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

Marina Vabistsevits[1,2]*, Tim Robinson[1,2], Ben Elsworth[1,2,3], Yi Liu[1,2], Tom Gaunt[1,2]

*[1]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, United Kingdom; [2]Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, United Kingdom; [3] Our Future Health, 2 New Bailey, 6 Stanley Street, Manchester M3 5GS.*

## Abstract

An increasing challenge in population health research is efficiently utilising the wealth of data available from multiple sources to investigate the mechanisms of disease 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 knowledge graph to build a comprehensive overview of risk factors for developing breast cancer.

We utilised MR-EvE ("Everything-vs-Everything") data to generate a list of causal risk factors for breast cancer, integrated this data with literature-mined relationships and identified potential mediators. We used multivariable MR to evaluate mediation and estimate the direct effects of these traits. We identified 213 novel and established lifestyle and molecular traits with evidence of an effect on breast cancer. We present the results of this evidence integration for four case studies (insulin-like growth factor I, cardiotrophin-1, childhood body size and age at menopause).

We demonstrate that using MR-EvE to identify disease risk factors is an efficient hypothesis-generating approach. Moreover, we show that integrating MR evidence with literature-mined data may identify causal intermediates and uncover the mechanisms behind disease.

# 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]–[8] 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) [9], [10] and literature-mined estimates for molecular and lifestyle traits, providing a valuable combination of data to systematically investigate casual 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 [9]. Studying disease causality with MR methods has been growing in popularity over the past decade and a half [10], [11], with the increasing availability of data from genome-wide association studies (GWAS) [12], [13] and well-developed analysis frameworks [14]. EpiGraphDB provides access to MR-EvE (Mendelian randomization "Everything-vs-Everything") pairwise causal estimates [15] between thousands of traits from the OpenGWAS database (https://gwas.mrcieu.ac.uk) [12]. 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 [16]. In addition to MR-EvE, EpiGraphDB also contains literature-mined relationships extracted from publications in PubMed [17] and mapped to specific phenotypes that have also been investigated using GWAS (and are therefore amenable to MR). Literature-based discovery entails connecting information that is explicitly stated in multiple publications across different research disciplines with the goal of deducing connections that have not been explicitly stated [18] [19]. Literature-mined evidence could be a rich resource for identifying new hypotheses or finding supporting evidence for potential causal effects identified in MR-EvE. It could also be used to identify mechanistic pathways or mediators for causal relationships identified using MR.

In this work, we apply evidence integration and triangulation to identify risk factors for breast cancer based on MR-EvE and literature-mined data from EpiGraphDB. Breast cancer is a useful exemplar case as it is a heterogeneous disease with a complex aetiology, affected by both genetic and lifestyle factors [20]–[23] and has a large body of research evidence well-suited for applying data mining methods. Here, we aim to use EpiGraphDB to build a comprehensive picture of the aetiology of breast cancer by combining evidence from MR and literature-mined data. Through this analysis, we seek to generate new hypotheses for causal relationships and gain some insight into their mechanistic detail to inform prevention and intervention strategies that reduce disease burden.

# Results

## Study overview

We present a summary of our study in Figure 1. First, we explored MR-EvE causal estimates in EpiGraphDB between various exposure traits and breast cancer to generate a list of potential causal risk factors and biomarkers. For the identified risk factor traits with validated MR effects, we also queried MR-EvE data to collect potential mediators of the observed effect (i.e. traits that are affected by the risk factor trait and also affect breast cancer). Then, we extracted literature-mined relationships from EpiGraphDB for the identified risk factors, aiming to use this data to assist with dissecting the risk factor/breast cancer relationships. We determined the intermediates between the individual traits and breast cancer from literature-mined data and were then able to test a selection of those as mediators in MR analysis. The potential mediators from MR-EvE and literature-mined data were then triangulated and the effects were validated with a multivariable MR (MVMR) [24]. Finally, we present detailed results for four candidate risk factors as case studies (insulin-like growth factor I (IGF-1), cardiotrophin-1, childhood body size, age at menopause).

**[Figure 1 here]**

## Collecting and validating evidence from MR-EvE

We searched EpiGraphDB for MR-EvE results for the available breast cancer GWAS as outcomes (N=21) (Supplementary Table 1). In total, 2332 traits were linked to breast cancer (with any MR result) (Supplementary Data 1). Of those, we kept the traits with some evidence of effect (defined as effect 95% confidence intervals not overlapping the null) on at least one of the outcomes (1643 traits). This initial filtering step discarded traits with very little evidence, and the rest of the traits were taken forward in the analysis with an acknowledgement that at this threshold, there are likely to be many false positives. At this hypothesis-generation stage we did not perform multiple-testing correction and instead applied filtering steps outlined in Supplementary Figure 1 that helped us to identify traits with stronger evidence.

The exposure traits were grouped into 12 trait categories. These included established cancer risk factor categories (anthropometric traits, reproductive traits, physical activity, smoking, alcohol), categories with a growing research interest (dietary traits, sleep traits, medication), and molecular traits (metabolites, lipids, proteins, and other biomarkers). The traits not falling into these categories were excluded from further analysis, but MR-EvE results for all are available in Supplementary Data 1.

We designed a R/Shiny app to facilitate the exploration of exposures that could be breast cancer risk or protective factors, available at: https://mvab.shinyapps.io/brca-miner/ (Supplementary Figure 2 shows a screenshot of the app for the anthropometric traits category). This exploration of pre-filtered MR-EvE results (905 traits) allowed us to evaluate each trait across multiple breast cancer outcomes and review the consistency and direction of effect. Overall, we found that the results were consistent across outcomes for the majority of exposure traits. Visualising MR-EvE results in this way provided a broad (but not exhaustive) overview of MR evidence of traits affecting breast cancer risk. Many of the identified effects have been reported in previous MR studies (particularly anthropometric, reproductive, alcohol, smoking and sleep traits). It should be noted that the extracted MR estimates were generated with the MR-Mixture-of-Experts method [15] (see Methods) and therefore some estimates may not result from the standard practice used for reporting in MR studies (top hits and inverse-variance weighted (IVW) [25] methods).

Whilst MR-EvE results allow rapid screening for risk factors, all potentially interesting traits should be validated using standard MR approaches with sensitivity analyses prior to reporting [26]. To validate the causal estimates for the traits presented in the app, we focused on the traits with consistent MR results across the three main outcome datasets (*meta-analysis*, *OncoArray, iCOGs 2017* – Supplementary Table 1). We kept the traits with evidence of an effect in the same direction in two out of three datasets (within an outcome subgroup, e.g. ER+ (oestrogen receptor-positive subtype) only). This approach identified 309 traits that had consistent evidence of an effect on breast cancer risk (Table 1, Supplementary Figure 1).

Next, we performed an IVW MR analysis using BCAC 2017 [27] and BCAC 2020 [28] (latest version of the data stratified by molecular subtypes, Supplementary Table 2) breast cancer outcomes, identifying evidence of effect from 213 risk factor traits (Figure 1a). Of these, 171 and 168 traits were found to affect BCAC 2017 and BCAC 2020, respectively (Supplementary Figure 1). The numbers of traits with evidence of an effect on each outcome subtype by trait category are available in Table 1. Our validation results (MR estimates) are available in Supplementary Data 4.

We also performed a range of sensitivity analyses for multi-instrument traits, including additional MR analyses that attempt to address pleiotropy – MR-Egger [29] and weighted median [30], and reviewed the Egger intercept [29], Cochran's Q for heterogeneity [31]. The sensitivity analysis results are available in Supplementary Data 2 and 3, where filters for heterogeneity and potential pleiotropy are provided. We did not use these analyses to make further trait exclusions from the final set of traits, but the results were taken into account when reviewing specific traits. Finally, to correct for multiple testing of many exposures and outcomes we applied the False Discovery Rate (FDR).

The MR results for the final set of traits (213 traits or 105 after FDR correction) with all BCAC breast cancer outcomes are presented in Figure 2. The sub-figures are also available in an interactive format (https://mvab.shinyapps.io/MR_heatmaps/), where hovering over each data point shows the MR estimate details. Figure 2 only includes trait traits that had evidence of effect in at least one outcome; the full results (309 traits) are available in Supplementary Data 2. Abbreviations used in Figure 2 and throughout the next section are defined in Supplementary Data 4.

**[Figure 2 here]**

*Lifestyle traits*

The majority of body size traits in the anthropometric trait group (e.g. *body mass index, obesity, childhood body size at age 10, hip/waist circumference*) (Figure 2a) have consistent protective effects across most breast cancer sybtypes, passing the FDR correction. Interestingly, '*impedance of whole body*' has a causal effect on most outcomes. There is also a risk-increasing effect from various measures of height.

Among the lifestyle traits, *'Type of transport used: cycle'* and *'Physical activity: strenuous sports'* show evidence of a protective effect on some outcomes. *'Skimmed milk intake'* appears to have a small protective effect across all breast cancer subtypes. There is also a protective effect from '*dried fruit intake'* on outcomes with larger sample sizes. '*Poultry intake'* has some evidence of an increased risk on varying breast cancer subtypes, although the confidence intervals are very wide for those estimates. 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 many outcomes. Higher '*age last used HRT'* (hormone-replacement therapy) also has a positive effect on the risk of triple-negative breast cancer (TNBC).

*Molecular traits*

The lipid traits make up a large proportion of measured serum metabolites (Figure 2b) and have a distinct pattern in their effect direction. There is a consistently positive (causal) effect from HDL particles measures and a negative (protective) effect from VLDL particles, largely passing the FDR correction. Very few of the lipid measures have an effect on ER- breast cancer, and evidence of effect is also sparse for HER2-enriched and TNBC (both are ER- subtypes). Interestingly, there is also little evidence of effect observed on Luminal B subtypes (even more consistently than for the ER- subtypes).
Among other metabolites (Figure 2b), 3-dehydrocarnitine, betaine, and X-11440 have a negative effect on the risk, with X-11440 also having a positive effect on HER2-enriched subtype. There is also evidence of a positive effect from 1-arachidonoyglycerophosphoinositol across different subtypes.

For display purposes, we grouped the protein traits (Figure 2c) into major biological pathways identified using the Reactome database [32]. Among the immune system proteins, there is evidence of negative effects from IL3Ra, IL6Ra, IL1RL1, and Filamin-A and positive effects from IL21, ICAM1, and ABO across various outcomes (all abbreviations are available in Supplementary Data 4). Alcohol dehydrogenase, TIE1, DCBLD2, diamine acetyltransferase-2, and CAMK1 have consistent FDR-corrected protective effects in both overall samples of breast cancer outcomes (BCAC 2017 and 2020). Calpastatin has a positive effect on the overall samples, ER+ and Luminal A subtypes. MAN1A2 has a positive effect on the overall samples, ER- and TNBC subtypes. Other proteins that passed FDR correction, but have no subtype-specific pattern are USP25, KMT5C, plexin-B2, C1QTNF5, PR domain zinc finger protein 1, and kallikrein-14.

**[Table 1 here]**

Lastly, we queried MR-EvE data to extract potential mediators of the 171 breast cancer risk factors (BCAC 2017 outcomes only) (Figure 1b) (Supplementary Data 5). We determined that 153 traits have potential mediators, which were manually validated with IVW and sensitivity analyses, as described above for the main exposure/outcome pairs (Supplementary Data 6). The largest number of potential mediators were found for anthropometric traits (mean = 29 in the category) and lipid traits (mean = 31). In other categories, the traits with the highest number of mediators included: skimmed milk intake (42), CUZD1 protein (40), cardiotrophin-1 (31), sleep duration (10), isovalerylcarnitine (13). The full table of counts of potential mediators for each trait is available in Supplementary Data 7.

## Extracting literature-mined relationships

We searched EpiGraphDB to extract literature-mined relationships between pairs of terms related to breast cancer, which together form a breast cancer "literature space" (Figure 1c). These pairs of terms connected by a predicate are also known as "literature triples", and represent semantic relationships between biological entities derived from published literature (e.g. protein A – stimulates – gene B) (see '**Literature-mined relationships in** EpiGraphDB**'** in Methods). The breast cancer literature space of triples (N = 18,848) is available in Supplementary Data 8.

We also extracted literature spaces for the risk factor traits identified after MR evidence validation (Figure 2). After the initial exclusion of traits that are unlikely to produce meaningful results (see '**Extracting literature space for a trait**' in Methods), we queried EpiGraphDB for 154 traits. Table 2 provides a summary of relative literature space sizes (<=5, <=50, >50 triples) by trait category. The literature spaces for available traits are in Supplementary Data 9.

We identified 49 traits with large literature spaces (i.e. > 50 triples), with some traits mapped to over 10,000 triples. These traits include many anthropometric traits (e.g. weight, body mass index, height), molecular traits (e.g. albumin, myeloperoxidase, IGF-1, iron, C-reactive protein, proline, triglycerides) and a few traits in other categories, e.g. medication (aspirin, allopurinol), age at menopause, and sleep duration. The traits with empty or small literature spaces (<=5 triples) were not usable (N=86). The lack of identified triples is likely a result of GWAS trait names being phrased in a non-standard way that differs from the way that the trait is normally referred to in the literature (hence not matching enough results), or could also be due to potential issues with the data mining steps and the source data in SemMedDB [17]. The traits with <= 50 triples (N=19) may be useful, but should be treated with caution, as an overall lower number of publications was identified to be linked to them. Therefore, from this point onwards, we focused on the traits with literature spaces with > 50 triples, as only those contain sufficient amount of data for identifying potential intermediates via literature spaces overlap with breast cancer. Supplementary Data 10 contains the counts of unique triples for all non-empty literature space traits and Supplementary Figure 4 shows the size distribution by categories.

**[Table 2 here]**

The method for literature-based discovery via overlapping literature spaces was devised as a part of this study and is an extension of the principle used in MELODI-Presto [34] and is conceptually based on Swanson linking [18] [35] (see **'Connecting traits via literature spaces'** in Methods for details). We extracted sequentially linked triples from literature spaces, anchoring the outer triples to the terms representing the trait (molecular traits only) and breast cancer outcome, to identify the overlapping terms between the two. This ensured that the trait term is always connected to breast cancer either directly or via chained triples (Figure 3). This allowed us to look for intermediates between a risk factor trait and breast cancer and use this data as another source of evidence for the potential mediators extracted from MR-EvE (Figure 1d). The literature overlap results are available in Supplementary Data 11 and as Sankey diagrams at https://mvab.shinyapps.io/literature_overlap_sankey/ for selected traits reviewed as case studies in the next section.

## MR and literature evidence integration case studies

In this section, we present the results summary for four case study risk factor traits (cardiotrophin-1 (from Sun *et al* 2018 [36]), IGF-1, childhood body size, age at menopause (from UK Biobank [37])), for which we integrated evidence from MR and literature-based discovery to identify potential mediators of their relationship with breast cancer (see '**MR and literature evidence integration** Methods).

The specific traits were selected based on two criteria: (1) evidence of effect in MR analysis with adequate performance in sensitivity analyses (2) suitably large literature space (> 50 triples). To demonstrate the workflow applied in different scenarios, we selected two pairs of lifestyle and molecular risk factors, with one in each pair having a known and/or well-studied effect on breast cancer (positive controls) and the other having an unexplained effect or a poorly understood mechanism (new hypotheses) (Figure 1e). We provide the full list of potential mediators for each trait in Supplementary Table 3, and the full results are available as HTML reports in Supplementary Data 12 or online at https://mvab.github.io/epigraphdb-breast-cancer/.

*Case study 1: Cardiotrophin-1*

Cardiotrophin-1 (CT-1) is an interleukin-6 family cytokine, mostly known for its role in inducing hypertensive heart disease [38]. MR provided some evidence that CT-1 reduces the overall breast cancer risk (OR – BCAC 2017: 0.97 [0.95:0.99], BCAC 2020: 0.98 [0.96:0.99] – odds of breast cancer per standard deviation (SD) increase in CT-1) and the risk of ER+ subtype: (OR 0.97 [0.95:0.99]).

From the MR-EvE mediator analysis, we identified 20 traits that attenuate the effect of CT-1 in MVMR (Supplementary Table 3, HTML report). Among the literature-mined potential intermediates, we identified GWAS data for 13 traits, out of which 2 had some evidence of an effect on breast cancer – leukemia inhibitory factor (LIF) (ER+ sample OR: 1.10 [1.02:1.18]) and LIF receptor (LIFR) (overall OR: 0.87 [0.82:0.91], ER+: 0.87 [0.82:0.92]). In the bidirectional analysis of CT-1 with both traits, we identified that CT-1 had a negative effect on LIF (effect per SD increase: -0.11 [-0.18:-0.04]) with little effect in the other direction, and with LIFR in both directions. In MVMR, the effect of CT-1 attenuated when accounted for LIF (OR: 0.98 [0.95:1.01]).

In the literature overlap data, CT-1 is directly connected to LIFR via ‘interacts with’ and ‘stimulates’ predicates. LIFR is further connected to breast cancer via several triples. The literature term LIF is not directly linked to CT-1, but it is linked to LIFR via other terms, and also the term ‘cytokine’ (CT-1 is a cytokine) is linked to the same terms as LIF.

In summary, there is some evidence that CT-1 is negatively associated with ER+ breast cancer via a negative effect on LIF, which itself separately increases breast cancer risk. This result is based on 3 instruments in CT-1 and 1 instrument in LIF, so further validation is important for these MVMR results.

*Case study 2: IGF-1*

Insulin-like growth factor-1 (IGF-1) has an important role in breast tissue differentiation and mammary gland function, and the molecular mechanism underlying its association with breast cancer has been extensively studied [39]. MR provided evidence of IGF-1 causal effect on the overall breast cancer risk (BCAC 2017: 1.07 [1.01:1.13]), ER+ subtype (1.07 [1.01:1.13]), and TNBC subtype (1.12 [1.01:1.23]). Female-only IGF-1 GWAS results reported in previous MR studies [40], [41] had similar positive effects with slightly higher magnitudes.

As the relationship of IGF-1 with breast cancer is well-studied, and IGF-1 literature space is large, we identified 291 intermediate terms, out of which 127 were instrumentable and 16 had evidence of an effect on breast cancer (Supplementary Table 3, HTML report). We found evidence of a bidirectional relationship of IGF-1 with two mediator traits: KIT ligand (nSNPs=1) and bioavailable testosterone (nSNPs=125) (IGF-1 as exposure: 0.21 [0.06:0.36] and 0.05 [0.01:0.10], and as outcome -0.03 [-0.07:-0.01] and 0.08 [0.01:0.15], respectively, effect per SD increase in the exposure trait). In MVMR of IGF-1 and KIT ligand the effects of both were attenuated, while in MVMR of IGF-1 and testosterone the effects of both remained almost unchanged with the full sample breast cancer as outcome. This suggests that IGF-1 and bioavailable testosterone effects on breast cancer are likely to be independent, while KIT ligand may be involved in the mechanism by which IGF-1 affects breast cancer risk. However, in a sensitivity analysis using female-only IGF-1 data as exposure, there was no evidence of an effect on KIT ligand, indicating likely sex-specific effects of the IGF-1 measure. In the MR-EvE mediators search, we did not identify any traits that attenuated the effect of female-only IGF-1 effect on breast cancer.

In the literature data, IGF-1 is linked to KIT ligand via CAMP, and is then linked to breast cancer via interaction with TP53. IGF-1 and testosterone are closely linked to each other and breast cancer in the literature, via multiple chains of triples, however, this part of the literature data is too noisy to try to decipher and select a single interesting connection.

*Case study 3: Childhood body size*

Childhood adiposity has previously been shown to have a protective effect on breast cancer risk in both observational [42], [43] and MR studies [40], [44] with the potential mechanisms and mediators of this effect

remaining unclear. Despite using a mixed-sex sample GWAS in this analysis ('body size at age 10'), there was similarly strong evidence of an effect on all breast cancer outcomes, e.g. OR 0.64 [0.56:0.73] (BCAC 2017 overall sample), in agreement with previous MR studies.

In the MR-EvE analysis, we identified 38 traits as potential mediators, many of which were adult anthropometric measures and lipids, and also several proteins (IGF-1, apolipoprotein A, fatty-acid binding protein-4) and two lifestyle traits (skimmed milk and fresh fruit intake) (Supplementary Table 3, HTML report). None of these traits affected the direct effect of childhood body size in MVMR. However, in a sensitivity analysis with female-only childhood body size, we observed a bidirectional relationship with skimmed milk intake (nSNPs=4): positive effect of body size on skimmed milk intake (0.08 [0.06:0.10], effect change per body size category increase) and positive effect of milk intake on body size (3.65 [1.35:5.96], effect per SD increase in milk intake). Skimmed milk intake itself is one of the protective factors identified in the main analysis. In MVMR, the direct effect of childhood body size on ER+ breast cancer when accounting for milk intake was attenuated (OR 0.86 [0.71:1.04]), suggesting a possible role of skimmed milk intake in mediation. However, this finding has to be treated with great caution as dietary traits instruments are often unspecific and pleiotropic (e.g. the top SNPs for skimmed milk are also associated with various anthropometric traits in OpenGWAS).

Among the literature-mined potential intermediates, we identified 8 traits that have an effect on breast cancer (IGF-1, bioavailable testosterone, total testosterone, C-reactive protein, Epidermal growth factor, fatty acid-binding protein-4, Ferritin, E-selectin), but only IGF-1, bioavailable testosterone, fatty acid-binding protein-4 were affected childhood body size. However, none of them attenuated the protective effect of childhood body size in MVMR. IGF-1 and fatty acid-binding protein-4 were identified as potential mediators by both MR-EvE and literature-based discovery.

*Case study 4: Age at menopause*

Age at menopause is an established risk factor for breast cancer in observational studies [45] and MR studies [40], [46]. In our analysis, we observed a strong causal effect on both versions of the overall breast cancer sample, ER+ sample, Luminal A and Luminal B1 subtype (ER+ subtypes), but little evidence of an effect on ER- sample, TNBC and HER-enriched (both ER- subtypes) and Luminal B2.

In the literature analysis, the overlap of 'menopause' and breast cancer literature spaces generated a very tangled network of literature triples, as the hypothesised biological mechanisms and hormonal pathways make these two research areas quite interconnected. We identified 326 literature terms that could be potential intermediates, 90 of which were available in OpenGWAS with at least one instrument. 19 of those traits also had an effect on breast cancer (Supplementary Table 3, HTML report), however, none of them were affected by age at menopause in MR analysis, and therefore could be considered potential mediators and tested in MVMR.

Through searching for potential mediators in MR-EvE, we identified that age at menopause showed some evidence of a positive effect on metabolite X-11445 (5-alpha-pregnane-3beta,20alpha-disulfate): 0.07 [0.002: 0.11] (SD change in metabolite per SD increase in age at menopause), which also had a negative effect on breast cancer (OR 0.82 [0.70:0.96]). However, the direct effect of age at menopause was not affected when accounting for this metabolite in MVMR analysis, suggesting that this analysis was not able to validate the mediating role of X-11445 in age at menopause effect.

# Discussion

In this study, we used breast cancer as an exemplar outcome to highlight the potential of evidence integration using a knowledge graph (EpiGraphDB) to systematically identify risk factors. We combined literature- and genetically-mined data to identify potential intermediates that may explain the relationship between risk factors and breast cancer. To demonstrate how our risk factor evidence integration works in more detail, we presented the results of four case studies.

We mined EpiGraphDB MR-EvE data to identify traits with evidence of an effect on breast cancer risk. The initial results presented in a R/Shiny app (https://mvab.shinyapps.io/brca-miner/) provided us with an interactive way to explore trait associations with a variety of outcomes. The identified causal relationships (N = 2332) were evaluated using rigorous filtering steps that produced a final set of traits with more reliable causal results (N = 213) (Figure 2, https://mvab.shinyapps.io/MR_heatmaps/). The final set of traits included lifestyle traits and molecular biomarkers, many of which have been linked to breast cancer in previous studies. Among the identified lifestyle traits, we observed trends that complement the risk factors summary in the World Cancer Research Fund (WCRF) [20] and International Agency for Research on Cancer (IARC) [23] cancer reports based on observational data: protective effects from strenuous sports/heavy DIY/cycling or walking as a means of transport (in the reports: vigorous physical activity), fresh/dried fruit intake (non-starchy vegetables or carotenoids in diet), skimmed milk consumption (dairy / high-calcium diet), childhood adiposity, adult adiposity, and risk-increasing effects from height, smoking, HRT use. Many traits have also been reported in individual MR studies (e.g. adult and childhood body size and other anthropometric measures [44], [47], [48] reproductive traits [40], sleep traits [49], height [50], IGF-1 [41], smoking [51], lipids [52], [53]), or in MR studies analysing multiple exposure traits as potential risk factors or biomarkers, e.g. identified with a traditional literature search [46] or focusing on a subset of molecular traits [54] [55]. The final set of traits also included many that have robust (FDR corrected) causal estimates of the effect on breast cancer but have not been reported in MR studies before. Among the lifestyle traits these include: increased risk from poultry intake, a protective effect from skimmed milk intake, dried fruit intake, never eating sugar, and among molecular traits some examples are: negative effects from of 3-dehydrocarnitine, betaine, C1QTNF5, diamine acetyltransferase-2, IL3Ra, TIE2, and positive effects of calpastatin, interleukin-21 and kallikrein−14.

We next focused on four case study traits (cardiotrophin-1, IGF-1, childhood body size, and age at menopause), which had varying levels of MR evidence of an effect on breast cancer. We explored literature spaces (i.e. triples of terms related to a phenotype) of these traits aiming to identify potential intermediates that may explain their connection to breast cancer and also extracted potential mediators for those traits from MR-EvE data. We compared the potential mediators of these traits between the two sources to identify if any of them have both sources of evidence. Finally, all potential mediators and literature intermediates (when instrumentable) were also tested in MVMR to assess if the estimated direct effect of the case study trait is changed when a mediator is considered. We determined that the negative effect of cardiotrophin-1 on ER+ breast cancer (OR 0.97 [0.95:0.99]) may be mediated via LIF, as it disrupts cardiotrophin-1 effect on breast cancer in MVMR and is also a clear intermediate link between and cardiotrophin-1 and breast cancer literature spaces. The IGF-1 effect on breast cancer (OR 1.07 [1.01:1.13]) was slightly challenging to study as we determined that the results differ if female-only IGF-1 data is used. Nevertheless, the IGF-1 relationship with breast cancer may be linked to KIT ligand, but this relationship needs further investigation. We identified 38 potential mediators of childhood body size protective effect on breast cancer risk (OR 0.64 [0.56:0.73]), but using MVMR we did not find strong evidence of mediation for any of these. Intriguingly, both literature-mined data and MR-EvE agreed on IGF-1 and fatty acid-binding protein 4 as potential mediators, suggesting that these need to be validated with other methods or data. Finally, we found that age at menopause and breast cancer literature spaces were tightly interconnected through various shared (predominantly hormonal) pathways, but despite that, all potential intermediates found in the literature showed no evidence of being affected by age at menopause in MR.

Having access to thousands of pre-generated MR estimates through an application programming interface (API) and R package (epigraphdb-r) [4] is a major strength of EpiGraphDB for studying risk factors for a disease of interest in a comprehensive way. This enables quick hypothesis testing or could be used for collecting candidate traits for a more targeted study. A similar approach to our breast cancer study has recently been taken to study the effect of chronotype on metabolic traits using MR-EvE data [56], and identify causal risk factors for severe COVID-19 using OpenGWAS directly [57].

However, it is important to note that MR-EvE data comes with limitations and challenges. Firstly, knowledge discovery is restricted only to the traits available in OpenGWAS and added to the MR-EvE data in EpiGraphDB. Secondly, a large proportion of GWAS traits in the database were traits analysed in mixed-sex samples, with the potential that female-only effects that are relevant for breast cancer risk could be overlooked in this

analysis. Moreover, many known risk factors (from observational or previous MR studies) may not come up in the MR-EvE review for two reasons, (1) some traits are not heritable/instrumentable or not in OpenGWAS, or (2) a trait was measured in another (larger) cohort or with additional covariates (e.g. BMI). In addition, the MR-MoE approach used for MR-EvE may have selected a different method from that used in published results, as it attempts to predict the most suitable method for the data. Potentially identified causal effects should therefore be manually validated.

Using literature-mined data to help strengthen the evidence of the observed causal associations or to generate hypotheses for the underlying mechanisms has great potential. We have devised a methodology (Figure 3) for working with the literature-mined relationships stored in EpiGraphDB to utilise them to complement MR causal inference results. The main strength of our approach is connecting traits using multiple triples from their literature spaces via overlapping terms, which is an extension of previously published methods [34], [58]. The identified intermediates that are directly or indirectly linked to the anchored terms provide new ideas for hypothesised mediators that could be tested in MR, or overall provide an insight into the biological pathway involved in linking the trait to breast cancer.

Despite the potential of literature-based knowledge discovery, it should be noted that literature-mined data is noisy and can be difficult to work with and interpret. The extraction of semantic triples from text and mapping terms to ULMS (Unified Medical Language System), and later to phenotypes, is challenging and could be imprecise, and therefore this data must be treated with caution. The extracted chains of intermediate terms between the trait of interest and breast cancer (presented as Sankey diagrams) should not be interpreted as true biological pathways. They simply represent relationships between terms that were mentioned together, and often a direct interaction is not assumed (e.g. relationship like: *coexists with*, *associated with*). Moreover, the extracted relationships may also be affected by publication bias (focusing on established knowledge). Therefore, any insights from literature data that are to be used as 'evidence' need to be traced back to the original publications to confirm the validity of extracted relationships.

Our evidence integration approach is based on identifying common potential mediators in both MR-EvE and literature-mined data, and subsequently using MVMR to assess whether a mediator may explain the trait's causal effect on breast cancer. Whilst we hoped to find a substantial number of mediators supported by two evidence sources, for the selected case studies we only found a few such mediators, and therefore we assessed all of the potential mediators (both with one and two sources of evidence) using MVMR. The use of MVMR to assess mediating effects is a widely used approach but using this as the main method for validating our findings may not have been optimal. Firstly, mediation analysis could be carried out with specialised mediation MR approaches [59] (e.g. product and difference methods), however, applying them to all mediators in each case study was beyond the scope of this paper. Secondly, as highlighted in the results, many mediators had a very limited number of instruments available, suggesting that MVMR analyses with those may not have been reliable and therefore interesting results should be replicated in future work using traits analysed in larger datasets. Lastly, using MR (MVMR or mediation MR) may not be the best way to validate the newly discovered causal effects and/or potential mediators, and the additional evidence for those should be sought from other analyses or sources such as observational studies or lab-based validation.

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 method for using literature-mined data to identify potential intermediates or mechanisms between traits of interest. Both are approaches for hypothesis generation and an attempt at bringing evidence from observational and genetic sources together. The resulting picture of breast cancer risk factors is not complete, but this work represents one of the most comprehensive analyses to date, subject to the discussed limitations. The risk factors identified in this work can be followed-up in future studies in more detail, and the overall approach can be used to study the aetiology of other diseases.

# Methods

## EpiGraphDB platform

EpiGraphDB (https://epigraphdb.org) is a Neo4j graph database and an analytical platform that contains integrated biomedical and epidemiological data, built to support data mining of risk factor/disease relationships. It contains trait relationships (causal, observational or 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 [12] (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 39,991; of these, 34,494 are integrated within EpiGraphDB.

EpiGraphDB (version 0.3.0) was accessed via R package *epigraphdb-r* (version 0.2.3) and with native Cypher queries used within R for more complex tasks. Examples of basic queries are provided in https://github.com/mvab/epigraphdb_mr_literature_queries.

## MR-EvE data

MR-EvE (Mendelian Randomization "Everything-vs-Everything") is one of the relationships between GWAS traits available in EpiGraphDB. Mendelian randomization (MR) estimates are available for many pairs of GWAS traits, bidirectionally. At the time of writing, 11,789 traits have MR-EvE relationships in the graph.

The MR-EvE estimates were generated using the MR Mixture-of-Experts (MR-MoE) method [15] prior to integration into EpiGraphDB. MR-MoE is a machine learning framework that automates the selection of SNPs and an MR method for use in any specific causal analysis. It tests over 20 strategies to selects the approach that is most likely to be correct for a specific MR analysis by predicting the model of pleiotropy [15]. This MR 'best estimate' is available as a relationship between GWAS nodes in EpiGraphDB.

## Breast cancer outcomes data

EpiGraphDB provides access to 24 breast cancer GWAS datasets (via OpenGWAS). The majority of available GWAS are derived from different versions of the Breast Cancer Association Consortium (BCAC) cohort (http://bcac.ccge.medschl.cam.ac.uk/) [27] and UK Biobank [37]. The versions of BCAC data include subsets of the cohort genotyped on separate arrays (*iCOGs* and *Oncoarray*), their *meta-analysis* (also referred to as BCAC 2017) and other cohort subsets (labelled as *GWASv1/2*) (Supplementary Table 1). In the MR-EvE risk factor discovery analysis, we included 21 outcomes, including ER+/ER- and overall samples.

The MR validation was done on BCAC 2017 (N=228,951; overall sample and ER+/ER- subtypes) and the latest release of the meta-analysed BCAC 2020 data (N=247,173; overall sample and five molecular subtypes: Luminal A, Luminal B1, Luminal B2, HER2-enriched, TNBC) [60] (not a part of OpenGWAS) (Supplementary Table 2).

## Identifying and validating risk factors trait from MR-EvE

The extraction and filtering of breast cancer risk factors identified in MR-EvE is summarised in Supplementary Figure 1 and in Supplementary Data 1. We reduced the initial 2332 identified traits to 309 lifestyle and molecular risk factors that had a consistent effect and direction across three main breast cancer outcomes (*meta-analysis* (BCAC 2017), *OncoArray, iCOGs 2017).* The traits were split into 12 categories: anthropometric, dietary, physical activity, reproductive, sleep, alcohol, smoking, medication, proteins, lipids, metabolites, and other biomarkers.

The MR-EvE estimates generated with the MR-MoE method should not be reported directly [15] and therefore we validated the observed effects using the standard MR practices for two-sample MR [26]. We extracted instruments (SNPs under P-value $< 5 \times 10^{-8}$ threshold, clumped with $r2 < 0.001$) and applied the inverse-variance weighted (IVW) [25] MR method. This validation analysis was performed on BCAC 2017 and BCAC 2020, to evaluate the potential differences between the two versions.

To investigate the potential violations of the MR assumptions and validate the robustness of MR results we performed additional MR analyses using MR-Egger [29] and weighted median [30] approaches, both of which are more robust to pleiotropy. The Egger intercept [29] was used to explore the potential for the presence of directional horizontal pleiotropy, and Cochran's Q statistic [31] was calculated to quantify the extent of

heterogeneity among SNPs, which is indicative of potential pleiotropy. All MR and sensitivity analyses were performed using the *TwoSampleMR* R package (version 0.5.6) [14].

To correct for multiple testing, we used the false-discovery rate (FDR) approach [61], applying it to the p-values from the IVW or Wald ratio results. A cut-off of 0.05 for the adjusted p-value was used to identify results with evidence of effect.

**Identifying potential mediators from MR-EvE**

We investigated potential mediators of the effect of the traits on breast cancer from the final set (Figure 1a) by querying MR-EvE data again. We used the EpiGraphDB confounder-search method (https://epigraphdb.org/confounder) to identify traits that are affected by the the exposure traits from the final set and that also have evidence of effect on breast cancer outcomes (Figure 1b). We focused only on the potential mediator traits that also fit into 12 trait categories that were used for selecting the main exposure traits. In addition, we performed manual validation and sensitivity analysis as described above for the main exposure/outcome pairs, using a two-step MR framework [62].

**Literature-mined relationships in EpiGraphDB**

EpiGraphDB contains literature-mined relationships between pairs of terms (also known as 'literature triples') that were derived from the published literature. These relationships were originally extracted by the MELODI-Presto tool [34], but are now fully integrated into EpiGraphDB literature nodes. The underlying data comes from SemMedDB [17], a well-established repository of literature-mined semantic triples, i.e. *'subject term 1 – predicate – object term 2'*, mined from titles and abstracts in PubMed. A subject/object term could be any biological entity (gene name, drug, phenotype, disease) and a predicate is a verb that represents a relationship between the two terms (affects, causes, inhibits, reduces, associated with, etc.). In EpiGraphDB, the mined triples of biomedical terms are mapped to GWAS traits, EFOs, and genes available in the graph.

A trait 'literature space' is a set of triples that are related to the trait. It is possible to extract a literature space for most GWAS traits in EpiGraphDB. A triple score indicates how frequently the triple occurs in the published literature within a defined time frame, and is an indication of how well-defined the relationship between two terms is. The matching of triples to a trait happens via the GWAS trait name, which is somewhat equivalent to performing a PubMed search for a keyword (i.e. the trait name). The success of matching depends on how the trait name is phrased, e.g. a trait name '*childhood obesity'* is likely to collect more relevant triples than '*comparative body size at age 10'*. Similarly, a trait name '*morning/evening person'* is less likely to find relevant triples than the name *'chronotype'*. Complex trait names, e.g. *'Types of physical activity in last 4 weeks: Heavy DIY (eg: weeding, lawn mowing, carpentry, digging)'* have a very low chance of matching meaningful results.

**Extracting literature space for a trait**

Breast cancer literature space was formed from the data linked to two GWAS nodes: '*Breast cancer'* and '*Malignant neoplasm of the breast*'. We identified 18,848 unique relationships (16,103 pairs of terms) based on 4989 unique terms derived from 23,809 publications, from the time frame 1954 - April 2020 (last update of literature relationships in EpiGraphDB). The triples unique to years before 1990 make up less than 2% of all literature-mined relationships for breast cancer. Supplementary Data 8 details this breast cancer literature space, with triple counts/scores and term types.

We also extracted literature spaces for the risk factor traits identified in MR estimate validation (Figure 2) (Supplementary Data 9). A subset of traits were excluded from this step: (1) traits with similar or duplicate names (e.g. *Body mass index* and *Body mass index (BMI)),* as they would collect the same triples; (2) traits with names that are unlikely to produce a meaningful set of results (e.g. *Comparative body size at age 10*); (3) the majority of lipid traits as they had very similar names; (4) metabolites with unknown names (i.e. starting with "X-").

Despite initial filtering of traits with similar names, there were still a few traits that represent the same phenotype. We combined literature spaces of those traits into broad concepts: height measures (*height, sitting height, standing height*), weight measures (*weight, body mass index (BMI)*), obesity (*overweight, obesity class 1, obesity class 2, extreme body mass index*), hip/waist measures (*hip circumference, waist circumference, waist-to-hip ratio*), fruit intake (*fresh fruit intake, dried fruit intake, cherry intake*), menopause (*had menopause, age at menopause*), HRT use (*age started HRT, ever used HRT*), childbirth (*number of live births, age at last live birth)*, lipids HDL, lipids VLDL.

The extracted literature spaces were cleaned using several strategies – protein/gene names were standardised (e.g. 'BRCA1 gene', 'BRCA1 protein', 'BRCA1, human' were all converted to simply 'BRCA1'); triples with the same item in both term1 and term2 were excluded; term names were standardised as much as possible to minimise redundancy (i.e. merging long/short names, name acronyms, synonyms, gene/protein names); terms with term types (e.g. aapp, gngm, dsyn) were filtered to include only the most common type to reduce redundancy.

**Connecting traits via literature spaces**

We designed an approach for investigating how two traits may be connected that is based on connecting their literature spaces (Figure 3a). The approach relies on extracting sequentially linked triples from both literature spaces and identifying the overlapping terms between them. This method is based on the concept used in MELODI-Presto [34], where the linkage of two triples was used to identify intermediates. We have expanded this basic principle by introducing multiple triples chaining to allow a multi-step connection between a trait of interest and disease outcome. This approach could also be considered a form of Swanson linking [18] [35], a concept widely known in literature-based discovery, in which two separately published results (e.g. A-B and B-C relationships) are combined to identify evidence for A-C relationship that is unknown or unexplored.

Our approach involves restricting (i.e. 'anchoring') the outer term in a chain of triples to a trait itself (e.g. in the IGF-1 literature space, term1 in triple1 is always IGF-1 or its name variations), or in breast cancer literature space, term2 of the outer (terminal) triple is restricted to the term representing breast cancer outcome (in practice, two terms were selected to represent it: *'breast diseases'* and '*malignant disease'*). This ensures that by overlapping two literature spaces we always start with the exposure trait and end with the breast cancer outcome, collecting all intermediate terms/paths between them. As anchoring is only possible for molecular traits (Figure 3b), for lifestyle traits (where no specific term is available) we had to rely on matching unlinked trait triples to any breast cancer space triples (Figure 3c).

For all traits, multiple filtering steps were performed: excluding terms that represent drugs/medications or other diseases (except 'obesity' and other anthropometrics-related terms), predicates that are unspecific or negative, e.g. *'coexists with'*, '*neg coexists with'*, '*neg isa'*, and for terms related bidirectionally, the triple with the direction with the higher score (i.e. more common) was retained. The resulting literature space overlaps are available for exploration as Sankey plots in the R/Shiny app (described in Results) for a selected number of traits, and an example for cardiotrophin-1 is shown in Figure 3d.

**[Figure 3 here]**

**MR and literature evidence integration**

The evidence integration workflow for the identified risk factors (Figure 1a) and their potential mediators from MR-EvE and literature-based discovery involves the steps outlined in Figure 1(b-d). We applied the workflow to the four case studies (Figure 1e) presented in the paper, but any other risk factor can be appraised in the same way given sufficient data.

For each risk factor / case study, the mediators extracted from MR-EvE were verified with the standard MR practices in two-step MR (Figure 1b). Next, we determined which of the literature overlap intermediate terms were available in OpenGWAS [12] and instrumentable with >= 1 genome-wide significant SNPs. We performed bidirectional MR [63] to establish the direction of the effect between the intermediate and the risk factor and identify the traits that are affected by the risk factor (i.e. likely mediator relationship) and the other way around (i.e. potential confounder relationship). This was followed by two-step MR [62] to select the potential mediators that have evidence of an effect on breast cancer (Figure 1c). Any mediators identified in both MR-EvE and literature analyses were flagged as having two sources of evidence for being a potential mediator, increasing our confidence in the finding (Figure 1d).

Then, for all mediators, we performed multivariable MR (MVMR) [24] as a validation step, with the risk factor and the mediator as exposures and breast cancer as the outcome (Figure 1d). With MVMR we were able to estimate the direct causal effect of each exposure, accounting for one another, in a single analysis. We used MVMR to identify if any of the potential mediators affect or attenuate the effect of the main risk factor trait, as those traits would be the most interesting candidates to explain the relationship between the risk factor and breast cancer.

## Data availability

The MR-EvE estimates and literature-mined relationships used in this work are available from EpiGraphDB (https://epigraphdb.org). The GWAS data used to represent literature-identified intermediates was taken from OpenGWAS (https://gwas.mrcieu.ac.uk). BCAC 2020 molecular subtype data is available at https://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/oncoarray-and-combined-summary-result/

## Code availability

Navigation page for web-based components of this article (apps and case study reports): https://mvab.github.io/epigraphdb-breast-cancer/

The main project development repository: https://github.com/mvab/epigraphdb-breast-cancer

The repository with basic EpiGraphDB queries example / minimal reproducible queries: https://github.com/mvab/epigraphdb_mr_literature_queries


## Acknowledgements

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_00011/4).

This work was also supported by a Cancer Research UK programme grant (the Integrative Cancer Epidemiology Programme) (C18281/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.


## Author contributions

T.R.G., B.E. and M.V. conceived the study. All authors contributed to study design and interpretation of results. M.V. cleaned the data, performed the analyses, interpreted the results, and wrote the initial draft of the manuscript as well as subsequent drafts with critical input on results interpretation from T.R., B.E., Y.L., T.R.G.

The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

## Competing interests

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

# References


[1] D. A. Lawlor, K. Tilling, and G. D. Smith, “Triangulation in aetiological epidemiology,” *Int. J. Epidemiol.*, vol. 45, no. 6, pp. 1866–1886, Dec. 2017.

[2] M. R. Munafò and G. Davey Smith, “Robust research needs many lines of evidence,” *Nature*, vol. 553, no. 7689. Nature Publishing Group, pp. 399–401, 23-Jan-2018.

[3] M. R. Munafò, J. P. T. Higgins, and G. D. Smith, “Triangulating evidence through the inclusion of genetically informed designs,” *Cold Spring Harb. Perspect. Med.*, vol. 11, no. 8, 2021.

[4] Y. Liu *et al.*, “EpiGraphDB: A database and data mining platform for health data science,” *Bioinformatics*, vol. 37, no. 9, pp. 1304–1311, Aug. 2021.

[5] D. Ochoa *et al.*, “Open Targets Platform: Supporting systematic drug-target identification and prioritisation,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1302–D1310, 2021.

[6] A. Struck *et al.*, “Exploring Integrative Analysis Using the BioMedical Evidence Graph,” *JCO Clin. Cancer Informatics*, no. 4, pp. 147–159, 2020.

[7] E. A. Coker *et al.*, “CanSAR: Update to the cancer translational research and drug discovery knowledgebase,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D917–D922, Jan. 2019.

[8] J. Piñero *et al.*, “DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D833–D839, 2017.

[9] S. Ebrahim and G. Davey Smith, “Mendelian randomization: Can genetic epidemiology help redress the failures of observational epidemiology?,” *Int. J. Epidemiol.*, vol. 32, no. 1, pp. 1–22, Feb. 2003.

[10] E. Sanderson *et al.*, “Mendelian randomization,” *Nat. Rev. Methods Prim.*, vol. 2, no. 1, Dec. 2022.

[11] J. Xu *et al.*, “Using Mendelian randomization as the cornerstone for causal inference in epidemiology,” *Environ. Sci. Pollut. Res.*, vol. 29, no. 4, pp. 5827–5839, 2022.

[12] B. Elsworth *et al.*, “The MRC IEU OpenGWAS data infrastructure,” *bioRxiv*, p. 2020.08.10.244293, Aug. 2020.

[13] A. Buniello *et al.*, “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1005–D1012, Jan. 2019.

[14] G. Hemani *et al.*, “The MR-base platform supports systematic causal inference across the human phenome,” *Elife*, vol. 7, May 2018.

[15] G. Hemani *et al.*, “Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome,” *bioRxiv*, no. 10, p. 173682, 2017.

[16] L. J. Jensen, J. Saric, and P. Bork, “Literature mining for the biologist: From information retrieval to biological discovery,” *Nature Reviews Genetics*, vol. 7, no. 2. pp. 119–129, 2006.

[17] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch, “SemMedDB: A PubMed-scale repository of biomedical semantic predications,” *Bioinformatics*, vol. 28, no. 23, pp. 3158–3160, 2012.

[18] T. Bekhuis, “Conceptual biology, hypothesis discovery, and text mining: Swanson’s legacy,” *Biomed. Digit. Libr.*, vol. 3, no. 1, pp. 1–7, Apr. 2006.

[19] G. Crichton, S. Baker, Y. Guo, and A. Korhonen, “Neural networks for open and closed Literature-based Discovery,” *PLoS One*, vol. 15, no. 5, p. e0232891, May 2020.

[20] 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. [Online]. Available: www.wcrf.org/matrix. [Accessed: 09-Feb-2022].

[21] K. L. Britt, J. Cuzick, and K. A. Phillips, “Key steps for effective breast cancer prevention,” *Nat. Rev. Cancer*, 2020.

[22] N. Harbeck *et al.*, *Breast cancer*, vol. 5, no. 1. 2019.

[23] C. P. Wild, E. Weiderpass, and B. W. Stewart, *World Cancer Report 2020: Cancer research for cancer prevention*. Lyon, France, 2020.

[24] S. Burgess and S. G. Thompson, “Multivariable Mendelian randomization: The use of pleiotropic genetic variants to estimate causal effects,” *Am. J. Epidemiol.*, vol. 181, no. 4, pp. 251–260, 2015.

[25] S. Burgess, A. Butterworth, and S. G. Thompson, “Mendelian randomization analysis with multiple genetic variants using summarized data,” *Genet. Epidemiol.*, vol. 37, no. 7, pp. 658–665, Nov. 2013.

[26] S. Burgess *et al.*, “Guidelines for performing Mendelian randomization investigations [version 2; peer review: 1 approved, 1 approved with reservations],” 2020.

[27] K. Michailidou *et al.*, “Association analysis identifies 65 new breast cancer risk loci,” *Nature*, vol. 551, no. 7678, pp. 92–94, Nov. 2017.

[28] H. Zhang *et al.*, “Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses,” *Nat. Genet.*

[29] J. Bowden, G. D. Smith, and S. Burgess, “Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression,” *Int. J. Epidemiol.*, vol. 44, no. 2, pp. 512–525, 2015.

[30] J. Bowden, G. Davey Smith, P. C. Haycock, and S. Burgess, “Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator,” *Genet. Epidemiol.*, vol. 40, no. 4, pp. 304–314, May 2016.

[31] J. Bowden *et al.*, “Improving the accuracy of two-sample summary-data Mendelian randomization: Moving beyond the NOME assumption,” *Int. J. Epidemiol.*, vol. 48, no. 3, pp. 728–742, 2019.

[32] B. Jassal *et al.*, “The reactome pathway knowledgebase,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D498–D503, 2020.

[33] L. A. C. Millard, N. M. Davies, T. R. Gaunt, G. D. Smith, and K. Tilling, “Software Application Profile:

PHESANT: a tool for performing automated phenome scans in UK Biobank," *Int. J. Epidemiol.*, vol. 47, no. 1, pp. 29–35, Feb. 2018.

[34] B. Elsworth and T. R. Gaunt, "MELODI Presto: A fast and agile tool to explore semantic triples derived from biomedical literature," *Bioinformatics*, Aug. 2020.

[35] N. R. Smalheiser, "Rediscovering Don Swanson:The Past, Present and Future of Literature-based Discovery," *J. Data Inf. Sci.*, vol. 2, no. 4, pp. 43–64, Dec. 2017.

[36] B. B. Sun *et al.*, "Genomic atlas of the human plasma proteome," *Nature*, vol. 558, no. 7708, pp. 73–79, Jun. 2018.

[37] C. Sudlow *et al.*, "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age," *PLoS Med.*, vol. 12, no. 3, Mar. 2015.

[38] A. González *et al.*, "Cardiotrophin-1 in hypertensive heart disease," *Endocrine*, vol. 42, no. 1, pp. 9–17, 2012.

[39] P. F. Christopoulos, P. Msaouel, and M. Koutsilieris, "The role of the insulin-like growth factor-1 system in breast cancer," *Molecular Cancer*, vol. 14, no. 1. BioMed Central Ltd., p. 43, 15-Feb-2015.

[40] M. Vabistsevits, G. Davey Smith, E. Sanderson, T. G. Richardson, B. Lloyd-Lewis, and R. C. Richmond, "Deciphering how early life adiposity influences breast cancer risk using Mendelian randomization," *Commun. Biol.*, vol. 5, no. 1, 2022.

[41] N. Murphy *et al.*, "Insulin-like growth factor-1, insulin-like growth factor-binding protein-3, and breast cancer risk: observational and Mendelian randomization analyses with ~430 000 women," *Ann. Oncol.*, vol. 31, no. 5, pp. 641–649, 2020.

[42] H. J. Baer, S. S. Tworoger, S. E. Hankinson, and W. C. Willett, "Body fatness at young ages and risk of breast cancer throughout life," *Am. J. Epidemiol.*, vol. 171, no. 11, pp. 1183–1194, 2010.

[43] A. Furer *et al.*, "Adolescent obesity and midlife cancer risk: a population-based cohort study of 2·3 million adolescents in Israel," *Lancet Diabetes Endocrinol.*, vol. 8, no. 3, pp. 216–225, 2020.

[44] T. G. Richardson, E. Sanderson, B. Elsworth, K. Tilling, and G. D. Smith, "Use of genetic variation to separate the effects of early and later life adiposity on disease risk: Mendelian randomisation study," *BMJ*, vol. 369, 2020.

[45] C. Group on Hormonal Factors in Breast Cancer, "Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies Collaborative Group on Hormonal Factors in Breast Cancer*," *Lancet Oncol.*, vol. 13, pp. 1141–1151, 2012.

[46] F. Chen *et al.*, "Mendelian randomization analyses of 23 known and suspected risk factors and biomarkers for breast cancer overall and by molecular subtypes," *Int. J. Cancer*, Apr. 2022.

[47] Y. Guo *et al.*, "Genetically Predicted Body Mass Index and Breast Cancer Risk: Mendelian Randomization Analyses of Data from 145,000 Women of European Descent," *PLoS Med.*, vol. 13, no. 8, p. e1002105, Aug. 2016.

[48] D. Freuer *et al.*, "Body Fat Distribution and Risk of Breast, Endometrial, and Ovarian Cancer: A Two-Sample Mendelian Randomization Study," *Cancers (Basel).*, vol. 13, no. 20, Oct. 2021.

[49] R. C. Richmond *et al.*, "Investigating causal relations between sleep traits and risk of breast cancer in women: Mendelian randomisation study," *BMJ*, vol. 365, 2019.

[50] M. Vithayathil, P. Carter, S. Karx, A. M. Mason, S. Burgess, and S. C. Larsson, "Body size and composition and risk of site-specific cancers in the UK Biobank and large international consortia: A mAU endelian: PleasenotethatasperP randomisation study," *PLoS Med.*, vol. 18, no. 7, pp. 1–20, 2021.

[51] N. Dimou *et al.*, "Causal effects of lifetime smoking on breast and colorectal cancer risk: Mendelian randomization study," *Cancer Epidemiol. Biomarkers Prev.*, vol. 30, no. 5, pp. 953–964, May 2021.

[52] K. E. Johnson *et al.*, "The relationship between circulating lipids and breast cancer risk: A Mendelian randomization study," *PLoS Med.*, vol. 17, no. 9, Sep. 2020.

[53] C. Nowak and J. Ärnlöv, "A Mendelian randomization study of the effects of blood lipids on breast cancer risk," *Nat. Commun.*, vol. 9, no. 1, Dec. 2018.

[54] T. Robinson, R. M. Martin, and J. Yarmolinsky, "Mendelian randomisation analysis of circulating adipokines and C-reactive protein on breast cancer risk," *Int. J. cancer*, vol. 147, no. 6, pp. 1597–1603, Sep. 2020.

[55] E. Bouras *et al.*, "Circulating inflammatory cytokines and risk of five cancers: a Mendelian randomization analysis," *BMC Med.*, vol. 20, no. 1, pp. 1–15, 2022.

[56] J. A. Williams *et al.*, "A causal web between chronotype and metabolic health traits," *Genes (Basel).*, vol. 12, no. 7, 2021.

[57] Y. Sun, J. Zhou, and K. Ye, "Extensive Mendelian randomization study identifies potential causal risk factors for severe COVID-19," *Commun. Med. 2021 11*, vol. 1, no. 1, pp. 1–11, Dec. 2021.

[58] B. Elsworth *et al.*, "MELODI: Mining enriched literature objects to derive intermediates," *Int. J. Epidemiol.*, vol. 47, no. 2, pp. 369–379, 2018.

[59] A. R. Carter *et al.*, "Mendelian randomisation for mediation analysis: Current methods and challenges for implementation," *European Journal of Epidemiology*. 2021.

[60] H. Zhang *et al.*, "Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses," *Nat. Genet.*, vol. 52, no. 6, pp. 572–581, Jun. 2020.

[61] K. Korthauer *et al.*, "A practical guide to methods controlling false discoveries in computational biology," *Genome Biol.*, vol. 20, no. 1, pp. 1–21, Jun. 2019.

[62] C. L. Relton and G. Davey Smith, "Two-step epigenetic mendelian randomization: A strategy for establishing the causal role of epigenetic processes in pathways to disease," *Int. J. Epidemiol.*, vol. 41, no. 1, pp. 161–176, 2012.

[63] J. Zheng *et al.*, "Recent Developments in Mendelian Randomization Studies," *Curr. Epidemiol. Reports*, vol. 4, no. 4, pp. 330–345, 2017.

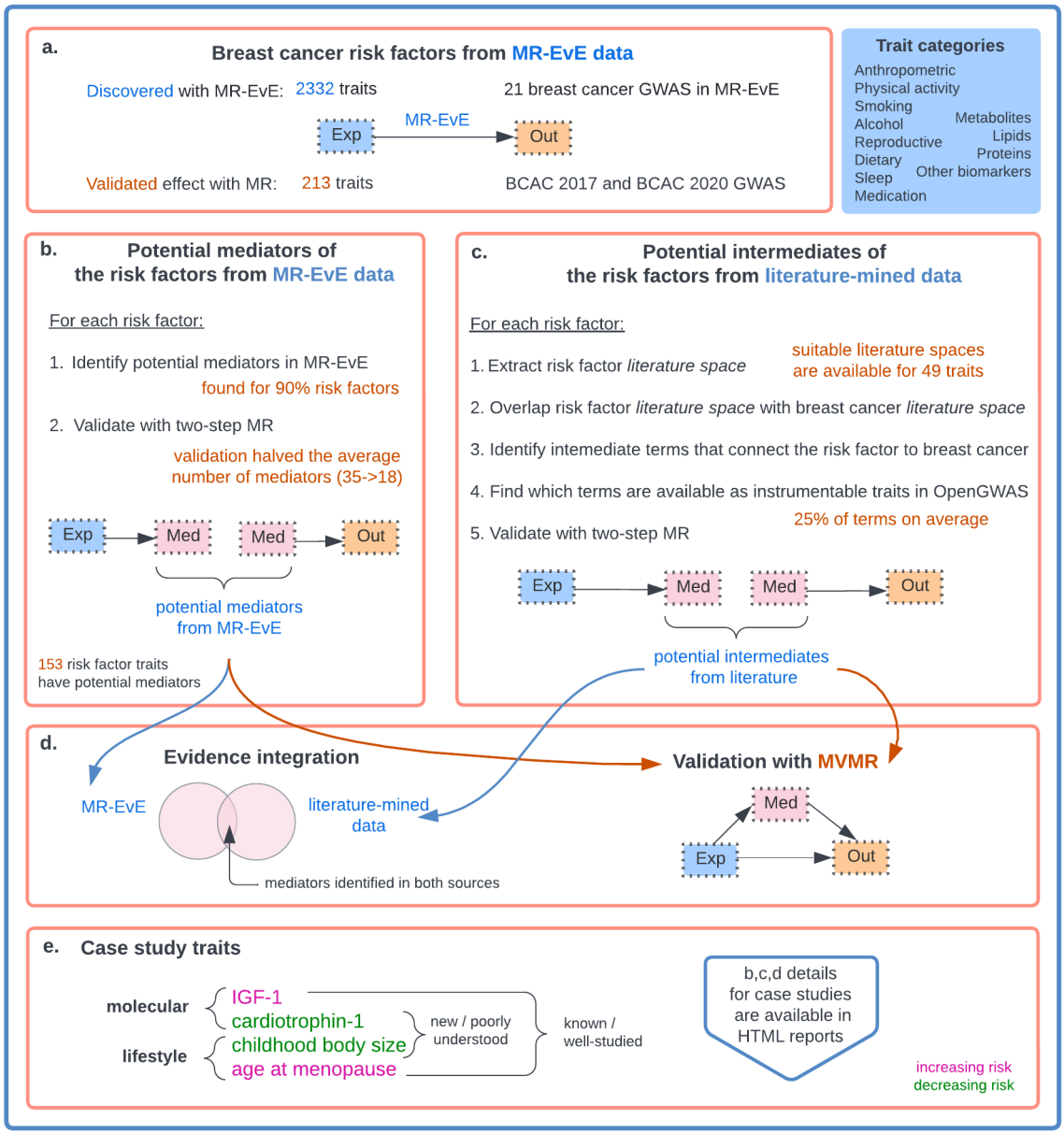


**Figure 1. Study overview.** (a) Mining MR-EvE data to identify breast cancer risk factors. (b) Mining MR-EvE data to identify potential mediators of the identified risk factors. (c) Using literature-mined data to generate another set of potential intermediates between the identified risk factors and breast cancer. (d) Integrating evidence for potential mediators from MR-EvE and literature-mined data, and validating interesting results with MVMR. (e) Presenting detailed results for four case studies. *Abbreviations:* 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); MVMR – multivariable Mendelian randomization; *literature space* – a collection of literature-mined relationships between pairs of biological terms related to a phenotype.

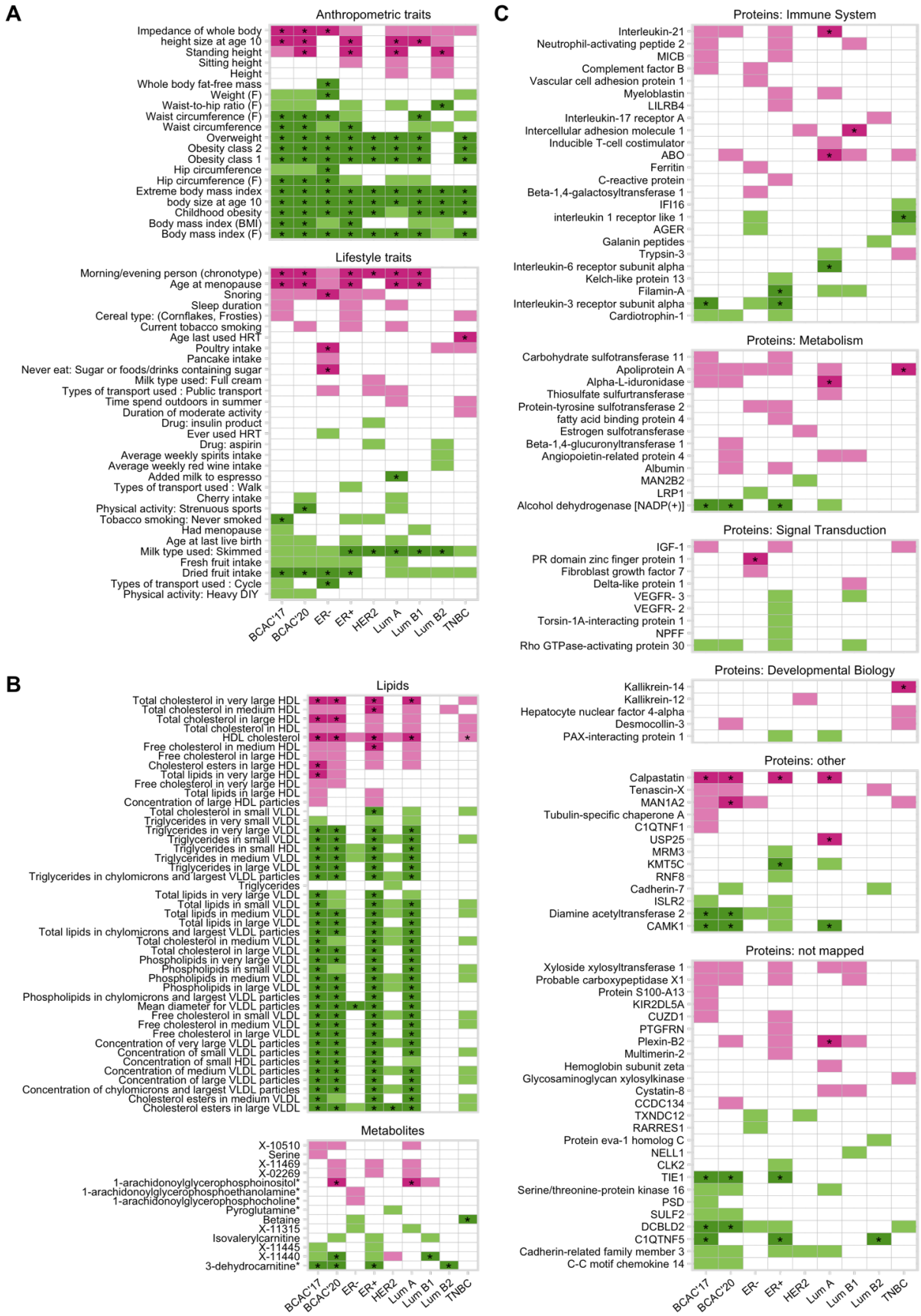


**Figure 2. Mendelian randomization effect direction heatmaps for (A) lifestyle and anthropometric traits, (B) metabolites and lipids, (C) proteins.** The traits in A/B/C groups are the exposures, and breast cancer (BCAC 2017 and 2020, including subtypes) are the outcomes. All included traits have validated evidence of an effect on at least one outcome. The colours represent effect direction: pink – positive (causal), green – negative (protective), white – null effect (based on 95% confidence intervals of odds of breast cancer per standard deviation increase in the exposure). The asterisk and darker colour shade indicate the results that passed FDR correction. The proteins are grouped by major pathways defined in the Reactome database. Proteins with long names are displayed as abbreviations or the corresponding gene name (defined in Supplementary Data 4). The exposure traits are from mixed-sex samples unless otherwise specified (F: female-only) or are female-specific reproductive traits. 'Never eat sugar' trait effect should be interpreted as protective, as the 'never' phenotype was treated as the baseline in the GWAS analysis as selected by Phesant [33]. Some anthropometric traits were excluded from the figure to reduce redundancy (details in Supplementary Data 1). Interactive version of the plot is available at: https://mvab.shinyapps.io/MR_heatmaps/

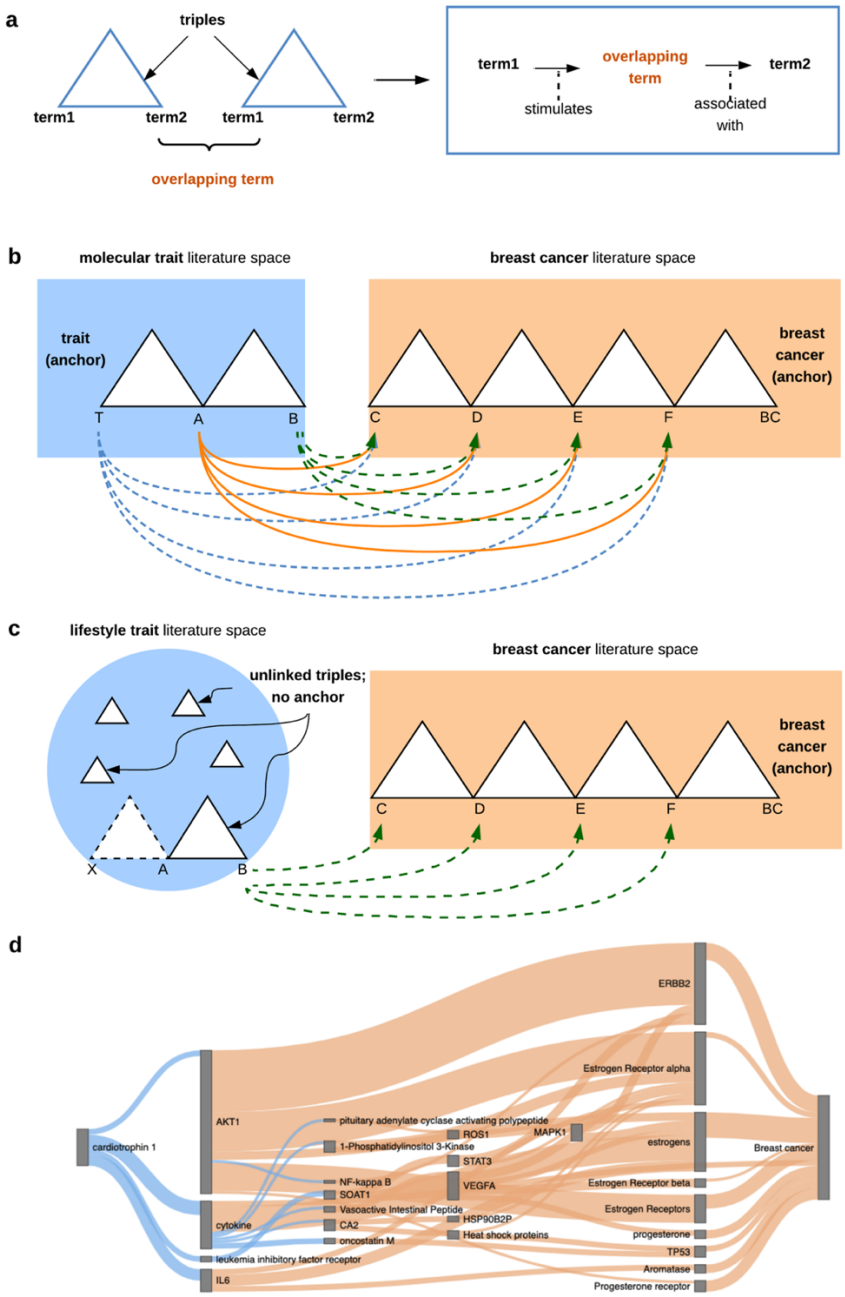


**Figure 3. Schemas of literature spaces overlap methods.** (a) A basic example of two triples linkage via an overlapping term. (b) Molecular traits literature space overlap method. Molecular trait literature space (blue box) and breast cancer literature space (purple box) both contain triples linked via overlapping terms. 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. Terms T, A, B can be overlapping terms with C, D, E, and F, which links the trait and breast cancer literature spaces. The arrows represent all possible paths from T to BC alternative to the full path going through all intermediates. (c) Lifestyle traits literature space overlap method. Lifestyle trait literature space (green circle) contains triples that cannot be anchored to any specific term representing the trait. The unlinked triples in the lifestyle trait space are matched to overlap (via B) with any triples in the breast cancer literature space with a path to breast cancer. The linked triples are then connected with any preceding triples in lifestyle trait space (via A), adding X-A triples into the spaces overlap. (d) *Example*: Sankey diagram of literature spaces overlap for cardiotrophin-1, showing how triples within and between cardioptrophin-1 and breast cancer spaces interconnect. The line thickness corresponds to the triple score (frequency) of the term pair. More examples are available at https://mvab.shinyapps.io/literature_overlap_sankey/

**Table 1. Trait counts by filtering stages, by trait categories, by breast cancer subtypes.** The term 'validated' refers to having evidence of effect in manual MR analysis (IVW or Wald ratio); 'consistent effect' = presence of evidence of effect and agreement in effect direction between 2 out of 3 main breast cancer datasets in OpenGWAS (*meta-analysis, OncoArray, iCOGs 2017*); BCAC 2017 is the same as *meta-analysis* from MR-EvE app; 'Total' refers to a unique number of traits in the category across all subtypes within data version; 'Overall' refers to a full sample breast cancer GWAS, i.e all subtypes together.

| Trait category | Traits in MR-EvE app | Traits with consistent effect | Traits validated in either BCAC | Traits validated in BCAC 2017 | | | | Traits validated in BCAC 2020 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Total** | **Total** | **Total** | **Total** | **Overall** | **ER+** | **ER-** | **Total** | **Overall** | **LumA** | **LumB1** | **LumB2** | **HER2** | **TNBC** |
| ***Lifestyle traits*** | | | | | | | | | | | | | | |
| Anthropometric | 56 | 29 | 28 | 27 | 22 | 23 | 19 | 26 | 21 | 19 | 17 | 17 | 10 | 12 |
| Dietary | 54 | 13 | 10 | 7 | 4 | 4 | 6 | 8 | 4 | 5 | 2 | 3 | 2 | 4 |
| Physical activity | 39 | 12 | 7 | 4 | 2 | 1 | 2 | 5 | 2 | 3 | 0 | 0 | 1 | 2 |
| Reproductive | 23 | 10 | 6 | 5 | 4 | 2 | 2 | 5 | 2 | 2 | 3 | 0 | 0 | 1 |
| Sleep | 14 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 2 | 2 | 1 | 0 | 2 | 0 |
| Medication | 45 | 3 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 2 | 0 |
| Alcohol | 15 | 3 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 |
| Smoking | 24 | 5 | 2 | 2 | 1 | 2 | 0 | 2 | 1 | 1 | 0 | 0 | 1 | 0 |
| ***Molecular traits*** | | | | | | | | | | | | | | |
| Proteins | 468 | 161 | 93 | 64 | 33 | 42 | 17 | 59 | 27 | 24 | 13 | 6 | 6 | 13 |
| Lipids | 83 | 49 | 48 | 47 | 46 | 45 | 5 | 46 | 43 | 43 | 0 | 2 | 17 | 20 |
| Metabolites | 77 | 20 | 15 | 12 | 5 | 6 | 4 | 10 | 7 | 5 | 3 | 1 | 2 | 1 |
| Other biomarkers | 7 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **TOTAL** | **905** | **309** | **213** | **171** | 120 | 128 | 57 | **168** | 109 | 104 | 39 | 32 | 43 | 53 |

**Table 2. Summary of querying literature-mined data from EpiGraphDB for GWAS traits of interest.** The results are split by trait category and literature space size in terms of unique triples (any number, <=5, <=50, >50).

| | Traits queried | Traits with non-empty lit space | Traits with <= 5 triples | Traits with <= 50 triples | Traits with > 50 triples |
|---|---|---|---|---|---|
| **Lifestyle traits** | | | | | |
| *Anthropometric* | 14 | 14 | 0 | 3 | 11 |
| *Dietary* | 12 | 5 | 3 | 1 | 1 |
| *Physical activity* | 7 | 1 | 0 | 0 | 1 |
| *Reproductive* | 5 | 4 | 0 | 2 | 2 |
| *Sleep* | 3 | 2 | 0 | 0 | 2 |
| *Medication* | 2 | 1 | 0 | 0 | 1 |
| *Smoking* | 2 | 1 | 0 | 0 | 1 |
| **Molecular traits** | | | | | |
| *Proteins* | 89 | 51 | 16 | 10 | 25 |
| *Lipids* | 12 | 9 | 2 | 2 | 5 |
| *Metabolites* | 8 | 2 | 1 | 1 | 0 |
| **TOTAL** | **154** | **90** | **22** | **19** | **49** |