Possible selection bias in Mendelian randomization studies: effect of body mass index on breast cancer as an example

Kezhen Fei[a], Chloe A Teasdale[a], Mary Beth Terry[b], C Mary Schooling[c,d]

a. Department of Epidemiology and Statistics, City University of New York (CUNY) Graduate School of Public Health and Health Policy, New York, NY, USA
b. Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY, USA
c. School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China
d. Graduate School of Public Health and Health Policy, The City University of New York, New York, NY, USA

Correspondence:
Kezhen Fei
Email: Kezhen.fei@gmail.com

Funding:
This project received no specific funding support.

Key words: adiposity, body mass index, breast cancer, Mendelian randomization, selection bias, survival bias

Word count: 3050 (not including abstract)

Number of Tables: 4

Number of Figures: 1

Abstract

The inverse association of body mass index (BMI) with postmenopausal breast cancer (BC) risk observed in Mendelian Randomization (MR) studies contradicts the International Agency for Research on Cancer (IARC) classification of high BMI as a risk factor for postmenopausal BC. We investigated whether this discrepancy stems from survival-driven selection bias in the underlying GWAS. Using summary statistics from three different GWAS, higher genetically predicted BMI was inversely associated with BC risk in the BCAC consortium (odds ratio (OR): 0.74, 95% CI: 0.64, 0.84) and the UK Biobank (OR: 0.70, 95% CI: 0.56, 0.87), but showed no relationship with sibling BC (BCs) risk (OR: 1.00, 95% CI: 0.99, 1.01). These findings demonstrate that previous MR studies may suffer from selection bias driven by survivorship bias. This highlights a critical methodological vulnerability in MR studies when both exposure and outcome cause premature mortality, and the underlying population is not fully captured by current GWAS.

Key messages

- Consistent with prior MR studies, higher BMI appeared protective against breast cancer (BC).
- Sibling-based analyses indicated no significant association between BMI and BC risk.
- This inverse association of genetically predicted BMI with BC is likely driven by survival bias, stemming from the exclusion of individuals who died from BC or high BMI prior to GWAS study enrolment.

Introduction

Breast cancer (BC) affects approximately one in every eight women, with projections indicating it will become the leading cause of cancer death among U.S. women by 2040.(1) While mortality rate from BC have declined in the US since 1989 due to improved detection and treatment,(2) reducing incidence remains critical and requires a deeper understanding of BC etiology. While the International Agency for Research on Cancer (IARC) has classified adult obesity as a risk factor for postmenopausal BC,(3) some recent Mendelian randomization (MR) studies suggest that higher body mass index (BMI), may inversely associate with or protect against BC in this population.(4-7) MR studies utilize genetic instruments as proxies to eliminate the confounding factors in observational studies.(8) However, MR findings showing an inverse association of BMI on BC risk conflict with IARC conclusions. Mounting evidence suggests that these MR studies may be biased.(9, 10) Importantly, while MR studies are designed to reduce confounding, they remain vulnerable to selection bias —particularly when participant selection is conditional on surviving both the exposure and the outcome,(11) as illustrated in Figure .

Genetic studies of BC are particularly vulnerable to survival bias. While the incidence of BC is rising among young women and it remains a leading cause of cancer-related death in this group,(12) most genome-wide association studies (GWAS) investigating causes of BC, risk factors like high BMI, have largely recruited women of middle age or older with cases often averaging in the late 50s.(13, 14) Failure to fully ascertain all deaths in the study population, including those occurring before recruitment, creates the potential for survival bias where analysis is restricted to survivors of a specific exposure and outcome. This bias can attenuate or even reverse estimates of harmful exposures that began before study entry. Given that higher

BMI is associated with lower survival,(15) and most GWAS participants are in late middle age at enrollment, the inverse association of genetically predicted BMI with BC risk may reflect the exclusion of women who died from high-BMI or BC prior to the study.

To assess whether previous MR studies showing an inverse association of genetically predicted BMI with BC risk were biased, we examined BC in the siblings (sibling BC [BCs]) of women who participated in GWAS studies. We assumed that BC among siblings is less susceptible to selection bias arising from missing outcomes in women who die middle age. These concepts are illustrated in the directed acyclic graphs in Figure . Panel A shows that selection bias arises when both the exposure (e.g. high BMI) and the outcome-related morbidity or mortality (e.g. participant BC [BCp]) influence an individual's likelihood of being recruited into a GWAS study. This creates an open pathway: Genetic Variant → BMI → Selection (Survival to recruitment) ← BCp (illustrated in Red). Conversely, Panel B illustrates that because a BCs does not influence the participant's likelihood of recruitment, this model is less prone to selection bias.

To determine whether the inverse association of adult BMI with BC observed in MR studies stems from selection bias from survivorship bias, we addressed three interrelated questions using two-sample MR (where possible), applied to the largest publicly available GWAS summary data.

- First, we investigated the association between adult BMI—our exposure—and overall survival, defined as lifespan.
- Second, we examined the association between genetic liability to BC—our outcome—and lifespan. To assess this, we utilized three distinct BC GWAS

summary statistics: two based on BCp (from the Breast Cancer Association Consortium [BCAC] and UK Biobank) and a third using BCs as a proxy. We hypothesized that BCp GWAS, which often exclude individuals who died young due to high BMI or aggressive BC, would show a weaker association with lifespan compared to the BCs proxy.

- Third, we analyzed the association of adult BMI with BC, using the three different BC GWAS summary statistics to highlight the differences in results.

Methods

Data sources

Body mass index (BMI)

Genetic predictors of BMI were obtained from a female-specific GWAS conducted by the Genetic Investigation of ANthropometric Traits (GIANT) consortium, encompassing 171,977 white women of European descent.(16) This cohort is unlikely to overlap with UK Biobank participants.(17) To mitigate potential bias from population stratification, the GWAS was adjusted for principal components of ancestry.(16)

Lifespan

Expected lifespan for participants was assessed based on mother's attained age (age at death or current age, n=412,937) sourced from the UK Biobank's cohort of approximately 500,000 individuals recruited between 2006 to 2010. Using maternal attained age as a proxy for women participants' lifespan is supported by evidence of the heritability of longevity,(18) and the correlation between parental lifespan and daughters' life expectancy.(19) Furthermore, maternal attained age is an established proxy for all-cause mortality.(20) In the UK Biobank, analyses of

attained age excluded adopted participants and mothers who died prematurely before age 57.(21) The GWAS of maternal attained age utilized Martingale residuals from Cox proportional hazard models, adjusting for age, assessment center, and genotyping array.(21) The estimates of association were converted into life years lost using an established approximation.(22)

Breast cancer

Three distinct BC GWAS were used to first to assess the relationship of genetic liability to BC with lifespan, and then to examine the impact of BMI on BC risk.

1. BCAC: The Breast Cancer Association Consortium (BCAC) GWAS, which includes a mix of case-control studies and cohort studies, has been utilized in numerous MR studies.(23-25) This dataset includes 122,977 BC cases and 105,974 controls of European-descent.(14) Genetic associations within BCAC were adjusted for up to ten principal components of ancestry and country.(14)
2. UK Biobank BCp Data: This GWAS of BCp included 462,933 participants of European descent (10,303 cases; 452,630 controls). To address potential confounding by population stratification, the GWAS analysis was adjusted for genotyping array and the first ten principal components of ancestry.(26)
3. UK Biobank BCs: This GWAS included 361,809 individuals of European-descent, comprising 16,586 BCs cases and 345,233 controls.(26) To minimize potential confounding by population stratification, the analysis was adjusted for genotyping array and the first ten principal components of ancestry.(26) Prior studies indicate that family history of BC is reliably reported within medical settings.(27, 28)

All UK Biobank GWAS for BC --including BCp cases and BCs cases–were conducted using linear regression. Stringent quality control was applied to exclude individuals with mismatched

sex or high relatedness. To ensure comparability, estimates of associations derived from linear regression were converted from the probability scale to odds ratios (ORs) using established methods.(29)

Table 1 provides further details on the GWAS data sources, including sex, age range, and geographic region.

Genetic instruments selection

Genome wide significant (p-value $<5x10^{-8}$) and independent ($r^2<0.001$) genetic variants were selected as instrumental variables for each exposure: BMI, liability to BCp from BCAC and the UK Biobank, and liability to BCs from the UK Biobank. Instrument strength was evaluated using the F-statistic, calculated using the approximation of genetic variant on exposure divided by its variance; the average F-statistic is reported.(30) An F-statistic threshold of 10 was considered as adequate to indicate strength of an instrument.(31, 32) Genetic associations between the instruments and the outcomes were obtained from relevant GWAS. Instruments were aligned across studies on the effect allele, with palindromic variants having a minor allele frequency (MAF) >0.42 removed to ensure accurate alignment.

Statistical analysis

Two-sample MR methods were employed to evaluate the potential causal impact of exposures (BMI and liability to BC) on outcomes (lifespan and BC). This instrumental variable approach relies on three core assumptions: relevance, independence, and exclusion restriction. To satisfy the relevance assumption, independent and genome-wide significant ($P < 5 \times 10^{-8}$) genetic

variants were selected as instruments for each exposure. The independence assumption was supported by the randomized nature of genetic alleles, reducing potential confounding.(33)

Two-sample MR estimates were derived using the Wald ratio method (genetic association with outcome divided by genetic association with exposure). Variant-specific estimates were synthesized using inverse-variance weighting (IVW) with a multiplicative random-effects model.(34) To address potential violations of the exclusion restriction caused by horizontal pleiotropy or selection bias (e.g., survival bias),(10) we conducted sensitivity analyses, including weighted median (WM) and MR-Egger (MRE). The WM estimator provides a consistent causal estimate if at least 50% of the weight comes from valid instruments.(35) MRE regression was used to evaluate pleiotropy with a significant intercept term suggesting possible invalidity of the main inverse-variance weighted (IVW) estimate.(30)

Given that GWAS of first-degree relatives (e.g., siblings) involve individuals sharing only 50% of their DNA, the genetic effect estimates of BMI on BCs were doubled, and their corresponding variance were quadrupled for the MR analysis.

Statistical heterogeneity

Statistical heterogeneity was assessed using the Cochran's Q (χ^2) and I^2 statistics.(30) Heterogeneity was considered substantial if the Cochrane's Q test yielded a p<0.10 or if $I^2 > 60\%$. Separately, we evaluate the I^2_{GX} statistics to assess the impact of no measurement error violation; in this context higher I^2_{GX} values indicate greater validity IVW estimates(30) particularly in two-sample MR, where I^2_{GX} greater than 97% is desirable.(36)

TwoSampleMR 0.5.6 R package was used to extract genetic instruments and align genetic variants across studies,(26) the MendelianRandomization R package was then used to obtain the causal effect estimates. All analyses were conducted using R (version 4.1.2).(37)

ETHICS

This study only uses genetic summary statistics created from information and materials previously collected with informed consent. City University of New York School of Public Health Human Research Protection Program reviewed the study proposal and deemed the study not a human subject research study. Thus, institutional review board approval was exempt for this study.

Results

BMI on lifespan

Higher BMI was strongly inversely associated with lifespan in women (**Table 2**). Each genetically predicted standard deviation (s.d.) increase in BMI was associated with a decrease of -1.50 (95% confidence interval (CI): -2.30, -0.70) years of life. The MR estimates were consistent across all sensitivity analyses, the F-statistics were adequate, and the MR-Egger intercept p-value of 0.70 showed no indication of instrument pleiotropy (a scenario where genetic variants influence lifespan through pathways other than BMI) (**Table 2**). This suggests that higher BMI is genetically linked to shorter lifespan in the studied female population.

Liability to breast cancer on lifespan

The estimated life years lost due to liability to BC using the BCAC GWAS was -0.53 (95% CI: -0.74, -0.31) years. The estimated life years lost due to liability to BCp from UK Biobank was -

1.12 (95% CI: -1.74, -0.51). The estimated life years lost due to liability to BCs was -1.74 (95% CI: -2.46, -1.03) years. The F-statistics were adequate. Sensitivity analysis gave a similar interpretation, with no evidence of genetic instrument pleiotropy from the non-significant MR-Egger intercepts (**Table 3**).

BMI on breast cancer

Genetically predicted higher BMI was inversely associated with BC risk in the BCAC consortium data (OR=0.74, 95% CI: 0.64, 0.84), and with BCp risk among UK Biobank participants (OR=0.70, 95% CI: 0.56, 0.87). Conversely, genetically predicted BMI showed no association with BCs risk within the UK Biobank (OR=1.00, 95% CI: 0.99, 1.01) (**Table 4**). The F-statistics were adequate. Sensitivity analysis revealed a significant MR-Egger intercept (p=0.007), indicated that the BMI-BC risk estimates from BCAC may be invalid (**Table 4**). This suggests the presence of genetic instrument pleiotropy or potential selection bias.

Discussion

In this MR study, we found no association of BMI with BC risk when using BCs risk as a proxy outcome, which is likely less affected by survival bias than BCp. The novel aim of our study was to investigate potential selection bias MR analyses of BMI and BC. Such bias can arise in GWAS, as participant recruitment usually occurs in late middle age, after deaths of early onset BC or high BMI have already occurred. To address this, we compared results using BCp (which exclude individual who died before recruitment), against BCs, which provides a more complete and less biased ascertainment of cases.

Consistent with previous MR studies, we observed an inverse association between higher BMI and BC risk using GWAS of BCp.(4-7) Among the three GWAS of liability to BC, BCs exhibited the strongest relationship with lifespan, suggesting lower susceptibility to survival bias compared to the other BC GWAS. Furthermore, sensitivity analysis indicated that the estimated association between BMI and BC right from BCAC GWAS might be invalid. This discrepancy is likely due to Neyman bias whereby exclusion of extremely sick individuals who have already died results in an underestimation of disease severity and a weak association with lifespan. Consequently, the BCAC GWAS finding regarding BMI and BC risk may be biased. These findings suggest selection bias may arise when examining exposures and outcomes that result in early mortality prior to study recruitment.(38)

Such an inverse relationship is biologically implausible. Strong evidence links high BMI (obesity) to increased risk and progression of breast cancer, particularly in postmenopausal women, through estrogen production, chronic inflammation, and insulin signaling.(39-42) Excess adipose tissue acts as an active endocrine organ, producing excessive aromatase, inducing inflammation through cytokines (IL-6, TNF-α), and promoting cancer cell proliferation, invasion, and tumor microenvironment remodeling.

This study is unique in its explicit consideration of selection bias arising from premature death related to both BMI and BC, distinguishing it from previous literature. Because selection on both exposure and outcome creates unrecoverable bias,(43) our analysis accounts for the fact that inclusion was conditioned on surviving both genetically predicted adult BMI and BC to reach recruitment in late middle-age (~57 years). Given our findings that both higher BMI and BC significantly reduce lifespan, we believe this “survivor bias” represents an important, yet previously overlooked, source of bias. Current solutions to this bias are largely limited to

sensitivity analysis,(44) whose parameters are difficult to estimate, or to models utilizing continuous outcomes.(45). Instead, we propose using sibling disease as a proxy outcome, which, due to its more complete ascertainment, is less susceptible to survival bias. Applying this approach, we found results conflicting with previous studies relied on BCp. Similar considerations regarding accounting for early death should be taken when assessing relationship between harmful exposures and outcomes, particularly when such deaths are not accounted for in GWAS.

Beyond addressing selection bias in prior Mendelian Randomization (MR) studies regarding BMI and breast cancer, this research serves as a model for investigating selection bias in other MR analyses of harmful exposures and outcomes. Special attention is required when both the exposure and the outcome are detrimental, as these factors can significantly influence biobank participation. By utilizing BCs as a proxy outcome to demonstrate bias, this study illustrates a methodology that can be applied in other contexts if suitable proxies are available.

Emerging evidence indicates that GWAS of ratio traits, such as BMI, are often denominator (height)-driven, leading to potentially spurious genetic associations that reflect stature more than adiposity.(46) This phenomenon may explain the paradoxical null or inverse relation observed between genetically predicted BMI and breast cancer in MR studies, which contradicts established observational findings. Furthermore, this phenomenon, likely acting in conjunction with selection/collider bias in prior MR studies, highlights the need for careful interpretation of genetic studies involving ratio phenotypes. Finally, the validity of these MR estimates depends on the gene-environment equivalence (GEE) assumption; however, satisfying GEE for BMI remains challenging, as genetic proxies may not perfectly recapitulate the timing or biological mechanisms of environmental BMI changes. Future research requires more comprehensive

GWAS recruitment and rigorous GEE evaluation.

The UK Biobank protocol specifies that with a 10% exposure prevalence, 5,000 cases (1% of the cohort) are required to detect a minimum odds ratio of 1.26 at a significance level of $\alpha = 10^{-4}$.(47) Given the substantial case count in the BCs cohort and the high instrument strength, this study likely possesses sufficient statistical power. However, these null findings may be influenced by potential selection bias, which could reduce statistical power and mask a true effect. Future work should investigate this, focusing specifically on the potential lack of power and/or selection bias.

Strengths and Limitations

While MR studies offer the advantage of providing less confounded estimates compared to other observational designs, several limitations remain. First, BCp and family history (BCs) are subject to measurement error. Such non-differential measurement error typically biases estimates towards the null hypothesis of no association. Second, the GIANT consortium data (mean age ≈ 56 years) may exclude individuals who died prematurely due to high BMI potentially leading to selection bias that attenuate the association. This underscores the need for more comprehensive recruitment in underlying GWAS studies. Third, UK Biobank BCp could not be analyzed by sub-type, as this information was not recorded, limiting our ability to examine the specific impact of BMI on distinct BC subtypes. Fourth, MR as instrumental variable analysis requires satisfying the core assumptions of relevance, independence, and exclusion-restriction. Relevance was ensured by using only genome-wide significant genetic instruments. To address potential violations of the exclusion restriction, we employed sensitivity analysis for genetic pleiotropy and utilizing BCs as an outcome to mitigate selection bias. Furthermore, genetic variants are

generally less strongly associated with potential confounders than observed exposures.(33) Fifth, the assessment of UK Biobank BC on lifespan is a one-sample MR study, which could potentially bias estimates towards confounded values; however, the large study size and high I^2_{GX} suggest this effect is minimal. Furthermore, as shown in previous studies, the impact on MR estimate might not be substantial.(48) Sixth, the use of two-sample MR methods assumes that the associations between exposure and outcome are linear.

Conclusion

Previous MR studies suggesting an inverse association between adult BMI with BC risk may be driven by selection bias. Supporting this, our analysis found no association between BMI and BCs risk, highlighting potential bias in earlier studies. Our findings are consistent with mounting evidence from IARC(3) and the 2018 World Cancer Research Fund (WCRF)/ American Institute for Cancer Research (AICR) reports, which indicate that adult weight gain and excess body fat increase the risk of post-menopausal breast cancer.(49) These findings highlight a critical limitation in GWAS and MR studies, particularly when both exposure and outcome are harmful and influence enrollment into the underlying study. The applicability of our proposed method to other MR studies depend on the availability of a suitable proxy outcome. Crucially, given the established risk of higher BMI for chronic diseases and mortality, null or inverse associations of adult BMI and BC risk should not be interpreted as suggesting a lack of a causal role, pending further evidence free from confounding and selection bias.

Data Availability

The datasets analyzed in this study were obtained from publicly available repositories and large-scale consortia. The specific sources, persistent web links, and unique database identifiers for each dataset are as follows:

- Locke (2015) / GIANT Consortium: Genome-wide association study (GWAS) summary statistics for body mass index (BMI) in females are available via the Broad Institute GIANT Consortium Data Files repository under dataset ID ieu-a-974.
- Michailidou (2017) / BCAC Consortium: Breast cancer GWAS summary statistics (including cases and controls) are accessible via the Centre for Cancer Genetic Epidemiology Breast Cancer Association Consortium portal under dataset ID ieu-a-1126.
- Elsworth (2018) / MRC-IEU: The participant breast cancer dataset consisting of 10,303 cases and 452,630 controls is archived in the Medical Research Council Integrative Epidemiology Unit OpenGWAS Database under dataset ID ukb-b-16890.
- Elsworth (2018) / MRC-IEU: The sibling breast cancer dataset consisting of 16,586 cases and 345,223 controls is archived in the Medical Research Council Integrative Epidemiology Unit OpenGWAS Database under dataset ID ukb-b-12227.
- Pilling (2017): Summary statistics relating to mother's attained age are hosted within the UK Biobank Resource Platform repository under dataset ID ebi-a-GCST006696.

No new primary data were generated during the current study.

References

1. Rahib L, Wehner M, Matrisian LM, Nead KT. Projection of cancer incidence and death to 2040 in the US: Impact of cancer screening and a changing demographic. Journal of Clinical Oncology. 2020;38(15_supply):1566-.
2. DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, et al. Breast cancer statistics, 2019. CA Cancer J Clin. 2019;69(6):438-51.
3. Lauby-Secretan B, Scoccianti C, Loomis D, Grosse Y, Bianchini F, Straif K. Body Fatness and Cancer--Viewpoint of the IARC Working Group. N Engl J Med. 2016;375(8):794-8.
4. Guo Y, Warren Andersen S, Shu XO, Michailidou K, Bolla MK, Wang Q, 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. 2016;13(8):e1002105.
5. Shu X, Wu L, Khankari NK, Shu XO, Wang TJ, Michailidou K, et al. Associations of obesity and circulating insulin and glucose with breast cancer risk: a Mendelian randomization analysis. Int J Epidemiol. 2019;48(3):795-806.
6. Amin HA, Kaewsri P, Yiorkas AM, Cooke H, Blakemore AI, Drenos F. Mendelian randomisation analyses of UK Biobank and published data suggest that increased adiposity lowers risk of breast and prostate cancer. Scientific Reports. 2022;12(1):909.
7. Hao Y, Xiao J, Liang Y, Wu X, Zhang H, Xiao C, et al. 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. 2022.
8. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32(1):1-22.
9. Smit RAJ, Trompet S, Dekkers OM, Jukema JW, le Cessie S. Survival Bias in Mendelian Randomization Studies: A Threat to Causal Inference. Epidemiology. 2019;30(6):813-6.
10. Schooling CM, Lopez PM, Yang Z, Zhao JV, Au Yeung SL, Huang JV. Use of Multivariable Mendelian Randomization to Address Biases Due to Competing Risk Before Recruitment. Front Genet. 2020;11:610852.
11. Boef AG, le Cessie S, Dekkers OM. Mendelian randomization studies in the elderly. Epidemiology. 2015;26(2):e15-6.
12. Fidler MM, Gupta S, Soerjomataram I, Ferlay J, Steliarova-Foucher E, Bray F. Cancer incidence and mortality among young adults aged 20-39 years worldwide in 2012: a population-based study. Lancet Oncol. 2017;18(12):1579-89.
13. Hewitt J, Walters M, Padmanabhan S, Dawson J. Cohort profile of the UK Biobank: diagnosis and characteristics of cerebrovascular disease. BMJ Open. 2016;6(3):e009161.
14. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92-4.
15. Bhaskaran K, Dos-Santos-Silva I, Leon DA, Douglas IJ, Smeeth L. Association of BMI with overall and cause-specific mortality: a population-based cohort study of 3.6 million adults in the UK. Lancet Diabetes Endocrinol. 2018;6(12):944-53.
16. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197-206.

17. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. Hum Mol Genet. 2018;27(20):3641-9.
18. Gudmundsson H, Gudbjartsson DF, Frigge M, Gulcher JR, Stefánsson K. Inheritance of human longevity in Iceland. Eur J Hum Genet. 2000;8(10):743-9.
19. Vagero D, Aronsson V, Modin B. Why is parental lifespan linked to children's chances of reaching a high age? A transgenerational hypothesis. SSM Popul Health. 2018;4:45-54.
20. Pilling LC, Kuo CL, Sicinski K, Tamosauskaite J, Kuchel GA, Harries LW, et al. Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. Aging (Albany NY). 2017;9(12):2504-20.
21. Pilling LC, Tamosauskaite J, Jones G, Wood AR, Jones L, Kuo CL, et al. Common conditions associated with hereditary haemochromatosis genetic variants: cohort study in UK Biobank. Bmj. 2019;364:k5222.
22. Timmers PR, Mounier N, Lall K, Fischer K, Ning Z, Feng X, et al. Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. Elife. 2019;8:e39856.
23. Vabistsevits M, Davey Smith G, Sanderson E, Richardson TG, Lloyd-Lewis B, Richmond RC. Deciphering how early life adiposity influences breast cancer risk using Mendelian randomization. Commun Biol. 2022;5(1):337.
24. Richardson TG, Sanderson E, Elsworth B, Tilling K, Davey Smith G. Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study. Bmj. 2020;369:m1203.
25. Ruth KS, Day FR, Tyrrell J, Thompson DJ, Wood AR, Mahajan A, et al. Using human genetics to understand the disease impacts of testosterone in men and women. Nat Med. 2020;26(2):252-8.
26. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7:e34408.
27. Augustinsson A, Ellberg C, Kristoffersson U, Borg A, Olsson H. Accuracy of self-reported family history of cancer, mutation status and tumor characteristics in patients with early onset breast cancer. Acta Oncol. 2018;57(5):595-603.
28. Ziogas A, Anton-Culver H. Validation of family history data in cancer family registries. Am J Prev Med. 2003;24(2):190-8.
29. Lloyd-Jones LR, Robinson MR, Yang J, Visscher PM. Transformation of Summary Statistics from Linear Mixed Model Association on All-or-None Traits to Odds Ratio. Genetics. 2018;208(4):1397-408.
30. Bowden J, Del Greco MF, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. International Journal of Epidemiology. 2016;45(6):1961-74.
31. Staiger D, Stock JH. Instrumental Variables Regression with Weak Instruments. Econometrica. 1997;65(3):557-86.

32. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med. 2008;27(8):1133-63.
33. Smith GD, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. PLoS Med. 2007;4(12):e352.
34. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol. 2013;37(7):658-65.
35. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet Epidemiol. 2016;40(4):304-14.
36. Minelli C, Del Greco MF, van der Plaat DA, Bowden J, Sheehan NA, Thompson J. The use of two-sample methods for Mendelian randomization analyses on single large datasets. Int J Epidemiol. 2021;50(5):1651-9.
37. R Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020.
38. Schooling CM, Fei K, Terry MB. Reassessing the causal role of early-life adiposity in breast cancer: could the apparent inverse associations be a manifestation of survival bias? Int J Epidemiol. 2023;52(4):1292-3.
39. Simone V, D'Avenia M, Argentiero A, Felici C, Rizzo FM, De Pergola G, et al. Obesity and Breast Cancer: Molecular Interconnections and Potential Clinical Applications. Oncologist. 2016;21(4):404-17.
40. Zhao C, Hu W, Xu Y, Wang D, Wang Y, Lv W, et al. Current Landscape: The Mechanism and Therapeutic Impact of Obesity for Breast Cancer. Front Oncol. 2021;11:704893.
41. Gershuni VM, Ahima RS, Tchou J. Obesity and Breast Cancer: A Complex Relationship. Curr Surg Rep. 2016;4.
42. Hu JJ, Zhang QY, Yang ZC. The correlation between obesity and the occurrence and development of breast cancer. Eur J Med Res. 2025;30(1):419.
43. Mohan K, Pearl J. Graphical Models for Processing Missing Data. JASA. 2021;116(534):1023-37.
44. Smith LH, VanderWeele TJ. Bounding Bias Due to Selection. Epidemiology. 2019;30(4):509-16.
45. Liu Z, Ye T, Sun B, Schooling M, Tchetgen Tchetgen E. On Mendelian Randomization Mixed-Scale Treatment Effect Robust Identification (MR MiSTERI) and Estimation for Causal Inference. arxir. 2020.
46. McCaw ZR, Dey R, Somineni H, Amar D, Mukherjee S, Sandor K, et al. Pitfalls in performing genome-wide association studies on ratio traits. HGG Adv. 2025;6(2):100406.
47. Biobank U. UK Biobank: Protocol for a large-scale prospective epidemiological resource (AMENDMENT ONE FINAL) https://www.ukbiobank.ac.uk/wp-content/uploads/2025/01/Main-study-protocol.pdf2007 [
48. Jiang T, Gill D, Butterworth AS, Burgess S. An empirical investigation into the impact of winner's curse on estimates from Mendelian randomization. Int J Epidemiol. 2023;52(4):1209-19.

49. World Cancer Research Fund / American Institute for Cancer Research. Continuous Update Project Expert Report 2018. Diet, nutrition, physical activity and breast cancer. dietandcancerreport.org.
50. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92-4.
51. Elsworth B, Lyon M, Alexander T, Liu Y, Matthews P, Hallett J, et al. The MRC IEU OpenGWAS data infrastructure. bioRxiv. 2020:2020.08.10.244293.
52. Pilling LC, Kuo C-L, Sicinski K, Tamosauskaite J, Kuchel GA, Harries LW, et al. Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. Aging. 2017;9(12):2504-20.

Table 1: Description of data sources

Variable	First author (Year)	Consortium	Sample Size	Population	Sex	Age range
Body mass index	Locke(16) (2015)	GIANT	171,977	95% European	Females	40-69 years
BCp	Michailidou(50) (2017)	BCAC	Cases=122,977; Controls=105,974	European	Females	58-66 years
BCp	Elsworth(51) (2018)	MRC-IEU	Cases=10,303; Controls=452,630	European	Females and males	40-69
BCs	Elsworth(51) (2018)	MRC-IEU	Cases=16,586; Controls=345,223	European	Females and males	40-69
Mother's attained age	Pilling(52) (2017)	NA	412,937	European	Females and males	40-69

Table 2: Mendelian Randomization estimates in women for body mass index (BMI) on life years in the UK Biobank

Exposure						Outcome – Life years (ebi-a-GCST006696)			
Units	Subject	Description	n	Method	Instrument #s (mean f-statistic)	Life years [1]	95% CI	p-value	MR-Egger intercept p-value, Q(p), I^2_{GX}
Female specific BMI per standard deviation	Self	Body mass index https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	1719 77	IVW	36 (67.1)	-1.50	-2.30 to -0.70	2.5e-04	
				WM		-1.48	-2.42 to -0.53	0.002	
				MRE		-1.10	-3.29 to 1.09	0.333	0.70, 68.8 (0), 87.5 %

IVW = Inverse variance weight; WM = Weighted median; MRE = MR-Egger.

[1] Years of life obtained from the log protection ratio multiplied by 2.5863 to account for children only sharing half their genetic endowment with their mothers and then by 10 as a verified actuarial rule of thumb (20)

Table 3: Mendelian Randomization estimates in women for liability to BCp from BCAC and from UK Biobank, and BCs from UK Biobank on mother's life years in the UK Biobank

Exposure						Outcome – Life years (ebi-a-GCST006696)			MR-Egger intercept p-value, Q(p), I^2_{GX}
Units	Subject	Description	Cases/participants	Instrument #s (mean f-statistic)	Method	Life years[3]	95% CI	p-value	
Liability to breast cancer in log odds	Participant	BCAC (ieu-a-1126)	122977/228951	122 (87.1)	IVW	-0.53	-0.74 to -0.31	2.8E-06	
					WWM	-0.43	-0.72 to -0.13	0.005	
					MRE	-0.61	-1.10 to -0.12	0.014	0.70, 193 (0), 93.1%
	Participant	UK Biobank (ukb-b-16890)	10303/462933[1]	17 (78.0)	IVW	-1.12	-1.74 to -0.51	0.0004	
					WWM	-1.16	-1.83 to -0.49	0.001	
					MRE	-1.75	-3.34 to -0.17	0.03	0.40, 31.4 (0.01), 90.7%
	Sibling	UK Biobank (ukb-b-12227)	16586/361809[2]	6 (49.2)	IVW	-1.74	-2.46 to -1.03	1.9E-06	
					WWM	-1.65	-2.47 to -0.83	7.9E-5	
					MRE	0.21	-2.63 to 3.05	0.89	0.17, 4.8 (0.31),

	59.2 %

IVW = Inverse variance weight; WM = Weighted median; MRE = MR-Egger.

[1] Given the self-reports included men, the denominator was halved when converting estimates of breast cancer obtained from linear regression to log odds for comparability

[2] Conservatively assumed that each person who reported on a sibling had one female sibling when converting estimates of breast cancer obtained from linear regression to log odds for comparability**.**

[3] Years of life obtained from the log protection ratio multiplied by 2.5863 to account for children only sharing half their genetic endowment with their mothers and then by 10 as a verified actuarial rule of thumb (20)

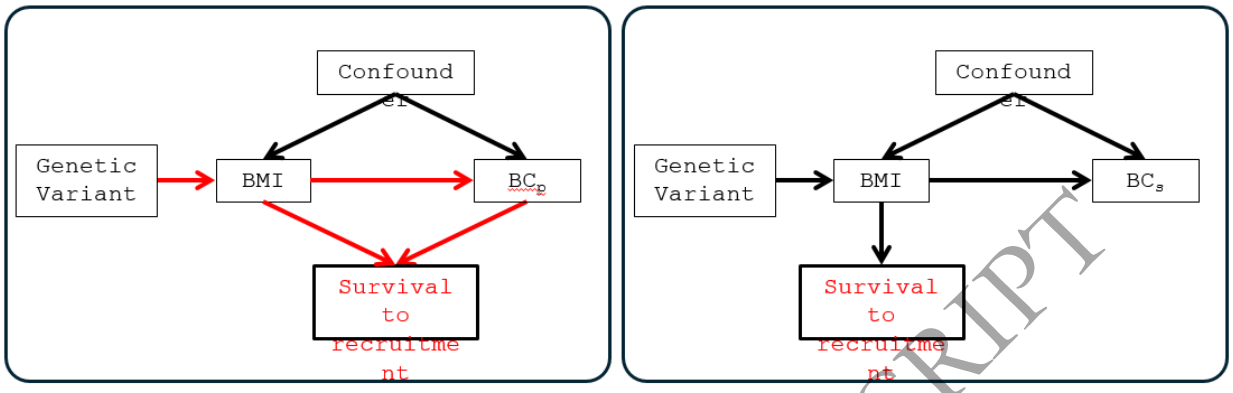
Table 4: Mendelian Randomization estimates in women for female specific BMI (n=111977) on BCp from BCAC and UK Biobank and BCs from UK Biobank

Exposure			Breast cancer Outcome for						
Units	Description	Instruments #s (mean f-statistic)	Subject	Study used	Method	Odds ratio	95% CI	p-value	MR-Egger intercept p-value, Q(p), I^2_{GX}
Female specific BMI per standard deviation	Body mass index https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	36 (65.5)	BCp	BCAC (ieu-a-1126)	IVW	0.74	0.64 to 0.84	6.3e-6	
					WM	0.73	0.65 to 0.83	3.9e-7	
					MRE	0.48	0.35 to 0.67	1.4e-4	0.007, 124 (0), 86.6%
		36 (66.4)	BCp[1]	UK Biobank (ukb-b-16890)[1]	IVW	0.70	0.56 to 0.87	0.002	
					WM	0.82	0.63 to 1.06	0.125	
					MRE	0.29	0.12 to 0.73	0.010	0.09, 60 (0.04), 87.2%

	37 (65.7)	BCs[2]	UK Biobank (ukb-b-12227)[2]	IVW	1.00	0.99 to 1.01	0.87	
				WM	0.99	0.98 to 1.00	0.20	
				MRE		0.95 to 1.02	0.40	0.39, 235 (0), 87.1 %

IVW = Inverse variance weight; WM = Weighted median; MRE = MR-Egger.
[1] Given the UK Biobank BCp included men, the denominator was halved when converted estimates of breast cancer obtained from linear regression to log odds for comparability
[2] Conservatively assumed that each person who reported on a sibling had one female sibling when converting estimates of breast cancer obtained from linear regression to log odds for comparability**.**

Figure 1. Directed acyclic graph (DAG) illustrating the presence or absence of selection bias for a harmful exposure such as high BMI across different types of outcomes



Panel A. A DAG showing selection bias may arise due to survival (to recruitment) caused by both the exposure (e.g. high BMI) and outcome (e.g. BCp)

Panel B. A DAG showing that selection bias cannot arise when the outcome is BC_s, since sibling survival does not affect recruitment of the participant

BMI=body mass index; BCp=participant breast cancer; BCs=sibling breast cancer