are the availability of both Mendelian randomization (MR) [8, 9] and literature-mined relationships between molecular and lifestyle traits and disease outcomes, providing a valuable combination of data to systematically investigate causal relationships.

MR is an approach to causal inference that uses genetic variants as instrumentable variables (IVs) to infer whether a modifiable health exposure influences a disease outcome [8]. Studying disease causality with MR methods has been growing in popularity over the past decade and a half [9, 10], with the increasing availability of data from genome-wide association studies (GWAS) [11, 12] and well-developed analysis frameworks [13]. EpiGraphDB provides access to MR-EvE (Mendelian randomization "Everything-vs-Everything") pairwise causal estimates [14] between thousands of traits from the OpenGWAS database (`https://gwas.mrcieu.ac.uk`) [11]. MR-EvE data can be used to explore potential causal links between thousands of traits for knowledge discovery or hypothesis generation.

Another aspect of knowledge discovery is extracting data from the biomedical literature, also known as literature-based discovery (LBD) [15]. In addition to MR-EvE, EpiGraphDB also contains literature-mined relationships extracted from

publications in PubMed [16] and mapped to specific phenotypes in the MR-EvE data. LBD entails connecting information that is explicitly stated in multiple publications across different research disciplines aiming to deduce connections that have not been explicitly stated [17, 18], offering the potential to identify mechanistic pathways or mediators for causal relationships identified using MR-EvE.

In this work, we use EpiGraphDB to build a comprehensive picture of the aetiology of breast cancer by combining evidence from MR and literature-mined data. Breast cancer is a useful exemplar, as it is a heterogeneous disease with a complex aetiology, affected by both genetic and lifestyle factors [19, 20, 21, 22] that also has a large body of published research evidence, suitable for data mining. We seek to generate new mechanistic hypotheses for causal relationships using evidence integration to inform prevention and intervention strategies that could reduce disease burden.

Statement of Significance

Problem or Issue	Limited integration of multiple sources of evidence to enhance causal inference and identify mechanisms in epidemiological research
What is already known	While MR studies are abundant, few utilise complementary biomedical data sources to enhance causal evidence or generate new hypotheses for underlying mechanisms
What this paper adds	By utilising MR-EvE and literature-mined data in EpiGraphDB, this study introduces a systematic, integrated approach for causal inference, applied to breast cancer risk factors. It demonstrates how combining MR and LBD enhances the identification of mediating traits, providing a model that can be applied to other complex diseases

2. Methods

2.1. Study Design

We present a summary of our EpiGraphDB data-focused study in Figure 1. First, we explored MR-EvE causal estimates between various exposure traits and breast cancer to generate a list of potential causal risk factors and biomarkers. Then, we used literature-mined relationships from EpiGraphDB to dissect the risk factor/breast cancer relationships. We designed an LBD method for overlapping literature spaces of traits to identify potential intermediates between individual traits and breast cancer.

We demonstrate evidence integration between MR and LDB on two case study breast cancer risk factors: *childhood body size* and *HDL-cholesterol*. We queried

MR-EvE data using a two-step MR approach [23, 24] to identify potential mediators (i.e. traits affected by each risk factor that in turn affect breast cancer risk). We also determined potential intermediates using LBD and evaluated them using MR and data from OpenGWAS. In this way, we triangulated evidence for the identified mediators from two independent data sources. Analysis code is available at `https://github.com/mvab/epigraphdb-breast-cancer` and `https://github.com/mvab/epigraphdb_mr_literature_queries`.

2.2. EpiGraphDB

EpiGraphDB (`https://epigraphdb.org`) is a graph database that integrates biomedical and epidemiological data, built to support data mining of risk factor/disease relationships. It contains trait relationships (observational and genetic), literature-mined relationships, biological pathways, protein-protein interactions (PPIs), drug–target relationships and other data sources [4]. EpiGraphDB is integrated with the OpenGWAS database data [11] (`https://gwas.mrcieu.ac.uk`), providing access to GWAS studies via the GWAS node of the graph. The number of GWAS studies in OpenGWAS at the time of writing is 50,040; of these, 34,494 are integrated within EpiGraphDB. EpiGraphDB (version 0.3.0) was accessed via R package *epigraphdb-r* (version 0.2.3).

2.3. Breast cancer data

EpiGraphDB provides access to multiple breast cancer GWAS datasets deposited in OpenGWAS. We used the most recent releases of the Breast Cancer Association Consortium (BCAC) (`http://bcac.ccge.medschl.cam.ac.uk/`) meta-analyses conducted in 2017 (BCAC 2017, N=228,951) [25] and 2020 (BCAC 2020, N=247,173) [26]. Since only BCAC 2017 is available in EpiGraphDB/OpenGWAS, this dataset was used for the MR-EvE risk factor discovery analysis. Both BCAC 2017 and 2020 were utilized for the validation stage. In addition to the overall sample, the datasets contain the ER+ and ER- sub-samples and five molecular subtypes (Supplementary Table 1). The study groups in the BCAC cohort do not include UK Biobank or other trait cohorts used in this study. The cohort includes European-ancestry individuals only.

2.4. Mendelian randomization and MR-EvE data

MR is an application of instrumental variable analysis where genetic variants are used as proxies (genetic instruments) to estimate the causal relationship between a modifiable health exposure and a disease outcome [8, 9]. MR-EvE (MR "Everything-vs-Everything") data is stored as bidirectional relationships between GWAS traits in EpiGraphDB, available for 11,789 GWAS traits at the time of writing. The MR-EvE estimates were generated using the MR Mixture-of-Experts

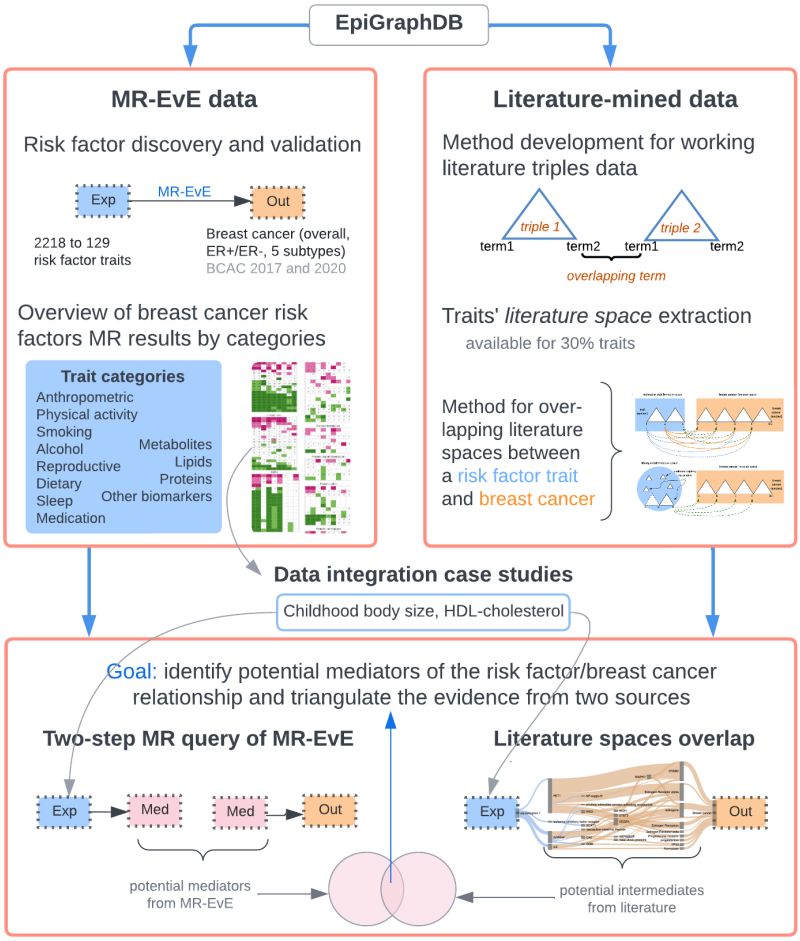


Figure 1: **Study overview**. Employing EpiGraphDB, MR-EvE data was used to identify breast cancer risk factors, and literature-mined data was used to design a literature-based discovery (LBD) method for identifying potential intermediates between risk factors and breast cancer. For two case studies, we present evidence integration between MR-EvE-guided and LBD-identified potential mediators. Abbreviations/definitions: MR-EvE – Mendelian Randomization "Everything-vs-Everything"; Exp – exposure; Med – mediator; Out – outcome; BCAC – Breast Cancer Association Consortium (refers to meta-analyses breast cancer cohort GWAS data); literature space – a collection of literature-mined relationships between pairs of biological terms related to a phenotype.

(MR-MoE) method by Hemani *et al* [14] before integration into EpiGraphDB. MR-MoE is a machine learning framework that automated the selection of instrument SNPs and the MR method for use in any specific causal analysis by predicting the model of pleiotropy and relating that prediction to the most appropriate model

[14]. The method selects the MR 'best estimate' of causal relationships between the exposures and the outcome traits.

The interconnection of many traits in the MR-EvE data makes it a valuable resource for identifying potential mediators and confounders [27] for a causal relationship of interest (`https://epigraphdb.org/confounder`) [4]. In this work, we make use of this functionality to generate hypotheses for breast cancer risk factors and potential mediators.

2.5. Identifying breast cancer risk factor traits from MR-EvE

2.5.1. Discovery

We queried EpiGraphDB to extract MR-EvE results for all available GWAS traits (N = 2218, Supplementary Table 2) with breast cancer as the outcome. Figure 2 summarises the discovery and validation stages of the analysis. The initial set of traits linked to breast cancer with any MR estimate was reduced to traits that passed False Discovery Rate (FDR), taking traits with the corrected p-value < 0.05 (N = 378) into the downstream analysis (Figure 2).

The traits were grouped into 12 categories, which included widely-known cancer risk factor categories (anthropometric traits, reproductive traits, physical activity, smoking, alcohol), categories with a growing research interest (dietary intake traits, sleep traits, medication), and molecular traits (metabolites, lipids, proteins, and other biomarkers) (N = 202). The traits outside of these categories were grouped as 'Other' and excluded. The full data from the discovery stage is available in Supplementary Table 3.

2.5.2. Validation

The results of the MR-EvE rapid screen of risk factors were validated by re-analysing the identified relationships using the robust practices in two-sample MR [28], because MR-EvE 'best estimates' selected by the MR-MoE method do not always correspond to the standard methods used in MR studies [14]. The validation results in this study are reported as per the STROBE-MR guidelines [29, 30].

For the majority of exposure traits, instruments were extracted using the genome-wide approach, i.e. SNPs with p-value $< 5 \times 10^{-8}$ and clumped with $r2 < 0.001$ using 1000 Genomes European LD reference panel. For protein traits, *cis*-region instruments were prioritised (within 1Mb on either side of the protein-encoding gene) [31]. The extracted *cis*-SNPs were filtered to p-value $< 5 \times 10^{-8}$ or a less stringent 'local' threshold of p-value $<$ 0.05/nSNPs, where nSNPs is the number of SNPs in the *cis* region. The clumping was done using $r2 < 0.01$, or $r2 < 0.001$ if the higher threshold produced ≥ 7 SNPs. If no instruments were available under these criteria, genome-wide extraction of instruments was applied. The details of

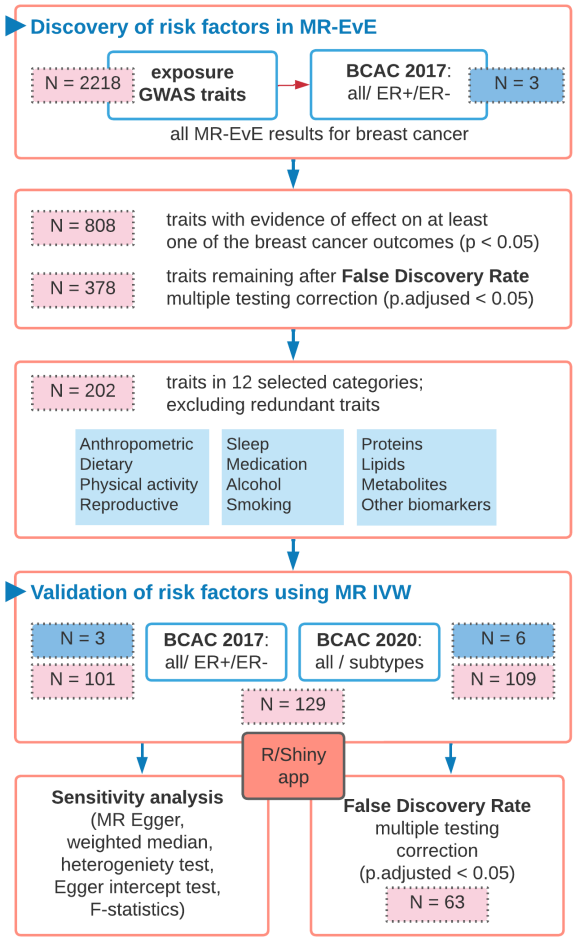


Figure 2: **MR-EvE evidence processing workflow, including Discovery and Validation stages.** The boxes show the number of exposures (pink) and breast cancer outcomes (blue) after/in each filtering step; MR-EvE results validation Shiny app: `https://mvab.shinyapps.io/MR_heatmaps/`. BCAC - Breast Cancer Association Consortium

instrument extraction for protein traits are provided in Supplementary Tables 4 and 5.

The validation analysis was performed on breast cancer molecular subtypes data (BCAC 2020) in addition to the overall sample and ER+/ER- subtypes (BCAC 2017). The inverse-variance weighted (IVW) [32] approach was used as the main two-sample MR method. To correct for multiple testing in the validation step, the FDR cut-off of 0.05 for the corrected p-value was applied to identify results with

more robust evidence of effect (Supplementary Table 4) (Figure 2).

To investigate potential violations of the MR assumptions and validate the robustness of MR results, additional analyses were performed. This included pleiotropy-robust sensitivity analyses such as MR-Egger [33] and weighted median [34]. The Egger intercept test [33] was used to test for directional horizontal pleiotropy, and Cochran's Q statistic [35] was calculated to quantify heterogeneity among SNPs, which is indicative of potential pleiotropy. F-statistic was used to assess instrument strength [36] (Supplementary Tables 4 and 6). Sensitivity analysis results were not used for additional filtering of all identified risk factor traits to avoid information loss due to dichotomisation [37], but were taken into account when evaluating individual traits. Additionally, the sensitivity results were included in the MR results heatmap visualisation to provide context on the performance in sensitivity tests for all traits. All MR and sensitivity analyses were performed using the *TwoSampleMR* R package (version 0.5.6) [13].

2.6. Using literature-mined relationships in EpiGraphDB

EpiGraphDB contains literature-mined relationships between pairs of terms (also known as 'literature triples') derived from the published literature [38]. The underlying literature data comes from SemMedDB (Semantic MEDLINE Database) [16], a well-established repository of literature-mined semantic triples, i.e. '*subject term 1 – predicate – object term 2*', mined from titles and abstracts of nearly 30 million biomedical articles in PubMed, using SemRep [39, 40] (further details are available in Appendix A). A subject/object term can be any biological entity (gene name, drug, phenotype, disease); a predicate is a verb that represents a relationship between the two terms (*affects, causes, inhibits, reduces, associated with*, etc.).

2.6.1. Literature space

Literature triples of biomedical terms are mapped to specific GWAS traits in EpiGraphDB. In this work, we introduce the concept of a '*literature space*'. The literature space for a trait (e.g. breast cancer) comprises a set of semantic triples extracted from the published literature for this trait. Literature space can be considered an 'automated literature overview' of conceptual relationships between biological entities related to the search term, i.e. breast cancer. In the literature space, each relationship triple has a score that indicates how frequently it occurs in the published literature within a defined time frame (linked to PubMed IDs), demonstrating how well-established this relationship is.

2.6.2. Literature spaces overlap method

In the field of LBD, the underlying principle of working with triple-based literature data is known as Swanson linking [41, 42]. In this approach, two separately

published results (e.g. A-B and B-C relationships) are combined to identify evidence for the A-C relationship that is unknown or unexplored. The discovery can be open and closed: in open, A is given to identify Bs linked to Cs, generating hypotheses; in closed, A and C are fixed, and Bs that link them are identified, testing hypotheses or finding intermediates [43, 18] (details in Appendix A).

In this work, we present an extension to the simple triple linking method used in MELODI-Presto [38], which focused on the overlap of two triples (i.e. A-B and B-C only) (3a). Here, we combine multiple triples from two literature spaces (exposure trait and breast cancer) into a so-called *'literature overlap'*. We apply multi-step triple linking within and between the literature spaces to generate hypotheses for the potential mediators of the exposure trait's effect on breast cancer. The literature data pre-processing steps and the method details are provided in Appendix A. The overview of our LBD method is presented in Figure 3, including closed (3b) and open (3c) discovery, and an example application (3d).

2.7. MR and literature evidence integration

We integrated the evidence from MR-EvE and literature mining for potential mediators of two breast cancer risk factors that we selected as case studies: *childhood body size* and *HDL-cholesterol*.

2.7.1. Potential mediators from MR-EvE data

We queried the MR-EvE data to identify potential mediators of each case study risk factor using the two-step MR framework [23, 24, 44]. The mediator search results were corrected for multiple testing using FDR at the individual case study level, for each step within two-step MR: *(step 1)* a risk factor trait effect on potential mediators, and *(step 2)* potential mediators' effects on breast cancer outcomes (BCAC 2017 data only).

The mediators passing the FDR-adjusted p-value < 0.05 filtering in both steps and restricted to the earlier specified 12 trait categories were taken into the validation step. The validation for two-step MR results included the same analysis and sensitivity methods used for the main risk factors (2.5.2). The validated two-step MR estimates were also corrected for FDR.

2.7.2. Potential mediators from the LBD literature overlap approach

The literature spaces for risk factor traits and breast cancer were constructed using the literature data linked to the following GWAS traits: breast cancer (combination of two traits: '*breast cancer*' - *ieu-a-1126* and '*malignant neoplasm of the breast*' - *finn-a-C3_BREAST*), HDL (*'HDL-cholesterol' - ukb-d-30760_irnt*), childhood body size (*'childhood obesity'- ieu-a-1096*) (Supplementary Tables 14-15).

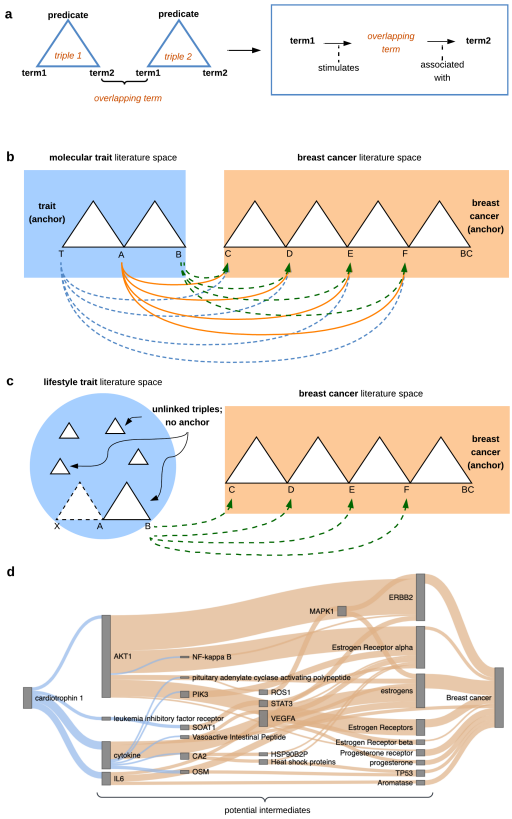


Figure 3: **Schemas of the literature-based discovery method, demonstrating the approach behind literature spaces overlap. (a)** A basic example of linking two triples via an overlapping term. **(b)** *Closed discovery*: overlap of two traits' literature spaces (blue - risk factor, orange - breast cancer) by chaining triples of terms between them. The outer terms are restricted ('i.e. 'anchored') to the term representing the trait (T: trait) and the term representing breast cancer outcome (BC: breast cancer). All intermediate terms are arbitrarily named A-F. The arrows represent all possible paths from T to BC, which are alternatives to the full path going through all intermediates. Closed discovery is applied to molecular traits. **(c)** *Open discovery*: Applied to non-molecular/lifestyle traits, where the trait is not available as a term and therefore cannot be anchored. The unlinked triples in the non-molecular trait space are first matched (via term B) to link with any triples in the breast cancer space. Then, the linked triples are connected to other triples in the non-molecular trait space (via term A), expanding the literature spaces overlap. **(d)** *Example of literature spaces overlap visualisation*: Sankey diagram showing how triples within and between a risk factor trait (here, cardiotrophin-1) and breast cancer spaces interconnect. The line width of each term relationship indicates how common it is in the literature (frequency score). Cardiotrophin-1 is presented here as an example. More examples, specifically the case study traits in this study, are available to view at `https://mvab.shinyapps.io/literature_overlap_sankey/`

In the literature overlap generated from triples in the risk factor trait and breast cancer literature spaces, all intermediate biological terms were evaluated as potential mediators. Each term was searched in OpenGWAS for representative GWAS data to be used in MR analysis, and genetic instruments for the identified traits were extracted as previously described (2.5.2).

We performed MR bidirectionally between the risk factor trait and each intermediate to identify the direction of effect, as it may not be correct in literature-mined relationships. Then, the intermediate traits were evaluated for an effect on breast cancer. The results were validated using the analysis and sensitivity methods described previously (2.5.2), and adjusted using FDR.

The evidence in both steps of the two-step MR analysis provided suggestive evidence of potential mediation between the risk factor and breast cancer via a trait identified by the LBD method.

3. Results

3.1. Breast cancer risk factors discovered from MR-EvE

We identified 129 traits across 12 categories that showed evidence of an effect on breast cancer risk from MR-EvE data (Figure 2). The discovery effect estimates were validated using the IVW MR analysis with sensitivity analyses, using 9 breast cancer outcome GWAS from two releases of BCAC data. The instruments were extracted as described in 2.5.2. To correct for multiple testing of many exposures and outcomes in the validation step, the FDR correction was applied. The validation results and sensitivity analyses are available in Supplementary Tables 4 and 6.

The MR results for the final set of traits (129 traits or 63 after FDR correction) are presented in Figure 4. The figure also provides an overview of sensitivity analysis results: potential pleiotropy "P", heterogeneity "H", and "X" for exposures with a low number of instruments ($\leq$ 2) where most sensitivity tests were not possible. The full results (202 traits) are available in the R/Shiny app (`https://mvab.shinyapps.io/MR_heatmaps/`) and Supplementary Table 4. The abbreviations for traits/biomarkers used throughout the section are defined in Supplementary Table 7. In the sections below, the results of several trait categories are summarised.

3.1.1. Non-molecular traits

The majority of anthropometric traits (e.g. *body mass index, obesity, waist-to-hip ratio, childhood body size*) (Figure 4a) have consistent inverse (risk-decreasing) effects across most breast cancer subtypes. Interestingly, *impedance of whole body* has a positive (risk-increasing) effect on most outcomes, potentially capturing a

Figure 4: **Mendelian randomization effect direction heatmaps for (A) lifestyle and anthropometric traits, (B) metabolites and lipids, (C) proteins (N=129) as exposures, and breast cancer (BCAC 2017 and 2020 overall samples and subtypes) as the outcomes.** The presented traits have evidence of an effect on at least one outcome, but the full set of traits analysed in the validation step (N=202) is available in the interactive version of the plot in the app: `https://mvab.shinyapps.io/MR_heatmaps/` and in Supplementary Table 4. [continued on the next page]

Figure 4: [continued] The proportions of the total size of each group displayed here are: Anthropometric – 95%, Lifestyle – 57%, Lipds -100%, Metabolites – 68%, Proteins – 54%. The colours represent effect direction: pink – positive (causal), green – negative (protective), white – null effect (based on 95% CIs of odds of breast cancer per standard deviation increase in the exposure). The darker colour shade indicates the results that passed the FDR correction. The letters summarise arbitrarily dichotomised sensitivity analysis results: "P" - potential pleiotropy (*abs*(egger intercept)>0.05); "H" - heterogeneity (Q-stat p-value < 0.05); "X" - low number of instruments (≤ 2) where sensitivity analyses were not possible; an empty cell indicates that no H/P/X issues were identified. For proteins, where possible, *cis*-only instruments were used (marked with an asterisk *). The proteins were grouped by major pathways defined in the Reactome database [45] for display purposes. Proteins are displayed by the encoding gene name (defined in Supplementary Table 7). The exposure traits are from mixed-sex samples unless otherwise specified (F: female-only) or are female-specific reproductive traits. The details on breast cancer outcome samples are provided in Supplementary Table 1.

different aspect of adiposity. The traits in this group show evidence of high instrument heterogeneity ('H' in Figure 4a) based on the arbitrary threshold used (Q-stat p-value < 0.05). This indicates the complexity of these phenotypes, suggesting that multiple underlying mechanisms are likely involved. Evaluating multiple related/correlated phenotypes is acceptable, as none of them are perfect measures of adiposity, and they likely proxy adiposity in different ways, e.g. BMI (subcutaneous fat) vs waist-hip ratio (central adiposity).

Among physical activity traits (UK Biobank traits based on a survey question *"Types of physical activity in last 4 weeks"*), *'Heavy DIY'* shows evidence of a protective effect in overall samples (with little indication for heterogeneity and pleiotropy). Among dietary traits, there is a protective effect from *'dried fruit intake'* on outcomes with larger sample sizes. The intake of *processed/sugary breakfast cereals (i.e. Frosties, Cornflakes)* appears to have a risk-increasing effect, but the instruments here are also likely highly pleiotropic as is commonly seen for dietary intake traits. The important considerations for dietary and physical activity exposures in MR analyses are addressed in Appendix C.

The *'never smoked'* phenotype appears to confer a decreased risk in selected breast cancer subtypes. *'Chronotype (evening preference)'* shows evidence of increasing the risk in outcomes with larger sample sizes. *'Snoring'* also has a causal effect on some outcomes. Higher *age at menopause* has a risk-increasing effect on most outcomes, while '*Had menopause*' phenotype (potentially capturing earlier menopause) is protective.

3.1.2. Molecular traits

The lipid traits have a distinct effect direction pattern (Figure 4b): a consistently positive (causal) effect from HDL particle measures and a negative (protective) effect from VLDL. For all, there is limited evidence of effect on ER- breast

cancer and several molecular subtypes. It should be noted that LDL measures were also available in the MR-EvE discovery data, (i.e. 2218 traits), but did not pass filtering thresholds in the validation step due to weak evidence of effect (data available in Supplementary Table 3). Among other metabolites (Figure 4b), 3-dehydrocarnitine, betaine, and X-11440 show the strongest evidence of a negative effect on breast cancer risk.

The proteins with the most robust evidence (FDR-adjusted p-value < 0.05 and based on *cis*-instruments only) are IL1RL1, ISLR2, SAT2 (negative effect) and PRTN3, B3GNT2, CPXM1, PLXNB2 (positive effect). The proteins that did not pass FDR correction, but are also based on *cis*-only instruments and have effect on more than one outcome, include IL6R, EREG, DCBLD2 (negative effect) and VTN, IDUA, DNAJB11, XXYLT1 (positive effect).

Among the proteins for which genome-wide instruments were used (which may be picking up horizontal pleiotropy), the most consistent evidence of effect was observed for FLNA, ADH6, CAMK1 (negative effect) and IL21, CAST (positive effect) with little heterogeneity and pleiotropy detected ('P' and 'H' not present in Figure 4c).

3.2. Evidence integration case studies

This section presents two exposure traits from the discovery of the MR-EvE risk factor, *childhood body size* and *HDL cholesterol*, as case studies. Although both have previously been investigated in relation to breast cancer, the underlying mechanisms remain unclear. We used MR-EvE data and data mined from literature to generate hypotheses for potential mediators and cross-validate evidence from both sources. Since these traits have opposite effects on breast cancer risk and represent both molecular and non-molecular risk factors, they provide a robust framework for demonstrating open and closed literature-based discovery, supported by large literature spaces (Appendix A).

3.2.1. Case study 1: Childhood body size

Childhood adiposity has been shown to have a protective effect on breast cancer risk in both observational [46, 47] and MR studies [48, 44], but the mechanism and mediators of this effect remain unclear. MR analysis shows strong evidence of an inverse effect from childhood body size (OpenGWAS ID: *ukb-b-4650*, female-only sample) on all breast cancer subtypes (Figure 4) (e.g. OR 0.60 [95% CI 0.53: 0.69] for BCAC 2017 overall sample).

MR-EvE mediators. In the initial MR-EvE scan, 125 putative mediators of the childhood body size effect on breast cancer were identified. Among these, several

were considered as potential mediators in earlier work [44]: IGF-1, age at menarche, age at menopause, age at last birth, glucose, glycated haemoglobin (Supplementary Table 8). Following two-step MR results validation, 71 traits had evidence of effect in both steps at nominal significance and 54 traits at the FDR-corrected level (Supplementary Tables 9 and 10). Among these, a large proportion were adult anthropometric traits (N=30). Since adult adiposity has been shown to be unlikely to mediate the childhood adiposity effect [48, 49], these traits were not reviewed further. Lipid traits (N=20) were also not considered further as they are unlikely to mediate substantial effect individually. The remaining candidate mediators traits were: *exercise to keep fit* (from UK Biobank physical activity questionnaire), *sleep duration, skimmed milk intake, and Apolipoprotein A (ApoA)* (Table 1).

mediator	outcome	step1: beta	step1: p-value	step2: OR	step2: p-value
Apolipoprotein A ukb-d-30630_irnt	Overall BC	-0.23 [-0.28:-0.17]	2.9e-15	0.77 [0.66:0.88]	2.2e-4
Skimmed milk ukb-d-1418_3	ER+ BC	0.09 [0.08:0.11]	2.1e-35	0.01 [0:0.15]	0.0014
Exercise to keep fit ukb-a-484	Overall BC	0.02 [0.001:0.04]	0.029	0.4 [0.21:0.75]	0.0044
Sleep duration ukb-b-4424	ER+ BC	-0.04 [-0.08:-0.01]	0.0053	1.49 [1.12:2.0]	0.0071

Table 1: **Two step-MR results for potential mediators of childhood body size/breast cancer relationship, identified in a hypothesis-free way via mining MR-EvE.** The results in the table are from the MR IVW validation analysis, showing beta [95%CI] for *step 1* (the effect of childhood body size on mediator, measured as SD change in mediator per body size category change) and OR [95%CI] for *step 2* (the odds of breast cancer per SD higher mediator). The table only includes the results that passed FDR in both steps. The table does not include anthropometric and lipid traits (the full results are available in Supplementary Table 9).

Higher childhood body size shows some evidence of an increasing effect on the amount of *exercise to keep fit* in adulthood (effect size: 0.02 [95% CI 0.001: 0.04], step 1), which in turn has a protective effect on breast cancer risk (OR: 0.4 [95% CI 0.21: 0.75], step 2) (in agreement with published MR study [50] and results in 3.1.1). The relationship between childhood adiposity and the amount of exercise in adulthood has not been explored with MR designs before, so this finding warrants further follow-up.

Higher childhood body size has a negative effect on *sleep duration* in adulthood (effect size -0.04 [95% CI -0.08: -0.01], step 1), while longer sleeping time increases breast cancer risk (stronger evidence for ER+, OR 1.49 [95% CI 1.12: 2.0], step 2), in agreement with published work [51].

Literature mediators. Next, we reviewed literature-mined terms that connect childhood adiposity to breast cancer. The Sankey diagram in Supplementary Figure B.6 shows the literature overlap between the two traits (Supplementary Table 17), performed using the open discovery approach (Figure 3c, 2.6.2) (view interactively at

`https://mvab.shinyapps.io/literature_overlap_sankey/`).

The intermediate terms of the literature overlap include traits considered in earlier hypothesis-driven work [44] as potential mediators: IGF-1, testosterone, SHBG, insulin, and oestradiol. IGF-1 was also found using MR-EvE, but did not pass the FDR correction (mixed-sex sample data). The overlap contains many other literature-mined terms that make interesting mediation hypotheses (e.g. leptin, prolactin, adiponectin).

Next, to assess the identified intermediate terms as potential mediators using MR analysis, we searched OpenGWAS for suitable GWAS data to represent them. Of the 112 potential intermediate traits identified, data was available for 49, with 38 traits having at least one SNP instrument (Supplementary Table 18).

We carried out MR bidirectionally between childhood body size (*ukb-b-4650*) and 38 potential intermediate traits (Supplementary Tables 19 and 20), followed by assessing their effect on breast cancer (Supplementary Table 21). A subset of combined results from the three MR analyses are presented in Table 2.

The criteria for a trait to have suggestive evidence of being a potential mediator (i.e. evidence of effect in *step1 forward* and *step2* (i.e. two-step MR), and preferably no effect in *step1 reverse*) was met by two traits: IGF-1 and bioavailable testosterone. These two traits have been previously evaluated in [44] two-step MR analysis, with equivalent results and sensitivity analyses.

3.2.2. Case study 2: HDL-cholesterol

HDL-cholesterol (HDL-C) has been shown to have a risk-increasing effect on breast cancer in several MR studies [52, 53, 54], which opposes the finding from a large meta-analysis of prospective studies that showed a modestly inverse association [55]. This makes HDL-C an interesting case study for hypothesis-free mediators search, as this may help to dissect the mechanism linking it to breast cancer.

The HDL-C trait in MR-EvE is based on mixed-sex sample data from UK Biobank (OpenGWAS ID: *ukb-d-30760_irnt*); the evidence of effect from IVW for BCAC 2017 overall sample is OR 1.09 [95% 1.04: 1.13] (no effect on certain subtypes) (Figure 4b), which is in agreement with previous studies and consistent between sensitivity analysis MR methods (Supplementary Tables 4 and 6).

MR-EvE mediators. In the initial MR-EvE scan, 102 potential mediators of HDL-C's effect on breast cancer were identified, primarily lipid traits, proteins, and anthropometric traits (Supplementary Table 11). After two-step MR validation, 60 traits showed nominal significance, and 44 remained significant after FDR correction (Supplementary Tables 12 and 13). Of these potential mediators, 36 were lipid traits, which we did not evaluate further due to their high correlation with HDL-C

Mediator trait	Step 1, forward (effect size, [95%CI])	Step 1, reverse (effect size,[95%CI])	Step 2 (OR, [95%CI])
Alanine transaminase ebi-a-GCST90013405	0.03 [0.02:0.04]*		
CRP ieu-b-35	0.39 [0.31:0.46]*		
CSNK1D prot-a-693			1.11 [1.01:1.22]
Cortisone met-a-354	-0.04 [-0.07:-0.02]*		
E selectin prot-b-77			0.95 [0.93:0.97]*
EGF prot-a-908		0.01 [0:0.02]	
FABP4 prot-b-71	0.59 [0.27:0.90]*		
Glycoprotein acetyls met-c-863	0.33 [0.20:0.45]*		
IGF-1 ukb-d-30770_irnt	-0.24 [-0.31:-0.16]*	-0.02 [-0.03:0]	1.07 [1.01:1.13]
Insulin receptor prot-a-1564			0.94 [0.92:0.97]*
IGF1R prot-a-1444			0.82 [0.76:0.88]*
Leptin receptor prot-a-1724		0.01 [0:0.01]	
Oestradiol (F) ieu-b-4872	-0.07 [-0.14:-0.01]	0.14 [0.08:0.21]*	
SHBG (F) ieu-b-4870	-0.20 [-0.29:-0.10]*		
TSHbeta levels ebi-a-GCST90010353		0.04 [0.01:0.08]	
TNFRSF11B prot-b-83			0.93 [0.87:0.99]
Bio Testosterone (F) ieu-b-4869	0.10 [0.04:0.17]*		1.12 [1.04:1.20]*
Total Testosterone (F) ieu-b-4864			1.14 [1.06:1.23]*

Table 2: **Subset of MR results (bidirectional and two-step) for potential mediators of childhood body size/breast cancer relationship, identified using LBD.** The table combines the results from three MR analyses: *step 1 forward* (childhood body size →mediator trait), *step 1 reverse* (mediator trait → childhood body size), and *step 2* (mediator trait → overall breast cancer). '*step 1 forward*' and '*step 1 reverse*' corresponded to bidirectional MR analysis; '*step 1 forward*' and '*step 2*' together correspond to two-step MR study design. The table only shows the results that pass the threshold of p.value < 0.05; the results that pass FDR are marked with an asterisk*. The full results are available in Supplementary Tables 19, 20, 21. *Abbreviations*: CRP - C-Reactive protein, TSHbeta - Thyrotropin subunit beta, TNFRSF11b – tumour necrosis factor receptor superfamily member, CSNK1D - Casein kinase I isoform delta, EGF – epidermal growth factor, FABP4 - fatty acid binding protein 4; (F) - female-only data

and, therefore, the need for multivariable MR (MVMR) analysis. The identified mediators also included several anthropometric traits, which may have a role in the effect of HDL-C on breast cancer risk, but also would require further investigation using MVMR for traits capturing different adiposity components (e.g., waist-hip ratio, BMI).

HDL-C was found to have an increasing effect on two protein traits, *ApoA* (ef-

fect size: 0.87 [95% 0.84: 0.90]) and *cathepsin F* (CTSF) (0.14 [95% 0.03: 0.24]) (Table 1. The relationship and interplay between HDL-C, ApoA, and cathepsins have been investigated before [56, 57]. ApoA is the major protein component of HDL, which has an inverse effect on breast cancer in MR (OR: 0.77 [0.66:0.88], using *cis*-instruments only), also previously studied in connection to breast cancer [58]. CTSF has a positive effect on breast cancer risk (OR: 1.14 [95% 1.08: 1.21], single *cis*-instrument). CTSF belongs to a class of evolutionally conserved cysteine proteases involved in various biological processes that have been linked to disease, including cancer progression [59, 60]. Cathepsins (F, S, K; each to various degree) degrade lipid-free molecules of ApoA, leading to a complete loss of the ApoA's function as a cholesterol acceptor, stopping its anti-inflammatory action [56], which may be related to CTSF's role in disease. Further investigation into the mechanism linking HDL-C, ApoA, and CTSF (and potentially other cysteine proteases), and breast cancer is needed.

mediator	outcome	step 1: beta	step1: p-value	step2: OR	step2: p-value
Apolipoprotein A ukb-d-30630_irnt	Overall BC	0.87 [0.84:0.9]	0	0.77 [0.66:0.88]	2.2e-4
Cathepsin F (CTSF) prot-a-722	Overall BC	0.14 [0.03:0.24]	0.014	1.14 [1.08:1.21]	7.5e-6
Childhood obesity ieu-a-1096	Overall BC	-0.19 [-0.35:-0.03]	0.023	0.82 [0.76:0.88]	2.4e-7
Hip circumference ieu-a-51	Overall BC	-0.07 [-0.11:-0.02]	0.0043	0.73 [0.59:0.9]	0.0039
Waist circumference ieu-a-63	Overall BC	-0.07 [-0.11:-0.02]	0.0039	0.73 [0.59:0.9]	0.0039
Weight ukb-b-12039	ER- BC	-0.04 [-0.07:-0.01]	0.0062	0.87 [0.79:0.95]	0.0029
Exercise to keep fit ukb-a-484	Overall BC	0.01 [0:0.02]	0.012	0.4 [0.21:0.75]	0.0044
Snoring ukb-a-14	ER- BC	0.01 [0:0.02]	0.0015	5.83 [1.9:17.87]	0.0021

Table 3: **Two step-MR results for potential mediators of HDL-cholesterol/breast cancer relationship, identified in a hypothesis-free way via mining MR-EvE.** The results in the table are from the MR IVW validation analysis, showing beta [95%CI] for *step 1* (the effect of HDL-C on mediator, measured as SD change in mediator per SD higher HDL-C) and OR [95%CI] for *step 2* (the odds of breast cancer per SD higher HDL-C). The table only includes the results that passed FDR in both steps. The table does not include lipid traits (the full results are available in Supplementary Table 12).

Literature mediators. Next, we reviewed literature-mined data connecting HDL and breast cancer to look for potential intermediates. Supplementary Figure B.7 shows the Sankey diagram of literature overlap between the traits (Supplementary Table 22), performed using the closed discovery approach (Figure 3b, 2.6.2) (view interactively at `https://mvab.shinyapps.io/literature_overlap_sankey/`).

The intermediate terms of this literature overlap include 120 unique (mostly

molecular) traits. Many of them are lipids-related traits (e.g., LDL, VLDL, ApoA, ApoB, Lipase); others include interesting hypotheses such as testosterone, IGF-1, insulin, CRP, and cathepsin D. Suitable GWAS datasets for MR validation were found for 50 of the 120 traits (Supplementary Table 23).

We carried out MR bidirectionally between HDL-C (*ukb-d-30760_irnt*) and 42 potential intermediate traits (>=1 SNP) (Supplementary Tables 24 and 25), followed by assessing their effect on breast cancer (Supplementary Table 26). A subset of combined results from these three MR analyses is presented in Table 4.

From these results, the traits that fit the criteria for suggestive evidence of mediation are ApoA and testosterone. ApoA has been identified as a potential mediator using MR-EvE as well. HDL-C has opposite effects on two types of testosterone: negative on bioavailable and positive on total (Table 4), potentially indicating different interaction mechanisms. Both testosterone measures have a risk-increasing effect on breast cancer (previously reported in [61, 44]). In the literature overlap, HDL is linked to testosterone via ApoE, HSD11B1, lipoproteins, and directly. In turn, testosterone is linked to breast cancer directly or via estrogen/progestin (Figure B.7).

Finally, cathepsin D (CTSD) has been identified as a potential intermediate in the literature overlap (via CETP/endopeptidases and lipoprotein/cyclosporine, Figure B.7). MR results show some evidence of HDL-C effect on CTSD (effect size -0.12 [95% CI -0.23:-0.003], but little evidence of CTSD effect on breast cancer (OR: 1.02 [95% CI 0.99:1.05], three *cis*-SNPs) (Table 4, Supplementary Table 26), suggesting an unlikely role of CTSD as a mediator. However, the discovery of a CTSD in literature overlap complements the finding of CTSDF in the MR-EvE data (Table 3), indicating that the cathepsins may have a role in the relationship of HDL-C and breast cancer, and the entire class of these enzymes should be explored in future work.

4. Discussion

In this study, we used the EpiGraphDB knowledge graph to systematically identify breast cancer risk factors. Additionally, we integrated literature-mined and genetic data to determine potential mediators of the identified causal relationships between these risk factors and breast cancer, which we presented via two case studies.

Through mining MR-EvE data in EpiGraphDB we identified 2218 traits affecting breast cancer risk, which were narrowed down to 129 traits with reliable causal evidence following rigorous validation (Figure 4, `https://mvab.shinyapps.io/MR_heatmaps/`). The identified risk/protective factors included both molecular biomarkers and non-molecular traits. Many of these traits

Mediator trait	**Step 1, forward (effect size,[95%CI])**	**Step 1, reverse (effect size, [95%CI])**	**Step 2 (OR, [95%CI])**
Adiponectin ieu-a-1	0.1 [0.05:0.14]*		
Albumin met-c-841		0.05 [0:0.1]	
Apolipoprotein A-I met-c-842	0.71 [0.63:0.78]*	0.69 [0.60:0.77]*	0.64 [0.50:0.80]*
Apolipoprotein B ieu-b-108	-0.21 [-0.29:-0.12]*	-0.16 [-0.28:-0.05]*	
Basigin prot-a-275		0.11 [0.08:0.14]*	
CRP ukb-d-30710_irnt		0.03 [0.001:0.07]	
CAMK1 prot-a-347		-0.03 [-0.05:-0.01]*	
CAMK1 prot-a-346			0.98 [0.97:0.99]*
Cathepsin D prot-b-51	-0.12 [-0.23:-0.003]		
Ephrin-B2 prot-a-905		-0.04 [-0.07:-0.01]	
Histone deacetylase 8 prot-a-1320		-0.24 [-0.27:-0.21]*	
IGF-1 ukb-d-30770_irnt			1.07 [1.01:1.13]*
IGF1R prot-a-1444		-0.1 [-0.13:-0.08]*	0.82 [0.76:0.88]*
Lipoprotein A ukb-d-30790_irnt		0.01 [0.001:0.02]*	
Lipoprotein lipase ebi-a-GCST90010149	0.24 [0.04:0.45]	0.27 [0.24:0.29]*	
Oestradiol (F) ieu-b-4872		0.28 [0.17:0.39]*	
STARD1 prot-a-2866	-0.13 [-0.24:-0.01]		
Superoxide dismutase prot-a-2799	0.12 [0.01:0.24]	-0.01 [-0.02:-0.01]*	
Bio Testosterone (F) ieu-b-4869	-0.07 [-0.11:-0.03]*	-0.11 [-0.17:-0.05]*	1.12 [1.04:1.20]*
Total Testosterone (F) ieu-b-4864	0.03 [0.001:0.05]		1.14 [1.06:1.23]
Total esterified cholesterol met-d-Total_CE	0.33 [0.27:0.39]*	0.30 [0.10:0.50]*	
VLDL cholesterol met-d-VLDL_C	-0.31 [-0.37:-0.24]*	-0.36 [-0.56:-0.17]*	

Table 4: **Subset of MR results (bidirectional and two-step) for potential mediators of HDL-C/breast cancer relationship, identified using LBD.** The table combines the results from three MR analyses: *step 1 forward* (HDL-C →mediator trait), *step 1 reverse* (mediator trait → HDL-C), and *step 2* (mediator trait → overall breast cancer). '*step 1 forward*' and '*step 1 reverse*' corresponded to bidirectional MR analysis; '*step 1 forward*' and '*step 2*' together correspond to two-step MR study design. The table only shows the results that pass the threshold of p.value < 0.05; the results that pass FDR are marked with an asterisk*. The full results are available in Supplementary Tables 24,25,26. *Abbreviations:* CRP - C-Reactive protein, CAMK1 - Calcium/calmodulin-dependent protein kinase type 1, STARD1- Steroidogenic acute regulatory protein; (F) - female-only data

have been reported before, e.g. anthropometric measures [48, 62, 63], reproductive traits [44], sleep traits [51, 64], height [65], smoking behaviour [66, 67], milk intake [68, 69], physical activity [70, 50, 71], lipids [52, 53, 54], metabolites [72], groups of proteins, e.g. adipokines [73], cytokines [74], inflammatory markers [75], metalloproteinases [76] and other circulating proteins [77, 78]; additionally, several studies evaluated multiple risk factor groups [44, 79, 80, 71]. Several molecular traits that have not previously been reported in MR studies include betaine, SAT2, PRTN3, PLXNB2, DCBLD2, IDUA. Appendix C provides an extended discussion of MR results of dietary and physical traits.

We presented two case study traits (*childhood body size* and *HDL-C)*, for which we identified potential mediators from MR-EvE data and literature data using our LBD method, aiming to triangulate the evidence from both sources. Two-step MR and sensitivity analyses were used to provide suggestive evidence of mediation through the identified traits. In the *childhood body size* case study, MR-EvE and literature-based mediator searches detected several overlapping traits, many of which have been considered as potential mediators in an earlier hypothesis-driven study [44] (e.g. IGF-1 and bioavailable testosterone), offering proof-of-concept evidence for this approach, in addition to identifying several traits not considered previously (e.g. physical activity and sleep duration). For *HDL-C*, the literature overlap method identified several intermediates that also appeared in the MR-EvE data (e.g. other lipid traits, ApoA, a cathepsin). HDL-C was also found to differentially affect total and bioavailable testosterone, which needs further investigation. Additionally, the entire class of cathepsin enzymes needs to be assessed for their mediating role in future work.

In our study, we took advantage of the pre-generated MR-EvE estimates in EpiGraphDB through the R package (*epigraphdb-r*) [4], and demonstrated how it enables fast hypothesis generation and prioritisation of candidate traits for more targeted investigation. Originally, MR-EvE was developed to provide a causal map of the human phenome, useful for supporting or rejecting specific hypotheses [14]. However, for most traits, patterns of horizontal pleiotropy and issues with invalid instruments were detected [14]. This highlighted the need for rigorous follow-up, including sensitivity analyses, biologically informed instrument selection, and triangulation with other sources of evidence to validate the associations [14, 81]. Hence, we validated our discoveries from MR-EvE using best practices in MR, accompanied by sensitivity analyses, *cis*-instrument extraction for protein traits, and also correction for multiple testing.

The MR-EvE data comes with several other limitations. Firstly, it is restricted to the traits available in OpenGWAS as of 2021, which were analysed in the everything-vs-everything way and subsequently added to EpiGraphDB. Secondly, a large proportion of those GWAS traits were measured in mixed-sex samples,

which suggests that traits with female-specific effects that are relevant for breast cancer risk could have been overlooked in the analysis (e.g. IGF-1 [44], testosterone [61]). Moreover, many known risk factors (from observational or previous MR studies) may not come up in the MR-EvE review for several reasons, (1) some traits are not heritable/instrumentable or not in OpenGWAS, (2) traits were measured in another (larger) cohort, or with additional covariates (e.g. BMI), (3) the filtering and FDR correction steps excluded traits with weaker evidence, which may have been reported by smaller studies (e.g. LDL-C).

Many traits identified using MR-EvE in this work have been published as individual MR studies over the recent years, despite MR-EvE data being available via EpiGraphDB since 2021 and earlier as a standalone resource (generated in 2017 by Hemani *et al* [14]) [82]. This highlights MR-EvE's value in generating hypotheses and guiding analyses, while also stressing the redundancy in the growing number of MR studies, which often report only two-sample MR results without providing further evidence, or attempting to dissect the underlying biological mechanisms [83, 84].

In our study, we address the need for evidence triangulation in MR studies by incorporating literature-mined data to strengthen the evidence for identified relationships and to generate hypotheses for underlying mechanisms. To achieve this, we developed the LBD method (Figure 3) that helps to identify mechanistic pathways or mediators for causal associations by mining literature data in EpiGraphDB. The identified intermediate terms may then be tested using MR (given suitable GWAS data is available). In both presented case studies, a third of the identified intermediates provided testable hypotheses for MR validation. In this work, we limited GWAS data search to OpenGWAS due to practical constraints, but future research could extend this search to other sources (e.g. GWAS catalog [85]) and the newly released molecular datasets (e.g. proteomics data (Olink) [86] and lipids/metabolites data (Nightingale NMR) [87]).

LBD and MR can mutually enhance each other: LBD can complement MR causal inference by identifying potential intermediates, while MR validation of LBD findings helps address challenges in evaluating discoveries and benchmarking the performance of LBD methods [88, 18]. Without proper evaluation, it's difficult to determine the usefulness or practicality of a discovery, which hinders the adoption of LBD in biomedical research [89]. Reviewing LBD discoveries with MR (if the terms are instrumentable and data is available) offers a relatively quick strategy for validation of the feasibility of the identified relationships. Certainly, MR is not a definitive answer for assessing traits' causal associations or validating a potential mediation mechanism. However, applying it to LBD discoveries may help to prioritise the most plausible mechanisms, which could be taken further into more costly and time-consuming validation.

Despite the potential gains of incorporating LBD into the biomedical/ epidemiological research process, the limitations and pitfalls of working with literature-mined data should be noted. Even though literature space extraction from EpiGraphDB is a straightforward process, the underlying raw triples data is noisy and may be difficult to interpret. The extraction of semantic triples from text, mapping terms to UMLS and later to GWAS phenotypes, is challenging and could be imprecise, and therefore this data must be treated with caution. Additionally, it is important to note that the chains of intermediate terms (presented as Sankey diagrams) should not be interpreted as true biological pathways. They simply represent relationships between biological entities mentioned in publications, and therefore direct interactions cannot be assumed. Moreover, the extracted relationships may also be affected by publication bias (i.e. due to abstracts often emphasising the established knowledge). Therefore, any insights drawn from literature-mined data must be traced back to the original publications to verify the validity of extracted relationships before being used as evidence.

5. Conclusion

In this article, we presented a novel and efficient approach to undertake a systematic investigation of breast cancer risk factors by mining EpiGraphDB, representing the largest MR study of breast cancer so far. We also introduced a novel LBD method for identifying potential intermediates or mechanisms between traits of interest. Our data-driven approach highlighted that MR and LBD can be used as complementary methods, providing supporting evidence to each other, as well as guiding the generation of hypotheses and bringing evidence from observational and genetic sources together. However, further validation through sensitivity tests, MVMR, and mediation analyses (where appropriate) is essential, as these are important steps in validation of causal relationships. The risk factors and mediators identified in this work can be followed up in future studies in more detail, seeking evidence from other analyses or sources, such as observational studies or lab-based validation. Lastly, the overall approach can be used to study the aetiology of other diseases.

6. CRediT authorship contribution statement

Marina Vabistsevits - Conceptualization, Formal analysis, Data curation, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing; **Yi Liu** - Supervision, Writing – review and editing, Methodology, Software; **Tim Robinson** - Supervision, Writing – review and editing; **Ben Elsworth** - Conceptualization, Supervision, Methodology, Software; **Tom**

R Gaunt - Conceptualization, Supervision, Methodology, Software, Writing – review and editing, Methodology, Project administration, Funding acquisition

7. Declaration of Competing Interest

T.R.G receives funding from Biogen and GSK for unrelated research.

8. Acknowledgments

M.V. is supported by the University of Bristol Alumni Fund (Professor Sir Eric Thomas Scholarship). T.R. is supported by NIHR Development and Skills Enhancement Award (NIHR 302363) and has received grants to attend educational workshops from Daiichi-Sankyo and Amgen. M.V., T.R., Y.L., T.R.G, work in the Medical Research Council Integrative Epidemiology Unit at the University of Bristol supported by the Medical Research Council (MC_UU_00032/03). This work was also supported by a Cancer Research UK programme grant (the Integrative Cancer Epidemiology Programme) (CC18281/A29019). This study was also supported by the NIHR Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

Appendix A. Literature overlap method details

Appendix A.1. Literature-mined relationships in EpiGraphDB - extended version

EpiGraphDB contains literature-mined relationships between pairs of terms (also known as 'literature triples') that were derived from the published literature. Figure A.5 summarises the tools/databases involved in making literature-mined data available in EpiGraphDB. Literature-mined relationships in EpiGraphDB were originally extracted by the MELODI-Presto tool [38]. The underlying literature data comes from SemMedDB (Semantic MEDLINE Database) [16], a well-established repository of literature-mined semantic triples, i.e. '*subject term 1 – predicate – object term 2*', mined from titles and abstracts of nearly 30 million biomedical articles in PubMed, using SemRep [39, 40]. SemRep is a natural language processing (NLP) system that parses biomedical text using linguistic principles and UMLS (Unified Medical Language System) domain knowledge [90] to extract semantic triples.

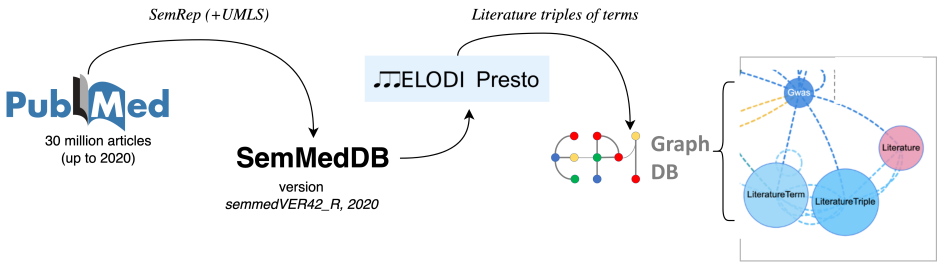


Figure A.5: **A workflow diagram showing the origins of literature data in EpiGraphDB and the tools/databases involved in making this data available.** EpiGraphDB literature term/triples nodes (on the right) is a subset of the full EpiGraphDB schema: `https://www.epigraphdb.org/about`

Appendix A.2. Literature space overlap method - extended version

Developing a method for connecting the literature spaces of exposure and outcome traits to explore potential links between them was one of the key aims of this work. The underlying principle of the literature space overlap approach is linking semantic triples of established biomedical knowledge into sequential 'chains', in order to uncover new connections. This principle of triple linking has previously been showcased in MELODI-Presto [38], and is an extension of the ABC paradigm [41, 43] with the expansion of a chain of Bs (B_1 to B_i): AB_1^iC (see 2.6.2). The details of the method are described below.

Triple cleaning. First, the extracted literature space for a trait is cleaned using several strategies – terms are standardised to minimise redundancy (i.e. merging long/short names, name acronyms, synonyms, gene/protein names); terms with different entity types (e.g. *aapp* - amino acid/protein, *gngm* - gene/genome) are filtered to include only the most common type; triples with the same *term1* and *term2* are excluded; among bidirectionally related terms the ones with more more common direction (indicated by triple score) is retained. Additionally, terms that represented drugs/medications or other diseases (except anthropometrics-related terms) as well as negative predicates (e.g. '*neg coexists with*') are excluded. Finally, a 'frequency score' is generated for each pair of terms, i.e. the sum of triple scores across all possible predicates linking the terms, which is later used for triple chaining visualisation (Sankey diagrams).

Term anchoring. To implement *closed discovery* from the ABC paradigm [43] (see 2.6.2), we perform 'term anchoring', which involves restricting the outer term in a chain of triples to be the trait itself. For example, for an exposure trait (e.g. HDL-C), we are interested in downstream entities, so *term1* in the first triple of a chain make from its literature space would be *'HDL'* (Figure 3b: **"T"** for trait). Inversely, in the breast cancer literature space, we are interested in entities upstream of breast cancer, so we would restrict *term2* of the terminal triple in the chain to be *'Breast cancer'* (Figure 3b: **"BC"** for breast cancer).

For non-molecular traits, it is more difficult to perform term anchoring, because specific terms representing such traits are often not available. Therefore, for those we rely on *open discovery* [43] (see 2.6.2): the spaces are connected by matching unlinked triples from the exposure trait's literature space to any breast cancer triples, and then extending chaining of the matched triples upstream into the trait's space (Figure 3c).

Triple chaining. Literature spaces contain a variety of triples (some examples in Table A.5), and only a small proportion of them contain the trait itself as a term. Using only the triples that include the trait would make the analysis quite limited. Therefore, 'triple chaining' is implemented as a part of the approach to extend the discovery space. Within a trait's literature space, we generate a basic set of triple chains, with the outer triple anchored to the trait term (see above), e.g. "**T**-A-B" and "C-D-E-F-**BC**", where letters A-E are terms forming chains (see Figure 3b). Then, the sets of triples from two literature spaces are "overlapped", which involves *term2* from the exposure trait's space triples to match with *term1* in triples from the breast cancer literature space. This can be done at any possible point, i.e. not necessarily as the longest multiple-step chain (e.g. "T-A-B-C-D-E-F-BC"), but can be of any variation (e.g. "T-A-B-BC", "T-D-E-F-BC").

term1	predicate	term2	triple score	PubMed ID (year)
estradiol	interacts_with	Estrogen Receptors	59	26466867 (2016) / 17975893 (2007) [...]
PTEN	interacts_with	Proto-Oncogene c-akt	21	23650412 (2013) / 27044817 (2016) [...]
estrogens	stimulates	Tamoxifen	18	10070945 (1999) / 28052658 (2017) [...]
Cyclin D1	interacts_with	estrogens	8	29935900 (2017) / 17638066 (2008) [...]
PGR	coexists_with	Estrogen Receptor alpha	4	29403307 (2018) / 26153859 (2015) [...]
CHEK1	interacts_with	BRCA1	2	31789405 (2019) / 12094328 (2002)
O-GlcNAc transferase	interacts_with	Cytoplasmic Protein	2	21255644 (2011) / 23576270 (2013)
polyglutamine	coexists_with	BRCA1	2	23483928 (2013) / 20422428 (2011)
Tretinoin	stimulates	Protein Kinase C	2	15516986 (2004) / 9284950 (1997)
Steroid Hormones	inhibits	SHBG	2	1320387 (1992) / 2173856 (1990)
AP2A1	stimulates	TCEAL1	1	16288208 (2006)
Protein Kinase C	stimulates	CTGF	1	16813525 (2006)
FOXM1	interacts_with	BIRC4	1	26404623 (2015)

Table A.5: **A subset of triples from the breast cancer literature space** (*term1 - predicate - term2*) (randomly selected for illustrative purposes). The triple score indicates the number of publications the triple appears in/was extracted from. The 'PubMed ID (year)' column contains (a subset of) PubMed IDs of those publications. Full data is in Supplementary Table 14.

Literature spaces. The literature overlap method performs best on literature spaces with $>$ 50 triples, as this is required for effective triple chaining. Small literature spaces are also less likely to contain the trait as a term, which is required for term anchoring. Therefore, literature space sizes were also considered when choosing case study traits, to ensure an effective application of the literature overlap method.

Literature spaces made of $<$ 50 triples indicate that fewer publications were found to be linked to them. This may be due to the trait being less studied (and published about), or issues with the trait name (i.e. non-standard phasing of a concept), or issues with the source data in SemMedDB [16] or trait mapping in MELODI-Presto [38] (Figure A.5). Literature spaces of this size are generally unusable in the literature space overlap method.

Appendix A.3. Literature spaces - extended details

Breast cancer. Breast cancer literature space was formed from data linked to two traits: '*Breast cancer*' (OpenGWAS ID: *ieu-a-1126*) and '*Malignant neoplasm of the breast*' (*finn-a-C3_BREAST*). 18,848 unique triples were identified, based on 4989 unique terms derived from 23,809 publications. The clean literature space is available in Supplementary Table 14. The breast cancer anchor term is '*Breast cancer*', which is combined from the terms '*breast diseases*' and '*malignant disease*'.

Case studies. The selected case studies include *HDL* (*'HDL-cholesterol' - ukb-d-30760_irnt*, 7808 unique triples) and *childhood body size* (*'childhood obesity'- ieu-a-1096*, 1255 unique triples) (Supplementary Tables 15-16). For childhood body size, we chose to use *'childhood obesity'* literature space over *'Comparative body size at age 10'* (*ukb-b-4650*, 211 triples) due to the larger space size and trait name specificity.

Appendix B. Sankey diagrams for case studies

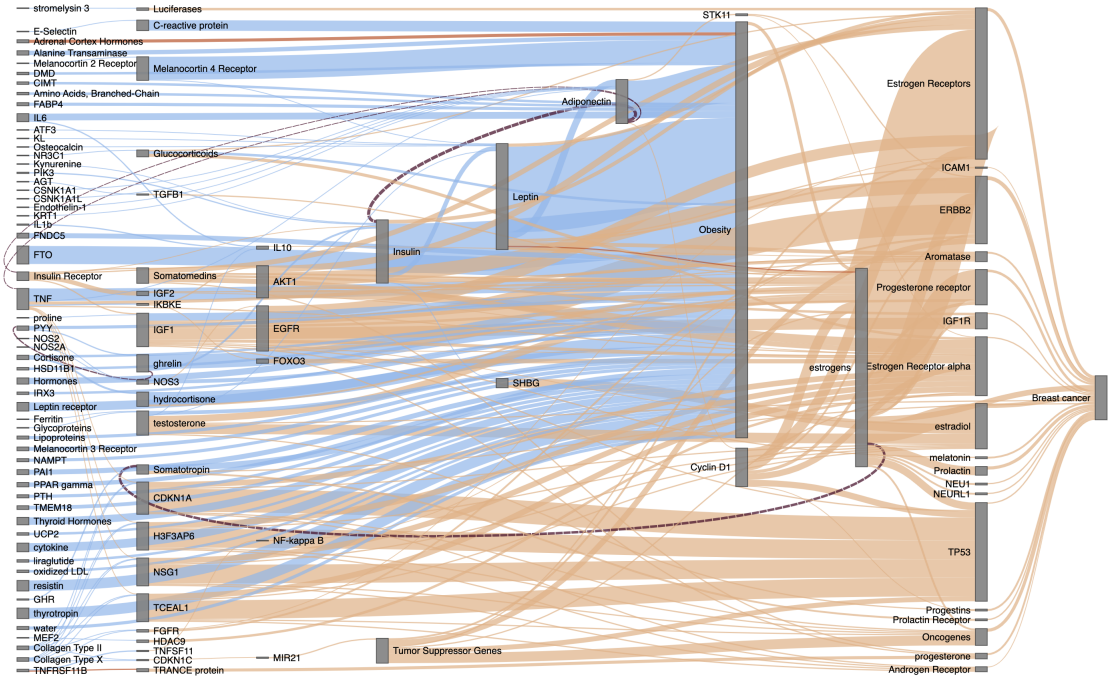


Figure B.6: **Sankey diagram of childhood obesity and breast cancer literature spaces overlap, showing the overlap of literature triples within and between the spaces.** The blue and orange relationships (triples) come from the childhood obesity and breast cancer literature spaces, respectively. The line width of each term relationship indicates how common it is in the literature (frequency score). Here, the open discovery approach was used, i.e. only one side of the overlap is anchored (*'Breast cancer'*); View interactively at `https://mvab.shinyapps.io/literature_overlap_sankey/`

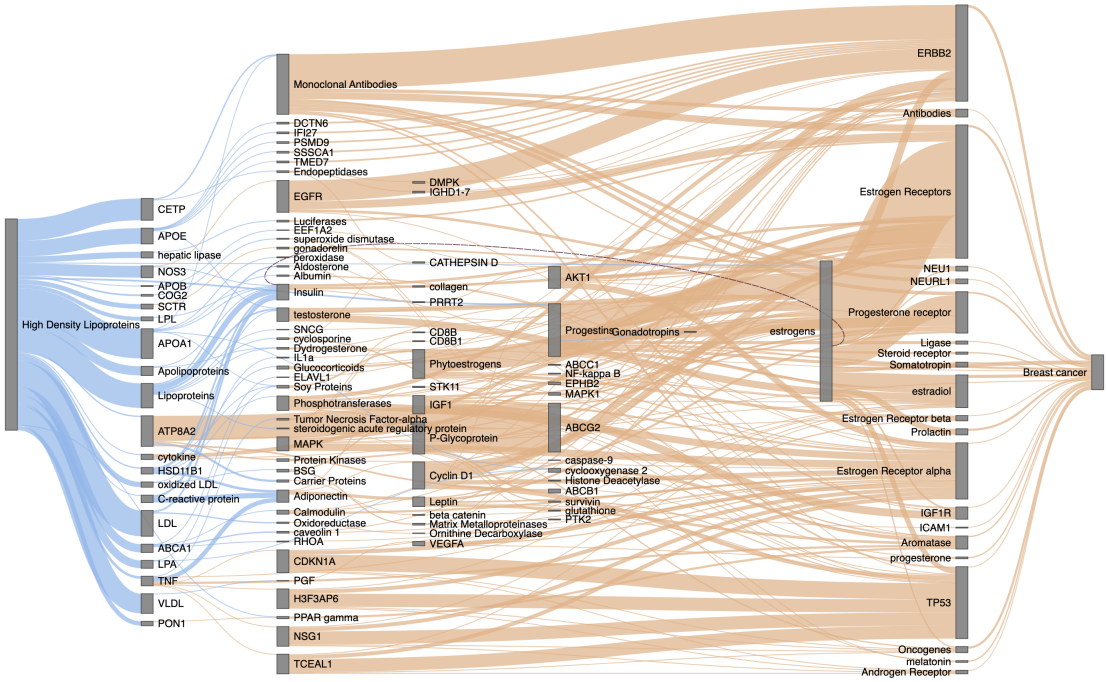


Figure B.7: **Sankey diagram of HDL and breast cancer literature spaces overlap, showing how triples within and between HDL and breast cancer spaces interconnect.** The blue and orange relationships (triples) come from the HDL and breast cancer literature spaces, respectively. The line width of each term relationship indicates how common it is in the literature (frequency score). Here, the closed discovery approach was used; the anchor terms are '*High Density Lipoproteins*' and '*Breast cancer*'. View interactively at `https://mvab.shinyapps.io/literature_overlap_sankey/`

Appendix C. Extended discussion of MR results for dietary and physical traits

The final set of 129 validated breast cancer risk factors identified from MR-EvE included several dietary traits. Genetically-proxied dietary intake exposures require careful consideration due to the distinct sources of bias and interpretation challenges they present. Dietary intake GWAS in UK Biobank can produce a considerable number of instruments due to their sample size, and the risk of horizontal pleiotropy among these SNPs is high [91]. Trait heterogeneity is also important to consider [92], i.e. the fact that genetic instruments associated with the dietary trait may represent different 'dimensions' of the same trait, e.g. intake of a certain food item and metabolism of its core component (e.g. coffee/caffeine [93]). The primary risk in over-interpreting dietary intake MR lies in the possibility that the

extracted instruments may be non-informative or associated with dietary patterns through socio-demographic and behavioural factors, potentially confounding the MR results [91]. Additionally, dietary trait instruments often include adiposity-related signals, meaning these variants could be affecting dietary patterns via their association with adiposity, leading to this association being captured in the MR results.

In this study, dietary traits with the most consistent evidence of effect were *'dried fruit intake'* (showing a protective effect) and *'processed/sugary cereal intake'* (associated with an increased risk). Despite those MR results showing good instrument strength and no strong indication for pleiotropy but some heterogeneity, we cannot claim that the captured effect is specific to those food items. The traits could be proxying two different environmental exposures, e.g. related to a broader class of foods that share similar characteristics, or overall dietary pattern, i.e. 'healthy' and 'unhealthy' food choices. The "gene-environment equivalence" assumption in MR states that the downstream effect of exposure modification is the same, regardless of it being genetically or non-genetically triggered. It could be hypothesised that the personal preference/choice to consume dried fruit vs sugary/processed cereal could be reflective of the other dietary patterns in those individuals' respective diets. Overall, the effect direction matches with the observational evidence of diet impact on breast cancer risk [94]. The 'healthy' dietary pattern includes a varied diet with high fruit and vegetable intake, which increases the overall fibre and micronutrient content in the diet (more likely to occur in those consuming dried fruit), while the 'unhealthy' pattern involves a larger amount of processed and less nutrient-rich food items (more likely to occur in those choosing to eat sweetened breakfast cereals). The genetic correlation analysis in the study by Pirastu *et al* [95] has shown a positive correlation of 'dried fruit intake' GWAS with other traits relating to a healthier diet (e.g. intake of fresh fruit, cooked veg, fish, salad, being vegetarian) and a negative correlation with dietary patterns considered less healthy (e.g. alcohol intake, processed meat, beef, spread on bread, salt, instant coffee), supporting the idea that *'dried fruit intake*' instruments may be proxying an overall healthier diet. The protective effect of '*dried fruit intake*' on breast cancer discovered from MR-EvE has been recently published in a separate MR study [96], in which the authors also reviewed the most likely confounders of this relationship (e.g. vitamin C, BMI, years of education) using MVMR, and found that the identified inverse relationship is not affected when adjusted for any of them.

Skimmed milk intake was another dietary trait identified from MR-EvE with evidence of a protective effect on breast cancer, but this result has to be carefully considered as it is misleading. Previous MR studies of milk intake effect [68, 69] selected instruments (nSNP=1) specifically within the lactase gene locus (*LCT*),

which encodes an enzyme that ensures tolerance of milk products. The variation in this locus affects the amount of milk consumed and, therefore, is often used as a proxy for this exposure. Both MR studies found a mild risk-increasing effect from milk intake. In MR-EvE data, the instruments for *'skimmed milk intake'* trait were extracted genome-wide (nSNP=3), none of them being in *LCT*. MR sensitivity tests identified likely horizontal pleiotropy and heterogeneity. This serves as a valuable negative example of dietary traits evaluation and highlights the importance of conducting more detailed investigations of these traits in individual studies.

Physical activity is one of the modifiable lifestyle factors that have been shown to have a protective effect on breast cancer risk, particularly vigorous activity [97]. MR studies using accelerometer-measured overall physical activity produced results to support this [70, 50]. Among the physical activity traits available in MR-EvE (UK Biobank traits based on a survey question *"Types of physical activity in last 4 weeks"*), several showed similar protective effects: *'Heavy DIY', 'strenuous sports'* and *'exercise to keep fit'* (the last identified in a case study), which may be proxying moderate-to-vigorous physical activity. Similarly, there was a risk-increasing effect from *'no physical activity'* trait, which is in agreement with 'sedentary time' found to increase the risk in [50] and the observationally-known detrimental effects of physical inactivity [98]. With survey-derived physical activity phenotypes (such as *'Heavy DIY'*) appearing unreliable, a potential follow-up to these findings would be to compare the genetic instruments identified for physical activity from survey phenotypes to those derived from accelerometer-measured data.

References

[1] D. A. Lawlor, K. Tilling, G. D. Smith, Triangulation in aetiological epidemiology, International Journal of Epidemiology 45 (6) (2017) 1866–1886. doi:10.1093/ije/dyw314.
URL https://academic.oup.com/ije/article/45/6/1866/2930550

[2] M. R. Munafò, G. Davey Smith, Robust research needs many lines of evidence (1 2018). doi:10.1038/d41586-018-01023-3.
URL https://www.nature.com/articles/d41586-018-01023-3

[3] M. R. Munafò, J. P. Higgins, G. D. Smith, Triangulating evidence through the inclusion of genetically informed designs, Cold Spring Harbor Perspectives in Medicine 11 (8) (2021). doi:10.1101/cshperspect.a040659.
URL https://pubmed.ncbi.nlm.nih.gov/33355252/

[4] Y. Liu, B. Elsworth, P. Erola, V. Haberland, G. Hemani, M. Lyon, J. Zheng, O. Lloyd, M. Vabistsevits, T. R. Gaunt, EpiGraphDB: A database and data mining platform for health data science, Bioinformatics 37 (9) (2021) 1304–1311. doi:10.1093/bioinformatics/btaa961.
URL https://doi.org/10.1101/2020.08.01.230193

[5] D. Ochoa, A. Hercules, M. Carmona, D. Suveges, A. Gonzalez-Uriarte, C. Malangone, A. Miranda, L. Fumis, D. Carvalho-Silva, M. Spitzer, J. Baker, J. Ferrer, A. Raies, O. Razuvayevskaya, A. Faulconbridge, E. Petsalaki, P. Mutowo, S. MacHlitt-Northen, G. Peat, E. McAuley, C. K. Ong, E. Mountjoy, M. Ghoussaini, A. Pierleoni, E. Papa, M. Pignatelli, G. Koscielny, M. Karim, J. Schwartzentruber, D. G. Hulcoop, I. Dunham, E. M. McDonagh, Open Targets Platform: Supporting systematic drug-target identification and prioritisation, Nucleic Acids Research 49 (D1) (2021) D1302–D1310. doi:10.1093/nar/gkaa1027.
URL https://academic.oup.com/nar/article/49/D1/D1302/5983621

[6] A. Struck, B. Walsh, A. Buchanan, J. A. Lee, R. Spangler, J. M. Stuart, K. Ellrott, Exploring Integrative Analysis Using the BioMedical Evidence Graph, JCO Clinical Cancer Informatics 4 (2020) 147–159. doi:10.1200/cci.19.00110.
URL https://ascopubs.org/doi/full/10.1200/CCI.19.00110

[7] E. A. Coker, C. Mitsopoulos, J. E. Tym, A. Komianou, C. Kannas, P. Di Micco, E. Villasclaras Fernandez, B. Ozer, A. A. Antolin, P. Workman, B. Al-Lazikani, CanSAR: Update to the cancer translational research and drug discovery knowledgebase, Nucleic Acids Research 47 (D1) (2019) D917–D922. doi:10.1093/nar/gky1129.
URL https://pubmed.ncbi.nlm.nih.gov/33219674/

[8] S. Ebrahim, G. Davey Smith, Mendelian randomization: Can genetic epidemiology help redress the failures of observational epidemiology?, International Journal of Epidemiology 32 (1) (2003) 1–22. doi:10.1007/s00439-007-0448-6.
URL https://link.springer.com/article/10.1007/s00439-007-0448-6

[9] E. Sanderson, M. M. Glymour, M. V. Holmes, H. Kang, J. Morrison, M. R. Munafò, T. Palmer, C. M. Schooling, C. Wallace, Q. Zhao, G. Davey Smith, Mendelian randomization, Nature Reviews Methods Primers 2 (1) (12 2022).

doi:10.1038/s43586-021-00092-5.
URL https://www.nature.com/articles/s43586-021-00092-5

[10] J. Xu, M. Li, Y. Gao, M. Liu, S. Shi, J. Shi, K. Yang, Z. Zhou, J. Tian, Using Mendelian randomization as the cornerstone for causal inference in epidemiology, Environmental Science and Pollution Research 29 (4) (2022) 5827–5839. doi:10.1007/s11356-021-15939-3.
URL https://pubmed.ncbi.nlm.nih.gov/34431050/

[11] B. Elsworth, M. Lyon, T. Alexander, Y. Liu, P. Matthews, J. Hallett, P. Bates, T. Palmer, V. Haberland, G. Davey Smith, J. Zheng, P. Haycock, T. R. Gaunt, G. Hemani, B. Elsworth, M. Lyon, T. Alexander, Y. Liu, P. Matthews, J. Hallett, P. Bates, T. Palmer, V. Haberland, G. D. Smith, J. Zheng, P. Haycock, T. R. Gaunt, G. Hemani, The MRC IEU OpenGWAS data infrastructure, bioRxiv (8 2020). doi:10.1101/2020.08.10.244293.
URL https://doi.org/10.1101/2020.08.10.244293

[12] E. Sollis, A. Mosaku, A. Abid, A. Buniello, M. Cerezo, L. Gil, T. Groza, O. Güneş, P. Hall, J. Hayhurst, A. Ibrahim, Y. Ji, S. John, E. Lewis, J. A. Macarthur, A. Mcmahon, D. Osumi-Sutherland, K. Panoutsopoulou, Z. Pendlington, S. Ramachandran, R. Stefancsik, J. Stewart, P. Whetzel, R. Wilson, L. Hindorff, F. Cunningham, S. A. Lambert, M. Inouye, H. Parkinson, L. W. Harris, The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource, Nucleic Acids Research 51 (D1) (2023) D977–D985. doi:10.1093/NAR/GKAC1010.
URL https://dx.doi.org/10.1093/nar/gkac1010

[13] G. Hemani, J. Zheng, B. Elsworth, K. H. Wade, V. Haberland, D. Baird, C. Laurin, S. Burgess, J. Bowden, R. Langdon, V. Y. Tan, J. Yarmolinsky, H. A. Shihab, N. J. Timpson, D. M. Evans, C. Relton, R. M. Martin, G. Davey Smith, T. R. Gaunt, P. C. Haycock, The MR-base platform supports systematic causal inference across the human phenome, eLife 7 (5 2018). doi:10.7554/eLife.34408.
URL https://pubmed.ncbi.nlm.nih.gov/29846171/

[14] G. Hemani, J. Bowden, P. Haycock, J. Zheng, O. Davis, P. Flach, T. Gaunt, G. D. Smith, Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome, bioRxiv (10) (2017) 173682. doi:10.1101/173682.
URL https://www.biorxiv.org/content/10.1101/173682v2

[15] L. J. Jensen, J. Saric, P. Bork, Literature mining for the biologist: From information retrieval to biological discovery (2006). doi:10.1038/nrg1768.
URL www.nature.com/reviews/genetics

[16] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, T. C. Rindflesch, SemMedDB: A PubMed-scale repository of biomedical semantic predications, Bioinformatics 28 (23) (2012) 3158–3160. doi:10.1093/bioinformatics/bts591.
URL https://pubmed.ncbi.nlm.nih.gov/23044550/

[17] T. Bekhuis, Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy, Biomedical Digital Libraries 3 (1) (2006) 1–7. doi:10.1186/1742-5581-3-2/METRICS.
URL https://bio-diglib.biomedcentral.com/articles/10.1186/1742-5581-3-2

[18] G. Crichton, S. Baker, Y. Guo, A. Korhonen, Neural networks for open and closed Literature-based Discovery, PLoS ONE 15 (5) (2020) e0232891. doi:10.1371/journal.pone.0232891.
URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0232891

[19] WCRF, World Cancer Research Fund/American Institute for Cancer Research. Continuous Update Project. Diet, Nutrition, Physical Activity and the Prevention of Cancer. Summary of evidence. (2018).
URL www.wcrf.org/matrix

[20] IARC, International Agency for Research on Cancer: Estimated cumulative risk of incidence in 2020, in females, in high-income countries, by cancer site; based on GLOBOSCAN 2020 data (2021).
URL https://gco.iarc.fr/today/online-analysis-multi-bars?v=2020&mode=cancer&mode_population=countries&population=900&populations=986&key=cum_risk&sex=2&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=14&nb_item

[21] K. L. Britt, J. Cuzick, K. A. Phillips, Key steps for effective breast cancer prevention, Nature Reviews Cancer (2020). doi:10.1038/s41568-020-0266-x.
URL http://dx.doi.org/10.1038/s41568-020-0266-x

[22] N. Harbeck, F. Penault-Llorca, J. Cortes, M. Gnant, N. Houssami, P. Poortmans, K. Ruddy, J. Tsang, F. Cardoso, Breast cancer, Nature Reviews Disease Primers 5 (1) (2019). doi:10.1038/s41572-019-0111-2.
URL https://pubmed.ncbi.nlm.nih.gov/31548545/

[23] C. L. Relton, G. Davey Smith, Two-step epigenetic mendelian randomization: A strategy for establishing the causal role of epigenetic processes in pathways to disease, International Journal of Epidemiology 41 (1) (2012) 161–176. doi:10.1093/ije/dyr233.
URL https://pubmed.ncbi.nlm.nih.gov/22422451/

[24] J. Zheng, D. Baird, M.-C. Borges, J. Bowden, G. Hemani, P. Haycock, D. M. Evans, G. D. Smith, Recent Developments in Mendelian Randomization Studies, Current Epidemiology Reports 4 (4) (2017) 330–345. doi:10.1007/s40471-017-0128-6.
URL https://doi.org/10.1007/s40471-017-0128-6

[25] K. Michailidou, S. Lindström, J. Dennis, J. Beesley, S. Hui, S. Kar, A. Lemaçon, P. Soucy, D. Glubb, A. Rostamianfar, M. K. Bolla, Q. Wang, J. Tyrer, E. Dicks, A. Lee, Z. Wang, J. Allen, R. Keeman, U. Eilber, J. D. French, X. Qing Chen, L. Fachal, K. McCue, A. E. McCart Reed, M. Ghoussaini, J. S. Carroll, X. Jiang, H. Finucane, M. Adams, M. A. Adank, H. Ahsan, K. Aittomäki, H. Anton-Culver, N. N. Antonenkova, V. Arndt, K. J. Aronson, B. Arun, P. L. Auer, F. Bacot, M. Barrdahl, C. Baynes, M. W. Beckmann, S. Behrens, J. Benitez, M. Bermisheva, L. Bernstein, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, B. Bonanni, A.-L. Børresen-Dale, J. S. Brand, H. Brauch, P. Brennan, H. Brenner, L. Brinton, P. Broberg, I. W. Brock, A. Broeks, A. Brooks-Wilson, S. Y. Brucker, T. Brüning, B. Burwinkel, K. Butterbach, Q. Cai, H. Cai, T. Caldés, F. Canzian, A. Carracedo, B. D. Carter, J. E. Castelao, T. L. Chan, T.-Y. David Cheng, K. Seng Chia, J.-Y. Choi, H. Christiansen, C. L. Clarke, M. Collée, D. M. Conroy, E. Cordina-Duverger, S. Cornelissen, D. G. Cox, A. Cox, S. S. Cross, J. M. Cunningham, K. Czene, M. B. Daly, P. Devilee, K. F. Doheny, T. Dörk, I. dos Santos-Silva, M. Dumont, L. Durcan, M. Dwek, D. M. Eccles, A. B. Ekici, A. Heather Eliassen, C. Ellberg, M. Elvira, C. Engel, M. Eriksson, P. A. Fasching, J. Figueroa, D. Flesch-Janys, O. Fletcher, H. Flyger, L. Fritschi, V. Gaborieau, M. Gabrielson, M. Gago-Dominguez, Y.-T. Gao, S. M. Gapstur, J. A. García-Sáenz, M. M. Gaudet, V. Georgoulias, G. G. Giles, G. Glendon, M. S. Goldberg, D. E. Goldgar, A. González-Neira, G. I. Grenaker Alnæs, M. Grip, J. Gronwald, A. Grundy, P. Guénel, L. Haeberle, E. Hahnen, C. A. Haiman, N. Håkansson, U. Hamann, N. Hamel, S. Hankin-

936 son, P. Harrington, S. N. Hart, J. M. Hartikainen, M. Hartman, A. Hein,
937 J. Heyworth, B. Hicks, P. Hillemanns, D. N. Ho, A. Hollestelle, M. J.
938 Hooning, R. N. Hoover, J. L. Hopper, M.-F. Hou, C.-N. Hsiung, G. Huang,
939 K. Humphreys, J. Ishiguro, H. Ito, M. Iwasaki, H. Iwata, A. Jakubowska,
940 W. Janni, E. M. John, N. Johnson, K. Jones, M. Jones, A. Jukkola-Vuorinen,
941 R. Kaaks, M. Kabisch, K. Kaczmarek, D. Kang, Y. Kasuga, M. J. Kerin,
942 S. Khan, E. Khusnutdinova, J. I. Kiiski, S.-W. Kim, J. A. Knight, V.-M.
943 Kosma, V. N. Kristensen, U. Krüger, A. Kwong, D. Lambrechts, L. Le Marc-
944 hand, E. Lee, M. Hyuk Lee, J. Won Lee, C. Neng Lee, F. Lejbkowicz,
945 J. Li, J. Lilyquist, A. Lindblom, J. Lissowska, W.-Y. Lo, S. Loibl, J. Long,
946 A. Lophatananon, J. Lubinski, C. Luccarini, M. P. Lux, E. S. K Ma, R. J.
947 MacInnis, T. Maishman, E. Makalic, K. E. Malone, I. Maleva Kostovska,
948 A. Mannermaa, S. Manoukian, J. E. Manson, S. Margolin, S. Mariapun,
949 M. Elena Martinez, K. Matsuo, D. Mavroudis, J. McKay, C. McLean,
950 H. Meijers-Heijboer, A. Meindl, P. Menéndez, U. Menon, J. Meyer, H. Miao,
951 N. Miller, N. Aishah Mohd Taib, K. Muir, A. Marie Mulligan, C. Mulot,
952 S. L. Neuhausen, H. Nevanlinna, P. Neven, S. F. Nielsen, D.-Y. Noh, B. G.
953 Nordestgaard, A. Norman, O. I. Olopade, J. E. Olson, H. Olsson, C. Olswold,
954 N. Orr, V. Shane Pankratz, S. K. Park, T.-W. Park-Simon, R. Lloyd, J. I.
955 A Perez, P. Peterlongo, J. Peto, K.-A. Phillips, M. Pinchev, D. Plaseska-
956 Karanfilska, R. Prentice, N. Presneau, D. Prokofyeva, E. Pugh, K. Pylkäs,
957 B. Rack, P. Radice, N. Rahman, G. Rennert, H. S. Rennert, V. Rhenius,
958 A. Romero, J. Romm, K. J. Ruddy, T. Rüdiger, A. Rudolph, M. Ruebner,
959 E. J. T Rutgers, E. Saloustros, D. P. Sandler, S. Sangrajrang, E. J. Sawyer,
960 D. F. Schmidt, R. K. Schmutzler, A. Schneeweiss, M. J. Schoemaker, F. Schu-
961 macher, P. Schürmann, R. J. Scott, C. Scott, S. Seal, C. Seynaeve, M. Shah,
962 P. Sharma, C.-Y. Shen, G. Sheng, M. E. Sherman, M. J. Shrubsole, X.-O. Shu,
963 A. Smeets, C. Sohn, M. C. Southey, J. J. Spinelli, C. Stegmaier, S. Stewart-
964 Brown, J. Stone, D. O. Stram, H. Surowy, A. Swerdlow, R. Tamimi, J. A.
965 Taylor, M. Tengström, S. H. Teo, M. Beth Terry, D. C. Tessier, S. Thana-
966 sitthichai, K. Thöne, R. A. E M Tollenaar, I. Tomlinson, L. Tong, D. Tor-
967 res, T. Truong, C.-C. Tseng, S. Tsugane, H.-U. Ulmer, G. Ursin, M. Untch,
968 C. Vachon, C. J. van Asperen, D. Van Den Berg, A. M. W van den Ouweland,
969 L. van der Kolk, R. B. van der Luijt, D. Vincent, J. Vollenweider, Q. Waisfisz,
970 S. Wang-Gohrke, C. R. Weinberg, C. Wendt, A. S. Whittemore, H. Wildiers,
971 W. Willett, R. Winqvist, A. Wolk, A. H. Wu, L. Xia, T. Yamaji, X. R. Yang,
972 C. Har Yip, K.-Y. Yoo, J.-C. Yu, W. Zheng, Y. Zheng, B. Zhu, A. Ziogas,
973 E. Ziv, S. R. Lakhani, A. C. Antoniou, A. Droit, I. L. Andrulis, C. I. Amos,
974 F. J. Couch, P. D. P Pharoah, J. Chang-Claude, P. Hall, D. J. Hunter, R. L.
975 Milne, M. García-Closas, M. K. Schmidt, S. J. Chanock, A. M. Dunning,

S. L. Edwards, G. D. Bader, G. Chenevix-Trench, J. Simard, P. Kraft, D. F. Easton, Association analysis identifies 65 new breast cancer risk loci, Nature (2017). doi:10.1038/nature24284.
URL www.nature.com/reprints.

[26] H. Zhang, T. U. Ahearn, J. Lecarpentier, D. Barnes, J. Beesley, G. Qi, X. Jiang, T. A. O'Mara, N. Zhao, M. K. Bolla, A. M. Dunning, J. Dennis, Q. Wang, Z. A. Ful, K. Aittomäki, I. L. Andrulis, H. Anton-Culver, V. Arndt, K. J. Aronson, B. K. Arun, P. L. Auer, J. Azzollini, D. Barrowdale, H. Becher, M. W. Beckmann, S. Behrens, J. Benitez, M. Bermisheva, K. Bialkowska, A. Blanco, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, B. Bonanni, D. Bondavalli, A. Borg, H. Brauch, H. Brenner, I. Briceno, A. Broeks, S. Y. Brucker, T. Brüning, B. Burwinkel, S. S. Buys, H. Byers, T. Caldés, M. A. Caligo, M. Calvello, D. Campa, J. E. Castelao, J. Chang-Claude, S. J. Chanock, M. Christiaens, H. Christiansen, W. K. Chung, K. B. Claes, C. L. Clarke, S. Cornelissen, F. J. Couch, A. Cox, S. S. Cross, K. Czene, M. B. Daly, P. Devilee, O. Diez, S. M. Domchek, T. Dörk, M. Dwek, D. M. Eccles, A. B. Ekici, D. G. Evans, P. A. Fasching, J. Figueroa, L. Foretova, F. Fostira, E. Friedman, D. Frost, M. Gago-Dominguez, S. M. Gapstur, J. Garber, J. A. García-Sáenz, M. M. Gaudet, S. A. Gayther, G. G. Giles, A. K. Godwin, M. S. Goldberg, D. E. Goldgar, A. González-Neira, M. H. Greene, J. Gronwald, P. Guénel, L. Häberle, E. Hahnen, C. A. Haiman, C. R. Hake, P. Hall, U. Hamann, E. F. Harkness, B. A. Heemskerk-Gerritsen, P. Hillemanns, F. B. Hogervorst, B. Holleczek, A. Hollestelle, M. J. Hooning, R. N. Hoover, J. L. Hopper, A. Howell, H. Huebner, P. J. Hulick, E. N. Imyanitov, C. Isaacs, L. Izatt, A. Jager, M. Jakimovska, A. Jakubowska, P. James, R. Janavicius, W. Janni, E. M. John, M. E. Jones, A. Jung, R. Kaaks, P. M. Kapoor, B. Y. Karlan, R. Keeman, S. Khan, E. Khusnutdinova, C. M. Kitahara, Y. D. Ko, I. Konstantopoulou, L. B. Koppert, S. Koutros, V. N. Kristensen, A. V. Laenkholm, D. Lambrechts, S. C. Larsson, P. Laurent-Puig, C. Lazaro, E. Lazarova, F. Lejbkowicz, G. Leslie, F. Lesueur, A. Lindblom, J. Lissowska, W. Y. Lo, J. T. Loud, J. Lubinski, A. Lukomska, R. J. MacInnis, A. Mannermaa, M. Manoochehri, S. Manoukian, S. Margolin, M. E. Martinez, L. Matricardi, L. McGuffog, C. McLean, N. Mebirouk, A. Meindl, U. Menon, A. Miller, E. Mingazheva, M. Montagna, A. M. Mulligan, C. Mulot, T. A. Muranen, K. L. Nathanson, S. L. Neuhausen, H. Nevanlinna, P. Neven, W. G. Newman, F. C. Nielsen, L. Nikitina-Zake, J. Nodora, K. Offit, E. Olah, O. I. Olopade, H. Olsson, N. Orr, L. Papi, J. Papp, T. W. Park-Simon, M. T. Parsons, B. Peissel, A. Peixoto, B. Peshkin, P. Peterlongo, J. Peto, K. A. Phillips, M. Piedmonte, D. Plaseska-Karanfilska, K. Prajzen-

danc, R. Prentice, D. Prokofyeva, B. Rack, P. Radice, S. J. Ramus, J. Rantala, M. U. Rashid, G. Rennert, H. S. Rennert, H. A. Risch, A. Romero, M. A. Rookus, M. Rübner, T. Rüdiger, E. Saloustros, S. Sampson, D. P. Sandler, E. J. Sawyer, M. T. Scheuner, R. K. Schmutzler, A. Schneeweiss, M. J. Schoemaker, B. Schöttker, P. Schürmann, L. Senter, P. Sharma, M. E. Sherman, X. O. Shu, C. F. Singer, S. Smichkoska, P. Soucy, M. C. Southey, J. J. Spinelli, J. Stone, D. Stoppa-Lyonnet, A. J. Swerdlow, C. I. Szabo, R. M. Tamimi, W. J. Tapper, J. A. Taylor, M. R. Teixeira, M. B. Terry, M. Thomassen, D. L. Thull, M. Tischkowitz, A. E. Toland, R. A. Tollenaar, I. Tomlinson, D. Torres, M. A. Troester, T. Truong, N. Tung, M. Untch, C. M. Vachon, A. M. van den Ouweland, L. E. van der Kolk, E. M. van Veen, E. J. vanRensburg, A. Vega, B. Wappenschmidt, C. R. Weinberg, J. N. Weitzel, H. Wildiers, R. Winqvist, A. Wolk, X. R. Yang, D. Yannoukakos, W. Zheng, K. K. Zorn, R. L. Milne, P. Kraft, J. Simard, P. D. Pharoah, K. Michailidou, A. C. Antoniou, M. K. Schmidt, G. Chenevix-Trench, D. F. Easton, N. Chatterjee, M. García-Closas, Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses, Nature Genetics 52 (6) (2020) 572–581. doi:10.1038/s41588-020-0609-2.
URL https://pubmed.ncbi.nlm.nih.gov/32424353/

[27] L. Darrous, N. Mounier, Z. Kutalik, Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics, Nature Communications 2021 12:1 12 (1) (2021) 1–15. doi:10.1038/s41467-021-26970-w.
URL https://www.nature.com/articles/s41467-021-26970-w

[28] S. Burgess, G. D. Smith, N. M. Davies, F. Dudbridge, D. Gill, M. M. Glymour, F. P. Hartwig, M. V. Holmes, C. Minelli, C. L. Relton, E. Theodoratou, G. Davey Smith, N. M. Davies, F. Dudbridge, D. Gill, M. M. Glymour, F. P. Hartwig, M. V. Holmes, C. Minelli, C. L. Relton, E. Theodoratou, Guidelines for performing Mendelian randomization investigations, Wellcome Open Research 4 (2023) 186. doi:10.12688/wellcomeopenres.15555.2.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7384151/

[29] V. W. Skrivankova, R. C. Richmond, B. A. Woolf, N. M. Davies, S. A. Swanson, T. J. Vanderweele, N. J. Timpson, J. P. Higgins, N. Dimou, C. Langenberg, E. W. Loder, R. M. Golub, M. Egger, G. D. Smith, J. B. Richards, Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): explanation and elaboration, BMJ

375 (10 2021). doi:10.1136/BMJ.N2233.
URL https://www.bmj.com/content/375/bmj.n2233

[30] V. W. Skrivankova, R. C. Richmond, B. A. Woolf, J. Yarmolinsky, N. M. Davies, S. A. Swanson, T. J. Vanderweele, J. P. Higgins, N. J. Timpson, N. Dimou, C. Langenberg, R. M. Golub, E. W. Loder, V. Gallo, A. Tybjaerg-Hansen, G. Davey Smith, M. Egger, J. B. Richards, Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization: The STROBE-MR Statement, JAMA 326 (16) (2021) 1614–1621. doi:10.1001/JAMA.2021.18236.
URL https://jamanetwork.com/journals/jama/fullarticle/2785494

[31] A. F. Schmidt, C. Finan, M. Gordillo-Marañón, F. W. Asselbergs, D. F. Freitag, R. S. Patel, B. Tyl, S. Chopade, R. Faraway, M. Zwierzyna, A. D. Hingorani, Genetic drug target validation using Mendelian randomisation, Nature Communications 11 (1) (2020). doi:10.1038/s41467-020-16969-0.
URL http://dx.doi.org/10.1038/s41467-020-16969-0

[32] S. Burgess, A. Butterworth, S. G. Thompson, Mendelian randomization analysis with multiple genetic variants using summarized data, Genetic Epidemiology 37 (7) (2013) 658–665. doi:10.1002/gepi.21758.
URL https://pubmed.ncbi.nlm.nih.gov/24114802/

[33] J. Bowden, G. Davey Smith, S. Burgess, Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression, International Journal of Epidemiology 44 (2) (2015) 512–525. doi:10.1093/ije/dyv080.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4849733/

[34] J. Bowden, G. Davey Smith, P. C. Haycock, S. Burgess, Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator, Genetic Epidemiology 40 (4) (2016) 304–314. doi:10.1002/gepi.21965.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4849733/

[35] J. Bowden, F. Del Greco M, C. Minelli, Q. Zhao, D. A. Lawlor, N. A. Sheehan, J. Thompson, G. Davey Smith, Improving the accuracy of two-sample summary-data Mendelian randomization: Moving beyond the NOME

assumption, International Journal of Epidemiology 48 (3) (2019) 728–742. doi:10.1093/ije/dyy258.
URL https://academic.oup.com/ije/article-abstract/48/3/728/5251908

[36] E. Sanderson, W. Spiller, J. Bowden, Testing and correcting for weak and pleiotropic instruments in two-sample multivariable Mendelian randomization, Stat Med. 40 (25) (2021) 5434–5452.
URL https://pubmed.ncbi.nlm.nih.gov/34338327/

[37] D. G. Altman, P. Royston, The cost of dichotomising continuous variables, British Medical Journal 332 (7549) (2006) 1080. doi:10.1136/bmj.332.7549.1080.

[38] B. Elsworth, T. R. Gaunt, MELODI Presto: A fast and agile tool to explore semantic triples derived from biomedical literature, Bioinformatics 37 (4) (2021) 583–585. doi:10.1093/bioinformatics/btaa726.
URL https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa726/5893950

[39] T. C. Rindflesch, M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, Journal of Biomedical Informatics 36 (6) (2003) 462–477. doi:10.1016/J.JBI.2003.11.003.
URL https://pubmed.ncbi.nlm.nih.gov/14759819/

[40] H. Kilicoglu, G. Rosemblat, M. Fiszman, D. Shin, Broad-coverage biomedical relation extraction with SemRep, BMC Bioinformatics 21 (1) (2020) 1–28. doi:10.1186/S12859-020-3517-7/TABLES/4.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3517-7

[41] N. R. Smalheiser, Rediscovering Don Swanson:The Past, Present and Future of Literature-based Discovery, Journal of Data and Information Science 2 (4) (2017) 43–64. doi:10.1515/JDIS-2017-0019.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5771422/

[42] D. R. Swanson, Fish oil, Raynaud's syndrome, and undiscovered public knowledge., Perspectives in biology and medicine 30 (1) (1986) 7–18. doi:10.1353/pbm.1986.0087.
URL https://pubmed.ncbi.nlm.nih.gov/3797213/

[43] M. Weeber, H. Klein, L. T. De Jong-Van Den Berg, R. Vos, Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries, Journal of the American Society for Information Science and Technology 52 (7) (2001) 548–557. doi:10.1002/asi.1104.
URL https://onlinelibrary.wiley.com/doi/10.1002/asi.1104

[44] M. Vabistsevits, G. Davey Smith, E. Sanderson, T. G. Richardson, B. Lloyd-Lewis, R. C. Richmond, Deciphering how early life adiposity influences breast cancer risk using Mendelian randomization, Communications Biology 5 (1) (2022). doi:10.1038/s42003-022-03272-5.

[45] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, The reactome pathway knowledgebase, Nucleic Acids Research 48 (D1) (2020) D498–D503. doi:10.1093/nar/gkz1031.
URL https://pubmed.ncbi.nlm.nih.gov/31691815/

[46] H. J. Baer, S. S. Tworoger, S. E. Hankinson, W. C. Willett, Body fatness at young ages and risk of breast cancer throughout life, American Journal of Epidemiology 171 (11) (2010) 1183–1194. doi:10.1093/aje/kwq045.
URL https://academic.oup.com/aje/article-abstract/171/11/1183/101419

[47] A. Furer, A. Afek, A. Sommer, L. Keinan-Boker, E. Derazne, Z. Levi, D. Tzur, S. Tiosano, A. Shina, Y. Glick, J. D. Kark, A. Tirosh, G. Twig, Adolescent obesity and midlife cancer risk: a population-based cohort study of 2·3 million adolescents in Israel, The Lancet Diabetes and Endocrinology 8 (3) (2020) 216–225. doi:10.1016/S2213-8587(20)30019-X.
URL http://dx.doi.org/10.1016/S2213-8587(20)30019-X

[48] T. G. Richardson, E. Sanderson, B. Elsworth, K. Tilling, G. Davey Smith, Use of genetic variation to separate the effects of early and later life adiposity on disease risk: Mendelian randomisation study, The BMJ 369 (2020). doi:10.1136/bmj.m1203.
URL http://dx.doi.org/10.1136/bmj.m1203

[49] Y. Hao, J. Xiao, Y. Liang, X. Wu, H. Zhang, C. Xiao, L. Zhang, S. Burgess, N. Wang, X. Zhao, P. Kraft, J. Li, X. Jiang, Reassessing the causal role

of obesity in breast cancer susceptibility: A comprehensive multivariable Mendelian randomization investigating the distribution and timing of exposure, International Journal of Epidemiology 52 (1) (2023) 58–70. doi:10.1093/ije/dyac143.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7614158/

[50] S. C. Dixon-Suen, S. J. Lewis, R. M. Martin, D. R. English, T. Boyle, G. G. Giles, K. Michailidou, M. K. Bolla, Q. Wang, J. Dennis, M. Lush, A. Investigators, T. U. Ahearn, C. B. Ambrosone, I. L. Andrulis, H. Anton-Culver, V. Arndt, K. J. Aronson, A. Augustinsson, P. Auvinen, L. E. Freeman, H. Becher, M. W. Beckmann, S. Behrens, M. Bermisheva, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, B. Bonanni, H. Brenner, T. Brüning, S. S. Buys, N. J. Camp, D. Campa, F. Canzian, J. E. Castelao, M. H. Cessna, J. Chang-Claude, S. J. Chanock, C. L. Clarke, D. M. Conroy, F. J. Couch, A. Cox, S. S. Cross, K. Czene, M. B. Daly, P. Devilee, T. Dörk, M. Dwek, D. M. Eccles, A. H. Eliassen, C. Engel, M. Eriksson, D. G. Evans, P. A. Fasching, O. Fletcher, H. Flyger, L. Fritschi, M. Gabrielson, M. Gago-Dominguez, M. García-Closas, J. A. García-Sáenz, M. S. Goldberg, P. Guénel, M. Gündert, E. Hahnen, C. A. Haiman, L. Häberle, N. Håkansson, P. Hall, U. Hamann, S. N. Hart, M. Harvie, P. Hillemanns, A. Hollestelle, M. J. Hooning, R. Hoppe, J. Hopper, A. Howell, D. J. Hunter, A. Jakubowska, W. Janni, E. M. John, A. Jung, R. Kaaks, R. Keeman, C. M. Kitahara, S. Koutros, P. Kraft, V. N. Kristensen, K. Kubelka-Sabit, A. W. Kurian, J. V. Lacey, D. Lambrechts, L. Le Marchand, A. Lindblom, S. Loibl, J. Lubiński, A. Mannermaa, M. Manoochehri, S. Margolin, M. E. Martinez, D. Mavroudis, U. Menon, A. M. Mulligan, R. A. Murphy, N. Collaborators, H. Nevanlinna, I. Nevelsteen, W. G. Newman, K. Offit, A. F. Olshan, H. Olsson, N. Orr, A. Patel, J. Peto, D. Plaseska-Karanfilska, N. Presneau, B. Rack, P. Radice, E. Rees-Punia, G. Rennert, H. S. Rennert, A. Romero, E. Saloustros, D. P. Sandler, M. K. Schmidt, R. K. Schmutzler, L. Schwentner, C. Scott, M. Shah, X. O. Shu, J. Simard, M. C. Southey, J. Stone, H. Surowy, A. J. Swerdlow, R. M. Tamimi, W. J. Tapper, J. A. Taylor, M. B. Terry, R. A. Tollenaar, M. A. Troester, T. Truong, M. Untch, C. M. Vachon, V. Joseph, B. Wappenschmidt, C. R. Weinberg, A. Wolk, D. Yannoukakos, W. Zheng, A. Ziogas, A. M. Dunning, P. D. Pharoah, D. F. Easton, R. L. Milne, B. M. Lynch, Physical activity, sedentary time and breast cancer risk: a Mendelian randomisation study, British Journal of Sports Medicine 56 (20) (2022) 1157–1170. doi:10.1136/bjsports-2021-105132.
URL https://bjsm.bmj.com/content/56/20/1157

[51] R. C. Richmond, E. L. Anderson, H. S. Dashti, S. E. Jones, J. M. Lane, L. B. Strand, B. Brumpton, M. K. Rutter, A. R. Wood, K. Straif, C. L. Relton, M. Munafò, T. M. Frayling, R. M. Martin, R. Saxena, M. N. Weedon, D. A. Lawlor, G. D. Smith, Investigating causal relations between sleep traits and risk of breast cancer in women: Mendelian randomisation study, The BMJ 365 (2019). doi:10.1136/bmj.l2327.
URL http://dx.doi.org/10.1136/bmj.l2327

[52] C. Nowak, J. Ärnlöv, A Mendelian randomization study of the effects of blood lipids on breast cancer risk, Nature communications 9 (1) (12 2018). doi:10.1038/S41467-018-06467-9.
URL https://pubmed.ncbi.nlm.nih.gov/30262900/

[53] A. Beeghly-Fadiel, N. K. Khankari, R. J. Delahanty, X. O. Shu, Y. Lu, M. K. Schmidt, M. K. Bolla, K. Michailidou, Q. Wang, J. Dennis, D. Yannoukakos, A. M. Dunning, P. D. Pharoah, G. Chenevix-Trench, R. L. Milne, D. J. Hunter, H. Per, P. Kraft, J. Simard, D. F. Easton, W. Zheng, A Mendelian randomization analysis of circulating lipid traits and breast cancer risk, International Journal of Epidemiology 49 (4) (2020) 1117–1131. doi:10.1093/ije/dyz242.
URL https://pubmed.ncbi.nlm.nih.gov/31872213/

[54] K. E. Johnson, K. M. Siewert, D. Klarin, S. M. Damrauer, K. M. Chang, P. S. Tsao, T. L. Assimes, K. N. Maxwell, B. F. Voight, The relationship between circulating lipids and breast cancer risk: A Mendelian randomization study, PLoS Medicine 17 (9) (2020) 1–21. doi:10.1371/JOURNAL.PMED.1003302.
URL https://pubmed.ncbi.nlm.nih.gov/32915777/

[55] M. Touvier, P. Fassier, M. His, T. Norat, D. S. Chan, J. Blacher, S. Hercberg, P. Galan, N. Druesne-Pecollo, P. Latino-Martel, Cholesterol and breast cancer risk: A systematic review and meta-analysis of prospective studies, British Journal of Nutrition 114 (3) (2015) 347–357. doi:10.1017/S000711451500183X.
URL https://pubmed.ncbi.nlm.nih.gov/26173770/

[56] L. Lindstedt, M. Lee, K. Öörni, D. Brömme, P. T. Kovanen, Cathepsins F and S block HDL3-induced cholesterol efflux from macrophage foam cells, Biochemical and Biophysical Research Communications 312 (4) (2003) 1019–1024. doi:10.1016/j.bbrc.2003.11.020.
URL https://pubmed.ncbi.nlm.nih.gov/14651973/

[57] N. Spielmann, D. M. Mutch, F. Rousseau, F. Tores, J. Hager, S. Bertrais, A. Basdevant, P. Tounian, B. Dubern, P. Galan, K. Clément, Cathepsin S genotypes are associated with Apo-A1 and HDL-cholesterol in lean and obese French populations, Clinical Genetics 74 (2) (2008) 155–163. doi:10.1111/j.1399-0004.2008.01043.x.
URL https://pubmed.ncbi.nlm.nih.gov/18565099/

[58] L. Ren, J. Yi, W. Li, X. Zheng, J. Liu, J. Wang, G. Du, Apolipoproteins and cancer, Cancer Medicine 8 (16) (2019) 7032–7043. doi:10.1002/cam4.2587.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6853823/

[59] O. C. Olson, J. A. Joyce, Cysteine cathepsin proteases: Regulators of cancer progression and therapeutic response, Nature Reviews Cancer 15 (12) (2015) 712–729. doi:10.1038/nrc4027.
URL https://pubmed.ncbi.nlm.nih.gov/26597527/

[60] M. Rudzińska, A. Parodi, S. M. Soond, A. Z. Vinarov, D. O. Korolev, A. O. Morozov, C. Daglioglu, Y. Tutar, A. A. Zamyatnin, The Role of Cysteine Cathepsins in Cancer Progression and Drug Resistance, International Journal of Molecular Sciences 20 (14) (7 2019). doi:10.3390/IJMS20143602.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6678516/

[61] K. S. Ruth, F. R. Day, J. Tyrrell, D. J. Thompson, A. R. Wood, A. Mahajan, R. N. Beaumont, L. Wittemans, S. Martin, A. S. Busch, A. M. Erzurumluoglu, B. Hollis, T. A. O'Mara, M. I. McCarthy, C. Langenberg, D. F. Easton, N. J. Wareham, S. Burgess, A. Murray, K. K. Ong, T. M. Frayling, J. R. Perry, Using human genetics to understand the disease impacts of testosterone in men and women, Nature Medicine 26 (2) (2020) 252–258. doi:10.1038/s41591-020-0751-5.
URL http://www.nature.com/articles/s41591-020-0751-5

[62] Y. Guo, S. Warren Andersen, X. O. Shu, K. Michailidou, M. K. Bolla, Q. Wang, M. Garcia-Closas, R. L. Milne, M. K. Schmidt, J. Chang-Claude, A. Dunning, S. E. Bojesen, H. Ahsan, K. Aittomäki, I. L. Andrulis, H. Anton-Culver, V. Arndt, M. W. Beckmann, A. Beeghly-Fadiel, J. Benitez, N. V. Bogdanova, B. Bonanni, A. L. Børresen-Dale, J. Brand, H. Brauch, H. Brenner, T. Brüning, B. Burwinkel, G. Casey, G. Chenevix-Trench, F. J. Couch, A. Cox, S. S. Cross, K. Czene, P. Devilee, T. Dörk, M. Dumont, P. A.

Fasching, J. Figueroa, D. Flesch-Janys, O. Fletcher, H. Flyger, F. Fostira, M. Gammon, G. G. Giles, P. Guénel, C. A. Haiman, U. Hamann, M. J. Hooning, J. L. Hopper, A. Jakubowska, F. Jasmine, M. Jenkins, E. M. John, N. Johnson, M. E. Jones, M. Kabisch, M. Kibriya, J. A. Knight, L. B. Koppert, V. M. Kosma, V. Kristensen, L. Le Marchand, E. Lee, J. Li, A. Lindblom, R. Luben, J. Lubinski, K. E. Malone, A. Mannermaa, S. Margolin, F. Marme, C. McLean, H. Meijers-Heijboer, A. Meindl, S. L. Neuhausen, H. Nevanlinna, P. Neven, J. E. Olson, J. I. Perez, B. Perkins, P. Peterlongo, K. A. Phillips, K. Pylkäs, A. Rudolph, R. Santella, E. J. Sawyer, R. K. Schmutzler, C. Seynaeve, M. Shah, M. J. Shrubsole, M. C. Southey, A. J. Swerdlow, A. E. Toland, I. Tomlinson, D. Torres, T. Truong, G. Ursin, R. B. Van Der Luijt, S. Verhoef, A. S. Whittemore, R. Winqvist, H. Zhao, S. Zhao, P. Hall, J. Simard, P. Kraft, P. Pharoah, D. Hunter, D. F. Easton, W. Zheng, Genetically Predicted Body Mass Index and Breast Cancer Risk: Mendelian Randomization Analyses of Data from 145,000 Women of European Descent, PLoS Medicine 13 (8) (2016) e1002105. doi:10.1371/journal.pmed.1002105.
URL http://dx.plos.org/10.1371/journal.pmed.1002105

[63] D. Freuer, J. Linseisen, T. A. O'mara, M. Leitzmann, H. Baurecht, S. E. Baumeister, C. Meisinger, Body Fat Distribution and Risk of Breast, Endometrial, and Ovarian Cancer: A Two-Sample Mendelian Randomization Study, Cancers 13 (20) (10 2021). doi:10.3390/CANCERS13205053.
URL https://pubmed.ncbi.nlm.nih.gov/34680200/

[64] B. L. Hayes, T. Robinson, S. Kar, K. S. Ruth, K. K. Tsilidis, T. Frayling, A. Murray, R. M. Martin, D. A. Lawlor, R. C. Richmond, Do sex hormones confound or mediate the effect of chronotype on breast and prostate cancer? A Mendelian randomization study, PLoS Genetics 18 (1) (2022) 1–28. doi:10.1371/journal.pgen.1009887.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8809575

[65] M. Vithayathil, P. Carter, S. Kar, A. M. Mason, S. Burgess, S. C. Larsson, S. Karx, A. M. Mason, S. Burgess, S. C. Larsson, Body size and composition and site-specific cancers in UK Biobank: a Mendelian randomisation study, PLoS Medicine 18 (7) (2021) 1–20. doi:10.1101/2020.02.28.970459.
URL http://dx.doi.org/10.1371/journal.pmed.1003706

[66] N. Dimou, J. Yarmolinsky, E. Bouras, K. K. Tsilidis, R. M. Martin, S. J.

Lewis, I. T. Gram, M. F. Bakker, H. Brenner, J. C. Figueiredo, R. T. Fortner, S. B. Gruber, B. van Guelpen, L. Hsu, R. Kaaks, S. S. Kweon, Y. Lin, N. M. Lindor, P. A. Newcomb, M. J. Sanchez, G. Severi, H. A. Tindle, R. Tumino, E. Weiderpass, M. J. Gunter, N. Murphy, Causal effects of lifetime smoking on breast and colorectal cancer risk: Mendelian randomization study, Cancer Epidemiology Biomarkers and Prevention 30 (5) (2021) 953–964. doi:10.1158/1055-9965.EPI-20-1218/71015/AM/CAUSAL-EFFECTS-OF-LIFETIME-SMOKING-ON-BREAST-AND.
URL https://aacrjournals.org/cebp/article/30/5/953/670780/Causal-Effects-of-Lifetime-Smoking-on-Breast-and

[67] H. A. Park, S. Neumeyer, K. Michailidou, M. K. Bolla, Q. Wang, J. Dennis, T. U. Ahearn, I. L. Andrulis, H. Anton-Culver, N. N. Antonenkova, V. Arndt, K. J. Aronson, A. Augustinsson, A. Baten, L. E. Beane Freeman, H. Becher, M. W. Beckmann, S. Behrens, J. Benitez, M. Bermisheva, N. V. Bogdanova, S. E. Bojesen, H. Brauch, H. Brenner, S. Y. Brucker, B. Burwinkel, D. Campa, F. Canzian, J. E. Castelao, S. J. Chanock, G. Chenevix-Trench, C. L. Clarke, A. L. Børresen-Dale, G. I. Grenaker Alnæs, K. K. Sahlberg, L. Ottestad, R. Kåresen, E. Schlichting, M. M. Holmen, T. Sauer, V. Haakensen, O. Engebråten, B. Naume, A. Fosså, C. E. Kiserud, K. V. Reinertsen, A. Helland, M. Riis, J. Geisler, D. M. Conroy, F. J. Couch, A. Cox, S. S. Cross, K. Czene, M. B. Daly, P. Devilee, T. Dörk, I. dos Santos-Silva, M. Dwek, D. M. Eccles, A. H. Eliassen, C. Engel, M. Eriksson, D. G. Evans, P. A. Fasching, H. Flyger, L. Fritschi, M. García-Closas, J. A. García-Sáenz, M. M. Gaudet, G. G. Giles, G. Glendon, M. S. Goldberg, D. E. Goldgar, A. González-Neira, M. Grip, P. Guénel, E. Hahnen, C. A. Haiman, N. Håkansson, P. Hall, U. Hamann, S. Han, E. F. Harkness, S. N. Hart, W. He, B. A. Heemskerk-Gerritsen, J. L. Hopper, D. J. Hunter, C. Clarke, D. Marsh, R. Scott, R. Baxter, D. Yip, J. Carpenter, A. Davis, N. Pathmanathan, P. Simpson, D. Graham, M. Sachchithananthan, D. Amor, L. Andrews, Y. Antill, R. Balleine, J. Beesley, I. Bennett, M. Bogwitz, L. Botes, M. Brennan, M. Brown, M. Buckley, J. Burke, P. Butow, L. Caldon, I. Campbell, D. Chauhan, M. Chauhan, A. Christian, P. Cohen, A. Colley, A. Crook, J. Cui, M. Cummings, S. J. Dawson, A. DeFazio, M. Delatycki, R. Dickson, J. Dixon, T. Edkins, S. Edwards, G. Farshid, A. Fellows, G. Fenton, M. Field, J. Flanagan, P. Fong, L. Forrest, S. Fox, J. French, M. Friedlander, C. Gaff, M. Gattas, P. George, S. Greening, M. Harris, S. Hart, N. Hayward, C. Hoskins, C. Hunt, P. James, M. Jenkins, A. Kidd, J. Kirk, J. Koehler, J. Kollias, S. Lakhani, M. Lawrence, G. Lindeman, L. Lipton, L. Lobb, G. Mann, D. Marsh, S. A. McLachlan, B. Meiser, S. Nightingale, S. O'Connell, S. O'Sullivan, D. G. Or-

tega, N. Pachter, B. Patterson, A. Pearn, K. Phillips, E. Pieper, E. Rickard, B. Robinson, M. Saleh, E. Salisbury, C. Saunders, J. Saunus, R. Scott, C. Scott, A. Sexton, A. Shelling, P. Simpson, A. Spurdle, J. Taylor, R. Taylor, H. Thorne, A. Trainer, K. Tucker, J. Visvader, L. Walker, R. Williams, I. Winship, M. A. Young, A. Jager, A. Jakubowska, E. M. John, A. Jung, R. Kaaks, P. M. Kapoor, R. Keeman, E. Khusnutdinova, C. M. Kitahara, L. B. Koppert, S. Koutros, V. N. Kristensen, A. W. Kurian, J. Lacey, D. Lambrechts, L. Le Marchand, W. Y. Lo, J. Lubiński, A. Mannermaa, M. Manoochehri, S. Margolin, M. E. Martinez, D. Mavroudis, A. Meindl, U. Menon, R. L. Milne, T. A. Muranen, H. Nevanlinna, W. G. Newman, B. G. Nordestgaard, K. Offit, A. F. Olshan, T. W. Park-Simon, P. Peterlongo, J. Peto, D. Plaseska-Karanfilska, P. Radice, G. Rennert, H. S. Rennert, A. Romero, E. Saloustros, E. J. Sawyer, M. K. Schmidt, R. K. Schmutzler, M. J. Schoemaker, L. Schwentner, C. Scott, M. Shah, X. O. Shu, J. Simard, A. Smeets, M. C. Southey, J. J. Spinelli, V. Stevens, A. J. Swerdlow, R. M. Tamimi, W. J. Tapper, J. A. Taylor, M. B. Terry, I. Tomlinson, M. A. Troester, T. Truong, C. M. Vachon, E. M. van Veen, J. Vijai, S. Wang, C. Wendt, R. Winqvist, A. Wolk, A. Ziogas, A. M. Dunning, P. D. Pharoah, D. F. Easton, W. Zheng, P. Kraft, J. Chang-Claude, Mendelian randomisation study of smoking exposure in relation to breast cancer risk, British journal of cancer 125 (8) (2021) 1135–1145. doi:10.1038/S41416-021-01432-8.
URL https://pubmed.ncbi.nlm.nih.gov/34341517/

[68] S. C. Larsson, A. M. Mason, S. Kar, M. Vithayathil, P. Carter, J. A. Baron, K. Michaëlsson, S. Burgess, Genetically proxied milk consumption and risk of colorectal, bladder, breast, and prostate cancer: a two-sample Mendelian randomization study, BMC Medicine 18 (1) (2020) 1–7. doi:10.1186/S12916-020-01839-9/FIGURES/1.
URL https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-020-01839-9

[69] A. L. Lumsden, A. Mulugeta, E. Hyppönen, Milk consumption and risk of twelve cancers: A large-scale observational and Mendelian randomisation study, Clinical Nutrition 42 (1) (2023) 1–8. doi:10.1016/j.clnu.2022.11.006.
URL http://www.clinicalnutritionjournal.com/article/S0261561422003910/fulltext

[70] N. Papadimitriou, N. Dimou, K. K. Tsilidis, B. Banbury, R. M. Martin, S. J. Lewis, N. Kazmi, T. M. Robinson, D. Albanes, K. Aleksandrova, S. I. Berndt, D. Timothy Bishop, H. Brenner, D. D. Buchanan, B. Bueno-de Mesquita, P. T.

Campbell, S. Castellví-Bel, A. T. Chan, J. Chang-Claude, M. Ellingjord-Dale, J. C. Figueiredo, S. J. Gallinger, G. G. Giles, E. Giovannucci, S. B. Gruber, A. Gsur, J. Hampe, H. Hampel, S. Harlid, T. A. Harrison, M. Hoffmeister, J. L. Hopper, L. Hsu, J. María Huerta, J. R. Huyghe, M. A. Jenkins, T. O. Keku, T. Kühn, C. La Vecchia, L. Le Marchand, C. I. Li, L. Li, A. Lindblom, N. M. Lindor, B. Lynch, S. D. Markowitz, G. Masala, A. M. May, R. Milne, E. Monninkhof, L. Moreno, V. Moreno, P. A. Newcomb, K. Offit, V. Perduca, P. D. Pharoah, E. A. Platz, J. D. Potter, G. Rennert, E. Riboli, M. J. Sánchez, S. L. Schmit, R. E. Schoen, G. Severi, S. Sieri, M. L. Slattery, M. Song, C. M. Tangen, S. N. Thibodeau, R. C. Travis, A. Trichopoulou, C. M. Ulrich, F. J. van Duijnhoven, B. Van Guelpen, P. Vodicka, E. White, A. Wolk, M. O. Woods, A. H. Wu, U. Peters, M. J. Gunter, N. Murphy, Physical activity and risks of breast and colorectal cancer: a Mendelian randomisation analysis, Nature communications 11 (1) (12 2020). doi:10.1038/S41467-020-14389-8.
URL https://pubmed.ncbi.nlm.nih.gov/32001714/

[71] H. Li, L. Hou, Y. Yu, X. Sun, X. Liu, Y. Yu, S. Wu, Y. He, Y. Wu, L. He, F. Xue, Lipids, Anthropometric Measures, Smoking and Physical Activity Mediate the Causal Pathway From Education to Breast Cancer in Women: A Mendelian Randomization Study, Journal of Breast Cancer 24 (6) (2021) 504–519. doi:10.4048/jbc.2021.24.e53.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8724372/

[72] Y. Feng, R. Wang, C. Li, X. Cai, Z. Huo, Z. Liu, F. Ge, C. Huang, Y. Lu, R. Zhong, J. Li, B. Cheng, H. Liang, S. Xiong, X. Mao, Y. Chen, R. Lan, Y. Wen, H. Peng, Y. Jiang, Z. Su, X. Wu, J. He, W. Liang, Causal effects of genetically determined metabolites on cancers included lung, breast, ovarian cancer, and glioma: a Mendelian randomization study, Translational lung cancer research 11 (7) (2022) 1302–1314. doi:10.21037/TLCR-22-34.
URL https://pubmed.ncbi.nlm.nih.gov/35958335/

[73] T. Robinson, R. M. Martin, J. Yarmolinsky, Mendelian randomisation analysis of circulating adipokines and C-reactive protein on breast cancer risk, International journal of cancer 147 (6) (2020) 1597–1603. doi:10.1002/IJC.32947.
URL https://pubmed.ncbi.nlm.nih.gov/32134113/

[74] E. Bouras, V. Karhunen, D. Gill, J. Huang, P. C. Haycock, M. J. Gunter,

M. Johansson, P. Brennan, T. Key, S. J. Lewis, R. M. Martin, N. Murphy, E. A. Platz, R. Travis, J. Yarmolinsky, V. Zuber, P. Martin, M. Katsoulis, H. Freisling, T. H. Nøst, M. B. Schulze, L. Dossus, R. J. Hung, C. I. Amos, A. Ahola-Olli, S. Palaniswamy, M. Männikkö, J. Auvinen, K. H. Herzig, S. Keinänen-Kiukaanniemi, T. Lehtimäki, V. Salomaa, O. Raitakari, M. Salmi, S. Jalkanen, CRUK, CAPS, PEGASUS, M. R. Jarvelin, A. Dehghan, K. K. Tsilidis, Circulating inflammatory cytokines and risk of five cancers: a Mendelian randomization analysis, BMC Medicine 20 (1) (2022) 1–15. doi:10.1186/s12916-021-02193-0.
URL https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-021-02193-0

[75] J. Yarmolinsky, J. W. Robinson, D. Mariosa, V. Karhunen, J. Huang, N. Dimou, N. Murphy, K. Burrows, E. Bouras, K. Smith-Byrne, S. J. Lewis, T. E. Galesloot, L. A. Kiemeney, S. Vermeulen, P. Martin, D. Albanes, L. Hou, P. A. Newcomb, E. White, A. Wolk, A. H. Wu, L. Le Marchand, A. I. Phipps, D. D. Buchanan, M. T. Landi, V. Stevens, Y. Wang, D. Albanes, N. Caporaso, P. Brennan, C. I. Amos, S. Shete, R. J. Hung, H. Bickeböller, A. Risch, R. Houlston, S. Lam, A. Tardon, C. Chen, S. E. Bojesen, M. Johansson, H. E. Wichmann, D. Christiani, G. Rennert, S. Arnold, J. K. Field, L. Le Marchand, O. Melander, H. Brunnström, G. Liu, A. Andrew, H. Shen, S. Zienolddiny, K. Grankvist, M. Johansson, M. D. Teare, Y. C. Hong, J. M. Yuan, P. Lazarus, M. B. Schabath, M. C. Aldrich, R. A. Eeles, C. A. Haiman, Z. Kote-Jarai, F. R. Schumacher, S. Benlloch, A. A. Al Olama, K. R. Muir, S. I. Berndt, D. V. Conti, F. Wiklund, S. Chanock, C. M. Tangen, J. Batra, J. A. Clements, H. Grönberg, N. Pashayan, J. Schleutker, S. J. Weinstein, C. M. West, L. A. Mucci, G. Cancel-Tassin, S. Koutros, K. D. Sørensen, E. M. Grindedal, D. E. Neal, F. C. Hamdy, J. L. Donovan, R. C. Travis, R. J. Hamilton, S. A. Ingles, B. S. Rosenstein, Y. J. Lu, G. G. Giles, R. J. MacInnis, A. S. Kibel, A. Vega, M. Kogevinas, K. L. Penney, J. Y. Park, J. L. Stanfrod, C. Cybulski, B. G. Nordestgaard, S. F. Nielsen, H. Brenner, C. Maier, C. J. Logothetis, E. M. John, M. R. Teixeira, S. L. Neuhausen, K. De Ruyck, A. Razack, L. F. Newcomb, D. Lessel, R. Kaneva, N. Usmani, F. Claessens, P. A. Townsend, J. E. Castelao, M. J. Roobol, F. Menegaux, K. T. Khaw, L. Cannon-Albright, H. Pandha, S. N. Thibodeau, D. J. Hunter, P. Kraft, W. J. Blot, E. Riboli, S. S. Zhao, D. Gill, S. J. Chanock, M. P. Purdue, G. Davey Smith, K. H. Herzig, M. R. Järvelin, C. I. Amos, A. Dehghan, M. J. Gunter, K. K. Tsilidis, R. M. Martin, Association between circulating inflammatory markers and adult cancer risk: a Mendelian randomization analysis, eBioMedicine 100 (2 2024). doi:10.1016/J.EBIOM.2024.104991/ATTACHMENT/

0C4432FD-1E6E-480A-A8CA-D4C73CF8C81C/MMC2.DOCX.
URL http://www.thelancet.com/article/S2352396424000264/fulltext

[76] Z. Zhao, Q. Cao, M. Zhu, C. Wang, X. Lu, Causal relationships between serum matrix metalloproteinases and estrogen receptor-negative breast cancer: a bidirectional mendelian randomization study, Scientific Reports 2023 13:1 13 (1) (2023) 1–25. doi:10.1038/s41598-023-34200-0.
URL https://www.nature.com/articles/s41598-023-34200-0

[77] X. Shu, L. Wu, N. K. Khankari, X. O. Shu, T. J. Wang, K. Michailidou, M. K. Bolla, Q. Wang, J. Dennis, R. L. Milne, M. K. Schmidt, P. D. Pharoah, I. L. Andrulis, D. J. Hunter, J. Simard, D. F. Easton, W. Zheng, Associations of obesity and circulating insulin and glucose with breast cancer risk: A Mendelian randomization analysis, International Journal of Epidemiology 48 (3) (2019) 795–806. doi:10.1093/ije/dyy201.
URL http://ccge.medschl.cam.ac.uk/research/consortia/

[78] K. Smith-Byrne, A. Hedman, M. Dimitriou, T. Desai, A. V. Sokolov, H. B. Schioth, M. Koprulu, M. Pietzner, C. Langenberg, J. Atkins, R. Cortez, J. McKay, P. Brennan, S. Zhou, B. J. Richards, J. Yarmolinsky, R. M. Martin, J. Borlido, X. J. Mu, A. Butterworth, X. Shen, J. Wilson, T. L. Assimes, R. J. Hung, C. Amos, M. Purdue, N. Rothman, S. Chanock, R. C. Travis, M. Johansson, A. Mälarstig, Identifying therapeutic targets for cancer: 2,094 circulating proteins and risk of nine cancers, medRxiv (2023) 2023.05.05.23289547doi:10.1101/2023.05.05.23289547.
URL https://www.medrxiv.org/content/10.1101/2023.05.05.23289547v1

[79] F. Chen, W. Wen, J. Long, X. Shu, Y. Yang, X. Shu, W. Zheng, Mendelian randomization analyses of 23 known and suspected risk factors and biomarkers for breast cancer overall and by molecular subtypes, International Journal of Cancer (4 2022). doi:10.1002/IJC.34026.
URL https://onlinelibrary.wiley.com/doi/10.1002/ijc.34026

[80] M. Escala-Garcia, A. Morra, S. Canisius, J. Chang-Claude, S. Kar, W. Zheng, S. E. Bojesen, D. Easton, P. D. P. Pharoah, M. K. Schmidt, Breast cancer risk factors and their effects on survival: a Mendelian randomisation study, BMC

Medicine 18 (1) (2020) 1–10. doi:10.1186/s12916-020-01797-2.
URL https://doi.org/10.1186/s12916-020-01797-2

[81] D. A. Lawlor, Commentary: Two-sample Mendelian randomization: Opportunities and challenges, International Journal of Epidemiology 45 (3) (2016) 908–915. doi:10.1093/ije/dyw127.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5005949/

[82] M. R. Munafo, J. Brown, M. Hefler, G. Davey Smith, Managing the exponential growth of mendelian randomization studies, Tobacco Control 33 (5) (2024) 559–560. arXiv:https://tobaccocontrol.bmj.com/content/33/5/559.full.pdf, doi:10.1136/tc-2024-058987.
URL https://tobaccocontrol.bmj.com/content/33/5/559

[83] G. D. Smith, S. Ebrahim, Mendelian randomisation at 20 years: how can it avoid hubris, while achieving more?, The Lancet Diabetes and Endocrinology 8587 (23) (2023) 10–12. doi:10.1016/S2213-8587(23)00348-0.
URL https://doi.org/10.1016/S2213-8587(23)00348-0

[84] S. Stender, H. Gellert-Kristensen, G. D. Smith, Reclaiming mendelian randomization from the deluge of papers and misleading findings, Lipids in Health and Disease 23 (1) (2024) 286. doi:10.1186/s12944-024-02284-w.
URL https://doi.org/10.1186/s12944-024-02284-w

[85] A. Buniello, J. A. Macarthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, H. Parkinson, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019, Nucleic Acids Research 47 (D1) (2019) D1005–D1012. doi:10.1093/nar/gky1120.
URL https://pubmed.ncbi.nlm.nih.gov/30445434/

[86] B. B. Sun, J. Chiou, M. Traylor, C. Benner, Y. H. Hsu, T. G. Richardson, P. Surendran, A. Mahajan, C. Robins, S. G. Vasquez-Grinnell, L. Hou, E. M. Kvikstad, O. S. Burren, J. Davitte, K. L. Ferber, C. E. Gillies, A. K. Hedman, S. Hu, T. Lin, R. Mikkilineni, R. K. Pendergrass, C. Pickering, B. Prins, D. Baird, C. Y. Chen, L. D. Ward, A. M. Deaton, S. Welsh, C. M. Willis,

N. Lehner, M. Arnold, M. A. Wörheide, K. Suhre, G. Kastenmüller, A. Sethi, M. Cule, A. Raj, H. M. Kang, L. Burkitt-Gray, E. Melamud, M. H. Black, E. B. Fauman, J. M. Howson, H. M. Kang, M. I. McCarthy, P. Nioi, S. Petrovski, R. A. Scott, E. N. Smith, S. Szalma, D. M. Waterworth, L. J. Mitnaul, J. D. Szustakowski, B. W. Gibson, M. R. Miller, C. D. Whelan, Plasma proteomic associations with genetics and health in the UK Biobank, Nature 622 (7982) (2023) 329–338. doi:10.1038/S41586-023-06592-6.
URL https://pubmed.ncbi.nlm.nih.gov/37794186/

[87] H. Julkunen, A. Cichońska, M. Tiainen, H. Koskela, K. Nybo, V. Mäkelä, J. Nokso-Koivisto, K. Kristiansson, M. Perola, V. Salomaa, P. Jousilahti, A. Lundqvist, A. J. Kangas, P. Soininen, J. C. Barrett, P. Würtz, Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank, Nature Communications 14 (1) (12 2023). doi:10.1038/S41467-023-36231-7.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9898515/

[88] E. Moreau, Literature-based discovery: Addressing the issue of the subpar evaluation methodology, Bioinformatics 39 (2) (2023) 1–3. doi:10.1093/bioinformatics/btad090.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9945845/

[89] E. Moreau, O. Hardiman, M. Heverin, D. O'Sullivan, Mining impactful discoveries from the biomedical literature, bioRxiv (2022) 2022.10.28.514184.
URL https://www.biorxiv.org/content/10.1101/2022.10.28.514184v1

[90] O. Bodenreider, The Unified Medical Language System (UMLS): Integrating biomedical terminology, Nucleic Acids Research 32 (DATABASE ISS.) (2004) 267–270. doi:10.1093/nar/gkh061.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308795/

[91] K. H. Wade, J. Yarmolinsky, E. Giovannucci, S. J. Lewis, I. Y. Millwood, M. R. Munafò, F. Meddens, K. Burrows, J. A. Bell, N. M. Davies, D. Mariosa, N. Kanerva, E. E. Vincent, K. Smith-Byrne, F. Guida, M. J. Gunter, E. Sanderson, F. Dudbridge, S. Burgess, M. C. Cornelis, T. G. Richardson, M. C. Borges, J. Bowden, G. Hemani, Y. Cho, W. Spiller, R. C. Richmond, A. R. Carter, R. Langdon, D. A. Lawlor, R. G. Walters, K. S. Vimaleswaran, A. Anderson, M. R. Sandu, K. Tilling, G. Davey Smith, R. M. Martin, C. L.

Relton, Applying Mendelian randomization to appraise causality in relationships between nutrition and cancer, Cancer Causes & Control 33 (2022). doi:10.1007/s10552-022-01562-1.
URL https://doi.org/10.1007/s10552-022-01562-1

[92] P. C. Haycock, S. Burgess, K. H. Wade, J. Bowden, C. Relton, G. D. Smith, Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies (4 2016). doi:10.3945/ajcn.115.118216.
URL https://pubmed.ncbi.nlm.nih.gov/26961927/

[93] M. C. Cornelis, T. Kacprowski, C. Menni, S. Gustafsson, E. Pivin, J. Adamski, A. Artati, C. B. Eap, G. Ehret, N. Friedrich, A. Ganna, I. Guessous, G. Homuth, L. Lind, P. K. Magnusson, M. Mangino, N. L. Pedersen, M. Pietzner, K. Suhre, H. Völzke, M. Bochud, T. D. Spector, H. J. Grabe, E. Ingelsson, M. Burnier, O. Devuyst, P. Y. Martin, M. Mohaupt, F. Paccaud, A. Pechere-Bertschi, B. Vogt, D. Ackermann, B. Ponte, M. Pruijm, Genome-wide association study of caffeine metabolites provides new insights to caffeine metabolism and dietary caffeine-consumption behavior, Human molecular genetics 25 (24) (2016) 5472–5482. doi:10.1093/HMG/DDW334.
URL https://pubmed.ncbi.nlm.nih.gov/27702941/

[94] S. E. Steck, E. A. Murphy, Dietary patterns and cancer risk, Nature Reviews Cancer 20 (2) (2020) 125–138. doi:10.1038/s41568-019-0227-4.
URL http://dx.doi.org/10.1038/s41568-019-0227-4

[95] N. I. Pirastu, C. McDonnell, E. J. Grzeszkowiak ID, N. I. Mounier, F. ImamuraID, J. MerinoID, F. R. DayID, J. ZhengID, N. TabaID, M. Pina ConcasID, L. RepettoID, K. A. KentistouID, A. RobinoID, T. EskoID, P. K. JoshiID, K. FischerID, K. K. OngID, T. R. GauntID, Z. Kutalik, J. R. B Perry, J. F. WilsonID, Using genetic variation to disentangle the complex relationship between food intake and health outcomes, PLOS Genetics 18 (6) (2022) e1010162. doi:10.1371/JOURNAL.PGEN.1010162.
URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1010162

[96] C. Jin, R. Li, T. Deng, Z. Lin, H. Li, Y. Yang, Q. Su, J. Wang, Y. Yang, J. Wang, G. Chen, Y. Wang, Association between dried fruit intake and pan-cancers incidence risk: A two-sample Mendelian randomization study, Frontiers in Nutrition 9 (July) (2022) 1–14. doi:10.3389/fnut.2022.899137.
URL https://pubmed.ncbi.nlm.nih.gov/30696460/

[97] H. K. Neilson, M. S. Farris, C. R. Stone, M. M. Vaska, D. R. Brenner, C. M. Friedenreich, Moderate-vigorous recreational physical activity and breast cancer risk, stratified by menopause status: A systematic review and meta-analysis (10 2016). doi:10.1097/GME.0000000000000745.
URL https://pubmed.ncbi.nlm.nih.gov/27779567/

[98] F. Chong, Y. Wang, M. Song, Q. Sun, W. Xie, C. Song, Sedentary behavior and risk of breast cancer: a dose-response meta-analysis from prospective studies, Breast cancer (Tokyo, Japan) 28 (1) (2021) 48–59. doi:10.1007/S12282-020-01126-8.
URL https://pubmed.ncbi.nlm.nih.gov/32607943/