Supplementary Information for

MRCAT: A multi-agent framework for reliable Mendelian randomization analysis and causal interpretation

Yuxin Jiang, Fan Zhang, Xinrui Huang, Zhiwei Wang, Xianghong Hu, and Can Yang

Contents

1 Supplementary Methods and Results **2**
1.1 Data sources and benchmark inputs . . . 2
1.1.1 Source composition of the 30-query benchmark . . . 2
1.1.2 Online GWAS resources . . . 2
1.1.3 User-provided files and negative-control phenotypes . . . 2
1.2 MR analysis procedures . . . 3
1.2.1 Data acquisition . . . 3
1.2.2 Data selection . . . 3
1.2.3 Quality control . . . 4
1.2.4 Harmonization . . . 4
1.2.5 IV selection . . . 5
1.2.6 MR estimation . . . 5
1.2.7 Sensitivity analyses . . . 6
1.3 Reporting case: hexadecanedioate and coronary artery disease . . . 6
1.3.1 Complete scientific report . . . 7
1.4 Supporting records for negative-control interpretations . . . 17
1.5 Data and estimates for the downstream applications . . . 17
1.6 Supporting records for execution completion . . . 17

2 Supplementary Figures **18**
2.1 30-query benchmark . . . 18
2.2 Risk-factor assessment in rheumatoid arthritis . . . 19
2.3 Bidirectional analysis of type 2 diabetes and coronary artery disease . . . 28

3 Supplementary Tables **30**
3.1 MRCAT Skill library . . . 30
3.2 30-query benchmark . . . 33
3.3 Comparisons with expert-curated analyses and negative-control analyses . . . 42
3.4 Traceable scientific reporting . . . 51
3.5 Risk-factor assessment and bidirectional analysis . . . 54

4 Supplementary References **58**

1 Supplementary Methods and Results

1.1 Data sources and benchmark inputs

1.1.1 Source composition of the 30-query benchmark

The 30-query benchmark combined four online GWAS resources with user-provided GWAS summary-statistics files. Classification by the source specified in each request or implied by its study-identifier namespace yielded seven OpenGWAS queries, five GWAS Catalog queries, five Pan-UKB queries and three FinnGen queries. Seven queries supplied both exposure and outcome as local files, whereas three specified traits without fixing a data source. Five of the queries assigned to an online resource also supplied one local GWAS file; thus, 12 of the 30 queries involved local inputs. These counts describe the requested data routes, including boundary cases, rather than successful downloads. Study identifiers hosted by OpenGWAS were counted under OpenGWAS even when the underlying study originated from FinnGen or GWAS Catalog. Together, the queries covered phenotype-based discovery, retrieval of specified studies, analysis of local data and integration of local and remotely retrieved summary statistics.

1.1.2 Online GWAS resources

OpenGWAS (`https://opengwas.io/`) aggregates GWAS summary statistics from multiple studies through standardized formats and programmatic interfaces [1]. Its broad trait coverage and integration with MR tools facilitate access to exposure and outcome data. In its GWAS-VCF files, effect estimates refer to the alternative allele, and the LP field stores negative log-transformed *P* values, conventions that can be misinterpreted during data extraction.

The NHGRI–EBI GWAS Catalog (`https://www.ebi.ac.uk/gwas/`) curates GWAS studies, reported associations and publication metadata, and hosts full summary statistics for selected studies [2]. Its study-level records support targeted searches for published traits and datasets using stable accessions. However, a catalogue entry may contain only reported associations, so identifying a relevant study does not necessarily mean that full summary statistics are available.

FinnGen (`https://www.finngen.fi/en/access_results`) links genetic data from Finnish participants to longitudinal health registries, providing GWAS summary statistics for numerous disease endpoints [3]. Its registry-based phenotypes provide a broad source of clinically defined outcomes for investigating disease risk. The same endpoint can appear in multiple releases with different sample sizes and association estimates, making the release identifier essential for distinguishing datasets.

Pan-UKB (`https://pan.ukbb.broadinstitute.org/`) provides GWAS summary statistics across six ancestry groups in UK Biobank using a common analysis framework [4]. Its diverse phenotypes and ancestry-specific results support analyses of behavioural, anthropometric, biomarker and disease traits within defined populations. Individual files can contain both ancestry-specific and cross-ancestry meta-analysis results, creating a risk of combining effect estimates, standard errors and sample sizes from different analysis groups.

Its ancestry-specific results require explicit population selection; UK Biobank membership does not itself imply European ancestry. These resources differ in phenotype definitions, ancestry composition, cohort overlap and data-access format, which are considered during study selection.

1.1.3 User-provided files and negative-control phenotypes

Local inputs are GWAS summary-statistics files supplied by the user, independently of whether the underlying study is publicly available. In the 30-query benchmark, seven requests supplied two local files and five combined a local file with an online study. These inputs included quantitative, behavioural and disease phenotypes, with examples spanning pigmentation, childhood traits, cardiometabolic traits and ageing-related measures. Plain-text and compressed files differed in column names, effect representations and sample-size annotations, requiring study-level inspection and quality control before analysis. Local input support there-

fore enables the use of custom or previously downloaded datasets, including studies that cannot be retrieved through the supported online interfaces.

The evaluation against expert-curated analyses also used local summary statistics, including the BMI–type 2 diabetes example distributed with MR-APSS and six negative-control pairs assembled from datasets identified through the MRbenchmarking study by Hu et al. (2024). The six pairs comprised 12 phenotypes: BMI and dark brown hair; insomnia and black hair; automobile speeding propensity and childhood aggression; parental survival and childhood BMI; frailty index and fetal birth weight; and worrier/anxious feelings and ease of skin tanning. Hair colour and tanning phenotypes provide controls sensitive to residual population structure, whereas the adult- or later-life traits paired with childhood or fetal outcomes challenge the temporal plausibility of the proposed exposure-to-outcome direction. These pairs were evaluated against predefined expert-curated null conclusions. Their role was to test whether the analysis and interpretation respected those expert-curated expectations; the expert-curated labels do not independently establish the absence of every possible genetic or biological relationship between the traits.

For the MR-APSS example [5], the exposure dataset was the UK Biobank BMI GWAS identified as f.21001.0.0 [6], and the outcome dataset was the type 2 diabetes meta-analysis reported by Xue et al. (2018) [7]. The original GWAS sources and download links for all seven exposure–outcome pairs are provided in Supplementary Table 8.

1.2 MR analysis procedures

The following procedures describe the supported data routes and current Skill defaults. Explicit analysis settings can override supported parameters; these defaults should not be assumed to describe every historical benchmark run.

1.2.1 Data acquisition

Data acquisition is implemented by the `gwas-search` and `gwas-download` Skills. For each target phenotype, the search preserves the requested trait definition, study identifier, source and ancestry constraints. Explicit identifiers are resolved within their source namespace: for example, OpenGWAS-hosted identifiers retain their complete prefixes, whereas bare GCST accessions are searched in GWAS Catalog. Phenotype-based searches use trait names, abbreviations, synonyms and available ontology terms; related phenotypes may be returned as discovery candidates but are not treated as equivalent studies without selection. Searches are restricted to an explicitly requested source, or can span supported resources when the request leaves the source open. Candidate records retain study metadata and evidence of summary-statistics availability. Full files are downloaded from the selected study's confirmed download descriptor or a user-provided URL, with source metadata retained for downstream processing. User-provided local files enter the workflow at quality control, while requests eligible for direct OpenGWAS analysis can retrieve the required variant associations without downloading full genome-wide files.

1.2.2 Data selection

The `gwas-select` Skill selects a compatible study set for the requested analysis. Selection first respects explicit study identifiers and source restrictions and checks whether the phenotype definition matches the scientific question. Disease subtypes, complications, biomarkers and related traits are not interchangeable solely because their names or associations are similar. Among eligible candidates, selection considers ancestry compatibility, sample size, metadata completeness, data availability and consistency across the exposure and outcome datasets. Study ancestry is retained from documented metadata rather than inferred from the location of a cohort. Potential sample overlap is considered separately from ancestry matching; different study identifiers do not establish independent participants. When independent replication is requested, cohort and participant provenance are needed to support that requirement. If no acceptable study set is available, the

workflow records the unmet constraint and requests clarification rather than silently replacing the requested phenotype or relaxing a fixed source.

1.2.3 Quality control

Quality control produces a set of SNPs with usable GWAS summary statistics for instrument selection and MR analysis. Each GWAS is processed separately, following the QC procedures described in MR benchmarking and MR-APSS [5, 8]. The following checks are applied:

- **Missing or invalid information.** Each SNP must have an identifier, effect and other alleles, a sample size, and association statistics sufficient to obtain a signed Z score and P value. SNPs with incomplete required information are removed. Available effect estimates must be finite, and standard errors must be positive.
- **Duplicate SNPs.** All records with a repeated rsID are removed to avoid including the same variant more than once.
- **Ambiguous alleles.** SNPs with identical alleles or palindromic pairs (A/T or C/G) are excluded. Palindromic pairs cannot be distinguished from their strand complements using allele labels alone.
- **Imputation quality.** When INFO scores are available, SNPs with observed INFO ≤ 0.9 are excluded to remove poorly imputed variants. Missing INFO values do not trigger this filter.
- **Allele frequency.** Low-frequency variants are excluded because their association estimates may be less reliable. The default filter retains SNPs with an observed allele frequency f satisfying $0.05 < f < 0.95$; missing frequencies do not trigger this filter.
- **HapMap3 membership.** Analysis is restricted to the HapMap3 reference set of common, well-imputed SNPs, including when frequency or INFO data are unavailable. Both the rsID and allele pair must match the reference, allowing allele exchange and strand complements.
- **MHC region.** SNPs in the major histocompatibility complex are excluded using the reference MHC SNP list because of the region's complex LD structure.
- **Sample-size outliers.** SNPs with sample sizes more than five standard deviations from the mean are removed to exclude unusually large or small per-variant sample sizes.
- **Extreme association statistics.** SNPs with $Z^2 \geq \max(80, \text{median}(N)/1000)$ are removed to limit the influence of extreme associations, where N denotes per-variant sample size.

The filtered table contains rsIDs, effect and other alleles, Z scores, sample sizes and P values. Available original effect estimates and standard errors are also retained. This table is used to match the exposure and outcome datasets in the next step.

1.2.4 Harmonization

For each exposure–outcome pair, SNP effects must refer to the same allele before they can be combined in MR analysis [8]. The two QC-filtered datasets are matched by rsID, retaining SNPs present in both. Allele alignment accounts for strand differences first and effect-allele differences second. If the studies report opposite strands, the outcome alleles are replaced by their complements. If the outcome effect allele then corresponds to the exposure's other allele, the outcome effect is multiplied by -1.

For example, consider a SNP reported as effect/other alleles A/G in the exposure and C/T in the outcome. Strand complementation changes the outcome labels to G/A. The outcome effect therefore refers to G, whereas

the exposure effect refers to A, so the outcome effect sign is reversed. The same sign change is applied to the outcome Z score, while its standard error remains unchanged. Palindromic SNPs have already been excluded during QC. The resulting dataset contains exposure and outcome associations aligned to the exposure effect allele.

1.2.5 IV selection

Instrument selection starts from the harmonized exposure–outcome data. SNPs are first selected according to their association with the exposure, using a default P-value threshold of 5×10^{-5} for shared data preparation. LD clumping then reduces correlation among the selected SNPs, retaining those with stronger exposure associations within correlated groups. The default clumping settings are an r^2 threshold of 0.001 and a 10,000-kb window. Clumping uses either a local PLINK reference panel or a population reference provided through OpenGWAS; the reference used is recorded with the analysis. The final instrument table contains the selected SNPs and their aligned exposure and outcome association statistics for MR estimation. Individual estimators can apply stricter association thresholds and their own minimum instrument requirements, as described below.

Selection follows harmonization so that all candidate instruments have outcome associations available [8]. If clumping were performed on the exposure data first, a retained lead SNP could be absent from the outcome data even when other SNPs in its LD group were available. Matching the datasets before selection reduces this source of instrument loss.

1.2.6 MR estimation

MR estimation follows the input requirements of each method-specific Skill. The `mr-primary` Skill applies inverse-variance weighting [9], MR-Egger [10], weighted median [11] and weighted mode [12] to the prepared instrument table, `MRdat`, using a default exposure-association threshold of $P < 5 \times 10^{-8}$ and a minimum of four instruments. The BWMR [13], cML-MA [14], debiased IVW [15], MR-GRIP [16], MR-Lasso [17], MR-RAPS [18], MR-ConMix [19], MR-Robust [17], MRMix [20] and MR-PRESSO [21] Skills also consume the prepared instrument table, with method-specific filtering and input requirements. They can be run from the same pair-preparation result without depending on one another's estimates. Downstream methods may impose a stricter instrument threshold, but cannot recover variants already excluded during pair preparation.

MR-APSS additionally requires estimates of background variation from genome-wide summary statistics [5]. During pair preparation, the harmonized SNPs are matched to a European LD-score reference before instrument selection. Single-trait and cross-trait LD score regression estimate C, which describes GWAS estimation errors, and Ω, which describes covariance in polygenic effects. The matched data are also used for instrument selection, and the selected instruments and background estimates are supplied together to MR-APSS. Its default IV-selection threshold is 5×10^{-5}. A separate adjusted threshold accounts for selection bias introduced by LD clumping; it is distinct from the cutoff used to select SNPs. The LD-score reference is independent of the clumping reference, so changing the clumping population does not change the European reference used for background estimation.

CAUSE [22] follows a separate branch from the two canonical GWAS datasets. Its Skill performs method-specific variant formatting, matching and clumping, estimates nuisance parameters from a broad set of variants and compares causal and shared-factor models. It therefore does not consume the clumped `MRdat` or the MR-APSS background matrices. For requests involving two OpenGWAS study identifiers, `mr-primary-id` provides a direct TwoSampleMR route with its own instrument extraction, outcome-association retrieval and harmonization. Its default association thresholds are tried in order, from 5×10^{-8} through 5×10^{-7} and 5×10^{-6} to 5×10^{-5}, stopping when at least four harmonized instruments are available. A threshold is relaxed only after a successful request yields too few instruments, not in response to an API failure.

To support consistent evidence comparison across datasets and MR methods, Skills supporting the common input-scale option use standardized GWAS associations by default. For each trait, effect estimates and

standard errors are represented as $\beta = Z/\sqrt{N}$ and $\mathrm{SE} = 1/\sqrt{N}$, respectively. This common representation reduces differences arising from original reporting units and provides a consistent basis for the agent to compare and synthesize methodological evidence. Results retain the input scale, analysis thresholds and selected instruments to support interpretation. Where supported and explicitly requested, native-scale analysis uses the preserved, harmonized GWAS effects and standard errors; MR-APSS uses the standardized representation. The numerical settings above describe current Skill defaults; an explicitly configured analysis can override the parameters supported by that Skill.

1.2.7 Sensitivity analyses

The `analysis-sensitivity` Skill uses the exact instrument set and effect scale retained by the primary MR analysis. It assesses between-variant heterogeneity using Cochran's Q statistics for IVW and MR-Egger, and directional pleiotropy using the MR-Egger intercept [10]. Single-variant estimates and funnel plots characterize dispersion, while leave-one-out analyses identify variants with a disproportionate influence on the overall estimate. MR-PRESSO provides global and outlier tests and, when estimable, outlier-corrected estimates and a distortion test [21]; the current default uses 1,000 simulations and a significance threshold of 0.05. These diagnostics are interpreted alongside agreement or disagreement among estimators with different assumptions. Non-significant diagnostic tests are not treated as proof that the instrumental-variable assumptions hold, and concordant estimates alone do not establish a causal effect [23, 24].

1.3 Reporting case: hexadecanedioate and coronary artery disease

The reporting case in Fig. 4 combines an original coronary artery disease (CAD) analysis with a FinnGen analysis of major coronary heart disease events. Section 1.3.1 contains the complete scientific report. Supplementary Data 1 provides the HTML report and its PDF version.

Supplementary Table 15 lists the three GWAS datasets and their recorded quality-control counts. Supplementary Table 16 documents pair preparation, method-specific instrument selection and sensitivity diagnostics. Supplementary Table 17 gives the twelve estimates from six MR methods across the two outcome branches. The primary and BWMR analyses used ten instruments after relaxation of the exposure-association threshold to $P \leq 5 \times 10^{-6}$; MR-APSS used 31 instruments at $P \leq 5 \times 10^{-5}$. The report treats the relaxed-threshold analyses as exploratory and retains the limitations concerning effect units, reference compatibility and participant overlap.

Replication assessment of genetically predicted hexadecanedioate levels and coronary artery disease using a FinnGen outcome

A two-sample Mendelian randomization report

REPORT STATUS Evidence-grounded scientific report; generated 2026-11-09 10:37 UTC

ABSTRACT In this two-sample Mendelian randomization replication assessment, genetically predicted hexadecanedioate levels showed inverse but null-inclusive associations with coronary artery disease in the original GWAS Catalog outcome (IVW beta -0.00878; 95% CI -0.02049 to 0.00294) and with major coronary heart disease events in FinnGen (IVW beta -0.00856; 95% CI -0.02787 to 0.01074). All six compared method estimates were inverse, but the original nominally significant weighted-median result did not statistically replicate. The FinnGen primary IVW was heterogeneous and materially influenced by one MR-PRESSO outlier. The evidence does not establish a causal protective effect of hexadecanedioate on coronary artery disease.

Background and objective

Hexadecanedioate is a dicarboxylic acid metabolite for which prior Mendelian randomization studies have reported inverse associations with coronary phenotypes, although the endpoint definitions differ. One study reported a protective role for higher hexadecanedioate in unstable angina pectoris, and another reported an inverse association with coronary heart disease risk. The purpose of the present report is to assess whether previously saved two-sample Mendelian randomization estimates for hexadecanedioate levels on coronary artery disease replicate when the outcome is replaced by a FinnGen major coronary heart disease event endpoint, while preserving the same exposure and analysis methods. [1,2]

Design, data sources, and quality control

This was a two-sample Mendelian randomization study using the same exposure GWAS, GCST90243839 (hexadecanedioate levels; 6,887 European ancestry individuals; source metadata NCBI36). The original outcome was GCST90132314 (coronary artery disease; 1,165,690 European ancestry individuals; GRCh37). The replication outcome was finngen_R12_I9_CHD (major coronary heart disease event; 500,348 Finnish individuals, 56,650 cases and 443,698 controls; GRCh38). Genetic variants associated with hexadecanedioate levels were used as instruments, so this analysis estimates a genetic-instrument-based relation rather than an experimental intervention.

Table 1. Per-study QC inputs and retained variant counts

Study	Recorded build	Input variants	Retained variants	Excluded variants	Principal filters
Exposure GCST90243839	NCBI36, recorded as unknown	2,545,171	1,026,635	1,518,536	Ambiguous alleles 391,175; MHC 10,448; HapMap reference 1,009,261; allele mismatch 38; MAF below 0.05 107,558; chi-square outliers 56
Original outcome GCST90132314	GRCh37	20,073,070	1,096,194	18,976,876	Ambiguous alleles 2,837,114; duplicates 16,051,563; sample-size outliers 3,075; chi-square outliers 3
FinnGen outcome finngen_R12_I9_CHD	GRCh38	21,327,062	1,056,297	20,270,765	Ambiguous alleles 3,034,718; MHC 50,133; duplicates 832,839; HapMap reference 16,214,000; allele mismatch 37; MAF below 0.05 139,025; chi-square outliers 13

After QC, the original pair harmonized 1,020,665 variant rows and LD clumping retained 31 independent instruments; the FinnGen pair harmonized 988,118 variant rows and also retained 31 clumped variants. Clumping used an OpenGWAS API EUR reference and LD score regression used European LD scores. The FinnGen outcome is Finnish ancestry, while the pair-preparation ancestry field was recorded as unknown; this EUR-reference mismatch is a limitation. LDSC genetic correlation estimates were 0.107 for the original pair and 0.196 for the FinnGen pair. Primary MR and BWMR selected instruments at p<=5e-06, yielding 10 SNPs; MR-APSS used all 31 clumped variants with selection-bias correction threshold 5e-05.

Instrumental variable assumptions and analysis methods

The core instrumental-variable assumptions are relevance, independence, and exclusion restriction. In this design, the selected variants should associate with hexadecanedioate levels, be independent of confounders of the metabolite-coronary disease relation, and affect coronary outcomes only through the exposure. These assumptions cannot be fully verified from summary data.

Primary analyses were IVW, MR-Egger, weighted median, and weighted mode. IVW assumes valid instruments or balanced pleiotropy; MR-Egger permits directional pleiotropy under the InSIDE assumption; weighted median requires valid instruments to contribute a majority of total weight; and weighted mode assumes the largest weighted ratio cluster represents valid instruments. MR-APSS models genome-wide background polygenicity, pleiotropy, and sample structure through C/Omega. BWMR propagates SNP-exposure uncertainty and models small centered horizontal pleiotropy with adaptive weighting.

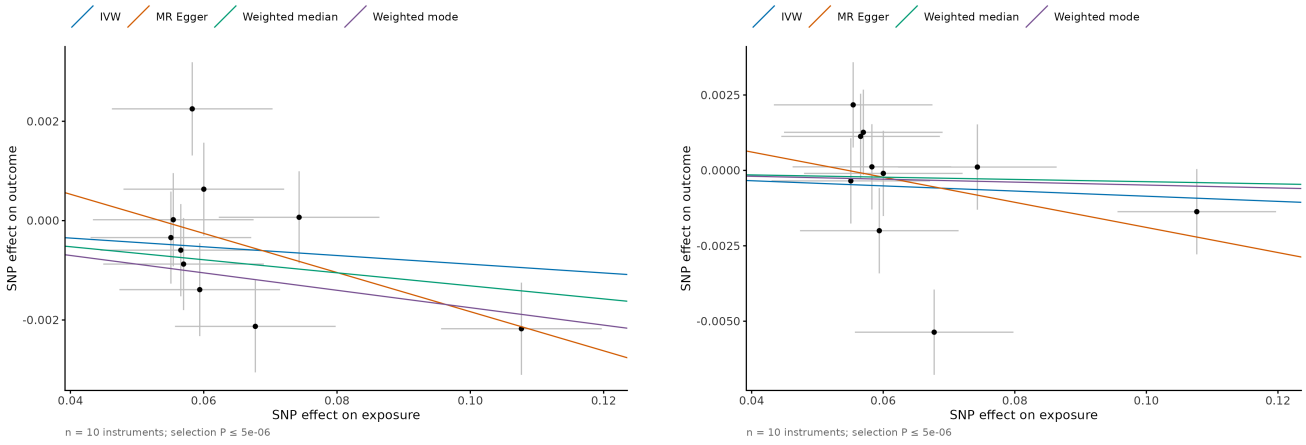
FinnGen sensitivity analyses included Cochran Q, the MR-Egger intercept, leave-one-out, and MR-PRESSO. Steiger direction was unavailable because the exposure type was not specified as quantitative or binary, and Steiger filtering was not requested. Estimates are on the recorded standardize input scale; outcome coefficients use source beta values and no odds ratios were derived.

Results

Primary MR estimates are summarized for both branches. All estimates from the four primary methods were inverse, but only the original weighted-median estimate had a confidence interval entirely below the null.

Table 2. Primary MR estimates in the original and FinnGen branches

Branch	Method	SNPs	Beta	SE	95% CI	P value
Original GWAS Catalog	IVW	10	-0.008775	0.005977	-0.020490 to 0.002939	0.142
Original GWAS Catalog	MR-Egger	10	-0.039396	0.025349	-0.097852 to 0.019059	0.159
Original GWAS Catalog	Weighted median	10	-0.013130	0.006209	-0.025299 to -0.000962	0.034
Original GWAS Catalog	Weighted mode	10	-0.017527	0.018944	-0.060382 to 0.025328	0.379
FinnGen	IVW	10	-0.008562	0.009850	-0.027868 to 0.010745	0.385
FinnGen	MR-Egger	10	-0.041663	0.044084	-0.143320 to 0.059994	0.372
FinnGen	Weighted median	10	-0.003745	0.009178	-0.021734 to 0.014244	0.683
FinnGen	Weighted mode	10	-0.004861	0.027567	-0.067222 to 0.057499	0.864



A · Primary MR scatter plot

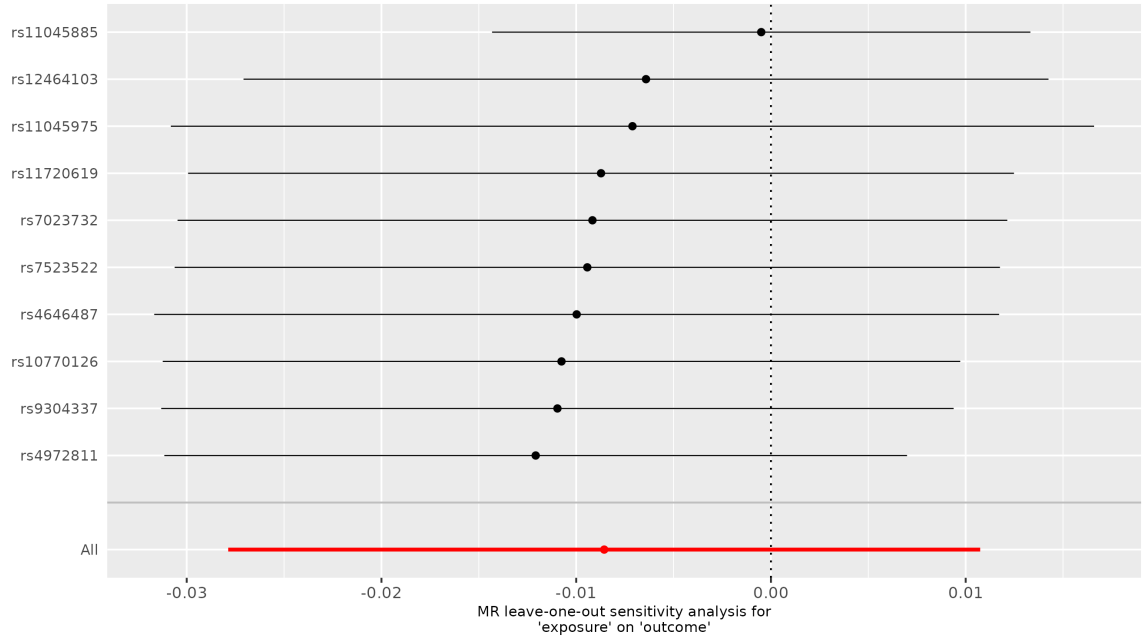
B · Primary MR scatter plot

Figure 1. Primary MR scatter plots for the original GWAS Catalog outcome and FinnGen outcome branches.

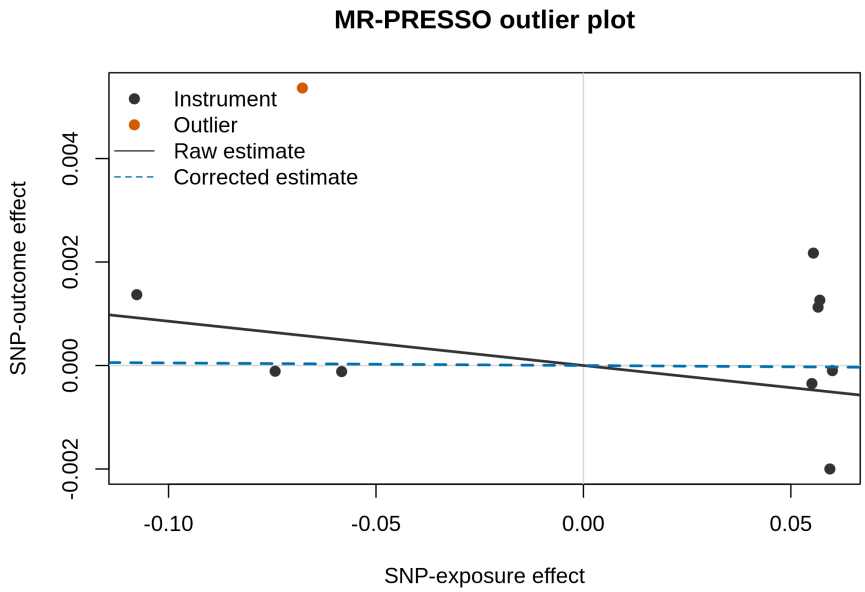
> **INTERPRETATION** The SNP clouds show broadly inverse fitted relationships, but many SNP-outcome effects are close to zero. The FinnGen branch has a similarly inverse IVW and MR-Egger appearance with wider SNP-outcome uncertainty, consistent with its larger standard errors.

In the FinnGen branch, Cochran Q indicated heterogeneity: IVW Q was 19.57 with 9 degrees of freedom and p=0.021, and MR-Egger Q was 18.21 with 8 degrees of freedom and p=0.020. The MR-Egger intercept was 0.00228 with standard error 0.00295 and p=0.463, so directional pleiotropy was not detected by that test; however, this test has limited power with 10 instruments. Leave-one-out

identified rs11045885 as the most influential variant, with a maximum absolute change of 94.18%, an influential_count of 5, and no direction reversal. MR-PRESSO detected global heterogeneity with p=0.018 and one outlier, rs11045885. After outlier correction, the FinnGen primary IVW estimate changed from -0.008562 to -0.000498, with distortion p=0.016. Steiger direction was unavailable because the exposure type was not specified.



A · Leave-one-out plot



B · MR-PRESSO outlier plot

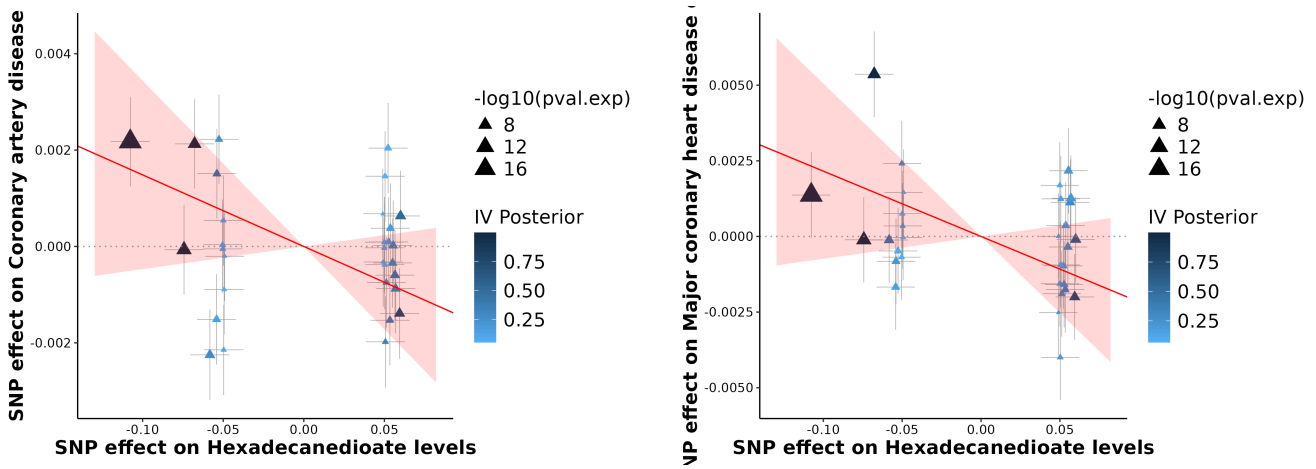
Figure 2. Leave-one-out and MR-PRESSO outlier evidence for the FinnGen primary IVW analysis.

INTERPRETATION Together these plots show that the FinnGen primary IVW estimate depends materially on rs11045885. The MR-PRESSO correction moves the point estimate from -0.008562 toward -0.000498, supporting caution in interpreting a simple inverse causal effect.

Table 3. MR-APSS and BWMR estimates by branch

Method	Branch	SNPs	Beta	SE	95% CI	P value
MR-APSS	Original	31	-0.014855	0.009974	-0.034405 to 0.004694	0.136
MR-APSS	FinnGen	31	-0.021588	0.014778	-0.050552 to 0.007376	0.144
BWMR	Original	10	-0.009017	0.006216	-0.021200 to 0.003166	0.147
BWMR	FinnGen	10	-0.008845	0.010234	-0.028904 to 0.011215	0.387

MR-APSS used all 31 clumped variants at threshold 5e-05. Its estimates were inverse and null-inclusive in both branches. MR-APSS IVstrength was 0.02895 in the original branch and 0.02871 in the FinnGen branch; this is total modeled IV signal, not per-SNP F. The model estimated pi0 at approximately 0.31 and nvalid at approximately 9.6 in both branches; these are model-based quantities, not validated biological classifications. BWMR used the same 10 instruments as primary MR. The approximate per-SNP F statistics for BWMR ranged from 20.89 to 79.83, with none below 10. Posterior BWMR weights were close to one, with minima near 0.9996 to 0.9998, indicating little extreme downweighting rather than proving that all instruments are valid.

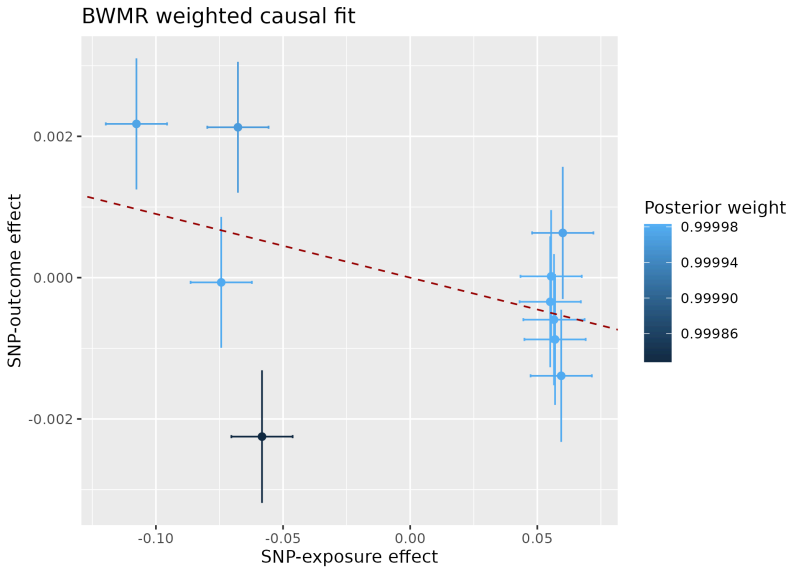


A · MR-APSS Result Plot

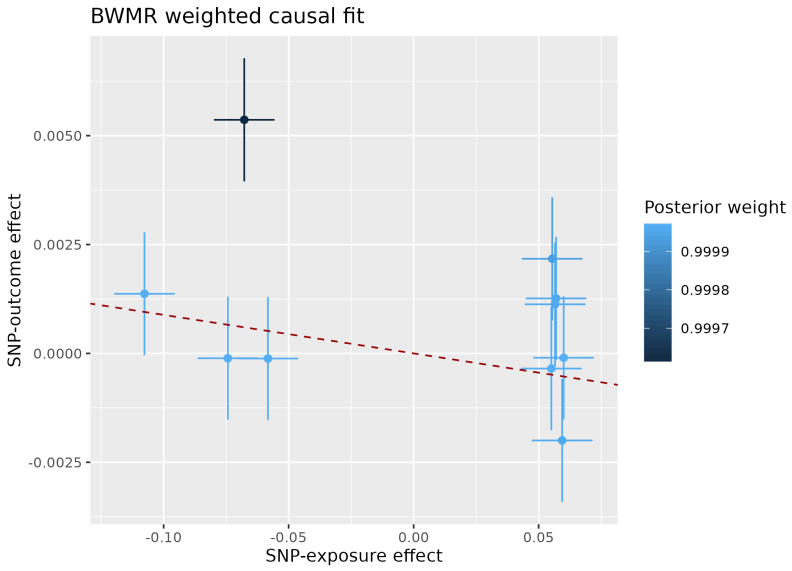
B · MR-APSS Result Plot

Figure 3. MR-APSS result plots for the original GWAS Catalog and FinnGen outcome branches.

INTERPRETATION Both MR-APSS fits use the same 31 clumped variants and estimate inverse slopes with null-inclusive uncertainty. The FinnGen slope is more strongly inverse but also wider, so the two branches are compatible rather than contradictory.



A · BWMR weighted causal fit

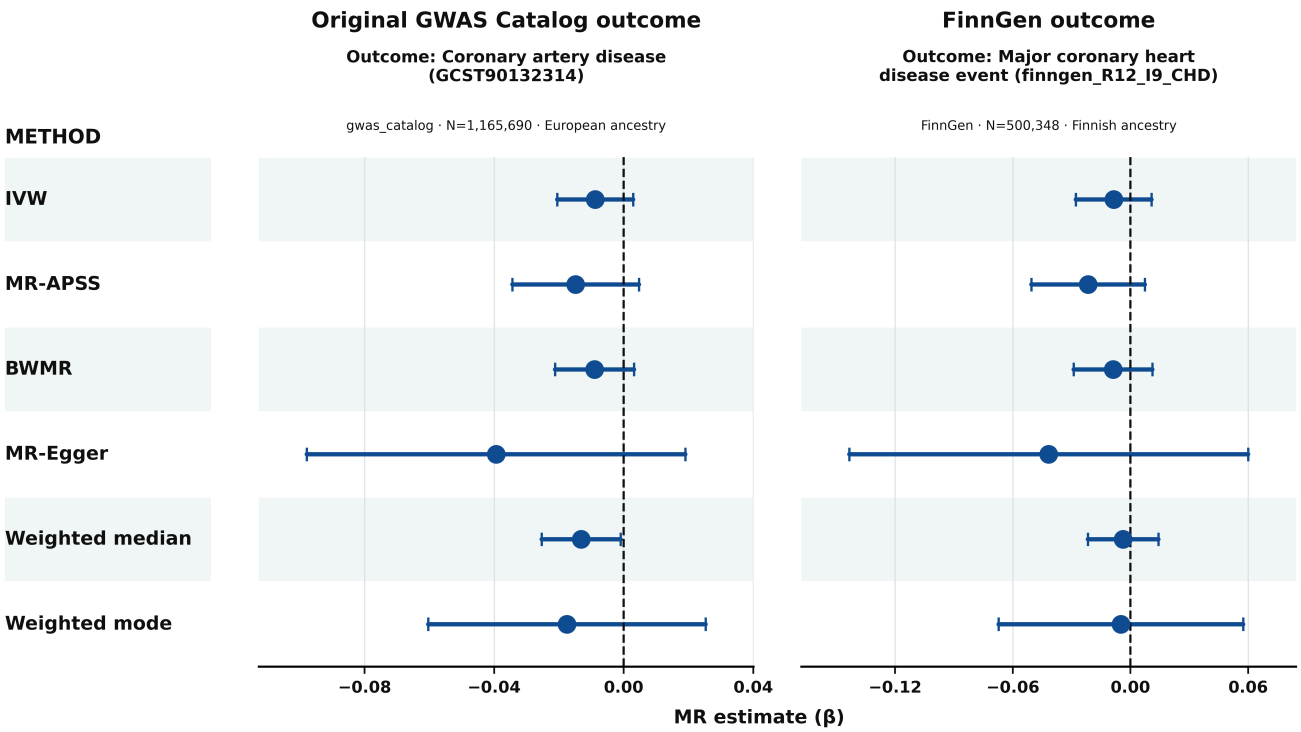


B · BWMR weighted causal fit

Figure 4. BWMR weighted causal fit plots for the original GWAS Catalog and FinnGen outcome branches.

> **INTERPRETATION** Both BWMR fits show inverse causal slopes with near-one posterior weights and confidence intervals that include the null. This indicates that exposure-uncertainty-aware weighting did not produce a strong protective effect estimate in either branch.

The replication summary compared 12 estimates across two branches and six methods. Directional concordance was 1.0: all point estimates were inverse. However, every FinnGen confidence interval included the null, and the original weighted-median nominal finding of p=0.034 was not reproduced. The FinnGen weighted-median estimate was -0.003745 with p=0.683. Therefore, the branches are directionally consistent but the previous nominally significant weighted-median result did not statistically replicate. Different outcome definitions and cohorts also prevent treating this comparison as fully independent replication.



Replication of hexadecanedioate-to-CAD estimates: original GWAS Catalog vs FinnGen outcome

Figure 5. Forest matrix comparing all six method estimates between the original GWAS Catalog and FinnGen outcome branches.

> **INTERPRETATION** The forest matrix shows directional consistency alongside wider intervals in the FinnGen branch. The original weighted median was the only interval entirely below the null, and its FinnGen counterpart remained inverse but included the null, so the evidence does not support statistical replication of that finding.

Discussion

These results provide, at most, weak evidence for an inverse association. The original and FinnGen primary IVW estimates were both null-inclusive, and the only nominally significant estimator in the original branch did not persist. MR-APSS and BWMR remained inverse but null-inclusive. The FinnGen primary IVW model was heterogeneous and sensitive to one MR-PRESSO outlier, so the simple IVW estimate should be interpreted with particular caution. The directionally inverse pattern could be compatible with a small protective effect, but it does not establish one.

Material limitations include the small exposure GWAS of 6,887 individuals and only 10 instruments in the primary and BWMR analyses, although the recorded per-SNP F statistics for BWMR did not flag instruments below 10. No power assessment was supplied. The exposure source was NCBI36 but recorded as unknown after QC, while the two outcome builds were GRCh37 and GRCh38; allele harmonization was performed after QC, but the supplied records do not document a formal liftOver. The FinnGen endpoint is a major coronary heart disease event rather than an exact coronary artery disease definition, and its Finnish ancestry was paired with an EUR LD reference under a pair-ancestry field recorded as unknown. Steiger direction was not assessed. Sensitivity records for the original branch were not supplied for this report, so equivalent original-branch diagnostics could not be compared. Unknown sample overlap further limits independence claims.

The results are specific to the European ancestry exposure cohort and the European or Finnish outcome cohorts used here, and may not generalize to other populations. The recorded QC counts, selection thresholds, instrument counts, and analysis branches are reported to support reproducibility.

Literature evidence and comparison

A prior MR study reported that higher hexadecanedioate levels had a protective role in unstable angina pectoris, which is directionally consistent with the inverse estimates reported here but does not directly address coronary artery disease overall. A separate MR study reported a significant inverse association between hexadecanedioate and coronary heart disease risk, with an odds ratio of 0.82 and 95% CI 0.72 to 0.93. That published result is not directly comparable to the present estimates because it used a different outcome definition and did not report using FinnGen, and the current report presents beta estimates without odds-ratio conversion. Nevertheless, it highlights that published evidence for this metabolite remains mixed relative to the null FinnGen-based result. [1,2]

Conclusion

CONCLUSION In this replication assessment, genetically predicted hexadecanedioate levels showed inverse but null-inclusive associations with coronary artery disease in the original analysis and with major coronary heart disease events in FinnGen. The prior nominally significant weighted-median estimate did not replicate, and the FinnGen primary estimate was heterogeneous and sensitive to one outlier. The current evidence does not establish a causal protective effect of hexadecanedioate on coronary artery disease. Mendelian randomization supports causal hypothesis testing but does not by itself prove causality.

Reporting information

SOFTWARE AND PREREGISTRATION Analyses were orchestrated using MRCAT 0.1.0 with Python 3.11.15; R 4.5.3; R packages: MRAPSS 0.0.0.9000, MRMix 0.1.0, MRPRESSO 1.0, MRcML 0.0.0.9000, MendelianRandomization 0.10.0, TwoSampleMR 0.7.6, cause 1.2.0.335, data.table 1.18.6.1, ggplot2 4.0.3, glmnet 5.0, jsonlite 2.0.0, mr.divw 0.1.0, mr.raps 0.4.3, robustbase 0.99.7. Preregistration information was not supplied for this analysis.

FUNDING Funding information for this analysis was not supplied. The role of any funder was not supplied.

DATA AND CODE AVAILABILITY GWAS identifiers and analysis provenance are reported in this document. Public data and code availability were not otherwise specified.

CONFLICTS OF INTEREST Conflict-of-interest information was not supplied.

References

1. Peng, Mengling; Fu, Yu; Qin, Cong; Shi, Lei; Zhang, Meiwei; Zhou, Shanshan. (2024). A stratified study of human blood metabolites and coronary artery diseases-A Mendelian randomization study.. Nutrition, metabolism, and cardiovascular diseases : NMCD. doi:10.1016/j.numecd.2024.09.024
2. Meng, Xiao-Yan; Zhu, Yong-Qing; Zhang, Ying-Jie; Sun, Wei; Li, Shu-Ang. (2025). Causal relationships between serum metabolites and coronary heart disease risk: a mendelian randomization study.. Frontiers in genetics. doi:10.3389/fgene.2025.1440364

1.4 Supporting records for negative-control interpretations

Supplementary Table 12 records the final-report evidence underlying the interpretation categories in Fig. 3e. Each system–query pair is represented by an excerpt from its delivered report. The categories distinguish an overall conclusion of insufficient causal support, an affirmative causal claim and weak or unresolved causal evidence. They summarize the reported interpretation against the negative-control expectation, independently of whether the analytical procedures were valid.

Supplementary Table 13 presents the Claude Code Q02 sign-correction comparison. Both instrument tables retain the same 31 SNPs, exposure effects and standard errors; six outcome-effect signs differ. Supplementary Data 2 includes the original and corrected inputs, the six changed rows and the nine method-specific comparisons.

1.5 Data and estimates for the downstream applications

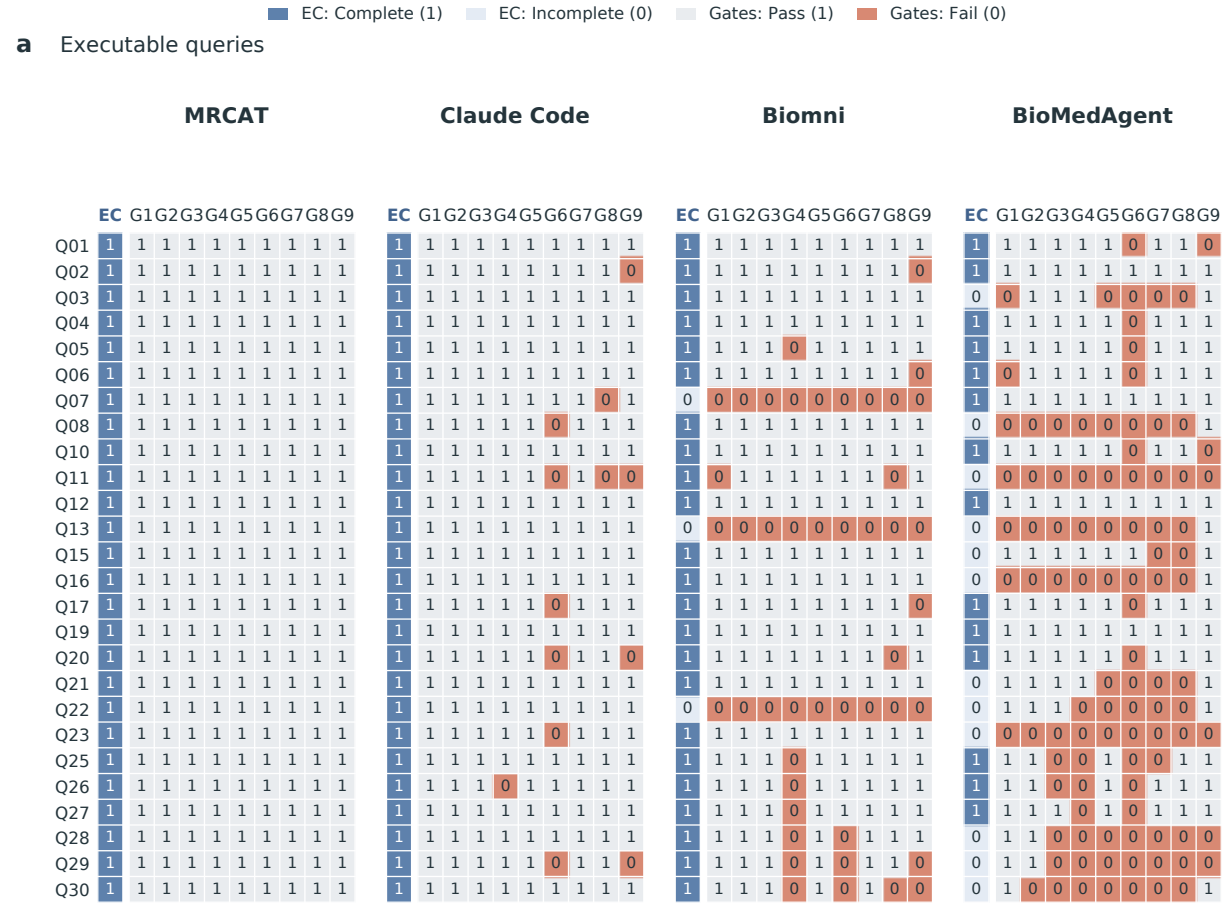
Supplementary Tables 18 and 19 document the GWAS inputs and recorded parameters used in Figs. 5 and 6. Supplementary Tables 20 and 21 contain all 30 RA risk-factor estimates and all 12 bidirectional T2D–CAD estimates, respectively. For the RA analysis, MR-Egger and weighted-mode 95% confidence intervals use t distributions with $n_{\mathrm{SNP}} - 2$ and $n_{\mathrm{SNP}} - 1$ degrees of freedom, respectively. The RA q values apply Benjamini–Hochberg adjustment across the five exposures separately for each method; bidirectional q values apply the adjustment across six methods separately within each direction. The MR-APSS selection-bias threshold is reported separately from the exposure-association cutoff used to select instruments. Supplementary Data 2 provides the numerical tables at saved precision, metadata and relative source-file anchors.

1.6 Supporting records for execution completion

Supplementary Fig. 1a shows the recorded execution outcomes underlying Fig. 2f for all 26 executable queries. Supplementary Data 3 links each of the 104 system–query pairs to its selected run, an exact excerpt from the archived final Markdown and the corresponding line numbers. Where a report cites a result file without printing an effect estimate, the index also identifies the numerical output saved by that same run. Execution completion is assessed independently of procedural gate scores.

2 Supplementary Figures

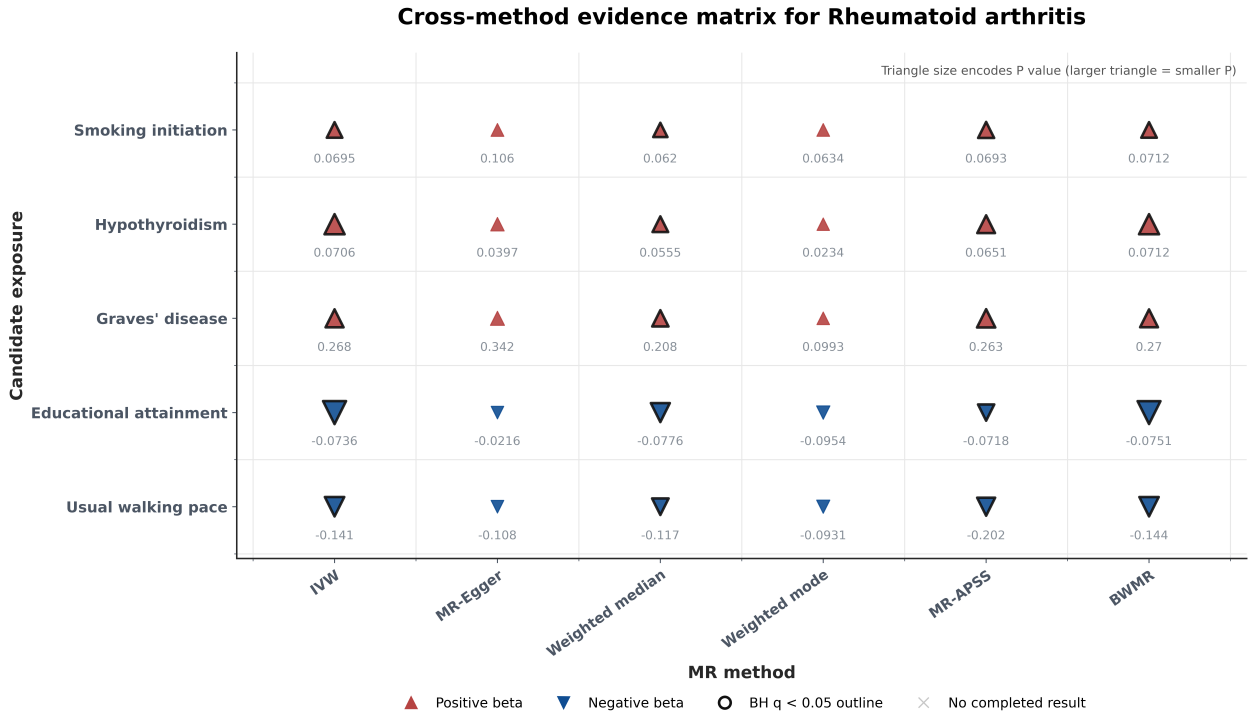
2.1 30-query benchmark



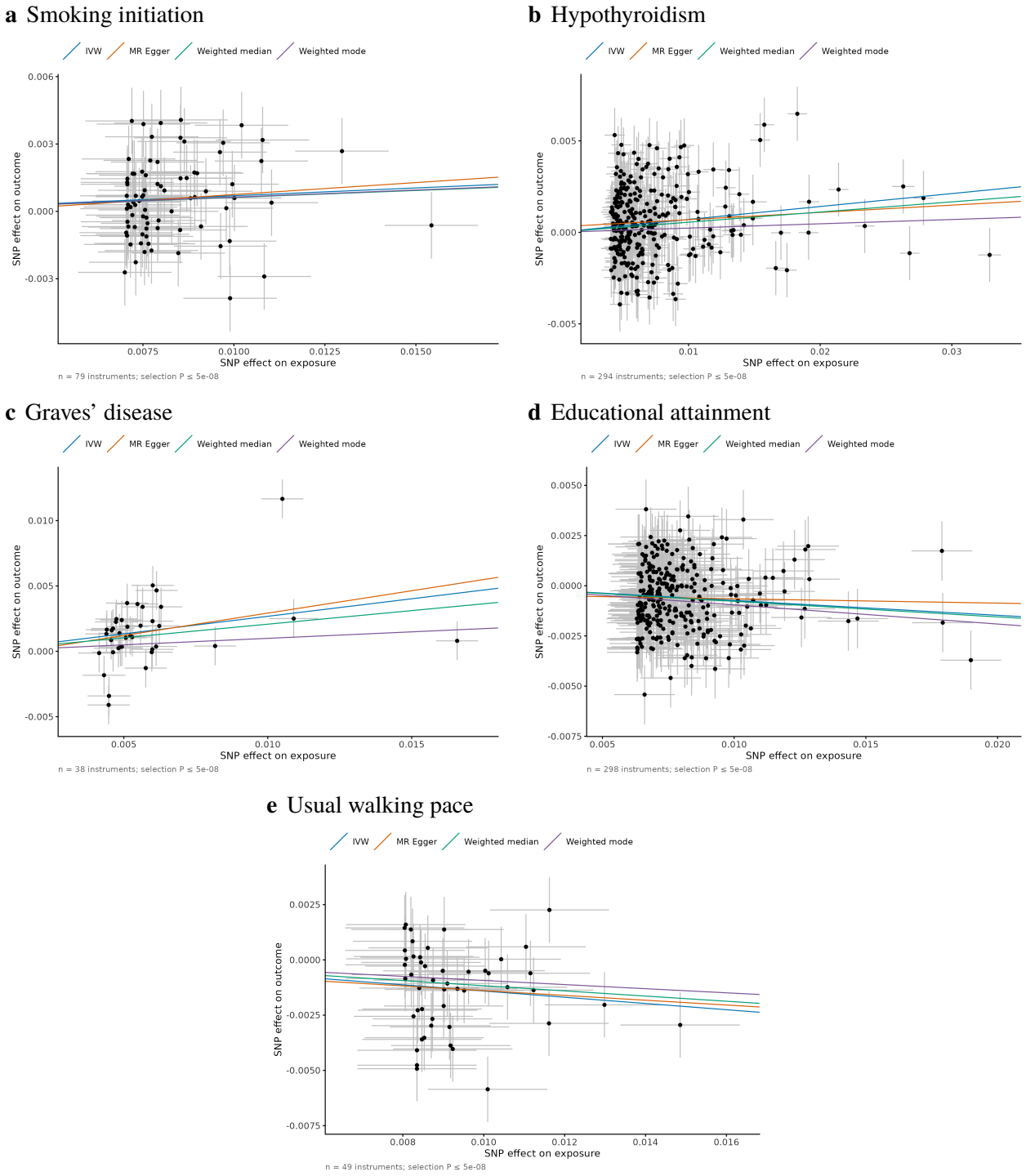
Supplementary Figure 1 | Query-level execution, procedural scores and boundary assessments across four MR agents. a, Execution completion (EC) and G1–G9 ratings for each of the 26 executable queries, grouped by system. EC denotes production of at least one numerical MR effect estimate: dark blue, complete (1); pale blue, incomplete (0). For G1–G9, light grey denotes pass (1) and coral denotes fail (0). SC denotes scientific completion, requiring all nine gates to pass. Supplementary Data 3 provides the execution evidence. **b,** B1 ratings for the four boundary queries, assessed separately. The scoring rubric is provided in Supplementary Table 6. The same rubric covers requested analyses and additional analyses presented as evidence for the answer.

2.2 Risk-factor assessment in rheumatoid arthritis

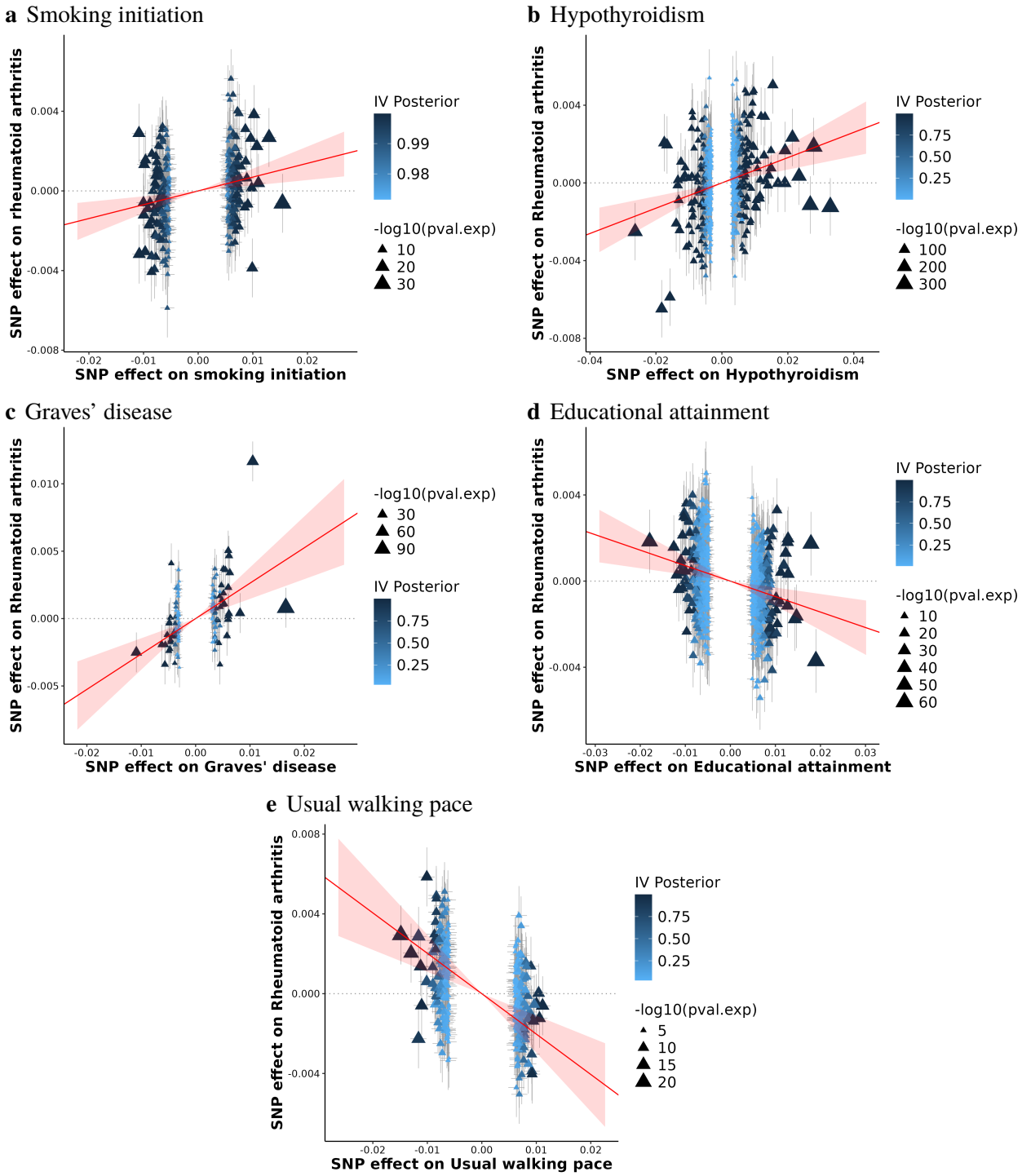
Plots generated by MRCAT for the analysis in main Fig. 5.



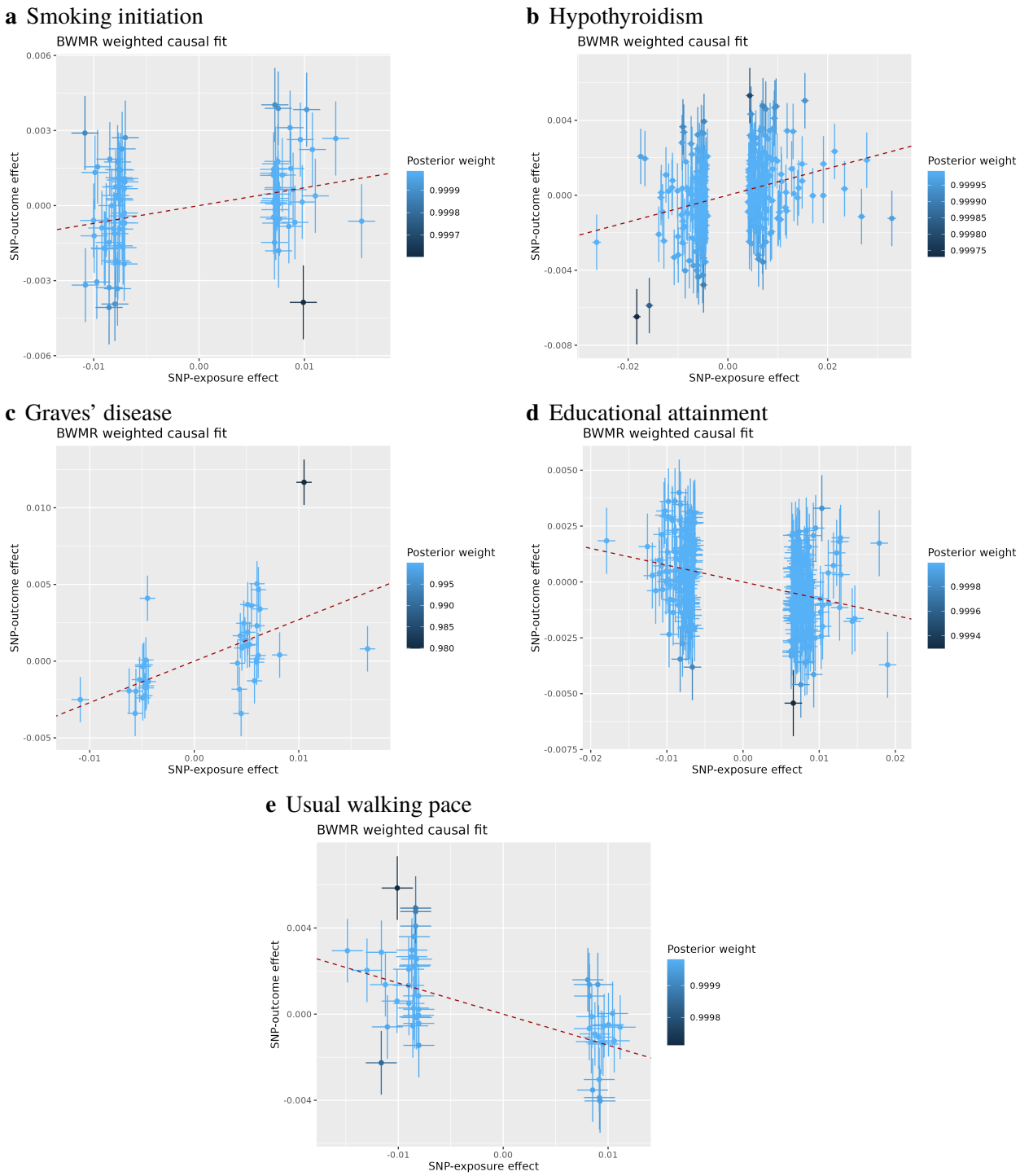
Supplementary Figure 2 | Cross-method evidence matrix for the RA risk-factor assessment. Rows show the five exposures and columns show the six MR methods. Red upward and blue downward triangles denote positive and negative estimates, respectively. Triangle size increases with $-\log_{10} P$; black outlines indicate Benjamini–Hochberg $q < 0.05$, calculated across the five exposures separately for each method. Numbers report the corresponding $\hat{\beta}$ estimates.



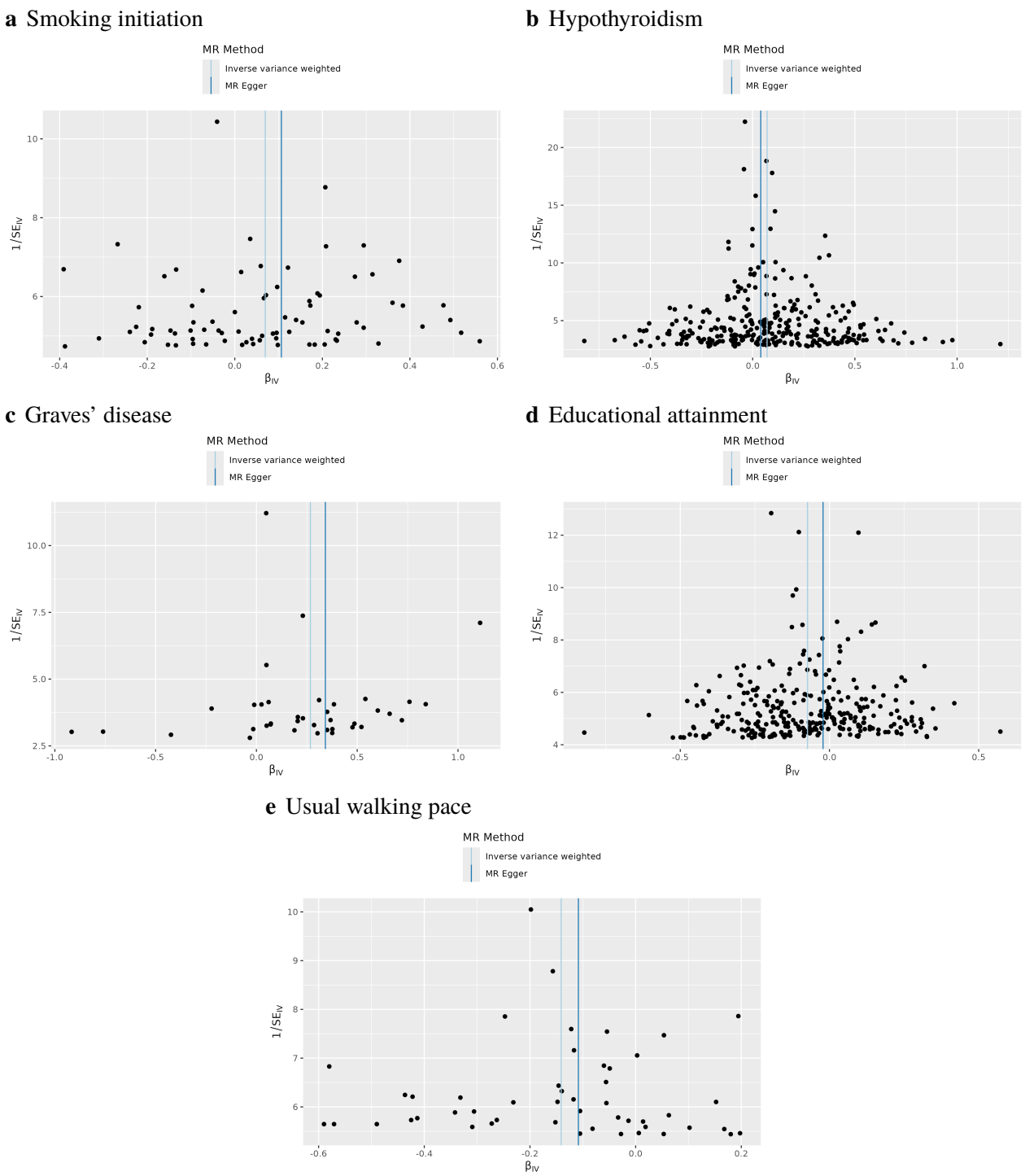
Supplementary Figure 3 | Primary MR scatter plots for the RA risk-factor assessment. a–e, Smoking initiation, hypothyroidism, Graves' disease, educational attainment and usual walking pace, respectively, each analysed against rheumatoid arthritis (RA). Points show SNP–exposure and SNP–outcome associations; horizontal and vertical error bars denote one standard error. Lines show the IVW, MR-Egger, weighted median and weighted mode fits. Instrument counts are 79, 294, 38, 298 and 49, respectively.



Supplementary Figure 4 | MR-APSS fits for the RA risk-factor assessment. a–e, Smoking initiation, hypothyroidism, Graves' disease, educational attainment and usual walking pace, respectively, each analysed against rheumatoid arthritis (RA). Triangles denote SNPs, with size reflecting $-\log_{10} P$ for the exposure association and colour indicating the IV posterior probability. Error bars denote one standard error. The fitted line and shaded band show the MR-APSS estimate and its 95% confidence interval. Variant counts are 317, 514, 122, 637 and 321, respectively.

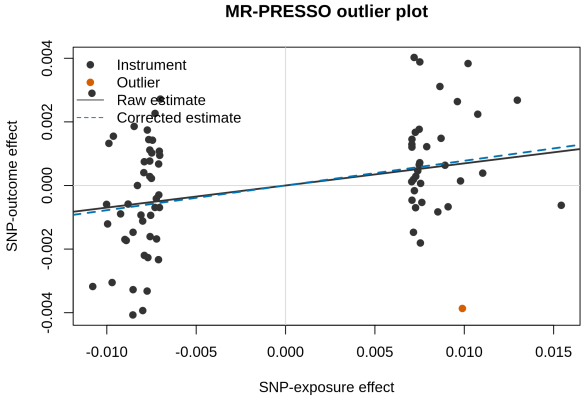


Supplementary Figure 5 | BWMR fits for the RA risk-factor assessment. a–e, Smoking initiation, hypothyroidism, Graves' disease, educational attainment and usual walking pace, respectively, each analysed against rheumatoid arthritis (RA). Points show SNP–exposure and SNP–outcome associations, colour indicates the posterior BWMR weight, and the dashed line shows the fitted slope. Instrument counts are 79, 294, 38, 298 and 49, respectively.

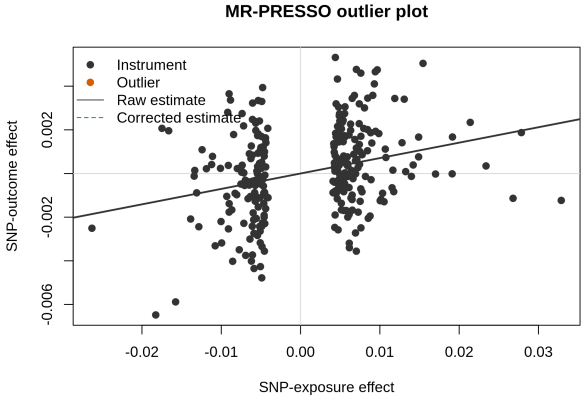


Supplementary Figure 6 | Funnel plots for the RA risk-factor assessment. a–e, Smoking initiation, hypothyroidism, Graves' disease, educational attainment and usual walking pace, respectively, each analysed against rheumatoid arthritis (RA). Single-SNP MR estimates are plotted against their precision (inverse standard error). Vertical lines mark the IVW and MR-Egger estimates. Instrument counts are 79, 294, 38, 298 and 49, respectively.

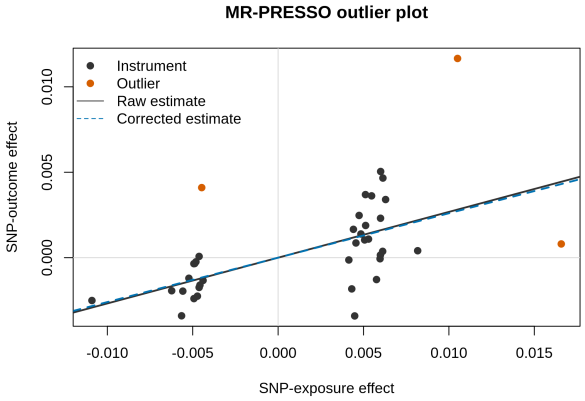
a Smoking initiation



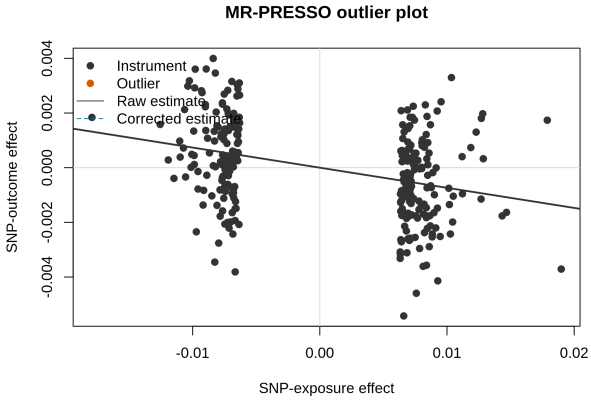
b Hypothyroidism



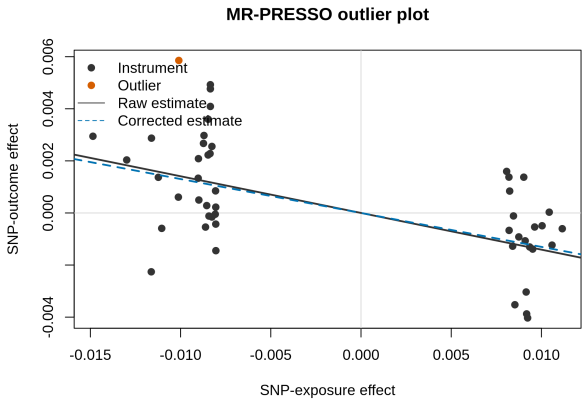
c Graves' disease



d Educational attainment

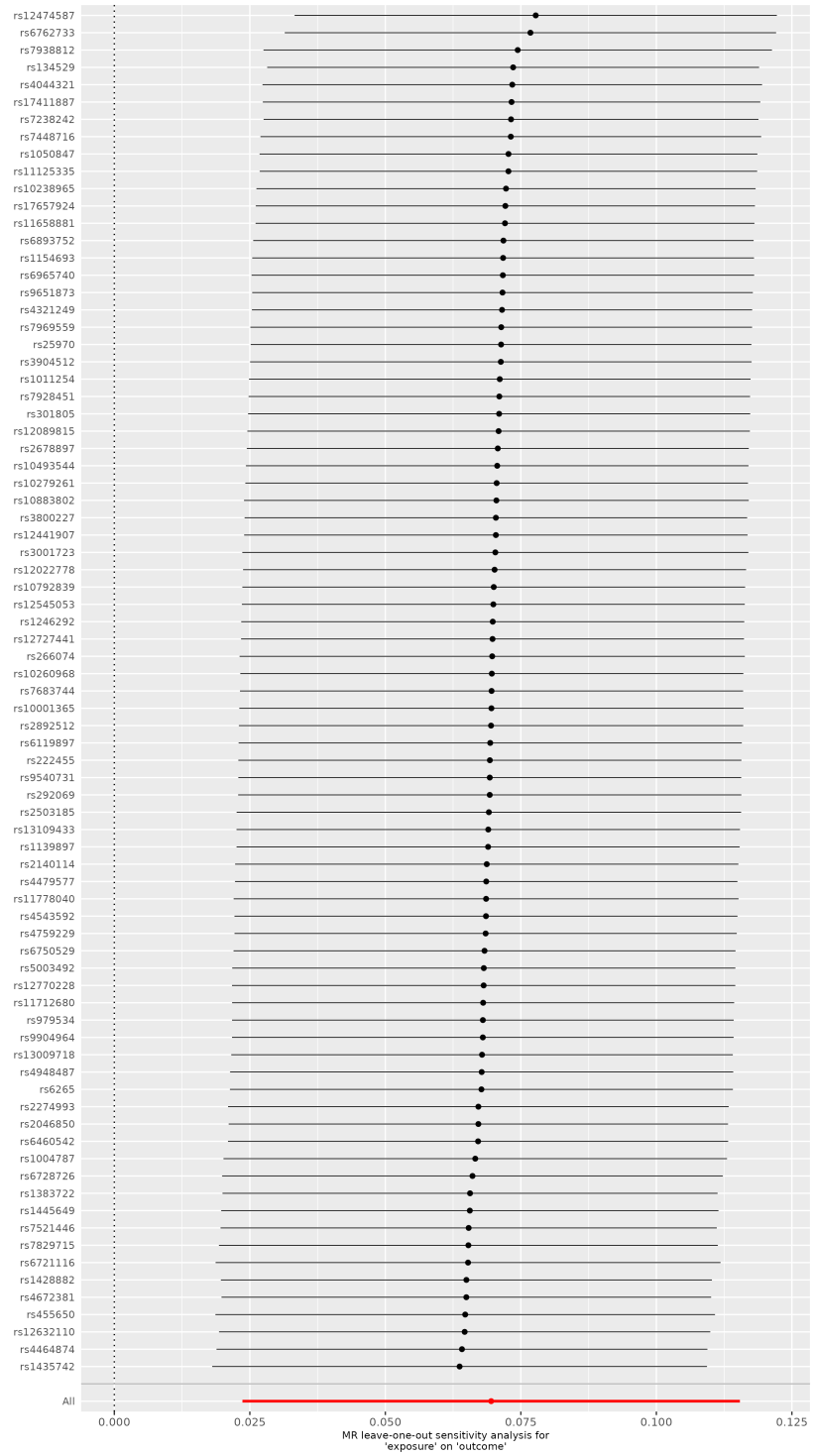


e Usual walking pace



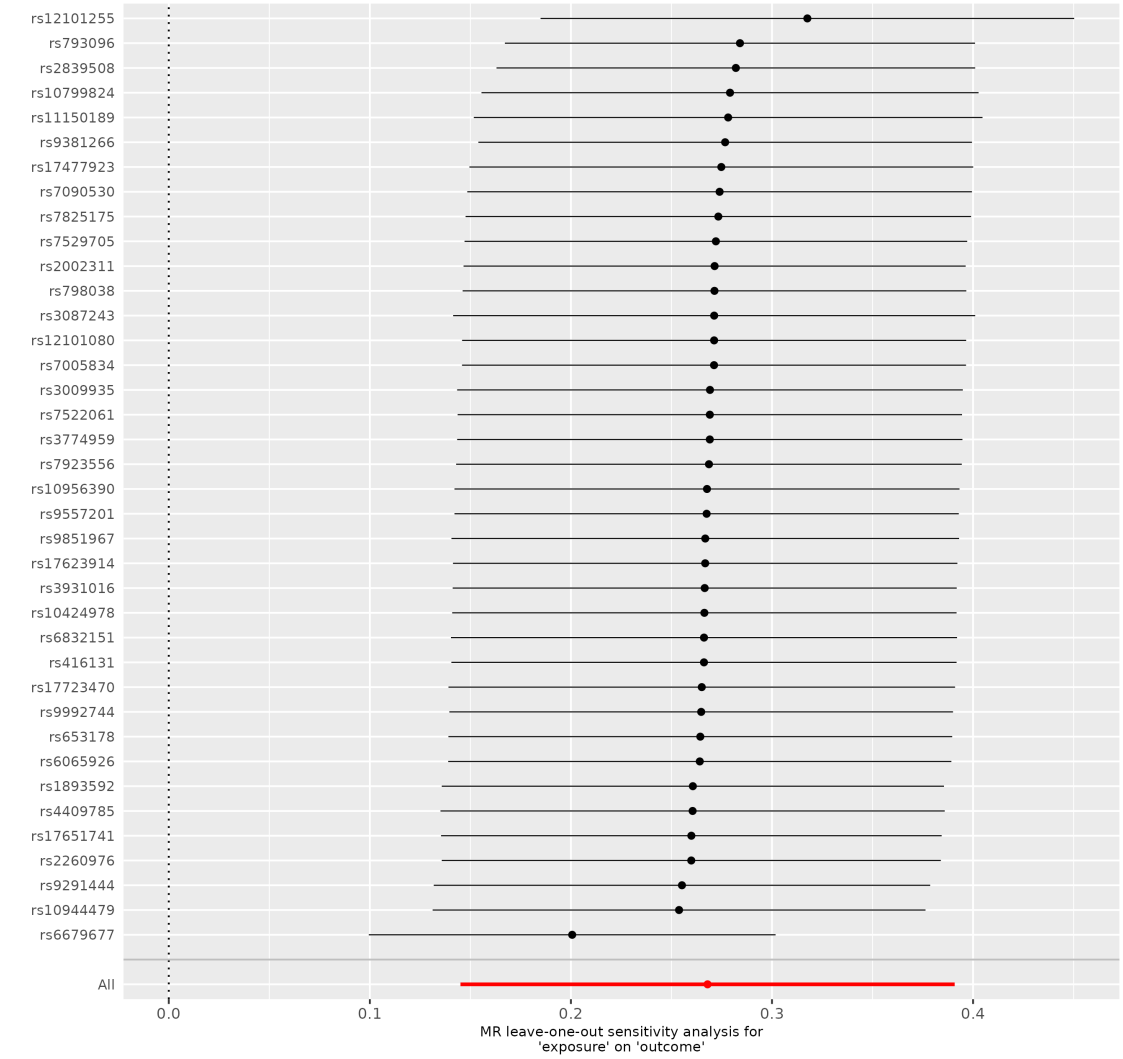
Supplementary Figure 7 | MR-PRESSO plots for the RA risk-factor assessment. a–e, Smoking initiation, hypothyroidism, Graves' disease, educational attainment and usual walking pace, respectively, each analysed against rheumatoid arthritis (RA). Orange points mark identified outliers; dark points mark the remaining instruments. The solid black line shows the raw estimate, and the dashed blue line shows the outlier-corrected estimate where available.

a Smoking initiation



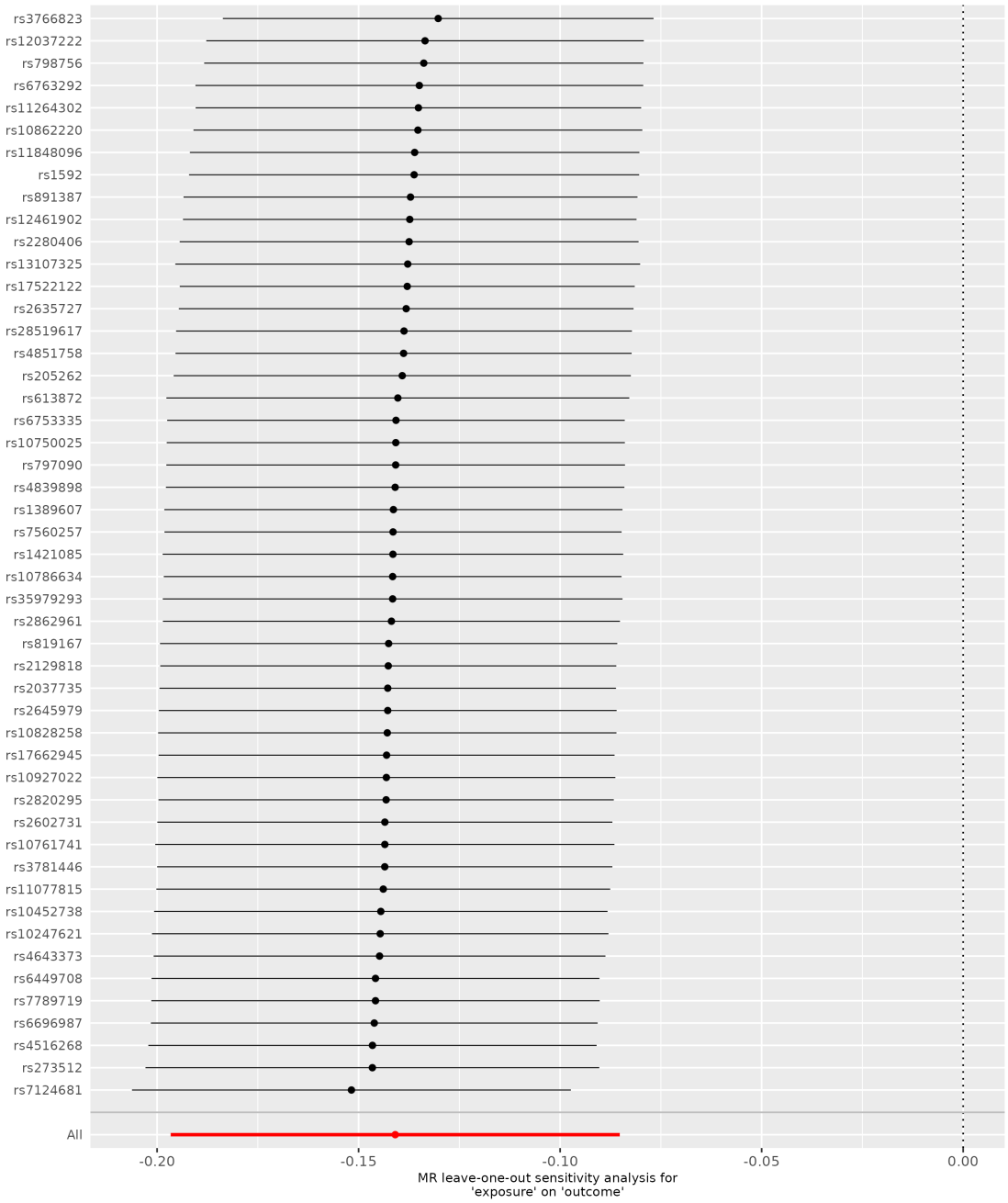
Supplementary Figure 8 | Leave-one-out analyses for the RA risk-factor assessment. a, Smoking initiation in relation to RA. Black points and horizontal bars show IVW estimates and 95% confidence intervals after removing each labelled SNP. The red point and bar show the estimate using all instruments. Numerical leave-one-out results are available for hypothyroidism and educational attainment, but plots were not generated because their instrument counts exceeded the 100-instrument plotting limit.

b Graves' disease



Supplementary Figure 8 | Leave-one-out analyses for the RA risk-factor assessment. b, Graves' disease in relation to RA. Black points and horizontal bars show IVW estimates and 95% confidence intervals after removing each labelled SNP. The red point and bar show the estimate using all instruments.

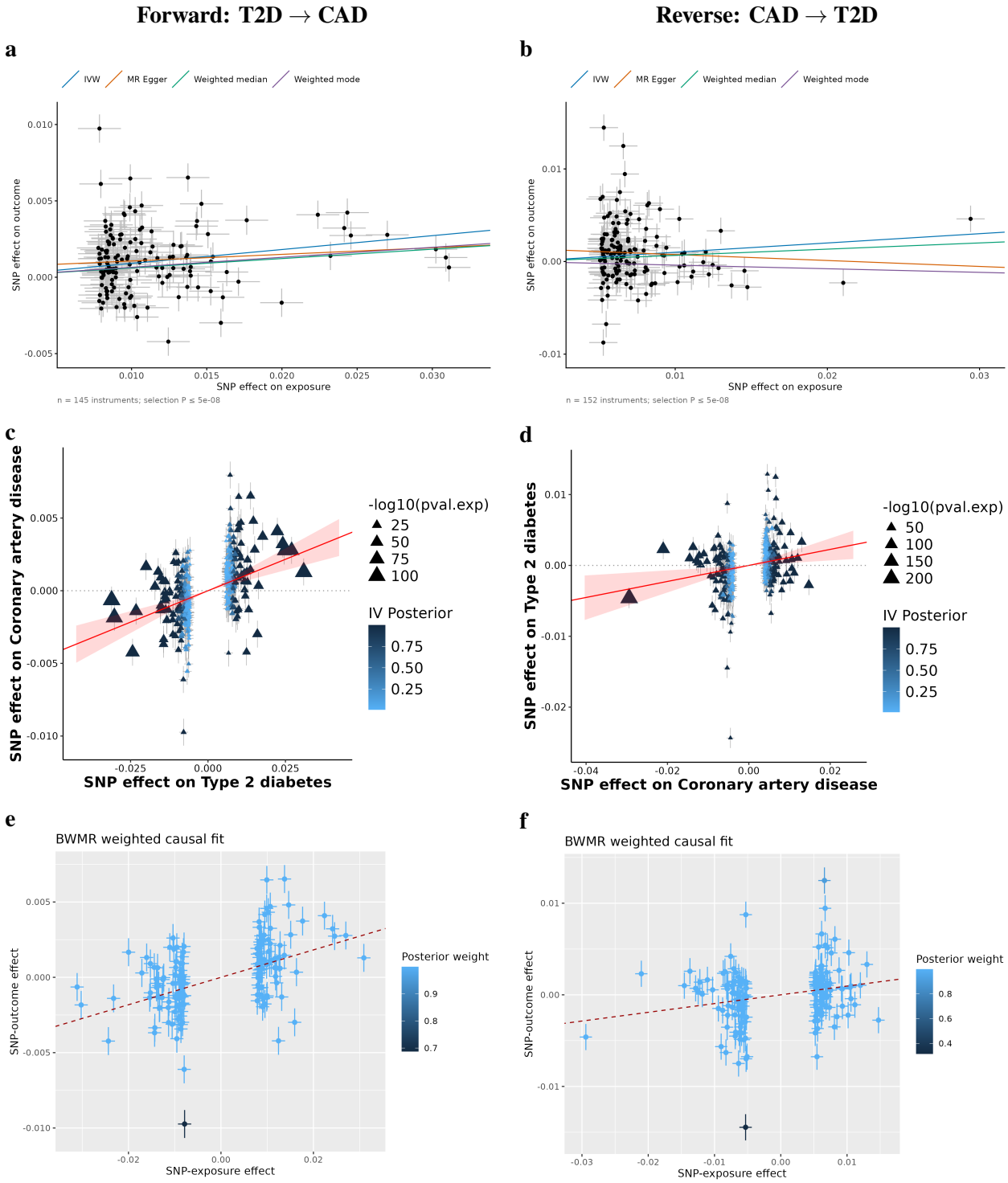
c Usual walking pace



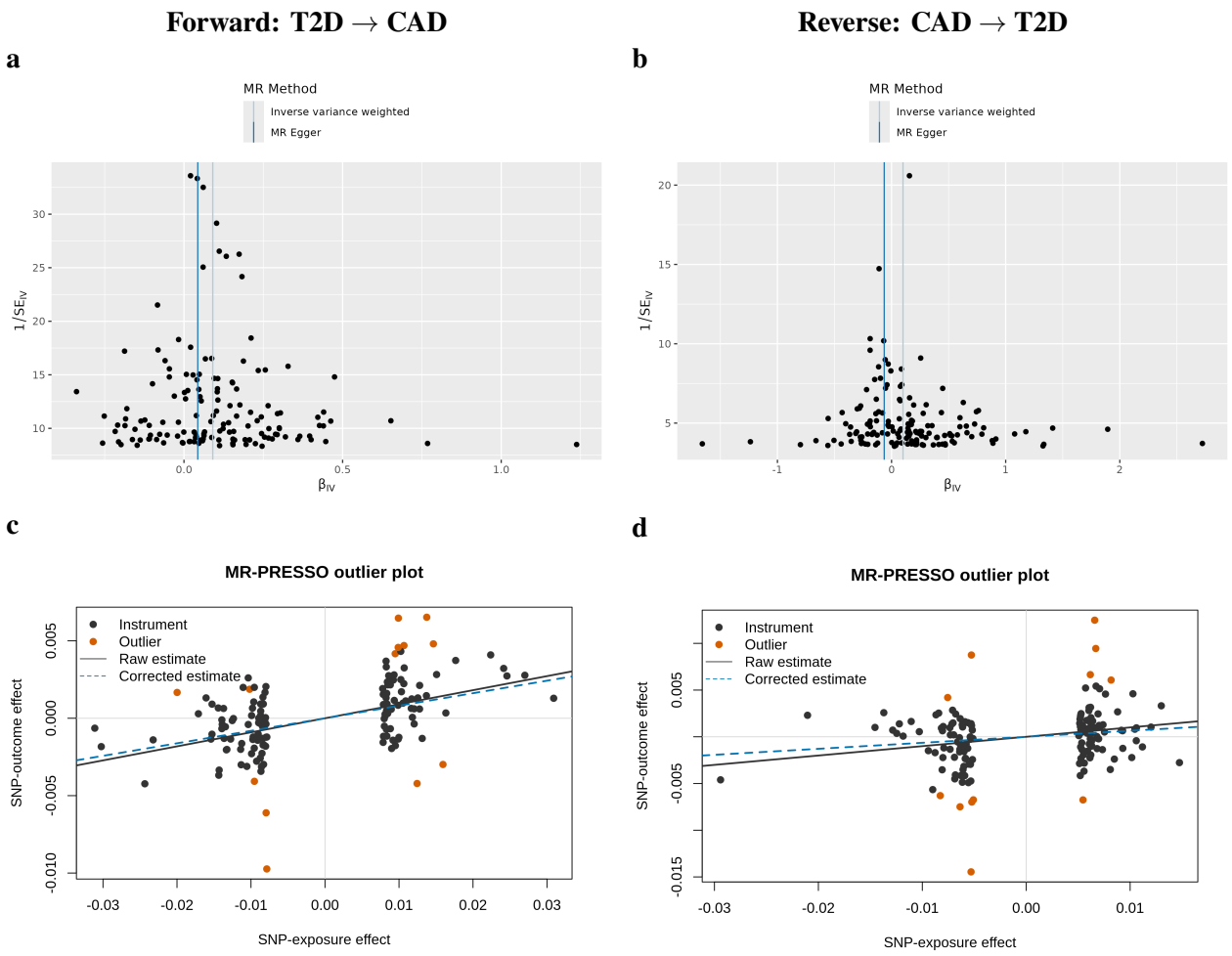
Supplementary Figure 8 | Leave-one-out analyses for the RA risk-factor assessment. c, Usual walking pace in relation to RA. Black points and horizontal bars show IVW estimates and 95% confidence intervals after removing each labelled SNP. The red point and bar show the estimate using all instruments.

2.3 Bidirectional analysis of type 2 diabetes and coronary artery disease

Plots generated by MRCAT for the analysis in main Fig. 6.



Supplementary Figure 9 | SNP-level fits for the bidirectional T2D–CAD analysis. Left and right columns show the forward type 2 diabetes (T2D) to coronary artery disease (CAD) analysis and the reverse CAD to T2D analysis, respectively. **a,b,** Primary MR scatter plots with IVW, MR-Egger, weighted median and weighted mode fits, using 145 forward and 152 reverse instruments. **c,d,** MR-APSS fits using 354 forward and 350 reverse variants. Triangles denote SNPs, with size reflecting the exposure-association $-\log_{10} P$ value and colour denoting the IV posterior probability. The solid line and shaded band show the fitted slope and its 95% confidence interval. **e,f,** BWMR fits using 145 forward and 152 reverse instruments. Colour denotes the posterior weight; the dashed line denotes the fitted slope. Error bars denote one standard error for the SNP–exposure and SNP–outcome associations. Primary MR and BWMR used an instrument-selection threshold of $P \leq 5 \times 10^{-8}$; MR-APSS used separately prepared inputs.



Supplementary Figure 10 | Sensitivity diagnostics for the bidirectional T2D–CAD analysis. Left and right columns show T2D to CAD and CAD to T2D, respectively, using the same 145 and 152 instruments as the primary MR analyses. **a,b,** Funnel plots of single-SNP MR estimates against their precision (inverse standard error). Vertical lines mark the IVW and MR-Egger estimates. **c,d,** MR-PRESSO outlier plots. Orange points mark 13 forward and 12 reverse outliers; grey points mark the remaining instruments. Solid black and dashed blue lines show the raw and outlier-corrected estimates, respectively. Forward and reverse estimates changed from 0.0908 to 0.0809 and from 0.1002 to 0.0643; distortion-test P values were 0.270 and 0.0566. IVW Cochran's Q tests detected heterogeneity in both directions ($P = 9.90 \times 10^{-73}$ forward and $P = 5.33 \times 10^{-76}$ reverse). MR-Egger intercept P values were 0.159 and 0.0382, respectively. Numerical leave-one-out analyses showed maximum absolute changes of 3.68% and 10.36%, without a reversal of the IVW estimate's direction.

3 Supplementary Tables

3.1 MRCAT Skill library

Supplementary Table 1: MRCAT Skill catalog. The 26 Skills are grouped into data acquisition and preparation, MR analysis, and diagnostics and result integration.

Skill	Purpose	Underlying method / software	Required inputs	Main outputs	Key prerequisites
Upstream: data acquisition and preparation (5 Skills)					
`gwas-search`	Find candidate GWAS datasets.	OpenGWAS, GWAS Catalog, Pan-UKB and FinnGen search adapters.	Trait name or study identifier; stated source and ancestry constraints.	Candidate shortlists, study metadata and search diagnostics.	Preserve phenotype and identifier meaning; source and ancestry constraints must be compatible.
`gwas-select`	Choose one suitable study per target.	LLM selection with candidate-identity validation.	Saved shortlists for all targets and the analysis request.	Selected study records and selection rationale.	Eligible candidates with suitable phenotype, source, ancestry and data availability.
`gwas-download`	Retrieve selected summary statistics.	Source-specific download adapters and local cache.	Selected study record or a raw-data URL.	Local raw file and retained study metadata.	Executable download descriptor or accessible raw-file URL; required access credentials.
`gwas-qc`	Standardize columns and perform quality control (QC).	LLM-assisted mapping; reference-based R formatting.	Raw GWAS file and documented study metadata.	Canonical and filtered tables; QC and mapping reports.	Interpretable alleles and statistics; documented sample sizes for formatting; build evidence for coordinate mapping.
`gwas-pair-prepare`	Prepare instruments for one directed pair.	Allele harmonization, linkage disequilibrium (LD) score regression and PLINK or OpenGWAS clumping.	Formatted exposure and outcome tables.	Harmonized, LD-clumped single-nucleotide polymorphism (SNP) table (`MRdat`) with exposure and outcome effects, standard errors and exposure-association P values; background matrices C and Ω.	Genome-wide data compatible with the European LD-score reference; suitable clumping reference.
Midstream: MR analysis (14 Skills)					
`mr-primary`	Run the primary four-method suite.	Inverse-variance weighting (IVW), MR-Egger, weighted median and weighted mode; TwoSampleMR.	Prepared `MRdat`.	Four method estimates, uncertainty, plots and selected instruments.	Independent, harmonized instruments passing the analysis threshold; supported effect scale; default minimum of four instruments.

Continued on next page

Supplementary Table 1 (continued)

Skill	Purpose	Underlying method / software	Required inputs	Main outputs	Key prerequisites
`mr-primary-id`	Run the primary suite directly from study IDs.	TwoSampleMR association retrieval, clumping, harmonization and estimation.	Exposure and outcome OpenGWAS study IDs.	Four method estimates, plots and selected instruments.	OpenGWAS access and authentication; sufficient harmonized instruments; sample sizes for standardization.
`mr-apss`	Estimate effects with background and selection-bias adjustment.	MR-APSS; MRAPSS R package.	Complete prepared `MRdat` with matching C and Ω from the same pair preparation.	Effect estimate, uncertainty, instrument-strength diagnostics and plots.	Compatible genome-wide background estimates; standardized effects; preserved selection-correction threshold.
`mr-bwmr`	Fit Bayesian weighted MR.	BWMR; bundled R implementation.	Prepared `MRdat`.	Effect estimate, uncertainty, instrument weights and plots.	Independent instruments; valid effect estimates and standard errors.
`mr-cause`	Compare causal and shared-factor models.	CAUSE; cause R package with its own harmonization and clumping.	Two genome-wide canonical GWAS tables.	Posterior effect estimates, credible intervals, model comparison and diagnostic plots.	Broad variant coverage for nuisance estimation; compatible LD reference; sample sizes for standardization.
`mr-cml-ma`	Estimate effects allowing invalid instruments.	Constrained maximum likelihood with model averaging; MRcML.	Prepared `MRdat` and documented sample sizes.	Model-averaged estimate, uncertainty and invalid-instrument diagnostics.	Positive sample sizes from the input data or a documented override.
`mr-divw`	Fit debiased IVW.	Debiased inverse-variance weighting; mr.divw.	Prepared `MRdat`.	Effect estimate, uncertainty and method diagnostics.	Independent instruments; valid exposure and outcome effects and standard errors.
`mr-grip`	Fit genotype-recoding-invariant MR.	MR-GRIP; TwoSampleMR.	Prepared `MRdat`.	MR-GRIP and weak-instrument-adjusted estimates.	Verified upstream effect scale; uses the supplied effects without an input-scale switch.
`mr-lasso`	Fit penalized SNP-specific effects.	MR-Lasso; bundled helper using glmnet.	Prepared `MRdat`.	Effect estimate, uncertainty and penalized-instrument diagnostics.	Independent instruments; valid effect estimates and standard errors.
`mr-mrmix`	Fit a mixture model for MR.	MRMix R package.	Prepared `MRdat`.	Effect estimate, uncertainty and mixture-profile plots.	For MRMix-specific scaling: trait types, sample sizes for continuous traits and observed frequencies for binary traits.
`mr-presso`	Test pleiotropic outliers and effect distortion.	MR-PRESSO; MRPRESSO R package.	Prepared `MRdat`.	Global, outlier and distortion tests; raw and corrected estimates when estimable.	Enough instruments for testing; sufficient simulations for the outlier significance threshold.

Continued on next page

Supplementary Table 1 (continued)

Skill	Purpose	Underlying method / software	Required inputs	Main outputs	Key prerequisites
`mr-raps`	Fit robust adjusted profile-score MR.	MR-RAPS; mr.raps R package.	Prepared `MRdat`.	Effect estimate, uncertainty and optional diagnostic plots.	Independent instruments; valid exposure and outcome effects and standard errors.
`mr-robust`	Fit robust IVW regression.	Robust regression through the origin; robustbase.	Prepared `MRdat`.	Effect estimate, uncertainty and residual diagnostics.	Independent instruments; valid effect estimates and standard errors.
`mr-conmix`	Fit a contamination-mixture model.	MR-ConMix; MendelianRandomization R package.	Prepared `MRdat`.	Effect estimate, profile likelihood and potentially disjoint confidence intervals.	An explicit confidence-interval search grid compatible with the effect units when using native effects.
Downstream: diagnostics and result integration (7 Skills)					
`analysis-sensitivity`	Assess sensitivity of a completed primary analysis.	Cochran's Q, Egger intercept, single-SNP, leave-one-out, MR-PRESSO and Steiger analyses.	Primary result record and its exact selected `MRdat`.	Diagnostic tests and plots; optional Steiger-filtered instruments and primary refit.	At least four instruments; unchanged primary scale and SNP set; verified trait metadata for Steiger analyses.
`analysis-method-comparison`	Compare methods for one directed pair.	Numerical comparison and cross-method visualization.	Completed method result records for the same pair.	Comparison table, effect plots and numerical diagnostics.	At least two usable method estimates; documented scales and units for effect comparison.
`analysis-bidirectional`	Compare forward and reverse estimates.	Direction-wise result integration and visualization.	Completed results for both directions of each trait pair.	Summary table, mirrored forest plots and quadrant plot.	At least one shared method across directions; documented direction-specific scales and units.
`analysis-mediation`	Summarize two-step mediation paths.	Product-of-coefficients estimation and delta-method uncertainty.	Completed $X \to Y$, $X \to M$ and $M \to Y$ results.	Total and indirect effects, mediation proportions and forest plots.	A shared method across all three paths; compatible effect scales and mediator units; zero component covariance assumed.
`analysis-replication`	Compare studies addressing the same trait pair.	Study-level comparison without pooling estimates.	Completed results and study identifiers for each combination.	Summary table, forest matrix and concordance diagnostics.	At least two study combinations; comparable traits and units; sample independence needs separate evidence.
`analysis-risk-factor-prioritization`	Compare candidate exposures for one outcome.	Within-method Benjamini–Hochberg adjustment and evidence matrices.	Completed results for multiple exposures and one outcome.	Summary table, adjusted q values, forest plot and method matrix.	At least two distinct exposures; raw effect magnitudes require compatible units for comparison.
`analysis-result-synthesis`	Integrate results across directed trait pairs.	Pair-level result aggregation and forest plotting.	Completed method results grouped by directed pair.	Combined summary table, forest plot and coverage diagnostics.	At least two distinct directed pairs; retain pair identities and method-specific effect scales.

3.2 30-query benchmark

Supplementary Table 2: Complete queries and classifications for the 30-query benchmark.

Query ID	Classification	Method demand	Data pathway	Complete query
Q01	Executable	Standard/default MR	Source specified	I want to investigate whether Vitamin D affects ease of skin tanning. Use UKB as the data source and run MR analysis.
Q02	Executable	Standard/default MR	Source specified	I want a quick check of whether grip strength affects astigmatism. Use OpenGWAS as the data source and run a simple MR analysis.
Q03	Executable	Standard/default MR	Source specified	I want to check whether red blood cell count affects appendicitis. Please find data from GWAS Catalog and run an MR analysis.
Q04	Executable	Standard/default MR	Source specified	I want to study whether gallstones increase the risk of acute pancreatitis. Use FinnGen as the data source and run an MR analysis.
Q05	Executable	Standard/default MR	Trait discovery	I want to use MR to analyze whether serum urate has a causal effect on gout.
Q06	Executable	Standard/default MR	Source specified	Does height cause black hair color? Use UKB data to run MR for me.
Q07	Executable	Multi-method analysis	Source specified	Does systolic blood pressure (SBP) affect CAD? Use OpenGWAS data and run IVW and MR-Egger.
Q08	Executable	Multi-method analysis	Source specified	I want to analyze whether diastolic blood pressure (DBP) affects CAD. Please find the corresponding data from GWAS Catalog and run MR with IVW and MR-Egger.
Q09	Boundary	Standard/default MR	Source specified	Does left-handedness tendency affect cataract risk? Use FinnGen as the data source and run an MR analysis.
Q10	Executable	Standard/default MR	Trait discovery	Does regular alcohol drinking cause liver cirrhosis? Use the default data source and run an MR analysis.
Q11	Executable	Multi-method analysis	ID specified	I have an outcome GWAS file at `./data/Q11/dataset_1f36c4cb.tsv.bgz` , sample size is 411,779. Use Pan-UKB phenotype 21002 as the body weight exposure. I want to check whether body weight causes this outcome, using IVW and MR-APSS.
Q12	Executable	Standard/default MR	Source specified	I want to check whether platelet count affects myopia. Use OpenGWAS as the data source and run a simple MR analysis.
Q13	Executable	Standard/default MR	Source specified	I want to explore whether age at menopause affects Dental caries. Please find data from GWAS Catalog and run MR.
Q14	Boundary	Standard/default MR	Source specified	I want to analyze whether atrial fibrillation increases the risk of cardioembolic stroke. Use FinnGen as the data source and run MR.

Continued on next page

Supplementary Table 2 (continued)

Query ID	Classification	Method demand	Data pathway	Complete query
Q15	Executable	Standard/default MR	Trait discovery	Does bone density affect osteoporotic fracture risk? Run MR directly with available data.
Q16	Executable	Specialized single-method	ID specified	I want to study whether right arm fat percentage affects black hair. The exposure UKB phenotype is 23119, the outcome is black hair, and the MR method is MR-Lasso.
Q17	Executable	Standard/default MR	ID specified	I want to check whether educational attainment affects otitis media. The exposure GWAS ID is ieu-a-1239, the outcome GWAS ID is finn-b-H8_OTIMEDNAS, and the data source is OpenGWAS. Please run an MR analysis.
Q18	Boundary	Multi-method analysis	Local file	I want to study whether the exposure affects the outcome. Use local data: the exposure file is `./data/Q18/dataset_7f3a2c1e.tsv.gz`, and the outcome file is `./data/Q18/dataset_b9d84e60.tsv.gz`. Please compare CAUSE, MR-APSS, and MR-ConMix results.
Q19	Executable	Standard/default MR	ID specified	I want a quick check of whether COPD affects COPD-related respiratory insufficiency. The exposure ID is finn-b-J10_COPD, the outcome ID is finn-b-COPD_INSUFFICIENCY, and I want a quick MR analysis.
Q20	Executable	Standard/default MR	ID specified	I want to investigate whether ApoB affects myocardial infarction. The exposure ID is ieu-b-108, the outcome ID is ieu-a-798, and I want an MR analysis.
Q21	Executable	Specialized single-method	ID specified	I want to study whether Crohn disease affects the outcome. The exposure ID is ebi-a-GCST004132, the outcome file is `./data/Q21/dataset_49885698.txt.gz`, and the method is cML-MA.
Q22	Executable	Specialized single-method	ID specified	I want to check whether BMI affects the outcome. Search for data in UKB; the exposure ID is 21001, the outcome file is `./data/Q22/dataset_0328e74d.gz`, and the method is MR-ConMix.
Q23	Executable	Specialized single-method	ID specified	I want to analyze whether insomnia affects the outcome. The exposure ID is GCST007387, the outcome file is `./data/Q23/dataset_1c1655f0.gz`, and the method is DIVW.
Q24	Boundary	Standard/default MR	ID specified	I want to investigate whether the exposure affects Anorexia, using the exposure data inside `./data/Q24/dataset_1e0e9c2f.txt.gz` and the outcome data with ID GCST008465.
Q25	Executable	Multi-method analysis	Local file	The exposure is at `./data/Q25/dataset_b1394586.txt.gz`, and the outcome is at `./data/Q25/dataset_85f96c8b.txt.gz`. Use MRMix and ivw, and run an MR analysis.
Q26	Executable	Multi-method analysis	Local file	I have two GWAS files: the exposure is the exposure trait: `./data/Q26/dataset_afa88e8a.txt`, and the outcome is the outcome trait: `./data/Q26/dataset_ce0a4ca2.txt`. For the exposure , total sample size = 404291; for the outcome , total sample size = 18988. Please run MR with cML-MA and ivw.

Continued on next page

Supplementary Table 2 (continued)

Query ID	Classification	Method demand	Data pathway	Complete query
Q27	Executable	Specialized single-method	Local file	I have an exposure file `./data/Q27/dataset_0e2182ae.tsv.gz` and an outcome file `./data/Q27/dataset_b9d84e60.tsv.gz`. I want to check whether the exposure affects the outcome, using MR-RAPS.
Q28	Executable	Multi-method analysis	Local file	I want to use local files to analyze whether the exposure affects the outcome. The exposure file is `./data/Q28/dataset_91d6e693.txt.gz`, and the outcome file is `./data/Q28/dataset_77a21037.txt.gz`. For the exposure, total sample size = 164610. Please provide ivw, MR-PRESSO, and MR-Robust results.
Q29	Executable	Multi-method analysis	Local file	I have an exposure GWAS file `./data/Q29/dataset_b5686fff.tsv.bgz` and an outcome file `./data/Q29/dataset_c48c054e.txt`. The exposure file already has n_complete_samples; the outcome total sample size is 185000. I want to check whether the exposure affects the outcome, using MRMix and ivw.
Q30	Executable	Specialized single-method	Local file	The exposure file is `./data/Q30/dataset_89666702.gz`, and the outcome file is `./data/Q30/dataset_47288f6a.gz`. Please analyze the effect of the exposure on the outcome with MR-APSS.

Classification. Executable (26 queries), the requested MR analysis can be completed under the specified conditions; Boundary (four queries), a data-access or analytical-input limitation requires recognition and an appropriate response. These categories describe the query design, rather than whether an individual agent completed the task.

Method demand. Standard/default MR, an MR analysis without a specified specialized method or an explicit request for multiple methods; Specialized single-method, analysis with one specified specialized MR method; Multi-method analysis, analysis with two or more specified MR methods.

Data pathway. Local file, both exposure and outcome GWAS datasets are supplied as local files; ID specified, at least one GWAS study or phenotype identifier is provided; Source specified, a data source is named without fixed study identifiers; Trait discovery, traits are named without fixed study identifiers or a specified data source. ID-specified queries may also supply a local file for the other trait.

Query IDs. The queries use the same IDs as Supplementary Fig. 1 and Supplementary Table 7.

Boundary conditions in the benchmark. No downloadable left-handedness GWAS (Q09) or executable downloadable cardioembolic-stroke GWAS (Q14) under the FinnGen-only constraint; missing exposure sample-size information required for MR-APSS (Q18); and no directly downloadable full summary statistics for GCST008465 through the benchmark acquisition route (Q24). Boundary handling is assessed separately using B1 (Supplementary Table 6).

Supplementary Table 3: MRAgent workflow scope and execution outcomes in the 30-query benchmark.

Query ID(s)	Scope	Reason / observed behavior
A. In-scope queries: observed execution		
Q02	✓	**Estimates and reports produced.** Five MR methods used 147 SNPs; all estimates were non-significant, with diagnostic plots saved.
Q05	✓	**Estimates produced; reporting errors.** Synonym expansion enabled MR with 304 SNPs but repeated the same GWAS pair. Missing funnel plots interrupted report assembly.
Q07	✓	**Requested methods executed.** Ran IVW, MR-Egger and three other methods (39 SNPs); reported discordant estimates and strong heterogeneity, with cautious interpretation.
Q10	✓	**No estimates: GWAS lookup error.** No catalogue matches were found for the supplied traits or returned synonyms; GWAS ID assignment raised an error.
Q12	✓	**Estimates produced; interpretation caveat.** Ran five methods (709 SNPs), reporting no significant effects. Noted heterogeneity but called the null finding robust in the final summary.
Q15	✓	**No estimates: outcome not matched.** Expanded both trait names and found exposure studies, but no matching outcome study or executable study pair.
B. Out-of-scope queries: grouped by limiting requirement		
Q01, Q06	×	Cannot enforce UKB-only data selection.
Q03, Q08, Q13	×	GWAS Catalog retrieval required.
Q04, Q09, Q14	×	Cannot enforce FinnGen-only data selection.
Q16	×	Specified UKB phenotype ID and MR-Lasso required.
Q17, Q19, Q20	×	Fixed GWAS study IDs required.
Q11, Q18, Q21–Q30	×	Local GWAS inputs required.

Supplementary Table 4: Tools available in the current BioMedAgent configuration.

Function	**Tools**	**Inputs and outputs**
OpenGWAS access	`opengwas_gwasinfo`, `opengwas_extract_instruments`, `opengwas_extract_outcome`	Study identifiers and SNP lists; returns study metadata, exposure instruments and outcome associations.
GWAS preparation	`gwas_format`, `gwas_clump`, `gwas_harmonise`	GWAS tables and specified settings; returns formatted, LD-clumped and allele-aligned data.
MR estimation	`mr_ivw`, `mr_egger`, `mr_weighted_median`, `mr_weighted_mode`, `mr_conmix`, `mr_robust`, `mr_raps`, `mr_mix`, `mr_cml`, `mr_divw`, `mr_radial_ivw`, `mr_radial_egger`	Harmonized association tables and method-specific inputs; returns effect estimates and uncertainty.
MR-APSS	`mrapss_format_gwas`, `mrapss_estimate_parameters`, `mrapss_clump`, `mr_apss`	GWAS data and supplied reference resources; returns formatted data, background parameters, clumped instruments and MR-APSS estimates through separate calls.
CAUSE	`cause_merge`, `cause_estimate_parameters`, `cause_fit`	Genome-wide GWAS tables and specified variants; returns merged data, nuisance parameters and fitted causal and sharing models.
Sensitivity analyses	`mr_heterogeneity`, `mr_pleiotropy`, `mr_presso`	Harmonized association tables; returns heterogeneity, MR-Egger intercept and MR-PRESSO test results.
Other biomedical tools	`cel2matrix`, `survival_curve`, `t_test`	Expression or sample-level data; returns expression matrices, survival curves or independent-samples t-test results.

Supplementary Table 5: Software and package versions in the analysis environment.

Software or package	**Version**
Python	3.11.15
R	4.5.3
numpy	2.4.6
pandas	3.1.0
scipy	1.17.1
data.table	1.18.6.1
jsonlite	2.0.0
TwoSampleMR	0.7.6
ieugwasr	1.2.0.9000
MRAPSS	0.0.0.9000
MendelianRandomization	0.10.0
MRcML	0.0.0.9000
mr.divw	0.1.0
MRMix	0.1.0
MRPRESSO	1.0
mr.raps	0.4.3
cause	1.2.0.335
loo	2.9.0
RadialMR	1.2.3
glmnet	5.0
robustbase	0.99.7
BWMR source code	Commit c00e818
ggplot2	4.0.3
ggrepel	0.9.8
gridExtra	2.3.1
matplotlib	3.11.2

Supplementary Table 6: Scoring rubric for procedural requirements and boundary honesty in the 30-query benchmark.

ID	Criterion	Scoring rule	Pass (1): example	Fail (0): example
Executable queries: procedural gates ($K_g = 9$)				
G1	Data Match	Use the requested exposure and outcome in the requested causal direction; do not use an unapproved proxy.	Body weight → CAD uses body-weight and CAD GWAS data.	BMI replaces body weight, or exposure and outcome are reversed.
G2	Exposure IV Extraction	Extract a non-empty candidate-IV set from the exposure data.	Candidate SNPs are selected using exposure associations.	SNPs are selected from the outcome, or no exposure candidates are produced.
G3	*P*-value Filter	Apply the requested or declared exposure *P*-value threshold exactly; do not assess instrument strength here.	Every retained candidate satisfies the declared $P < 1 \times 10^{-5}$.	A candidate with $P = 1 \times 10^{-6}$ is retained under $P < 5 \times 10^{-8}$.
G4	LD Clumping	Perform actual LD clumping when more than one IV is used; do not assess reference-panel ancestry here.	Multiple IVs are LD-clumped at a declared r^2 and window; a single-IV analysis also passes because clumping is unnecessary.	Multiple IVs are used without LD clumping, or only physical-distance pruning is performed.
G5	Variant Match	Match associations by valid variant identity and compatible genome build, or use a documented LD-proxy retrieval and target-mapping procedure. Proxy use alone is not a failure; allele-direction handling is assessed in G6.	Exact rsID/build-compatible matching, or documented OpenGWAS/TwoSampleMR proxy retrieval with its LD threshold and target mapping.	An association is assigned to the wrong target, an unsupported substitution is made, or matching is not completed.
G6	Allele Alignment	Align effect alleles, flip beta signs when required, and enforce exclusion flags in every evaluated estimator input.	A reversed pair is corrected and rows marked unsuitable are excluded before the estimator runs.	Opposite alleles are used without sign correction, or an estimator uses rows already marked incompatible or unresolved.
G7	Effect Scale	Supply association estimates to MR on the correct scale.	An odds ratio is converted to $\log(\mathrm{OR})$ before use as beta.	$\mathrm{OR} = 1.20$ is used directly as $\beta = 1.20$.
G8	Method Execution	Run the requested MR method with all method-required inputs and settings valid for that method, and produce a numerical result.	The requested method receives valid required inputs and returns an estimate.	A mandatory method input or setting has the wrong meaning or value, the method fails, or a different method is used without permission.
G9	Result Consistency	Match the reported method, status, and numbers to the execution result; narrative length is not scored.	The reported estimate, SE, and *P*-value match the result artifact.	The trace shows failure, but the report claims success or gives different numbers.

Continued on next page

Supplementary Table 6 (continued)

ID	Criterion	Scoring rule	Pass (1): example	Fail (0): example
Boundary queries: criterion B1				
B1	Boundary Honesty	Satisfy either route A (supported stop) or route B (supported and disclosed adaptation), defined below.	Pause the affected analysis with a supported explanation and request for the missing information/decision; OR execute an appropriate alternative and disclose its scope in the final solution.	Neither route is satisfied: an unexplained crash, unsupported boundary claim, uncompleted alternative, unjustified input approximation, or misleadingly reported substitution.
B1 passing conditions ($K_b = 2$): either condition is sufficient				
b_{q1}	Supported stop	Identify a concrete, evidence-supported scientific or data limitation; pause the affected analysis; explain what is unavailable and what information, resource or decision is needed to continue. A final blocked report requesting that input is sufficient; an interactive question mark or a particular runner status is not required.		
b_{q2}	Supported adaptation	Take a scientifically appropriate action to address the limitation, execute the corresponding analysis, and disclose the action, consequences and result scope in the final solution. The trace must support the action and executed result. An explicitly limited alternative analysis does not establish completion of an unavailable exact-study analysis.		

Supplementary Table 6: scoring notes.

Scoring and notation. Each gate is scored 1 for pass and 0 for fail. Each observed defect is assigned to its corresponding gate. For executable query q, g_{qk} indicates whether gate k was passed; the procedure score is the G1–G9 total (0–9). The system-level mean is calculated across 26 executable queries. Scientific completion requires all nine gates to pass. For boundary query q, b_{q1} and b_{q2} denote the supported-stop and supported-adaptation routes, respectively. B1 passes when either route is satisfied; $K_b = 2$ counts these alternative conditions. Boundary scores are reported separately.

Evidence and method scope. Codex (runtime version 0.164.0) with GPT-6.1 Sol assessed each system's execution traces, intermediate result files and final reports. It assigned binary scores to G1–G9 for executable queries and B1 for boundary queries, using the same rubric for every system. The assessment covers requested methods and additional methods presented as evidence for the answer; findings are attributed to the affected method. G9 compares the reported methods, status and numerical values with execution outputs. Results and resource measurements are matched by session.

Supplementary Table 7: Recorded LLM calls and total tokens for each selected query run.

	Q01		Q02		Q03		Q04		Q05		Q06		Q07		Q08		Q09		Q10	
Agent	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens
MRCAT	25	481,037	10	79,102	24	341,773	25	351,665	25	436,433	25	478,379	10	84,148	21	282,120	10	83,146	25	405,252
Claude Code	142	8,939,044	18	206,798	84	3,903,242	163	9,972,031	37	686,500	60	1,578,957	15	117,818	48	1,581,080	67	4,573,148	20	251,056
Biomni	52	1,507,826	17	146,533	33	532,246	26	330,036	64	2,183,346	18	182,751	2	18,077	29	476,781	42	1,059,173	29	570,702
BioMedAgent	159	738,837	120	453,924	144	686,463	157	738,946	154	686,754	159	817,293	117	401,709	157	784,176	146	691,956	167	818,827

	Q11		Q12		Q13		Q14		Q15		Q16		Q17		Q18		Q19		Q20	
Agent	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens
MRCAT	20	383,205	9	67,556	24	356,237	9	72,570	9	68,183	23	442,917	25	427,540	7	48,194	11	91,917	25	414,358
Claude Code	172	18,281,712	32	482,784	59	2,077,526	76	3,214,682	53	1,179,886	17	177,111	25	264,537	107	11,613,053	5	18,715	43	1,065,191
Biomni	80	4,091,599	19	152,030	11	81,232	40	765,172	29	550,355	52	1,052,791	14	106,788	53	1,814,573	15	130,885	12	114,609
BioMedAgent	220	1,456,838	129	528,830	150	798,901	166	742,154	164	649,556	125	407,148	139	606,380	250	1,651,627	119	451,895	142	574,048

	Q21		Q22		Q23		Q24		Q25		Q26		Q27		Q28		Q29		Q30	
Agent	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens
MRCAT	18	247,983	19	329,607	21	270,986	10	118,453	17	266,953	15	199,755	14	176,431	17	217,356	16	218,905	14	183,322
Claude Code	38	1,085,805	14	165,694	114	7,068,504	70	3,298,591	42	961,971	73	2,265,365	37	1,330,168	34	458,581	45	1,197,024	67	2,751,247
Biomni	17	235,709	20	227,171	47	863,904	72	3,390,811	18	211,947	16	142,733	12	102,258	27	264,805	20	253,388	69	2,565,650
BioMedAgent	138	554,999	131	520,868	163	580,404	215	1,091,510	119	409,413	119	445,191	110	368,128	203	906,100	552	1,079,267	126	595,633

Token and call counts cover the selected sessions used for scoring and include recorded cache tokens. Paired resource plots include executable queries with all nine gates passed and complete usage records for both systems: 17 MRCAT–Claude Code pairs, 11 MRCAT–Biomni pairs and 4 MRCAT–BioMedAgent pairs. Shaded comparator entries are paired with MRCAT for the same query; MRCAT entries are shaded when included in at least one comparison. Boundary runs are included in this table and excluded from the paired plots.

3.3 Comparisons with expert-curated analyses and negative-control analyses

Supplementary Table 8: Original GWAS sources for the MR-APSS example and six negative-control tasks.

Task	Exposure dataset	Outcome dataset	Exposure download URL	Outcome download URL
MR-APSS example	Body mass index (BMI); UK Biobank (UKB), f.21001.0.0.	Type 2 diabetes; Xue et al. (2018).	https://atlas.ctglab.nl/ukb2_sumstats/f.21001.0.0_res.EUR.sumstats.MACfilt.txt.gz	https://cnsgenomics.com/data/t2d/Xue_et_al_T2D_META_Nat_Commun_2018.gz
Positive control	High blood pressure; UKB, 6150_4.	Angina; UKB, 20002_1074.	http://atlas.ctglab.nl/ukb2_sumstats/6150_4_logistic.EUR.sumstats.MACfilt.txt.gz	http://atlas.ctglab.nl/ukb2_sumstats/20002_1074_logistic.EUR.sumstats.MACfilt.txt.gz
Positive control	LDL cholesterol (LDL-C); UKB, Neale Lab (2018), 30780_raw.	Coronary artery disease (CAD); CARDIoGRAMplusC4D (2015).	https://broad-ukb-sumstats-us-east-1.s3.amazonaws.com/round2/additive-tsvs/30780_raw.gwas.imputed_v3.both_sexes.varorder.tsv.bgz	http://www.cardiogramplusc4d.org/media/cardiogramplusc4d-consortium/data-downloads/cad.additive.Oct2015.pub.zip
Negative Q01	BMI; UKB, f.21001.0.0.	Dark brown hair; UKB, 1747_4.	https://atlas.ctglab.nl/ukb2_sumstats/f.21001.0.0_res.EUR.sumstats.MACfilt.txt.gz	https://atlas.ctglab.nl/ukb2_sumstats/1747_4_logistic.EUR.sumstats.MACfilt.txt.gz
Negative Q02	Sleeplessness / insomnia; UKB, f.1200.0.0.	Black hair; UKB, 1747_5.	http://atlas.ctglab.nl/ukb2_sumstats/f.1200.0.0_res.EUR.sumstats.MACfilt.txt.gz	https://atlas.ctglab.nl/ukb2_sumstats/1747_5_logistic.EUR.sumstats.MACfilt.txt.gz
Negative Q03	Automobile speeding propensity; $N = 404{,}291$.	Childhood aggression; EAGLE.	https://thessgac.com/papers/2	https://assets.vu.nl/fdcfbfb4-ea5f-0080-b339-861bb5cb584d/2c7e7d91-c95b-4617-8211-6e7fa1990bab/EAGLE_aggression_METAL_RESULTS.zip
Negative Q04	Parental survival; LifeGen phase 2.	Childhood BMI; GCST90002409.	https://datashare.ed.ac.uk/bitstreams/d3d9edc2-ade5-453c-82ab-0f7721b3dcc6/download	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90002001-GCST90003000/GCST90002409/GCST90002409_buildGRCh37.tsv.gz
Negative Q05	Frailty index; Atkins et al. (2020), UKB participants of European ancestry aged 60–70 years, $N = 164{,}610$.	Fetal birth weight; European meta-analysis (2019).	https://ndownloader.figshare.com/files/23063870	https://egg-consortium.org/BW5/Fetal_BW_European_meta.NG2019.txt.gz
Negative Q06	Worrier / anxious feelings; UKB field 1980, Neale Lab, both sexes.	Ease of skin tanning; UKB field 1727, Neale Lab, both sexes.	https://broad-ukb-sumstats-us-east-1.s3.amazonaws.com/round2/additive-tsvs/1980.gwas.imputed_v3.both_sexes.tsv.bgz	https://broad-ukb-sumstats-us-east-1.s3.amazonaws.com/round2/additive-tsvs/1727.gwas.imputed_v3.both_sexes.tsv.bgz

Supplementary Table 9: Complete queries and input mappings for the comparisons with expert-curated analyses.

Case ID	Exposure–outcome mapping	Methods (n)	Complete query
MR-APSS example	BMI → Type 2 diabetes	1	I have two local GWAS summary-statistic files. Treat `./data/mrapss/mrapss-exposure.gz` as the exposure and `./data/mrapss/mrapss-outcome.gz` as the outcome. Please use MR-APSS to analyze whether the exposure causally affects the outcome. Both datasets are from participants of European ancestry. Follow the standard MR-APSS workflow, using an exposure-association threshold of 5e-5 and the OpenGWAS European reference panel to clump variants at r2=0.001 within a 1 Mb window.
Positive control	High blood pressure → Angina	2	The exposure is at `./data/positive_controls/dataset_b1394586.txt.gz`, and the outcome is at `./data/positive_controls/dataset_85f96c8b.txt.gz`. Use MRMix and ivw, and run an MR analysis.
Positive control	LDL-C → CAD	2	I have an exposure GWAS file `./data/positive_controls/dataset_b5686fff.tsv.bgz` and an outcome file `./data/positive_controls/dataset_c48c054e.txt`. The exposure file already has n_complete_samples; the outcome total sample size is 185000. I want to check whether the exposure affects the outcome, using MRMix and ivw.
Negative Q01	BMI → Dark brown hair	7	Use `./data/negative_controls/Q01/dataset_9c5a2e10.gz` as the exposure and `./data/negative_controls/Q01/dataset_0328e74d.gz` as the outcome GWAS file. Analyze whether the exposure causally affects the outcome using MR-Egger, weighted median, weighted mode, MR-APSS, MRMix, MR-ConMix, and CAUSE. Before MR, standardize both GWAS using beta = Z/sqrt(N) and SE = 1/sqrt(N), where Z is the SNP Z score and N is the GWAS sample size. Select exposure instruments using P < 5e-8 and use 1mb window to clump.
Negative Q02	Insomnia → Black hair	13	Use the local exposure GWAS `./data/negative_controls/Q02/dataset_ff1939bc.gz` and `./data/negative_controls/Q02/dataset_1c1655f0.gz` as the outcome GWAS file. Extract the archive before analysis. Analyze whether the exposure causally affects the outcome using IVW, MR-Egger, weighted median, weighted mode, MR-PRESSO, MR-Robust, MR-APSS, MR-RAPS, MR-Lasso, MRMix, MR-ConMix, dIVW, and CAUSE. Before MR, standardize both GWAS using beta = Z/sqrt(N) and SE = 1/sqrt(N), where Z is the SNP Z score and N is the GWAS sample size. Select exposure instruments using P < 5e-8 and use 1mb window to clump.
Negative Q03	Automobile speeding propensity → Childhood aggression	6	Treat `./data/negative_controls/Q03/dataset_afa88e8a.txt` as the exposure GWAS and `./data/negative_controls/Q03/dataset_ce0a4ca2.txt` as the outcome GWAS. The exposure total sample size is 404291. Analyze whether the exposure causally affects the outcome using MR-Egger, weighted median, weighted mode, MRMix, MR-ConMix, and CAUSE. Before MR, standardize both GWAS using beta = Z/sqrt(N) and SE = 1/sqrt(N), where Z is the SNP Z score and N is the GWAS sample size. Select exposure instruments using P < 5e-8 and use 1mb window to clump.
Negative Q04	Parental survival → Childhood BMI	12	Treat `./data/negative_controls/Q04/dataset_0e2182ae.tsv.gz` as the exposure GWAS and `./data/negative_controls/Q04/dataset_b9d84e60.tsv.gz` as the outcome GWAS. Analyze whether the exposure causally affects the outcome using IVW, MR-Egger, weighted median, weighted mode, MR-PRESSO, MR-Robust, MR-APSS, MR-RAPS, MR-Lasso, MR-ConMix, dIVW, and CAUSE. Before MR, standardize both GWAS using beta = Z/sqrt(N) and SE = 1/sqrt(N), where Z is the SNP Z score and N is the GWAS sample size. Select exposure instruments using P < 5e-8 and use 1mb window to clump.

Continued on next page

Supplementary Table 9 (continued)

Case ID	Exposure–outcome mapping	Methods (n)	Complete query
Negative Q05	Frailty index → Fetal birth weight	13	Treat `./data/negative_controls/Q05/dataset_91d6e693.txt.gz` as the exposure GWAS and `./data/negative_controls/Q05/dataset_77a21037.txt.gz` as the outcome GWAS. The exposure total sample size is 164610, and the outcome file contains its sample-size column. Analyze whether the exposure causally affects the outcome using IVW, MR-Egger, weighted median, weighted mode, MR-PRESSO, MR-Robust, MR-APSS, MR-RAPS, MR-Lasso, MRMix, MR-ConMix, dIVW, and CAUSE. Before MR, standardize both GWAS using beta = Z/sqrt(N) and SE = 1/sqrt(N), where Z is the SNP Z score and N is the GWAS sample size. Select exposure instruments using P < 5e-8 and use 1mb window to clump.
Negative Q06	Worrier / anxious feelings → Ease of skin tanning	2	Treat `./data/negative_controls/Q06/dataset_c8197ae4.bgz` as the exposure GWAS and `./data/negative_controls/Q06/dataset_3a6d9f20.bgz` as the outcome GWAS. Both datasets are from participants of European ancestry. Analyze whether the exposure causally affects the outcome using MR-APSS and CAUSE. Before MR, standardize both GWAS using beta = Z/sqrt(N) and SE = 1/sqrt(N), where Z is the SNP Z score and N is the GWAS sample size. Select exposure instruments using P < 5e-8 and use 1mb window to clump.

Case identities and inputs. Negative Q01–Q06 are the six negative-control cases, independently numbered from the 30-query benchmark. The exposure–outcome column identifies the phenotypes for readers; the requests supplied to the agents used file references without these trait labels. Original GWAS sources are listed in Supplementary Table 8. The six negative-control cases request 53 method-specific analyses in total.

Supplementary Table 10: Positive-control results for three exposure–outcome pairs.

Analysis	Method	SNPs (n)	β	SE	95% Wald CI	P value
Body mass index → Type 2 diabetes						
Reference	MR-APSS	1,227	0.2090	0.0330	[0.1443, 0.2737]	2.40×10^{-10}
MRCAT	MR-APSS	1,227	0.2090	0.0330	[0.1443, 0.2737]	2.41×10^{-10}
Claude Code	MR-APSS	1,228	0.2127	0.0266	[0.1606, 0.2648]	1.39×10^{-15}
Biomni	MR-APSS	7,442	−0.0835	0.00456	[−0.0924, −0.0746]	6.23×10^{-75}
High blood pressure → Angina						
Reference	IVW (fixed-effect)	197	0.1520	0.0109	[0.1306, 0.1733]	3.87×10^{-44}
MRCAT	IVW (random-effects)	167	0.1516	0.0160	[0.1203, 0.1829]	2.50×10^{-21}
Claude Code	IVW (fixed-effect)	207	0.3286	0.0229	[0.284, 0.373]	8.6×10^{-47}
Claude Code	IVW (random-effects)	207	0.3286	0.0313	[0.267, 0.390]	9.4×10^{-26}
Biomni	IVW	17,129	0.3674	0.003378	[0.3608, 0.3740]	$< 10^{-300}$
Reference	MRMix	197	0.1300	0.01836	[0.0940, 0.1660]	1.42×10^{-12}
MRCAT	MRMix	167	0.1300	0.0177	[0.0953, 0.1647]	2.06×10^{-13}
Claude Code	MRMix	207	0.3200	0.0591	[0.204, 0.436]	6.3×10^{-8}
Biomni	MRMix	17,129	0.3600	0.005606	[0.3490, 0.3710]	$< 10^{-300}$
LDL cholesterol → Coronary artery disease						
Reference	IVW (fixed-effect)	142	0.1444	0.01195	[0.1210, 0.1679]	1.29×10^{-33}
Reference	IVW (random-effects)	142	0.1444	0.02585	[0.0938, 0.1951]	2.30×10^{-8}
MRCAT	IVW (random-effects)	121	0.1510	0.0277	[0.0967, 0.2053]	4.94×10^{-8}
Claude Code	IVW (fixed-effect)	157	0.367	0.031	[0.307, 0.428]	1.8×10^{-32}
Claude Code	IVW (random-effects)	157	0.367	0.080	[0.211, 0.524]	4.1×10^{-6}
Biomni	IVW (random-effects)	22,790	0.258	0.007	[0.244, 0.272]	2.25×10^{-297}
Reference	MRMix	142	0.1800	0.03274	[0.1158, 0.2442]	3.84×10^{-8}
MRCAT	MRMix	121	0.2000	0.0367	[0.1281, 0.2719]	5.01×10^{-8}
Claude Code	MRMix	157	0.245	0.098	[0.052, 0.438]	0.013
Biomni	MRMix	22,790	0.400	0.0143	[0.372, 0.428]	8.18×10^{-173}

Supplementary Table 11: Method-specific expert-curated and agent results for the six negative-control queries.

		Expert-curated results		MRCAT		Claude Code		Biomni	
Query	**Method**	β	P	β	P	β	P	β	P
Q01: BMI → dark brown hair									
Q01	MR-Egger	-0.02665	0.5733	-0.02776	0.5581	-0.02498	0.5312	-0.023	0.552
Q01	Weighted median	-0.02345	0.05129	-0.02447	0.04294	-0.02422	0.0455	-0.021	0.077
Q01	Weighted mode	-0.02147	0.5256	-0.02144	0.7053	-0.009948	0.6848	-0.01	0.682
Q01	MR-APSS	-0.008133	0.7626	-0.007301	0.7949	-0.01661	0.4591	-0.031	0.02537
Q01	MRMix	-0.02	0.3456	-0.03	0.1521	-0.04	0.01799	-0.03	0.05166
Q01	MR-ConMix	-0.02	0.06171	-0.02	0.04537	-0.02813	0.01367	-0.03	0.0233
Q01	CAUSE	-0.01313	0.785	-0.01152	0.8041	-0.004896	0.6426	-0.01921	0.4389
Q02: Insomnia → black hair									
Q02	IVW	0.05714	0.1939	0.06324	0.1693	0.1457	5.762×10^{-5}	0.1263	0.001809
Q02	MR-Egger	0.1155	0.6581	0.2372	0.5837	0.2783	0.4513	0.1067	0.6407
Q02	Weighted median	0.09479	0.05314	0.08001	0.09789	0.1506	0.001045	0.1014	0.03801
Q02	Weighted mode	0.09838	0.2205	0.1114	0.3989	0.1515	0.1744	0.09551	0.6802
Q02	MR-PRESSO	0.0563	0.1681	0.04171	0.3475	0.1457	3.591×10^{-4}	0.1263	0.004659
Q02	MR-Robust	0.05531	0.2132	0.06221	0.2108	0.1325	0.003337	0.119	0.005
Q02	MR-APSS	0.1665	0.111	0.2047	0.1005	0.1977	3.117×10^{-4}	0.1597	0.004387
Q02	MR-RAPS	0.05552	0.2225	0.06354	0.1884	0.1469	1.863×10^{-4}	0.1197	0.005904
Q02	MR-Lasso	0.05699	0.1078	0.06419	0.09079	0.1457	5.762×10^{-5}	0.109	0.005
Q02	MRMix	0.07	0.9044	2.009×10^{-17}	1	2.009×10^{-17}	1	0.01	0.9165
Q02	MR-ConMix	0.05	0.144	0.14	0.1861	0.253	9.745×10^{-4}	0.22	0.19
Q02	dIVW	0.05864	0.1858	0.06499	0.1616	0.1498	4.368×10^{-5}	0.129	0.001
Q02	CAUSE	0.05735	0.7201	0.07646	0.538	0.01713	0.8115	0.09009	0.2935
Q03: Automobile speeding propensity → childhood aggression									
Q03	MR-Egger	-0.6682	0.6213	-0.6682	0.6213	-0.1567	0.9133	-0.157	0.912
Q03	Weighted median	-0.3419	0.1013	-0.3419	0.1077	-0.3808	0.1449	-0.396	0.133
Q03	Weighted mode	0.114	0.7988	0.114	0.8431	-0.3643	0.616	-0.364	0.369
Q03	MRMix	-0.09	0.8783	-0.09	0.8783	-0.42	0.5071	-0.42	0.507
Q03	MR-ConMix	-0.19	0.302	-0.19	0.302	-0.24	0.2917	-0.24	0.292
Q03	CAUSE	-0.5227	0.3051	-0.357	0.4531	-0.5793	0.2807	-0.5692	0.263

Continued on next page

Supplementary Table 11 (continued)

		Expert-curated results		MRCAT		Claude Code		Biomni	
Query	**Method**	β	P	β	P	β	P	β	P
Q04: Parental survival → childhood BMI									
Q04	IVW	-0.2784	0.1109	-0.2784	0.1109	-0.4306	0.2198	-0.1481	0.4764
Q04	MR-Egger	0.1339	0.7623	0.1339	0.7623	0.5894	0.4599	0.2342	0.4314
Q04	Weighted median	-6.911×10^{-4}	0.9974	-6.911×10^{-4}	0.9974	-0.09349	0.6273	-0.05105	0.7845
Q04	Weighted mode	0.02413	0.9194	0.02413	0.9769	0.0033	0.9863	0.04193	0.8266
Q04	MR-PRESSO	-0.2784	0.142	-0.2784	0.142	-0.4457	0.02045	-0.3664	0.05782
Q04	MR-Robust	-0.2762	0.2314	-0.2762	0.2314	-0.2151	0.3532	-0.1481	0.4764
Q04	MR-APSS	-0.05525	0.8388	-0.05217	0.848	0.02069	0.9256	-0.3688	0.3098
Q04	MR-RAPS	-0.2717	0.1491	-0.2717	0.1491	-0.2315	0.07941	-0.2122	0.3746
Q04	MR-Lasso	-0.2784	0.1109	-0.2784	0.1109	-0.2109	0.1714	-0.1593	0.3003
Q04	MR-ConMix	-0.89	0.175	-0.89	0.175	-0.59	—	0	1
Q04	dIVW	-0.2826	0.09442	-0.2826	0.09442	-0.4373	0.2028	-0.3868	0.263
Q04	CAUSE	-0.2708	0.5689	-0.2377	0.6639	-0.3185	0.6886	-0.2005	0.7143
Q05: Frailty index → fetal birth weight									
Q05	IVW	-0.05698	0.4344	-0.05698	0.4344	-0.0839	0.0806	-0.06562	0.1484
Q05	MR-Egger	-2.055	0.1556	-2.055	0.1556	-0.3697	0.6142	-0.1114	0.7554
Q05	Weighted median	-0.0364	0.5684	-0.0364	0.556	-0.09946	0.05422	-0.08906	0.03936
Q05	Weighted mode	-0.00277	0.9829	-0.00277	0.9853	-0.138	0.5758	-0.1481	0.1258
Q05	MR-PRESSO	-0.08931	0.1599	-0.08931	0.1599	-0.08233	0.05172	-0.083	0.03746
Q05	MR-Robust	-0.05798	0.4246	-0.05798	0.4246	-0.09776	0.03078	-0.06728	0.1385
Q05	MR-APSS	-0.119	0.6634	-0.1277	0.7354	-0.1555	0.08403	-0.09875	0.1563
Q05	MR-RAPS	-0.06158	0.4319	-0.06158	0.4319	-0.09352	0.0757	-0.06437	0.1796
Q05	MR-Lasso	-0.05698	0.4344	-0.05406	0.407	-0.1299	5.193×10^{-4}	-0.06467	0.1081
Q05	MRMix	2.009×10^{-17}	1	2.009×10^{-17}	1	-0.07	0.6673	-0.19	0.3441
Q05	MR-ConMix	-0.29	0.09628	-0.29	0.09628	-0.1874	—	-0.2154	0.04928
Q05	dIVW	-0.05864	0.4159	-0.05864	0.4159	-0.0864	0.08712	-0.06742	0.1417
Q05	CAUSE	-0.05737	0.8788	-0.006003	1	-0.09	0.4596	-0.03308	0.8679
Q06: Worrier / anxious feelings → ease of skin tanning									
Q06	MR-APSS	0.1852	0.1155	0.1929	0.1201	0.09715	0.04942	0.2082	0.01661
Q06	CAUSE	0.1598	0.1883	0.159	0.1789	0.05213	0.03793	0.1214	0.3554

Sources and reporting. Query IDs and requested methods follow Supplementary Table 9. The 53 expert-curated results were compiled from the published MRbenchmarking result tables at an instrument-selection threshold of 5×10^{-8}. Only requested methods are included. Values are displayed to four significant figures. Dashes indicate P values omitted from Claude Code's final solution.

Supplementary Table 12: Final-report evidence for interpretation classifications in the six negative-control queries.

System	Query	Consistency with expert-curated conclusions	Excerpt from the final report
MRCAT	Q01	Consistent	"Overall, these results do not provide consistent support for a causal effect."
MRCAT	Q02	Consistent	"The estimates are consistent with no strong evidence of a causal effect"
MRCAT	Q03	Consistent	"Overall, there is no consistent evidence across methods for a causal effect of the exposure on the outcome."
MRCAT	Q04	Consistent	"there was no clear evidence of a causal exposure-to-outcome effect under the requested standardized scale and models."
MRCAT	Q05	Consistent	"the outlier- and model-sensitive methods do not provide consistent evidence of a causal effect."
MRCAT	Q06	Consistent	"Overall, the results are directionally positive but do not establish robust causal evidence."
Claude Code	Q01	Uncertain	"Overall, the evidence for a causal effect of this exposure on the outcome is weak/suggestive at best"
Claude Code	Q02	Uncertain	"The apparent IVW-type signal should therefore be read cautiously."
Claude Code	Q03	Consistent	"No evidence that the exposure causally affects the outcome."
Claude Code	Q04	Consistent	"Overall, there is no convincing evidence that the exposure causally affects the outcome"
Claude Code	Q05	Uncertain	"Weak suggestive evidence of a negative causal effect, not conclusive."
Claude Code	Q06	Inconsistent	"Both methods indicate a positive causal effect of the exposure on the outcome"
Biomni	Q01	Consistent	"A conservative interpretation is insufficient evidence for a causal effect of the exposure on the outcome."
Biomni	Q02	Inconsistent	"the overall evidence favors a positive causal effect."
Biomni	Q03	Consistent	"No significant causal effect of the exposure on the outcome was detected."
Biomni	Q04	Consistent	"There is no robust evidence that the exposure causally affects the outcome"
Biomni	Q05	Consistent	"there is no robust evidence that the exposure causally affects the outcome."
Biomni	Q06	Uncertain	"Taken together, the evidence for a causal effect of the exposure on the outcome is weak/uncertain."

Notes. Classifications summarize the final interpretation delivered to the user, as shown in Fig. 3e. Consistency with expert-curated conclusions denotes an overall conclusion that causal support is insufficient or absent; inconsistency with expert-curated conclusions denotes an affirmative causal claim; uncertain denotes suggestive, weak or unresolved causal evidence. Classification uses the overall conclusion of the complete delivered report.

Supplementary Table 13: Comparison before and after correcting six outcome-effect signs in Claude Code negative-control Q02.

Method	**Original β**	**Corrected β**	**Original P**	**Corrected P**
IVW	0.145698	0.0785529	5.76198×10^{-5}	0.0651276
Weighted median	0.150564	0.0694963	0.00104493	0.151435
MR-PRESSO	0.145698	0.0785529	0.000359082	0.0750232
MR-Robust	0.132524	0.0633169	0.00333725	0.18578
MR-APSS	0.197687	0.110876	0.00031166	0.074551
MR-RAPS	0.146924	0.074019	0.000186265	0.108028
MR-Lasso	0.145698	0.0592458	5.76198×10^{-5}	0.124887
MR-ConMix	0.253039	-0.0201711	0.000974526	0.682604
dIVW	0.14978	0.0807539	4.36785×10^{-5}	0.0608342

Notes. The original and corrected input tables contain the same 31 SNPs. Only the outcome-effect signs change for `rs4767643`, `rs6561715`, `rs4886860`, `rs12049261`, `rs1592757`, `rs8180817`. Exposure effects and both traits' standard errors are unchanged. All nine originally nominally significant results have corrected $P > 0.05$. Values are shown with up to six significant figures. Full-precision results, the six changed variants and both 31-SNP input tables are included in Supplementary Data 2.

Supplementary Table 14: Execution and reporting outcomes of BioMedAgent on the six negative-control queries.

Queries	**Observed outcome**	**Numerical results**
Q01, Q02, Q06	Execution terminated with errors before the requested analyses and final reports were completed.	No completed final report was produced.
Q04, Q05	Data-processing and clumping errors resulted in empty instrument sets, preventing the requested MR analyses.	No MR estimates were produced.
Q03	Method-specific outputs were generated but were not incorporated into the final report. The final analysis also omitted the requested exposure standardization.	Numerical outputs existed in intermediate files, but the final analysis did not satisfy the query requirements.

Notes. Query IDs follow Supplementary Table 9. Empty instrument sets in Q04 and Q05 arose from processing errors.

3.4 Traceable scientific reporting

Supplementary Table 15: GWAS datasets and quality-control counts for the hexadecanedioate–CAD reporting case.

Dataset or processing stage	Shared exposure	Original outcome	FinnGen outcome
A. Dataset identities and recorded study metadata			
Trait	Hexadecanedioate levels	Coronary artery disease	Major coronary heart disease event
Study identifier	`GCST90243839`	`GCST90132314`	`finngen_R12_I9_CHD`
Source	GWAS Catalog	GWAS Catalog	FinnGen, release R12
Cohort	TwinsUK and KORA	Multi-cohort meta-analysis including UK Biobank	FinnGen
Ancestry	European	European	Finnish
Study sample size	6,887	1,165,690; 181,522 cases and 984,168 controls	500,348; 56,650 cases and 443,698 controls
Recorded source genome build	NCBI36	GRCh37	GRCh38
B. Variants retained after each recorded quality-control stage			
Input variants	2,545,171	20,073,070	21,327,062
Ambiguous-variant exclusion	2,153,996	17,235,956	18,292,344
MHC exclusion	2,143,548	17,235,956	18,242,211
Duplicate exclusion	2,143,548	1,184,393	17,409,372
HapMap3 matching	1,134,287	1,184,393	1,195,372
HapMap3 allele-compatibility check	1,134,249	—	1,195,335
MAF filtering (minimum 0.05)	1,026,691	1,099,272	1,056,310
Invalid-*P* exclusion	1,026,691	1,099,272	1,056,310
Missing-value exclusion	1,026,691	1,099,272	1,056,310
Sample-size outlier exclusion	1,026,691	1,096,197	1,056,310
Extreme-association exclusion; final retained variants	1,026,635	1,096,194	1,056,297
Sample sizes used in formatted data	Fixed at 6,887	Per-variant *N*: 1,037,860–1,165,720; median 1,159,040	Fixed at 500,348

Notes. Panel B gives the variants remaining after each QC stage. A dash denotes a stage without a separate count. The exposure dataset was shared between outcome branches. Study-level sample sizes are from source metadata; the original CAD outcome retained its per-variant sample sizes. Genome builds are source-reported. MHC, major histocompatibility complex; MAF, minor allele frequency.

Supplementary Table 16: Pair preparation, instrument selection and sensitivity diagnostics for the hexadecanedioate–CAD reporting case.

Operation or diagnostic	Original CAD branch	FinnGen branch
A. Pair preparation and method-specific instruments		
Harmonized variants	1,020,665	988,118
Exposure candidates at $P \leq 5 \times 10^{-5}$	207	202
OpenGWAS API clumping	$r^2 = 0.001$; 10,000-kb window; clumping $P = 0.999$	$r^2 = 0.001$; 10,000-kb window; clumping $P = 0.999$
Recorded clumping-reference label	EUR	EUR
Recorded LD-score reference	`eur_w_ld_chr`	`eur_w_ld_chr`
Recorded pair ancestry	European	Unknown
LD-clumped variants	31	31
Primary/BWMR instruments at $P \leq 5 \times 10^{-8}$	3	3
Primary/BWMR instruments at $P \leq 5 \times 10^{-7}$	3	3
Primary/BWMR instruments at $P \leq 5 \times 10^{-6}$; selected analysis	10	10
Primary analysis minimum instrument count	4	4
MR-APSS instruments; selection-bias correction threshold	31; 5×10^{-5}	31; 5×10^{-5}
B. Sensitivity diagnostics using the ten primary-analysis instruments		
IVW Cochran's Q	$Q = 16.6493$; df= 9; $P = 0.0545$	$Q = 19.566$; df= 9; $P = 0.0208$
MR-Egger Cochran's Q	$Q = 13.9617$; df= 8; $P = 0.0828$	$Q = 18.2118$; df= 8; $P = 0.0197$
MR-Egger intercept	0.0021; SE= 0.0017; $P = 0.2498$	0.0023; SE= 0.0029; $P = 0.4627$
MR-PRESSO global test	$P = 0.057$	$P = 0.018$
MR-PRESSO outliers	Not tested	One: `rs11045885`
MR-PRESSO raw and corrected estimates	Raw $\beta = -0.0088$; corrected estimate unavailable	Raw $\beta = -0.0086$; corrected $\beta = -0.0005$
MR-PRESSO distortion test	Unavailable	$P = 0.016$
Leave-one-out most influential SNP	`rs11045975`	`rs11045885`
Leave-one-out maximum absolute percentage change	45.7966%	94.1809%
Leave-one-out reversal of IVW direction	No	No

Supplementary Table 17: Complete method-specific estimates for the original and FinnGen branches of the hexadecanedioate–CAD reporting case.

Outcome branch	Method	Exposure P threshold	SNPs	β	SE	95% CI lower	95% CI upper	P
Original CAD	IVW	5×10^{-6}	10	-0.0088	0.006	-0.0205	0.0029	0.1421
Original CAD	MR-Egger	5×10^{-6}	10	-0.0394	0.0253	-0.0979	0.0191	0.1588
Original CAD	Weighted median	5×10^{-6}	10	-0.0131	0.0062	-0.0253	-0.001	0.0344
Original CAD	Weighted mode	5×10^{-6}	10	-0.0175	0.0189	-0.0604	0.0253	0.379
Original CAD	MR-APSS	5×10^{-5}	31	-0.01485545	0.009974351	-0.03440518	0.004694275	0.13639
Original CAD	BWMR	5×10^{-6}	10	-0.009016685	0.006215747	-0.02119955	0.003166178	0.1468858
FinnGen	IVW	5×10^{-6}	10	-0.0086	0.0099	-0.0279	0.0107	0.3847
FinnGen	MR-Egger	5×10^{-6}	10	-0.0417	0.0441	-0.1433	0.06	0.3723
FinnGen	Weighted median	5×10^{-6}	10	-0.0037	0.0092	-0.0217	0.0142	0.6833
FinnGen	Weighted mode	5×10^{-6}	10	-0.0049	0.0276	-0.0672	0.0575	0.8639
FinnGen	MR-APSS	5×10^{-5}	31	-0.02158823	0.01477763	-0.05055238	0.007375919	0.14405
FinnGen	BWMR	5×10^{-6}	10	-0.00884466	0.01023446	-0.0289042	0.01121488	0.3874758

Notes. The exposure was `GCST90243839`; the original and FinnGen outcomes were `GCST90132314` and `finngen_R12_I9_CHD`, respectively (Supplementary Table 15). The threshold column gives the exposure-association cutoff for instrument selection; the final column gives the method-specific MR P value. Values are shown with up to seven significant figures. P values are unadjusted. IVW, inverse-variance weighting; BWMR, Bayesian weighted Mendelian randomization.

3.5 Risk-factor assessment and bidirectional analysis

Supplementary Table 18: GWAS datasets used in the multi-exposure RA and bidirectional T2D–CAD applications.

Application	Phenotype	Study identifier	Resource	Sample size	Build	Variants after QC
RA risk factors	smoking initiation	ieu-b-4877	OpenGWAS	607,291	GRCh37	1,015,051
RA risk factors	Hypothyroidism	GCST90627749	GWAS Catalog	1,786,062	GRCh38	1,062,008
RA risk factors	Graves' disease	GCST90627744	GWAS Catalog	1,881,665	GRCh38	1,065,958
RA risk factors	Educational attainment	GCST90566696	GWAS Catalog	753,152	GRCh37	1,065,259
RA risk factors	Usual walking pace	ukb-b-4711	OpenGWAS	459,915	GRCh37	1,073,175
RA risk factors	Rheumatoid arthritis	GCST90837201	GWAS Catalog	457,512	GRCh37	1,071,703
Bidirectional T2D–CAD	Type 2 diabetes	ebi-a-GCST90018926	OpenGWAS	490,089	GRCh37	1,100,030
Bidirectional T2D–CAD	Coronary artery disease	GCST90132314	GWAS Catalog	1,165,690	GRCh37	1,096,194

Notes. The RA outcome was shared across five exposure analyses; the same T2D and CAD datasets were used in both directions. Genome builds are source-reported. Full metadata are provided in Supplementary Data 2. QC, quality control.

Supplementary Table 19: Recorded instrument-selection settings for the downstream applications.

Exposure	**Outcome**	**Pair-preparation *P* cutoff**	**Primary/BWMR *P* cutoff**	**Primary/BWMR SNPs**	**MR-APSS SNPs**	**MR-APSS selection-bias threshold**
Smoking initiation	Rheumatoid arthritis	5×10^{-5}	5×10^{-8}	79	317	2.06581×10^{-5}
Hypothyroidism	Rheumatoid arthritis	5×10^{-5}	5×10^{-8}	294	514	6.88287×10^{-7}
Graves' disease	Rheumatoid arthritis	5×10^{-5}	5×10^{-8}	38	122	4.61244×10^{-5}
Educational attainment	Rheumatoid arthritis	5×10^{-5}	5×10^{-8}	298	637	2.30789×10^{-6}
Usual walking pace	Rheumatoid arthritis	5×10^{-5}	5×10^{-8}	49	321	2.79222×10^{-5}
Type 2 diabetes	Coronary artery disease	5×10^{-5}	5×10^{-8}	145	354	1.87077×10^{-5}
Coronary artery disease	Type 2 diabetes	5×10^{-5}	5×10^{-8}	152	350	1.41622×10^{-5}

Notes. Every pair used LD clumping through the OpenGWAS API EUR reference, with $r^2 = 0.001$, a 10,000-kb window and clumping parameter $P = 0.999$.

Supplementary Table 20: Complete estimates for five candidate risk factors in relation to rheumatoid arthritis.

Exposure	Method	SNPs	β	SE	95% CI lower	95% CI upper	P	q
Smoking initiation	IVW	79	0.0695249	0.0234123	0.0236367	0.115413	0.00298204	0.00298204
Smoking initiation	MR-Egger	79	0.106419	0.136242	-0.164872	0.377711	0.437135	0.667642
Smoking initiation	Weighted median	79	0.0619904	0.0296177	0.00393966	0.120041	0.0363476	0.0363476
Smoking initiation	Weighted mode	79	0.0633906	0.0936395	-0.123031	0.249813	0.500431	0.789028
Smoking initiation	MR-APSS	317	0.0693384	0.0214319	0.0273318	0.111345	0.0012152	0.0012152
Smoking initiation	BWMR	79	0.0711639	0.0241097	0.023909	0.118419	0.00316054	0.00316054
Hypothyroidism	IVW	294	0.0706286	0.0133479	0.0444666	0.0967905	1.21417×10^{-7}	3.03541×10^{-7}
Hypothyroidism	MR-Egger	294	0.0397274	0.0275588	-0.0145117	0.0939665	0.150501	0.376254
Hypothyroidism	Weighted median	294	0.0555379	0.018437	0.0194014	0.0916745	0.00259272	0.0032409
Hypothyroidism	Weighted mode	294	0.0233957	0.0897641	-0.153268	0.20006	0.794556	0.794556
Hypothyroidism	MR-APSS	514	0.0650941	0.0156894	0.0343429	0.0958453	3.3404×10^{-5}	5.56733×10^{-5}
Hypothyroidism	BWMR	294	0.071229	0.0135197	0.0447303	0.0977277	1.37529×10^{-7}	3.43821×10^{-7}
Graves' disease	IVW	38	0.267963	0.0627027	0.145065	0.39086	1.92393×10^{-5}	2.40491×10^{-5}
Graves' disease	MR-Egger	38	0.342169	0.174069	-0.0108602	0.695198	0.0570828	0.285414
Graves' disease	Weighted median	38	0.207668	0.0614291	0.0872665	0.328069	0.000723266	0.00120544
Graves' disease	Weighted mode	38	0.0993005	0.205157	-0.316386	0.514987	0.631222	0.789028
Graves' disease	MR-APSS	122	0.262768	0.0593938	0.146356	0.37918	9.6817×10^{-6}	4.57625×10^{-5}
Graves' disease	BWMR	38	0.270495	0.0644046	0.144262	0.396728	2.66992×10^{-5}	3.3374×10^{-5}
Educational attainment	IVW	298	-0.0735753	0.0112018	-0.0955308	-0.0516198	5.09337×10^{-11}	2.54668×10^{-10}
Educational attainment	MR-Egger	298	-0.0215722	0.0501893	-0.120345	0.0772008	0.667642	0.667642
Educational attainment	Weighted median	298	-0.0776467	0.0162965	-0.109588	-0.0457055	1.89208×10^{-6}	9.46041×10^{-6}
Educational attainment	Weighted mode	298	-0.0953747	0.0709757	-0.235054	0.0443043	0.18005	0.596938
Educational attainment	MR-APSS	637	-0.071849	0.0213908	-0.113775	-0.029923	0.00078261	0.000978262
Educational attainment	BWMR	298	-0.0750652	0.0115279	-0.0976599	-0.0524704	7.4363×10^{-11}	3.71815×10^{-10}
Usual walking pace	IVW	49	-0.140908	0.0284219	-0.196615	-0.0852016	7.13147×10^{-7}	1.18858×10^{-6}
Usual walking pace	MR-Egger	49	-0.108346	0.192718	-0.496044	0.279352	0.576651	0.667642
Usual walking pace	Weighted median	49	-0.116945	0.033262	-0.182138	-0.0517513	0.000438318	0.0010958
Usual walking pace	Weighted mode	49	-0.0931203	0.0780627	-0.250076	0.0638353	0.238775	0.596938
Usual walking pace	MR-APSS	321	-0.202475	0.0472563	-0.295097	-0.109853	1.8305×10^{-5}	4.57625×10^{-5}
Usual walking pace	BWMR	49	-0.144444	0.029282	-0.201836	-0.087051	8.1037×10^{-7}	1.35062×10^{-6}

Notes. Effects are shown with 95% confidence intervals. P values are unadjusted; q values use Benjamini–Hochberg adjustment across the five exposures separately for each method. Values are shown with up to six significant figures; full-precision estimates are in Supplementary Data 2. Input datasets and settings are listed in Supplementary Tables 18 and 19.

Supplementary Table 21: Complete estimates for the bidirectional T2D–CAD analysis.

Direction	**Method**	**SNPs**	β	**SE**	**95% CI lower**	**95% CI upper**	P	q
Forward	IVW	145	0.0907556	0.0137663	0.0637737	0.117738	4.32221×10^{-11}	1.47248×10^{-10}
Forward	MR-Egger	145	0.0434574	0.0361501	-0.0273968	0.114312	0.231298	0.277557
Forward	Weighted median	145	0.0611671	0.0123143	0.0370311	0.0853031	6.79391×10^{-7}	1.01909×10^{-6}
Forward	Weighted mode	145	0.0653258	0.0634412	-0.0590188	0.18967	0.304874	0.304874
Forward	MR-APSS	354	0.0864624	0.0153161	0.0564428	0.116482	1.65×10^{-8}	3.3×10^{-8}
Forward	BWMR	145	0.0910193	0.013846	0.0638811	0.118157	4.90828×10^{-11}	1.47248×10^{-10}
Reverse	IVW	152	0.100238	0.0323383	0.0368545	0.163621	0.00193745	0.00802188
Reverse	MR-Egger	152	-0.0641181	0.0848457	-0.230416	0.10218	0.451013	0.541215
Reverse	Weighted median	152	0.06694	0.0281718	0.0117232	0.122157	0.0174952	0.0262428
Reverse	Weighted mode	152	-0.0389303	0.151161	-0.335205	0.257345	0.797111	0.797111
Reverse	MR-APSS	350	0.113058	0.0396868	0.0352717	0.190844	0.0043891	0.0087782
Reverse	BWMR	152	0.0954092	0.0317721	0.033136	0.157683	0.00267396	0.00802188

Notes. Effects are shown with 95% confidence intervals. P values are unadjusted; q values use Benjamini–Hochberg adjustment across the six methods separately within each direction. Values are shown with up to six significant figures; full-precision estimates are in Supplementary Data 2. Input datasets and settings are listed in Supplementary Tables 18 and 19. Forward denotes T2D to CAD; reverse denotes CAD to T2D.

4 Supplementary References

[1] Ben Elsworth, Matthew Lyon, Tessa Alexander, Yi Liu, Peter Matthews, Jon Hallett, Phil Bates, Tom Palmer, Valeriia Haberland, George Davey Smith, Jie Zheng, Philip Haycock, Tom R Gaunt, and Gibran Hemani. The MRC IEU OpenGWAS data infrastructure. *bioRxiv*, 2020. doi: 10.1101/2020.08.10.244293. URL `https://doi.org/10.1101/2020.08.10.244293`. Preprint.

[2] Maria Cerezo, Elliot Sollis, Yue Ji, Elizabeth Lewis, Ala Abid, Karatuğ Ozan Bircan, Peggy Hall, James Hayhurst, Sajo John, Abayomi Mosaku, Santhi Ramachandran, Amy Foreman, Arwa Ibrahim, James McLaughlin, Zoë Pendlington, Ray Stefancsik, Samuel A Lambert, Aoife McMahon, Joannella Morales, Thomas Keane, Michael Inouye, Helen Parkinson, and Laura W Harris. The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Research*, 53(D1):D998–D1005, 2025. doi: 10.1093/nar/gkae1070. URL `https://doi.org/10.1093/nar/gkae1070`.

[3] Mitja I. Kurki, Juha Karjalainen, Priit Palta, Timo P. Sipilä, Kati Kristiansson, Kati M. Donner, Mary P. Reeve, Hannele Laivuori, Mervi Aavikko, Mari A. Kaunisto, Anu Loukola, Elisa Lahtela, Hannele Mattsson, Päivi Laiho, Pietro Della Briotta Parolo, Arto A. Lehisto, Masahiro Kanai, Nina Mars, Joel Rämö, Tuomo Kiiskinen, Henrike O. Heyne, Kumar Veerapen, Sina Rüeger, Susanna Lemmelä, Wei Zhou, Sanni Ruotsalainen, Kalle Pärn, Tero Hiekkalinna, Sami Koskelainen, Teemu Paajanen, Vincent Llorens, Javier Gracia-Tabuenca, Harri Siirtola, Kadri Reis, Abdelrahman G. Elnahas, Benjamin Sun, Christopher N. Foley, Katriina Aalto-Setälä, Kaur Alasoo, Mikko Arvas, Kirsi Auro, Shameek Biswas, Argyro Bizaki-Vallaskangas, Olli Carpen, Chia-Yen Chen, Oluwaseun A. Dada, Zhihao Ding, Margaret G. Ehm, Kari Eklund, Martti Färkkilä, Hilary Finucane, Andrea Ganna, Awaisa Ghazal, Robert R. Graham, Eric M. Green, Antti Hakanen, Marco Hautalahti, Åsa K. Hedman, Mikko Hiltunen, Reetta Hinttala, Iiris Hovatta, Xinli Hu, Adriana Huertas-Vazquez, Laura Huilaja, Julie Hunkapiller, Howard Jacob, Jan-Nygaard Jensen, Heikki Joensuu, Sally John, Valtteri Julkunen, Marc Jung, Juhani Junttila, Kai Kaarniranta, Mika Kähönen, Risto Kajanne, Lila Kallio, Reetta Kälviäinen, Jaakko Kaprio, FinnGen, Nurlan Kerimov, Johannes Kettunen, Elina Kilpeläinen, Terhi Kilpi, Katherine Klinger, Veli-Matti Kosma, Teijo Kuopio, Venla Kurra, Triin Laisk, Jari Laukkanen, Nathan Lawless, Aoxing Liu, Simonne Longerich, Reedik Mägi, Johanna Mäkelä, Antti Mäkitie, Anders Malarstig, Arto Mannermaa, Joseph Maranville, Athena Matakidou, Tuomo Meretoja, Sahar V. Mozaffari, Mari E. K. Niemi, Marianna Niemi, Teemu Niiranen, Christopher J. O'Donnell, Ma'en Obeidat, George Okafo, Hanna M. Ollila, Antti Palomäki, Tuula Palotie, Jukka Partanen, Dirk S. Paul, Margit Pelkonen, Rion K. Pendergrass, Slavé Petrovski, Anne Pitkäranta, Adam Platt, David Pulford, Eero Punkka, Pirkko Pussinen, Neha Raghavan, Fedik Rahimov, Deepak Rajpal, Nicole A. Renaud, Bridget Riley-Gillis, Rodosthenis Rodosthenous, Elmo Saarentaus, Aino Salminen, Eveliina Salminen, Veikko Salomaa, Johanna Schleutker, Raisa Serpi, Huei-yi Shen, Richard Siegel, Kaisa Silander, Sanna Siltanen, Sirpa Soini, Hilkka Soininen, Jae Hoon Sul, Ioanna Tachmazidou, Kaisa Tasanen, Pentti Tienari, Sanna Toppila-Salmi, Taru Tukiainen, Tiinamaija Tuomi, Joni A. Turunen, Jacob C. Ulirsch, Felix Vaura, Petri Virolainen, Jeffrey Waring, Dawn Waterworth, Robert Yang, Mari Nelis, Anu Reigo, Andres Metspalu, Lili Milani, Tõnu Esko, Caroline Fox, Aki S. Havulinna, Markus Perola, Samuli Ripatti, Anu Jalanko, Tarja Laitinen, Tomi P. Mäkelä, Robert Plenge, Mark McCarthy, Heiko Runz, Mark J. Daly, and Aarno Palotie. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613(7944):508–518, 2023. doi: 10.1038/s41586-022-05473-8. URL `https://doi.org/10.1038/s41586-022-05473-8`.

[4] Konrad J. Karczewski, Rahul Gupta, Masahiro Kanai, Wenhan Lu, Kristin Tsuo, Ying Wang, Ray-

mond K. Walters, Patrick Turley, Shawneequa Callier, Nirav N. Shah, Nikolas Baya, Duncan S. Palmer, Jacqueline I. Goldstein, Gopal Sarma, Matthew Solomonson, Nathan Cheng, Sam Bryant, Claire Churchhouse, Caroline M. Cusick, Timothy Poterba, John Compitello, Daniel King, Wei Zhou, Cotton Seed, Hilary K. Finucane, Mark J. Daly, Benjamin M. Neale, Elizabeth G. Atkinson, and Alicia R. Martin. Pan-UK Biobank genome-wide association analyses enhance discovery and resolution of ancestry-enriched effects. *Nature Genetics*, 57(10):2408–2417, 2025. doi: 10.1038/s41588-025-02335-7. URL https://doi.org/10.1038/s41588-025-02335-7.

[5] Xianghong Hu, Jia Zhao, Zhixiang Lin, Yang Wang, Heng Peng, Hongyu Zhao, Xiang Wan, and Can Yang. Mendelian randomization for causal inference accounting for pleiotropy and sample structure using genome-wide summary statistics. *Proceedings of the National Academy of Sciences*, 119 (28):e2106858119, 2022. doi: 10.1073/pnas.2106858119. URL https://doi.org/10.1073/pnas.2106858119.

[6] Kyoko Watanabe, Sven Stringer, Oleksandr Frei, Maša Umićević Mirkov, Christiaan de Leeuw, Tinca J. C. Polderman, Sophie van der Sluis, Ole A. Andreassen, Benjamin M. Neale, and Danielle Posthuma. A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51(9):1339–1348, 2019. doi: 10.1038/s41588-019-0481-0. URL https://doi.org/10.1038/s41588-019-0481-0.

[7] Angli Xue, Yang Wu, Zhihong Zhu, Futao Zhang, Kathryn E. Kemper, Zhili Zheng, Loic Yengo, Luke R. Lloyd-Jones, Julia Sidorenko, Yeda Wu, eQTLGen Consortium, Allan F. McRae, Peter M. Visscher, Jian Zeng, and Jian Yang. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature Communications*, 9(1):2941, 2018. doi: 10.1038/s41467-018-04951-w. URL https://doi.org/10.1038/s41467-018-04951-w.

[8] Xianghong Hu, Mingxuan Cai, Jiashun Xiao, Xiaomeng Wan, Zhiwei Wang, Hongyu Zhao, and Can Yang. Benchmarking Mendelian randomization methods for causal inference using genome-wide association study summary statistics. *The American Journal of Human Genetics*, 111(8):1717–1735, 2024. doi: 10.1016/j.ajhg.2024.06.016. URL https://doi.org/10.1016/j.ajhg.2024.06.016.

[9] Stephen Burgess, Adam Butterworth, and Simon G Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665, 2013. doi: 10.1002/gepi.21758. URL https://doi.org/10.1002/gepi.21758.

[10] J. Bowden, G. Davey Smith, and S. Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2): 512–525, 2015. doi: 10.1093/ije/dyv080. URL https://doi.org/10.1093/ije/dyv080.

[11] Jack Bowden, George Davey Smith, Philip C. Haycock, and Stephen Burgess. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314, 2016. doi: 10.1002/gepi.21965. URL https://doi.org/10.1002/gepi.21965.

[12] Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 46(6):1985–1998, 2017. doi: 10.1093/ije/dyx102. URL https://doi.org/10.1093/ije/dyx102.

[13] Jia Zhao, Jingsi Ming, Xianghong Hu, Gang Chen, Jin Liu, and Can Yang. Bayesian weighted Mendelian randomization for causal inference based on summary statistics. *Bioinformatics*, 36(5):1501–1508, 2020. doi: 10.1093/bioinformatics/btz749. URL https://doi.org/10.1093/bioinformatics/btz749.

[14] Haoran Xue, Xiaotong Shen, and Wei Pan. Constrained maximum likelihood-based Mendelian randomization robust to both correlated and uncorrelated pleiotropic effects. *The American Journal of Human Genetics*, 108(7):1251–1269, 2021. doi: 10.1016/j.ajhg.2021.05.014. URL https://doi.org/10.1016/j.ajhg.2021.05.014.

[15] Ting Ye, Jun Shao, and Hyunseung Kang. Debiased inverse-variance weighted estimator in two-sample summary-data Mendelian randomization. *The Annals of Statistics*, 49(4):2079–2100, 2021. doi: 10.1214/20-AOS2027. URL https://doi.org/10.1214/20-AOS2027.

[16] Frank Dudbridge, Bethany Voller, Ruby M. Woodward, Katie L. Saxby, Timothy M. Frayling, Luke C. Pilling, and Jack Bowden. Getting to GRIPS with MR-Egger: Modelling directional pleiotropy independently of allele coding. *PLOS Genetics*, 21(12):e1011967, 2025. doi: 10.1371/journal.pgen.1011967. URL https://doi.org/10.1371/journal.pgen.1011967.

[17] Jessica M. B. Rees, Angela M. Wood, Frank Dudbridge, and Stephen Burgess. Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates. *PLOS ONE*, 14(9): e0222362, 2019. doi: 10.1371/journal.pone.0222362. URL https://doi.org/10.1371/journal.pone.0222362.

[18] Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S. Small. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *The Annals of Statistics*, 48(3):1742–1769, 2020. doi: 10.1214/19-AOS1866. URL https://doi.org/10.1214/19-AOS1866.

[19] Stephen Burgess, Christopher N Foley, Elias Allara, James R Staley, and Joanna M. M. Howson. A robust and efficient method for mendelian randomization with hundreds of genetic variants. *Nature Communications*, 11:376, 2020. doi: 10.1038/s41467-019-14156-4. URL https://doi.org/10.1038/s41467-019-14156-4.

[20] Guanghao Qi and Nilanjan Chatterjee. Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nature Communications*, 10:1941, 2019. doi: 10.1038/s41467-019-09432-2. URL https://doi.org/10.1038/s41467-019-09432-2.

[21] Marie Verbanck, Chia-Yen Chen, Benjamin Neale, and Ron Do. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*, 50(5):693–698, 2018. doi: 10.1038/s41588-018-0099-7. URL https://doi.org/10.1038/s41588-018-0099-7.

[22] Jean Morrison, Nicholas Knoblauch, Joseph H Marcus, Matthew Stephens, and Xin He. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, 52(7):740–747, 2020. doi: 10.1038/s41588-020-0631-4. URL https://doi.org/10.1038/s41588-020-0631-4.

[23] Eleanor Sanderson, M. Maria Glymour, Michael V. Holmes, Hyunseung Kang, Jean Morrison, Marcus R. Munafò, Tom Palmer, C. Mary Schooling, Chris Wallace, Qingyuan Zhao, and George Davey Smith. Mendelian randomization. *Nature Reviews Methods Primers*, 2:6, 2022. doi: 10.1038/s43586-021-00092-5. URL https://doi.org/10.1038/s43586-021-00092-5.

[24] Stephen Burgess, George Davey Smith, Neil M Davies, Frank Dudbridge, Dipender Gill, M Maria Glymour, Fernando P Hartwig, Zoltán Kutalik, Michael V Holmes, Cosetta Minelli, Jean V Morrison, Wei Pan, Caroline L Relton, and Evropi Theodoratou. Guidelines for performing Mendelian randomization investigations: update for summer 2023. *Wellcome Open Research*, 4:186, 2023. doi: 10.12688/wellcomeopenres.15555.3. URL `https://doi.org/10.12688/wellcomeopenres.15555.3`. Version 3; updated 4 August 2023; originally published in 2019.