# Biobank-scale exposome-wide risk factors in cardiometabolic disease: observational, predictive, and causal evidence

Sivateja Tangirala[1], Arjun K Manrai[1], John PA Ioannidis[2]*, Chirag J Patel[1]*

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts.
[2]Department of Prevention Research, Stanford University School of Medicine, Stanford, CA.

*Co-corresponding authors

## Abstract

Cardiovascular disease and diabetes are intricately related and influenced by factors within the "exposome". Distinguishing between correlational and causal risk associations is challenging, especially at exposome scale. Here, we triangulate observational Exposure-Wide Association Study (*ExWAS*) evidence with "randomized" evidence for the exposome using mendelian randomization (MR) for almost 500 exposures. First, the *ExWAS* identified 144 significant factors for coronary artery disease (CAD) and 237 for type 2 diabetes (T2D), with 120 shared between both. These factors had modest predictive ability (variance explained) for both phenotypes. However, genetic-based causality was deduced for only 14 factors in CAD and 16 in T2D, with seven implicated in both. Additionally, we found strong concordance of MR-validated findings between prevalent and incident disease associations (85.7% [12/14] for CAD and 87.5% [14/16] for T2D). Most correlational findings pertain to lifestyle factors (particularly diet), but social educational factors are more prominently highlighted among those with causal support.

## Introduction

Type 2 diabetes (T2D) and coronary artery disease (CAD) are causally related[1] and carry a tremendous burden of disease worldwide. Environmental factors are hypothesized to have a major role in explaining their risk [2]. The exposome [3] is a comprehensive way to assess the collective contribution of environmental and behavioral factors in cardiometabolic outcomes. One complexity of the exposome includes that there are multiple domains to examine, including *social* (e.g., education and income), *behavior* (dietary and lifestyle), *physical-chemical* (e.g., nutrients and chemicals), *biological* (e.g., infection), and *ecosystem* (e.g. air pollution) variables. In the past, exposure association studies have mostly analyzed a few variables from a handful of domains at a time with one or few health outcomes of interest, leading to potentially fragmented, false positive, biased, and irreproducible literature [4], that does not do justice to how multiple exposures are associated with the outcomes. Genetic epidemiology studies have arguably solved some of these issues. Genetic variables, however, are very different from time-varying, non-static, and densely correlated exposure variables [5]. Nevertheless, the large-scale massive assessments of genetic epidemiology can be applied to some extent also in assessing multiple exposures and multiple correlated and related outcomes[6,7].

Exposome-wide association studies (ExWASs), have been proposed in order to systematically analyze hundreds of exposome factors in multiple phenotypes that attempt to analyze across the diverse categories of exposure as a whole, while taking into account testing multiplicity [8,9]. However, the degree by which instrumental variables can guide inference across the exposome is understudied [5]. Exposures of the exposome are known to be densely correlated among themselves [5,10,11]. With the advent of biobank data with both measured genotypes, exposures, and

longitudinal phenotypic outcomes one can harness mendelian randomization (MR)[1,12,13] as a way to triangulate evidence for potentially causal relationships independent of their dense correlation.

We combine ExWAS and MR to identify the exposomic associations with T2D and CAD (Figure 1). First, we conducted an ExWAS systematically testing each of 495 individual factors of the exposome using a sample of 472,240 white European participants from the UK Biobank (UKB). and benchmarked the findings using two-sample MR with a FinnGen sample comprised of 218,957 participants. This allowed us to characterize which among a wide array of exposures have strongest evidence for causality.

## Results

### Baseline characteristics

We report the baseline characteristics of clinical and demographic variables including age, sex, average household income, HbA1c, systolic blood pressure, diastolic blood pressure, BMI, LDL, HDL, triglycerides, total cholesterol, family history for CAD, family history for T2D, prevalent CAD, prevalent T2D, and smoking history of the UK Biobank (UKB) cohort participants in Table 1. Briefly, the average age of participants was 56.76, 54.48% were female participants, and ~64% of participants had an average household income of less than 52,000 Euros. Out of the 472,240 European White individuals we utilize for our analysis, 7,426 have developed CAD and 12,050 have developed T2D (after baseline visit) while considering those who had already developed CAD (n=10,772) or T2D (n=17,303) at baseline separately. Additionally, out of the 218,957 participants in the FinnGen sample we utilize to benchmark our findings via two sample MR, 29,193 have developed T2D and 21,012 have developed CAD [14].

### Distribution of observational association sizes in incident CAD and T2D

We associated 495 exposures (Extended Data Table 1) with CAD and T2D. With FDR<0.05, we identified 144 significant exposures for CAD (Supplemental Table 1, Figure 2, Supplemental Results) and 237 significant exposures for T2D (Supplemental Table 2, Supplemental Figure 1, Supplemental Results). Association per exposomic category (as defined[3]) for each disease are in

Extended Data Table 2. For CAD, exposomic category-wise representation among significant associations range from 2.33% (Ecosystems; exposomic factors pertaining to Ecosystems comprise 2.33% of significant associations) to 58.9% (Lifestyle). Similarly for T2D, category-wise representation among significant associations range from 2.59% (Ecosystems) to 52.8% (Lifestyle) .

For CAD, the IQR of HRs for significant continuous variables (for a 1 standard deviation [SD] change) that are protective factors is [0.89,0.94] and for risk factors it is [1.06, 1.11] while the IQR of HRs for significant binary variables that are protective factors is [0.75, 0.83] and for risk factors it is [1.25, 1.55]. For T2D, respectively, the IQR of HRs for significant continuous variables (for a 1 standard deviation [SD] change) are [0.86, 0.92] and [1.07, 1.13] and the IQR of HRs for significant binary variables are [0.64, 0.80] and [1.23, 1.53].

We found 120 significant exposures shared between T2D and CAD (119 of which are concordant and one discordant in direction of effect). Additionally, we found 117 exposures specific to T2D and 24 exposures specific to CAD. The Pearson correlation between the beta coefficients of CAD and T2D is 0.56 and 0.30 among continuous and binary exposure factors, respectively (Figure 2).

**Concordance of findings between prevalent and incident disease ExWAS**

Additionally, we sought to assess the concordance of findings from ExWAS-identified associations for prevalent disease with those of incident disease. We associated 495 exposures

with prevalent CAD and T2D. With FDR<0.05, we identified 232 significant exposures for prevalent CAD (Supplemental Table 3) and 309 significant exposures for prevalent T2D (Supplemental Table 4, Supplemental Results).

We found 137 significant exposures shared between prevalent and incident CAD (134 of which are concordant and 3 discordant in direction of effect). Furthermore, we found 62 exposures specific to prevalent CAD and 24 exposures specific to incident CAD. The Pearson correlation between the beta coefficients of prevalent and incident CAD is 0.53 and 0.38 among continuous and binary exposure factors, respectively (Supplemental Figure 2). Additionally, considering the prevalent disease ExWAS as our "diagnostic test", we sought to compare the concordance of the prevalent CAD ExWAS with incident CAD ExWAS by computing the sensitivity and specificity of the prevalent CAD ExWAS test. We estimate the sensitivity to be 84.8% and specificity to be 65.2% (Supplemental Table 5).

Similarly, we compare the concordance between exposure associations for prevalent and incident T2D. We found 189 significant exposures shared between prevalent and incident T2D (184 of which are concordant and 5 discordant in direction of effect). Furthermore, we found 48 exposures specific to prevalent T2D and 24 exposures specific to incident T2D. The Pearson correlation between the beta coefficients of prevalent and incident T2D is 0.79 and 0.75 among continuous and binary exposure factors, respectively. We estimate the sensitivity and specificity of the prevalent T2D ExWAS test relative to incident T2D ExWAS to be 88.5% and 61.9%, respectively (Supplemental Table 6).

**MR-based assessment of observational ExWAS associations**

We sought evidence for genetic-based causality and performed bi-directional MR between each exposure-outcome pair for which Genome-wide Association Study [GWAS] summary statistics were available (for 123/144 (85.4%) of CAD FDR significant associations and 182/237 (76.8%) of T2D FDR significant associations) to test whether the observational associations from the ExWAS are potentially causal. In order to enhance comparability, we use odds ratios (ORs) of exposures computed from logistic regressions after assessing the concordance in estimates between hazard ratios computed from Cox proportional hazard models and odds ratios computed from logistic regressions (Supplemental Figures 3 and 4). We visualize the concordance (and/or discordance) between ExWAS and MR ORs among these exposure-disease pairs in Figure 3 for CAD (and Supplemental Figure 5 for T2D).

We identified 14 associations as being nominally significant (p-value less than 0.05) in forward MR and *not* p-value significant in the reverse direction in CAD while also being found to be concordant in direction to the corresponding ExWAS associations. The IQR of the ORs for CAD among MR-validated associations was [0.38, 0.58] and [1.79, 6.15] for protective and risk factors, respectively. The IQR of ORs for T2D among MR-validated associations was [0.25, 0.63] and [1.86, 2.38] for protective and risk factors, respectively. We also formally tested the difference in effect sizes (absolute value of beta estimates) of MR-validated factors and non-MR-validated factors with the Mann-Whitney test. MR-validated factors had stronger effects than non-MR-validated ones for CAD (Mann-Whitney statistic (W) = 571, p-value = 0.0044) and possibly also for T2D (W= 935, p-value = 0.021).

The MR results for CAD and T2D are made available in Supplemental Tables 7 and 8, respectively. Additionally, we found 12 out of the 14 total (85.7%) prevalent ExWAS-identified associations validated by MR to be the same as the MR-validated ones for incident ExWAS for CAD. Similarly, we found 14 out of the 16 (87.5%) total prevalent ExWAS-identified associations validated by MR to be the same as the MR-validated ones for incident ExWAS for T2D.

We describe the associations that were not only identified as significant from the ExWAS but also found to be significant (p-value less than 0.05) from the MR analysis for CAD (while also concordant in direction). The top 3 MR-validated protective factor associations (in order of increasing MR p-value) were having the educational qualification of A levels/AS levels or equivalent (MR OR: 0.39, MR p-value: $1.12 \times 10^{-5}$, ExWAS OR: 0.71, ExWAS FDR: $1.92 \times 10^{-26}$), having the educational qualification of college or university degree (MR OR: 0.57, MR p-value: $3.51 \times 10^{-5}$, ExWAS OR: 0.66, ExWAS FDR: $3.57 \times 10^{-44}$), and the age participant first had sexual intercourse (MR OR: 0.75, MR p-value: $4.08 \times 10^{-5}$, ExWAS OR: 0.82, ExWAS FDR: $3.44 \times 10^{-45}$). The top 3 MR-validated risk factors were Townsend deprivation index at recruitment (MR OR: 1.75, MR p-value: $5.86 \times 10^{-3}$, ExWAS OR: 1.14, ExWAS FDR: $9.23 \times 10^{-20}$), having no educational or other professional qualifications (e.g. nursing, teaching) (MR OR: 1.83, MR p-value: 0.0105, ExWAS OR: 1.51, ExWAS FDR: $6.22 \times 10^{-49}$), and no type of physical activity in the last four weeks (MR OR: 10.5, MR p-value: 0.0313, ExWAS OR: 2.07, ExWAS FDR: $2.37 \times 10^{-73}$).

We also report the proportion of MR-validated exposures by exposomic category for CAD.

Among social factors, 7.87% (7/89) were MR-validated and among lifestyle factors 5.3% (7/132) were MR-validated. More specifically, among education factors, 62.5% (5/8) were MR-validated and among dietary factors only 4.1% (3/73) were MR-validated. None were MR-validated among ecosystems and physical-chemical factors. Similarly, we report the proportion of MR-validated exposures by exposomic category for T2D. Among social factors, 6.9% (6/87) were MR-validated and among lifestyle factors 6.8% (9/132) were MR-validated. More specifically, among education factors, 62.5% (5/8) were MR-validated and among dietary factors only 4.1% (3/73) were MR-validated. 4.8% (1/21) were MR-validated among physical-chemical factors and none were MR-validated among ecosystems factors.

Next, considering ExWAS as our "diagnostic test", we sought to better understand the concordance of our ExWAS test with respect to MR by computing the sensitivity and specificity of ExWAS. Therefore, we extended our MR analysis to compute associations for all ExWAS tested disease-exposure pairs for which GWAS summary statistics were available [518 (52.3% of 990 total ExWAS-tested disease-exposure pairs)]. More specifically, we estimated the degree to which findings were both ExWAS significant and MR significant, or sensitivity (77.8% for CAD and 94.1% for T2D). Additionally, we estimated the degree to which ExWAS non-significant associations were also found to be MR non-significant, or specificity (37.3 % for CAD and 26.3% for T2D) (Extended Data Tables 3 and 4).

We observed a difference in the sensitivity and specificity of ExWAS between the lifestyle and social exposomic variable categories. For lifestyle factors, we estimate 70% sensitivity for CAD and 75% for T2D and (31.3% specificity for CAD and 21.4% for T2D) [Supplemental Tables 9

and 10]. For social factors, the sensitivity was 87.5% for CAD and 85.7% for T2D and specificity was 45% for CAD and 34.6% for T2D [Supplemental Tables 11 and 12]. The correlation between ExWAS and MR beta estimates were 0.146 and -0.0271 among binary and continuous variables, respectively.

Additionally, we report evidence of reverse causal associations among p-value less than 0.05 significant associations identified from MR for both CAD and T2D (~21%). For example, we identified major dietary changes in last 5 years due to illness (Reference category: no major dietary changes due to illness) in relation to CAD (MR OR: $2.159 \times 10^{3}$, MR p-value: $1.37 \times 10^{-2}$, MR reverse OR: 1.01, MR reverse p-value: $4.73 \times 10^{-8}$) and T2D (MR OR: $3.07 \times 10^{5}$, MR p-value: $5.69 \times 10^{-5}$, MR reverse OR: 1.02, MR reverse p-value: $1.5 \times 10^{-17}$).

We identified 7 MR associations (5 education-related variables and 2 lifestyle factors [age first had sexual intercourse and usual walking pace]) in both T2D and CAD (Table 2). These variables included age first had sexual intercourse, educational qualification of college or university degree, usual walking pace, having no educational or other professional qualifications (e.g. nursing, teaching), educational qualification of A levels /AS levels or equivalent, educational qualification of O levels/GCSEs or equivalent, age completed full time education. For example, we identified the educational qualification of college or university degree to be among the top associations for both T2D (MR OR: 0.407, MR p-value: $5.49 \times 10^{-10}$, ExWAS OR: 0.54, ExWAS FDR: $2.64 \times 10^{-137}$) and CAD (MR OR: 0.566, MR p-value: $3.51 \times 10^{-5}$, ExWAS OR: 0.657, ExWAS FDR: $3.57 \times 10^{-44}$). Furthermore, we found all of the factors to be protective

except for having no educational or other professional qualifications (e.g. nursing, teaching) which we found to be a risk conferring factor for both CAD (MR OR: 1.83, MR p-value: 0.0105, ExWAS OR: 1.51, ExWAS FDR: $6.22 \times 10^{-49}$) and T2D (MR OR: 2.59, MR p-value: $2.68 \times 10^{-4}$, ExWAS OR: 1.81, ExWAS FDR: $3.92 \times 10^{-163}$).

**Assessing the variance of cardiometabolic diseases explained by exposomic, demographic, genetic and clinical risk factors**

We contextualize the variance (computed as Nagelkerke $R^2$) explained by validated exposomic factors with demographic risk factors (Extended Data Table 5). Briefly, for each incident disease we run two major sets of models: a) regressing incident disease status on demographic factors and b) regressing disease status on demographic factors and exposomic factors. Additionally, we ran sets of models where we constrained the space of exposomic factors to just MR-validated exposures in addition to separate sets of models considering a wider array of 196 exposures measured (albeit limited by number of complete cases), cumulatively leading to the specification of models A-I (Extended Data Table 5).

For incident CAD, we find that 42.3% of the variance explained by exposomics and demographics can be explained by exposomics alone. Additionally, 15.8% of the variance explained by the model including MR validated exposures and demographics is explained by MR-validated exposures alone. Moreover, we observed how the variance explained by the models change after adjusting for T2D PRS and BMI. Exposomic variables alone account for

60.5% of the variance explained by the full model. Similarly, MR validated exposures alone account for 18.9% of the variance explained by the full model. We describe the T2D results in the Supplemental Results section.

Additionally, we computed the Nagelkerke $R^2$ from logistic regression models that were run by serially adding exposomic factors to the previous model in order of increasing FDR-corrected p-values in addition to the baseline demographic and clinical risk factors (Figure 4). The IQR of the absolute value of pairwise correlations of the 196 exposures was [0.01, 0.08] (Supplemental Figure 6). For CAD and T2D, the Nagelkerke $R^2$ for models ceased to increase when incorporating beyond the top 175 exposures by FDR-corrected p-values.

**Contextualization of MR-validated exposures in risk for cardiometabolic disease with established clinical risk factors**

Finally, we run clinical risk factor models for CAD and T2D consisting of age, sex, family history, BMI, systolic blood pressure, diastolic blood pressure, HbA1c, LDL, HDL, triglycerides, total cholesterol, and smoking history (inspired by prior literature [15,16] **).** We compare the effects of these factors with those of seven MR-validated exposures commonly implicated in CAD and T2D (Supplemental Table 13). As shown, the magnitude of the effects of the MR-validates exposures matches or sometimes even exceeds the magnitude of the effects of established clinical risk factors.

## Discussion

Our analysis integrates ExWAS and MR methods to characterize the exposomic architecture of factors associated with cardiometabolic disease, made possible by use of biobank data, such as UK Biobank and FinnGen. Using ExWAS, we identified 144 (out of 495 [29.1%]) and 237 factors (out of 495 [47.9%]) for CAD and T2D, respectively. The larger number of factors identified for T2D may reflect, at least in part, the larger power available for T2D which was more common an outcome than CAD. Overall, statistically significant associations for these major outcomes seem to be too numerous. This may explain also the plethora of significant associations published in the epidemiological literature, where exposures are usually tested one at a time or a few at a time [17].

However, for both CAD and T2D the magnitude of the effect sizes of ExWAS-identified factors was modest. Furthermore, the vast majority of them showed no causal evidence on MR assessment. MR is an approach that uses genetic variants as "instruments" to test for a potential causal association between exposures and disease [12]. From our MR analysis, the validated factors tended to have somewhat larger effect sizes than those seen with simple observational correlation analyses in ExWAS. Only 14 (9.72%) and 16 (6.75%) of ExWAS findings were concordant with MR findings for the two phenotypes, CAD and T2D, respectively. We observed higher causality rates among social and more specifically, education-related factors than for lifestyle factors, even though the latter accounted for the vast majority of ExWAS associations. While ExWAS may be fairly sensitive when a large database like the UK Biobank is analyzed, specificity is quite low for T2D and CAD relative to MR, if MR is seen as the gold standard.

The strong presence of education-related variables among the apparently causally validated ones is in line with other studies, including MR-based ones. For example, Tillmann et al. found that educational attainment is causal and protective (OR: 0.67 for a 1 SD increase in higher education roughly corresponding to 3.6 years of additional schooling[18]) risk factor for coronary artery disease in the Coronary Artery Disease Genetics Consortium (CARDIoGRAMplusC4D), also suggesting increased levels of education have a protective effect [18]. Additionally, we identified some factors suggestive of behavior including the age first had sexual intercourse in association with CAD and T2D risk and this is also in accordance with prior evidence[19]. Focused and intense "lifestyle" interventions that target weight loss for individuals with elevated glucose have been developed that have causally led to decreases in T2D incidence[20]. Interventions may need to be tailored and tested targeting social and, in particular, educational factors. Such interventions would require a broader community outlook. Life course studies that examine education early in life may also be needed.

Conversely, among many lifestyle dietary factors, none survived MR scrutiny for causality in our data. While some factors may not have been detected in MR analyses due to low power, the overall pattern suggests a disjoint between correlational and causative structures of different types of factors associated with these two major disease phenotypes. With rare exceptions, detected associations with lifestyle factors (most prominent among them being dietary factors may reflect plain correlations rather than causal relationships. This is congruent with the evidence that while millions of published articles present or discuss observational associations with specific dietary and other lifestyle factors, the far smaller corpus of interventional randomized trials relevant to these exposures yields mostly null results[21, 22]. Moreover, other

investigators using MR methods in large genetic consortia for CAD, T2D and ischemic stroke found no causal evidence for any of the dietary factors that they tested and found only one causal association (with fat) for heart failure[23].

Overall, we provide an overview of the currently measured exposomic risk architecture shared between T2D and CAD. Specifically, we can explain up to ~8% and ~12% in CAD and T2D outcomes respectively with current biobank scale measurements leaving much to be ascribed to newer measurements and instruments [3]. However, we find low concordance between MR and ExWAS for both CAD and T2D. Restricting the variables to just MR-implicated findings, we find that these variables explain 2 and 3% of variation, in CAD and T2D respectively. Although a crude comparison, heritability ranges have been reported from 30-50%, leaving much to be explained with the exposures considered in this study. The main reason for lack of concordance includes confounding with ExWAS findings reflecting false positive associations. Additionally, we suggest potential classes of exposures to assay for a larger number of participants and with greater resolution in the future. For example, the physical-chemical exposome (e.g., chemical pollutants, etc.) needs to be more comprehensively assayed at biobank scale [24].

We also note limitations of our study. First, given the small sample size of individuals of non-White ethnic groups, we expect the results to have high uncertainty in MR approaches; therefore, we have excluded other ethnic groups from our analysis. A systematic review of MR studies for CAD suggests that direction of effects of modifiable risk factors tends to be similar in different ethnic groups, but differences in magnitude of effects may not be uncommon[25]**.** Second, MR methods may also have bias and flaws, and they are not a perfect gold standard by any means.

Limited power may lead to false-negative findings for causality[26], but some biases may actually also create larger causal effects[27]. Third, it is unclear if we have addressed confounding for all domains of exposures as there is no consensus on which adjustment factors to use for the different domains; future work could look at optimizing the difference between observational and MR associations as a function of covariate selection. Results may vary depending on what analytical choices are made[28].

Finally, prevalent disease outcome data coupled with genetic variant data as instrumental variables for MR may suffice to identify exposures, especially when working with cohorts without sufficient longitudinal outcome follow-up data. This is evidenced by the high degree of commonality of MR-validated exposures between prevalent and incident disease ExWAS-identified associations (85.7% [12 out of 14 total] for CAD and 87.5% [14 out 16 total] for T2D). However, prevalent and incident outcome data may each have its sets of distinct biases.

In conclusion, MR-validated factors for CAD and T2D are far fewer than the rich range of factors identified by ExWAS. Causality also seems more commonly documented for educational factors, while lifestyle factors have many ExWAS signals, but these rarely have MR causal support. Additional similar analyses in more cohorts with relevant information would be useful to expand evidence on the robustness of these findings.

## Online Methods

### Study population

The UK Biobank cohort is a prospective cohort including over 500,000 participants of ages 40-69 during recruitment from 2006-2010 [13]. Differences between the UK Biobank cohort individuals and the general UK population were studied by Fry et al. in order to better understand sampling uncertainty [14]. Their study suggested that nonparticipants are more likely to be male, younger, and live in more socioeconomically deprived areas than UK Biobank participants [14]. Information regarding how the UK Biobank data is maintained and validated can be found at https://biobank.ndph.ox.ac.uk/~bbdatan/Data_cleaning_overall_doc_showcase_v1.pdf.

The National Research Ethics Service Committee North West Multi-Centre Haydock has approved the UKB cohort research and written informed consent to participate in the study was provided by all participants[15]. We analyzed $n$ = 472,240 European White individuals (out of total $N$ = 502,628 participants in the cohort). The other top three ethnicities represented (by sample size) in the UKB cohort (Indian, Caribbean, and African) were low in sample size (Indian n = 5,951, Caribbean n = 4,517, African n = 3,394) when integrating exposure data and therefore we may not have adequate power for detection of associations. Given the small sample size of individuals of non-white ethnic groups, we expect the results to have high uncertainty and thus the correlations would be weaker; therefore, we have excluded other ethnic groups from our analysis. Approval for the use of this data was approved by the UK Biobank (project ID: 22881). The Harvard internal review board (IRB) deemed the research as non-human subjects research (IRB: IRB16-2145). Formal consent was obtained by the UK Biobank (https://biobank.ctsu.ox.ac.uk/ukb/ukb/docs/Consent.pdf).

## Cardiometabolic disease outcomes

In our study, we investigate two major cardiometabolic incident outcomes in the UKB, including coronary artery disease (CAD) and type 2 diabetes as ascertained by using ICD-10 and self-reported disease status information (Supplemental Table 14). We identify cases of CAD and T2D as ones that occur after the baseline visit (incident cases) while considering individuals who have the disease at baseline (prevalent cases) separately.

## Categorizing ecosystems, lifestyle, social, and physical-chemical exposures

We considered 538 total variables of which we had an adequate number of complete cases to investigate 495 "exposures" that can be categorized (as per Vermeulen et al. [3] ) as ecosystems, lifestyle, social, and physical-chemical factors throughout the paper. We utilize the data for these exposures collected during the participants' baseline visits (2006-2010). These 495 exposures spanned 17 UK Biobank-defined categories (e.g., education, smoking, greenspace and coastal proximity, sun exposure, estimated nutrients yesterday) (Extended Data Table 1).

We averaged quantitative factors (e.g., infectious antigens [25 exposures]) across measurements from multiple visits. For exposures that did not have many observations in subsequent instances, we used only the data from the baseline visit (first instance of measurement collected during 2006–2010) (e.g., environmental factors from the estimated nutrients yesterday category). We also performed rank-based inverse normal transformation (INT) of these factors (as was suggested by Millard et al. [29]). For categorical variables (which were also collected from multiple visits of a participant to the assessment center), we used data from the baseline visit (first

instance of measurement collected during 2006–2010) as this contained the highest number of observations. Additionally, categorical variables with multiple levels were converted to sets of binary variables where each binary variable indicates whether a participant has a given value of this variable (as was suggested by Millard et al. [29]). Ordinal categorical environmental factor variables were analyzed by treating such variables as continuous variables and real-valued quantitative environmental factor variables were scaled.

**Data-driven identification of exposure-disease associations**

We used Cox proportional hazard models to associate each of the 495 factors and each of the two cardiometabolic diseases: CAD (coronary artery disease), T2D (type 2 diabetes) [individually], while adjusting for sex, age, assessment center, ethnicity, average total household income after tax, and 40 genetic principal components (computed and provided by the UK Biobank). We adjust resulting p-values of associations for multiple comparisons using the false discovery rate (FDR) approach [30]. We report hazard ratios and FDR-adjusted *p*-values for the associations. Additionally, we run logistic regression to enhance comparison for downstream analyses (i.e. MR).

**Concordance between observational and MR-based exposure-disease associations**

We performed two-sample bidirectional mendelian randomization for all associations for which there were available GWAS summary statistics for the corresponding exposures (n = 292 exposures). We perform bidirectional MR to account for potential reverse causality [31]. We used GWAS summary statistics for identifying instruments for each exposure from GWAS summary

statistics generated by the Neale Lab[32] and MRC IEU OpenGWAS [33] and made freely accessible on the MR-Base platform [33]. We use the TwoSampleMR R package [33] to test the identified instruments for each exposure for each of the two outcomes using summary statistics of GWAS from the FinnGEN cohort (https://www.finngen.fi/en). Summary statistics from FinnGen were adjusted by sex, age, genotyping batch, and ten principal components. We utilize the FinnGen sample of 218,957 participants to validate our findings via two sample MR. Among these participants, 29,193 have developed T2D and 21,012 have developed CAD [14]. Additionally, we used the following thresholds to ascertain genetic instruments including p-value < $5x10^{-8}$, linkage disequilibrium (LD) $R^2$: 0.001, and clumping distance of 10000 kb. We report causal estimates computed using the inverse-variance weighting (IVW) method between each exposure-disease pair.

**Assessing the variance of cardiometabolic diseases explained by exposomic and demographic risk factors**

We estimated the variance (computed as Nagelkerke $R^2$) explained by validated exposomic factors with demographic risk factors (Extended Data Table 6). Briefly, for each disease we run two major sets of logistic regression models: a) regressing disease status on demographic, genetic risk score (for CAD and T2D) and clinical (BMI for T2D) factors and b) regressing disease status on demographic factors and exposomic factors. Additionally, we ran sets of models where we constrained the space of exposomic factors to just MR-validated ones in addition to separate sets of models considering a wider array of 196 exposures measured (albeit limited by number of complete cases). Finally, we ran logistic regression models serially adding

exposomic factors to the previous model in order of increasing FDR-corrected p-values until we ran the final model including the full wider array of exposomic factors (in addition to the baseline demographic and clinical risk factors). Additionally, pairwise correlations between categorical and numeric exposure variables were computed using the "polycor" R package.

**Acknowledgements**

We thank HMS Research Computing for computing support. The UK Biobank project number is 22881. Funding sources include NIEHS R01 ES032470, NIDDK R01DK137993, NIEHS U24ES036819, Vranos Foundation. The funders had no role in the study design or drafting of the manuscript(s).

## Figure Legends

**Figure 1. Schematic overview of analysis of exposome-wide risk architecture of T2D and CAD.** This schematic diagram depicts our analytic workflow. (a) Exposure-wide association study (ExWAS) was performed to discover exposure-disease associations for T2D and CAD in the UK Biobank. (b) Two Sample bi-directional mendelian randomization (MR) was performed to replicate ExWAS-identified associations and test for causality. Briefly, genetic instruments for exposures were identified from the UK Biobank sample and tested on the FinnGen sample with respect to CAD and T2D

**Figure 2. Exposomic architecture of CAD.** This multipanel figure depicts exposome-wide findings for CAD. (a) **Volcano plot of ExWAS associations for CAD.** We visualize the FDR-corrected p-values on the negative log 10 scale versus hazard ratios of exposures. The red color indicates FDR < 0.05 significant associations and blue color indicates FDR > 0.05 associations. Top five associations are labeled. (b) **Distribution of hazard ratios of ExWAS associations for CAD and T2D stratified by significance.** The distribution of the hazard ratios (HR) computed from the absolute value of the regression beta coefficients for CAD and T2D are depicted with empirical cumulative distribution function plots and are colored by ExWAS significance. The distribution of HRs for exposures that are ExWAS FDR less than 0.05 significant are depicted in red. The distribution of HRs for exposures that are ExWAS FDR greater than 0.05 are depicted in blue. For the underlying dataset, n=472,240. (c) **Distribution of p-values of ExWAS and MR associations for CAD and T2D.** The distributions of all p-values derived from ExWAS for CAD and T2D are depicted with empirical cumulative distribution function plots. For the underlying dataset, n=472,240. (d) **Association size of CAD versus T2D for 495 exposures.**

The hazard ratios (HR) of CAD versus T2D for 495 different exposures. Exposures that are FDR less than $5x10^{-4}$ significant for both CAD and T2D deviate and have HRs either less than 0.5 or greater than 2 for either disease are labeled and shown with error bars corresponding to the 95% confidence intervals for both CAD and T2D. The black dashed line on the scatterplot represents a linear regression line. For the underlying dataset, n=472,240.

**Figure 3. ExWAS association size versus MR causal estimate for CAD.** The odds ratios (OR) computed from ExWAS versus the OR computed from MR for 260 different exposures in association with CAD. The blue color indicates evidence for potential reverse causality and the red color indicates no evidence for potential reverse causality. Triangles indicate associations with significant MR p-values. Horizontal and vertical black lines demarcate OR of 1 (null association). For the underlying datasets, n=472,240 for the UK Biobank sample and n = 218,957 for the FinnGen sample.

**Figure 4. Visualizing the change in variance explained by step-wise addition of exposomic variables to baseline model of demographic and clinical risk factors compared to exposure-specific univariate models.** We visualize the Nagelkerke $R^2$ of each model constructed by step-wise addition of exposomic factors in descending order of FDR significance to the original baseline model comprised of demographic, genetic risk score (for CAD and T2D) and clinical (BMI for T2D) factors. Black dot indicates variance explained by the baseline model. Top exposures (by FDR) for the multivariable analysis are labeled. We also overlay the univariate (exposure) model results. We visualize the cumulative exposure-specific Nagelkerke $R^2$ as we iterate through each univariate (exposure) model in descending order of FDR significance with

smaller point size for both CAD and T2D, separately.

## References

1. Ross, S. *et al.* Mendelian randomization analysis supports the causal role of dysglycaemia and diabetes in the risk of coronary artery disease. *Eur. Heart J.* **36**, 1454–1462 (2015).
2. Münzel, T., Sørensen, M., Hahad, O., Nieuwenhuijsen, M. & Daiber, A. The contribution of the exposome to the burden of cardiovascular disease. *Nat. Rev. Cardiol.* **20**, 651–669 (2023).
3. Vermeulen, R., Schymanski, E. L., Barabási, A.-L. & Miller, G. W. The exposome and health: Where chemistry meets biology. *Science* **367**, 392–396 (2020).
4. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
5. Ioannidis, J. P. A., Loy, E. Y., Poulton, R. & Chia, K. S. Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. *Sci. Transl. Med.* **1**, 7ps8 (2009).
6. Ioannidis, J. P. A. Exposure-wide epidemiology: revisiting Bradford Hill. *Stat. Med.* **35**, 1749–1762 (2016).
7. Fedak, K. M., Bernal, A., Capshaw, Z. A. & Gross, S. Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerg. Themes Epidemiol.* **12**, 14 (2015).
8. Manrai, A. K., Ioannidis, J. P. A. & Patel, C. J. Signals Among Signals: Prioritizing Nongenetic Associations in Massive Data Sets. *Am. J. Epidemiol.* **188**, 846–850 (2019).
9. Wild, C. P. The exposome: from concept to utility. *Int. J. Epidemiol.* **41**, 24–32 (2012).
10. Patel, C. J. & Ioannidis, J. P. A. Placing epidemiological results in the context of

multiplicity and typical correlations of exposures. *J. Epidemiol. Community Health* **68**, 1096–1100 (2014).

11. Patel, C. J. & Manrai, A. K. Development of exposome correlation globes to map out environment-wide associations. *Pac. Symp. Biocomput.* **20**, 231–242 (2015).
12. Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* **362**, k601 (2018).
13. Emdin, C. A., Khera, A. V. & Kathiresan, S. Mendelian Randomization. *JAMA* **318**, 1925–1926 (2017).
14. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
15. Meigs, J. B. *et al.* Genotype Score in Addition to Common Risk Factors for Prediction of Type 2 Diabetes. (2008) doi:10.1056/NEJMoa0804742.
16. He, Y. *et al.* Comparisons of Polyexposure, Polygenic, and Clinical Risk Scores in Risk Prediction of Type 2 Diabetes. *Diabetes Care* **44**, 935–943 (2021).
17. Janiaud, P. *et al.* Validity of observational evidence on putative risk and protective factors: appraisal of 3744 meta-analyses on 57 topics. *BMC Med.* **19**, 157 (2021).
18. Tillmann, T. *et al.* Education and coronary heart disease: mendelian randomisation study. *BMJ* **358**, j3542 (2017).
19. Mills, M. C. *et al.* Identification of 371 genetic variants for age at first sex and birth linked to externalising behaviour. *Nat Hum Behav* **5**, 1717–1730 (2021).
20. Knowler, W. C. *et al.* Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N. Engl. J. Med.* **346**, 393–403 (2002).
21. Trepanowski, J. F. & Ioannidis, J. P. A. Perspective: Limiting dependence on

nonrandomized studies and improving randomized trials in human nutrition research: Why and how. *Adv. Nutr.* **9**, 367–377 (2018).

22. Young, S. S. & Karr, A. Deming, data and observational studies. *Signif. (Oxf.)* **8**, 116–120 (2011).
23. Niu, Y.-Y., Aierken, A. & Feng, L. Unraveling the link between dietary factors and cardiovascular metabolic diseases: Insights from a two-sample Mendelian Randomization investigation. *Heart Lung* **63**, 72–77 (2024).
24. Patel, C. J., Bhattacharya, J. & Butte, A. J. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS One* **5**, e10746 (2010).
25. Silva, S., Fatumo, S. & Nitsch, D. Mendelian randomization studies on coronary artery disease: a systematic review and meta-analysis. *Syst. Rev.* **13**, 29 (2024).
26. Brion, M.-J. A., Shakhbazov, K. & Visscher, P. M. Calculating statistical power in Mendelian randomization studies. *Int. J. Epidemiol.* **42**, 1497–1501 (2013).
27. Burgess, S., Thompson, S. G. & CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. *Int. J. Epidemiol.* **40**, 755–764 (2011).
28. Klau, S., Hoffmann, S., Patel, C. J., Ioannidis, J. P. & Boulesteix, A.-L. Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. *Int. J. Epidemiol.* **50**, 266–278 (2021).
29. Millard, L. A. C., Davies, N. M., Gaunt, T. R., Davey Smith, G. & Tilling, K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *Int. J. Epidemiol.* **47**, 29–35 (2018).
30. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and

Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

31. Richmond, R. C. & Davey Smith, G. Commentary: Orienting causal relationships between two phenotypes using bidirectional Mendelian randomization. *International journal of epidemiology* vol. 48 907–911 (2019).
32. GitHub - Nealelab/UK_Biobank_GWAS: Overview of the data QC, code, and GWAS summary output from the 2017 UK Biobank data release. *GitHub* https://github.com/Nealelab/UK_Biobank_GWAS.
33. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, (2018).

## Tables

| Variable (units) | Label | Mean (SD) or Percentage |
|---|---|---|
| Age (years) | Age | 56.76 (8.028) |
| Sex (%) | Female | 54.48 |
| Sex (%) | Male | 45.52 |
| Avg. household income (%) | < 18,000 | 19.07 |
| Avg. household income (%) | 18,000 to 30,999 | 21.74 |
| Avg. household income (%) | 31,000 to 51,999 | 22.43 |
| Avg. household income (%) | 52,000 to 100,000 | 17.55 |
| Avg. household income (%) | > 100,000 | 4.65 |
| HbA1c (mmol/mol) | HbA1c | 35.96 (6.517) |
| Systolic pressure (mm Hg) | Systolic pressure | 139.9 (19.68) |
| Diastolic pressure (mm Hg) | Diastolic pressure | 82.16 (10.68) |
| BMI (kg/m2) | BMI | 27.41 (4.783) |
| LDL (mmol/L) | LDL | 3.57 (0.87) |
| HDL (mmol/L) | HDL | 1.45 (0.383) |
| Triglycerides (mmol/L) | Triglycerides | 1.75 (1.02) |
| Total cholesterol (mmol/L) | Total cholesterol | 5.71 (1.14) |
| Family history for CAD (%) | Family history for CAD | 39.8 |
| Family history for T2D (%) | Family history for T2D | 17.1 |
| Smoking (%) | Never | 53.94 |
| Smoking (%) | Previous | 35.55 |
| Smoking (%) | Current | 10.5 |
| Prevalent CAD (%) | Prevalent CAD | 2.28 |
| Prevalent T2D (%) | Prevalent T2D | 3.66 |

**Table 1. Sample baseline characteristics.**

| xposure | CAD MR OR | CAD MR p-value | CAD ExWAS OR | CAD ExWAS FDR | T2D MR OR | T2D MR p-value | T2D ExWAS OR | T2D ExW FDR |
|---|---|---|---|---|---|---|---|---|
| ualifications: A vels/AS levels or uivalent | 0.386 | 1.12E-05 | 0.705 | 1.92E-26 | 0.419 | 7.02E-04 | 0.627 | 2.97E- |
| ualifications: ollege or niversity degree | 0.566 | 3.51E-05 | 0.657 | 3.57E-44 | 0.406 | 5.49E-10 | 0.54 | 2.64E-1 |
| ge first had xual intercourse | 0.749 | 4.08E-05 | 0.821 | 3.44E-45 | 0.626 | 1.50E-10 | 0.826 | 6.85E- |
| ge completed full ne education | 0.598 | 3.59E-04 | 0.896 | 9.84E-11 | 0.551 | 1.55E-03 | 0.876 | 3.42E- |
| ualifications: one of the above | 1.825 | 1.05E-02 | 1.51 | 6.22E-49 | 2.591 | 2.68E-04 | 1.81 | 3.92E-1 |
| ualifications: O vels/GCSEs or uivalent | 0.398 | 1.38E-02 | 0.797 | 8.78E-18 | 0.21 | 7.12E-04 | 0.731 | 2.40E- |
| sual walking pace | 0.592 | 1.81E-02 | 0.74 | 7.65E-115 | 0.233 | 1.34E-06 | 0.625 | < 1E-163 |

**Table 2. Exposomic variables significant in both MR and ExWAS and in both phenotypes.**

## Table Legends

**Table 1. Sample baseline characteristics.** This table shows the mean (and standard deviation) or percentage of key clinical and demographic variables for the sample of White UKB participants that were primarily used for our analysis including age, sex, average household income, HbA1c, systolic blood pressure, diastolic blood pressure, BMI, LDL, HDL, triglycerides, total cholesterol, family history for CAD, family history for T2D, prevalent CAD, prevalent T2D, and smoking history.

**Table 2. Exposomic variables significant in both MR and ExWAS and in both phenotypes**. This table includes MR-derived odds ratios (ORs) and p-values and ExWAS-derived ORs and FDR-corrected p-values for the seven exposomic variables we found significant in both MR and ExWAS in both phenotypes.

## Extended Data Table Legends

**Extended Data Table 1. Exposure variable breakdown.** This table shows the exposure categories as defined by Vermeulen et al.[3] as well as the UKB, the number of variables within each category, the class of the variable category (i.e. whether the category contains biomarker (B), geographic (G) or questionnaire (Q) variables), and some example variables within each category.

**Extended Data Table 2. ExWAS-identified enrichment of exposomic categories (as defined by Vermeulen et al.[3]) for each disease.** This table shows the breakdown of the percentage of ExWAS-identified FDR < 0.05 associations by exposure categories as defined by Vermeulen et al for CAD and T2D.

**Extended Data Table 3. Observational versus MR significance classification table for CAD.** This table includes estimates of false positives, false negatives, true positives, true negatives when comparing ExWAS findings with respect to gold-standard causal estimates obtained from MR for CAD. True Positives were ascertained as associations that had observational and MR beta estimates concordant in direction in addition to the aforementioned observational and MR significance criteria.

**Extended Data Table 4. Observational versus MR significance classification table for T2D.** This table includes estimates of false positives, false negatives, true positives, true negatives when comparing ExWAS findings with respect to gold-standard causal estimates obtained from

MR for T2D. True Positives were ascertained as associations that had observational and MR beta estimates concordant in direction in addition to the aforementioned observational and MR significance criteria.

**Extended Data Table 5. Breakdown of variance (R2) explained by exposomic and MR-validated exposomic factors with demographic risk factors.** This table contextualizes the variance (computed as $R^2$) explained by exposomic, MR-validated exposomic factors with demographic risk factors. Delta $R^2$ is computed as the difference between $R^2$ computed from full model and $R^2$ computed model only including demographic and clinical risk factor covariates.

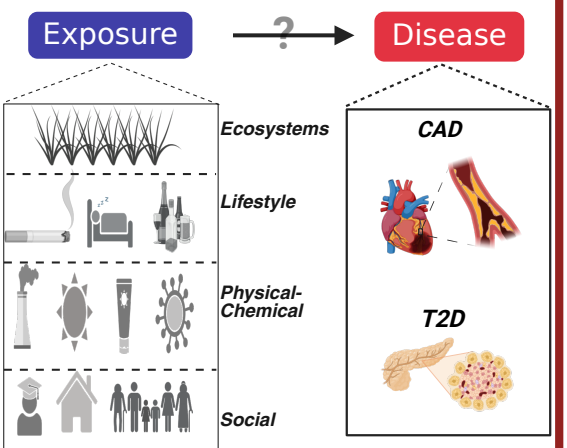
**Extended Data Table 6. Baseline demographic breakdown of samples used for models A-I.** This table shows the breakdown of the mean and shows the mean or percentage of key demographic variables for the samples of White UKB participants that were primarily used for our models A-I including age, sex, average household income, and assessment center.

# Exposomic Risk for Cardiometabolic Disease

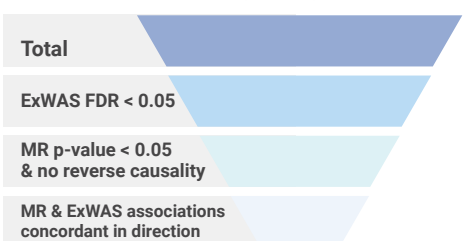
## a. ExWAS

## b. Causal MR

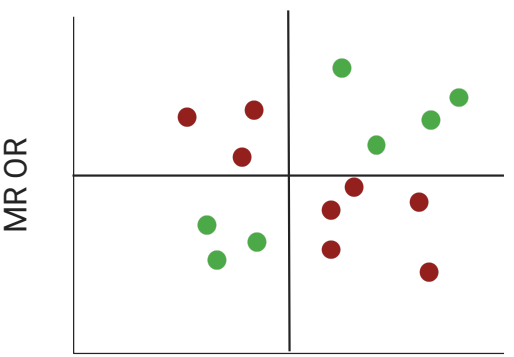
### Discovery of exposure-disease associations




biobank uk



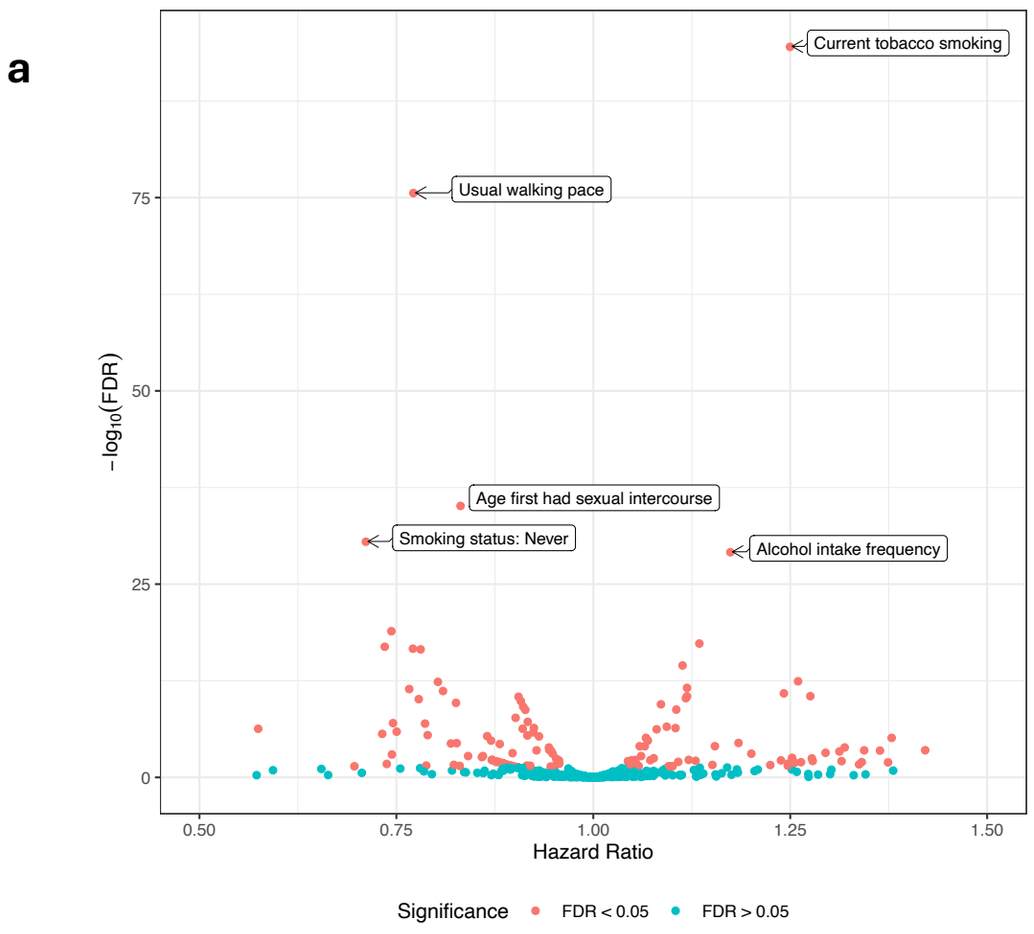

### Replication and causality testing of identified associations

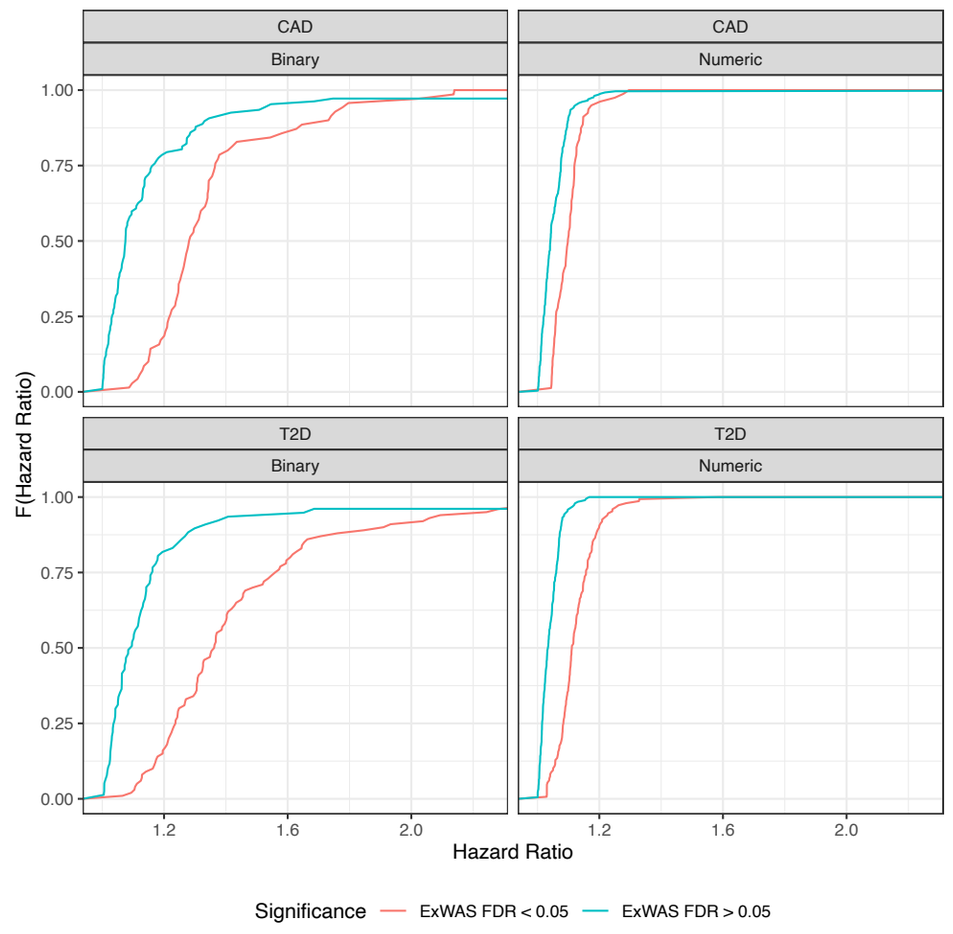


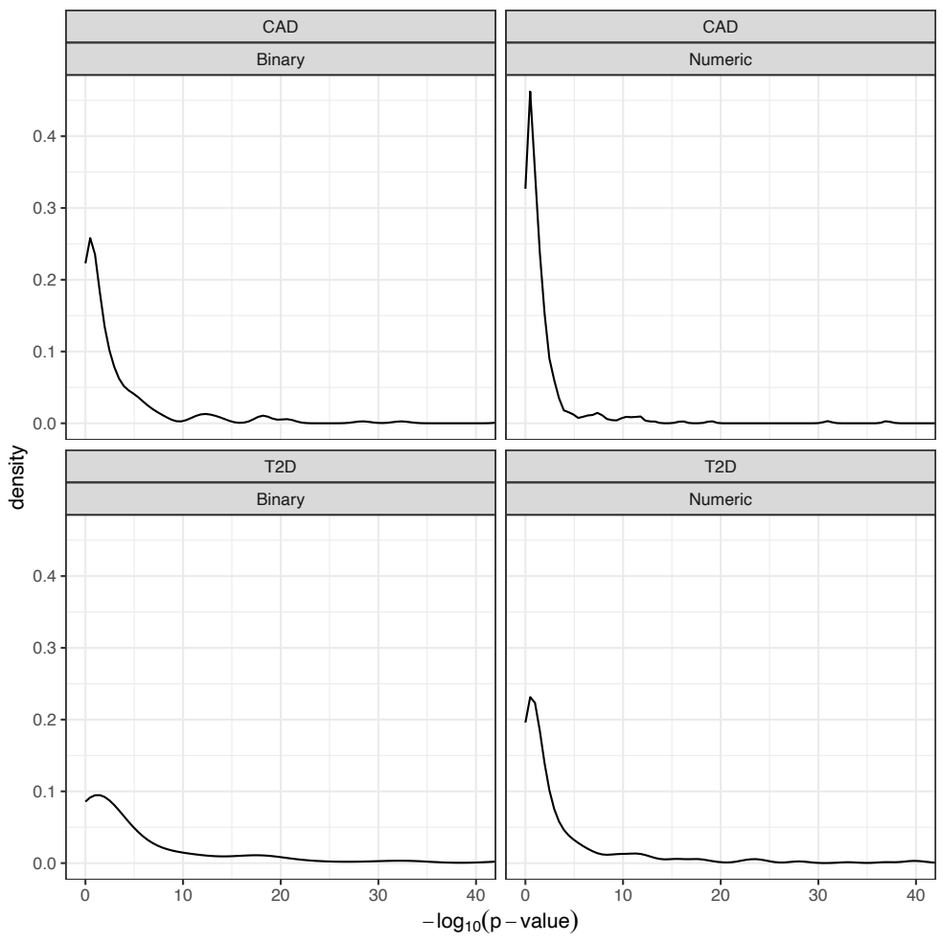

MR BASE

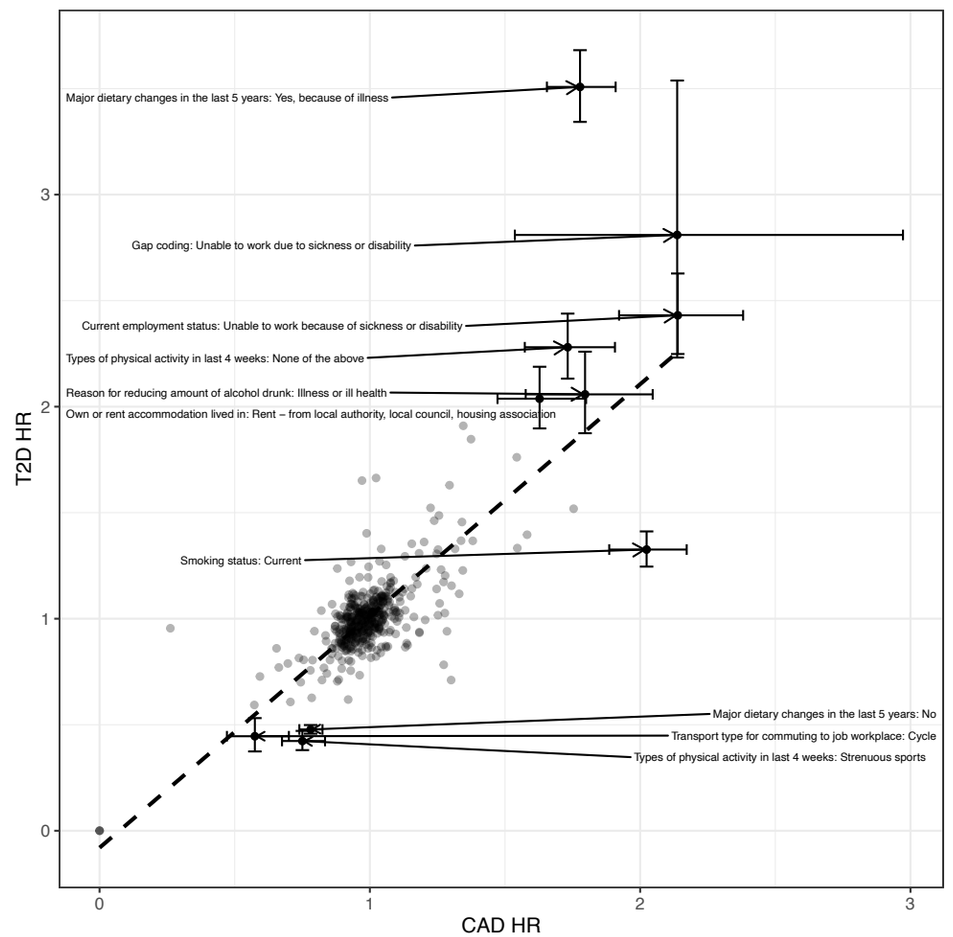
biobank uk

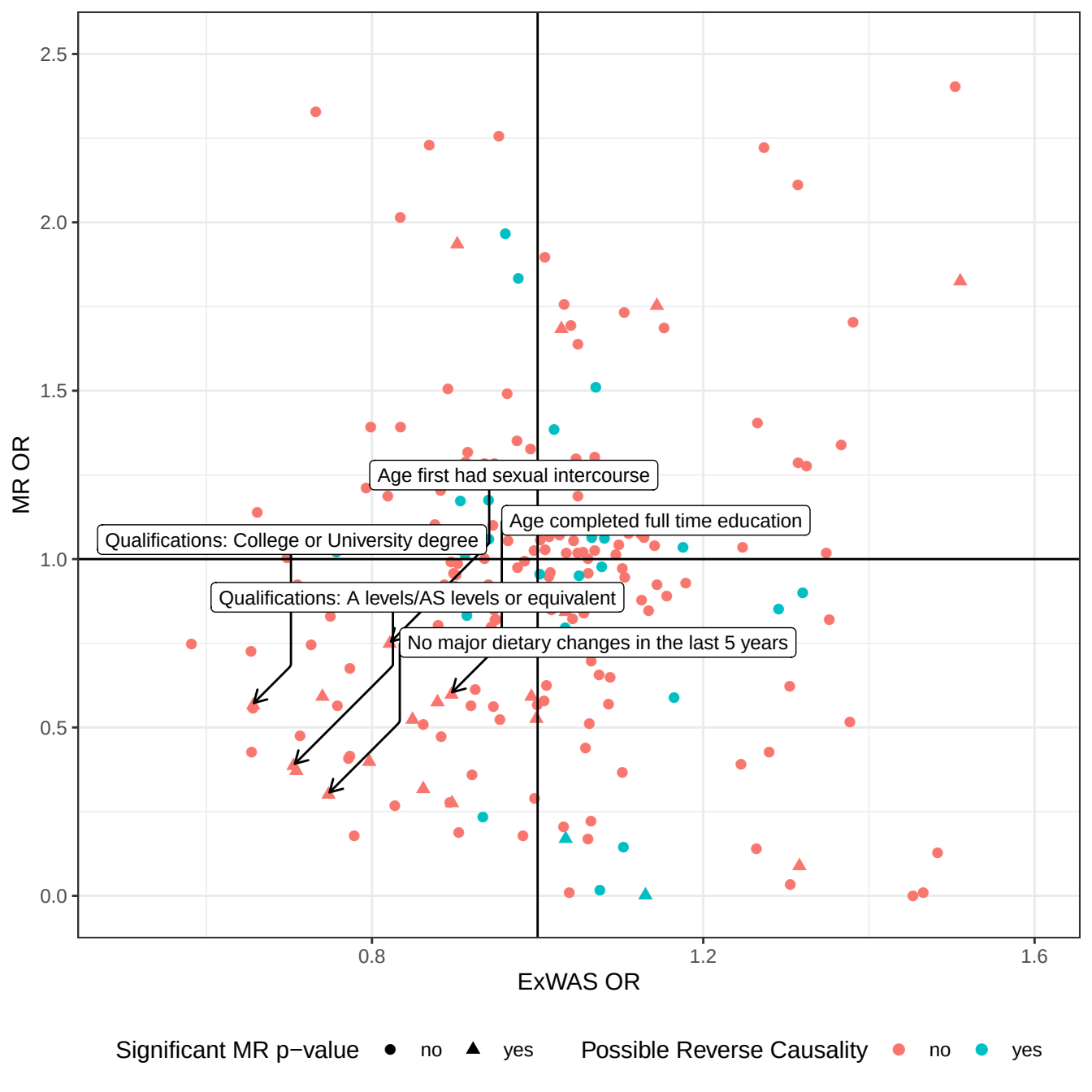
FINNGEN

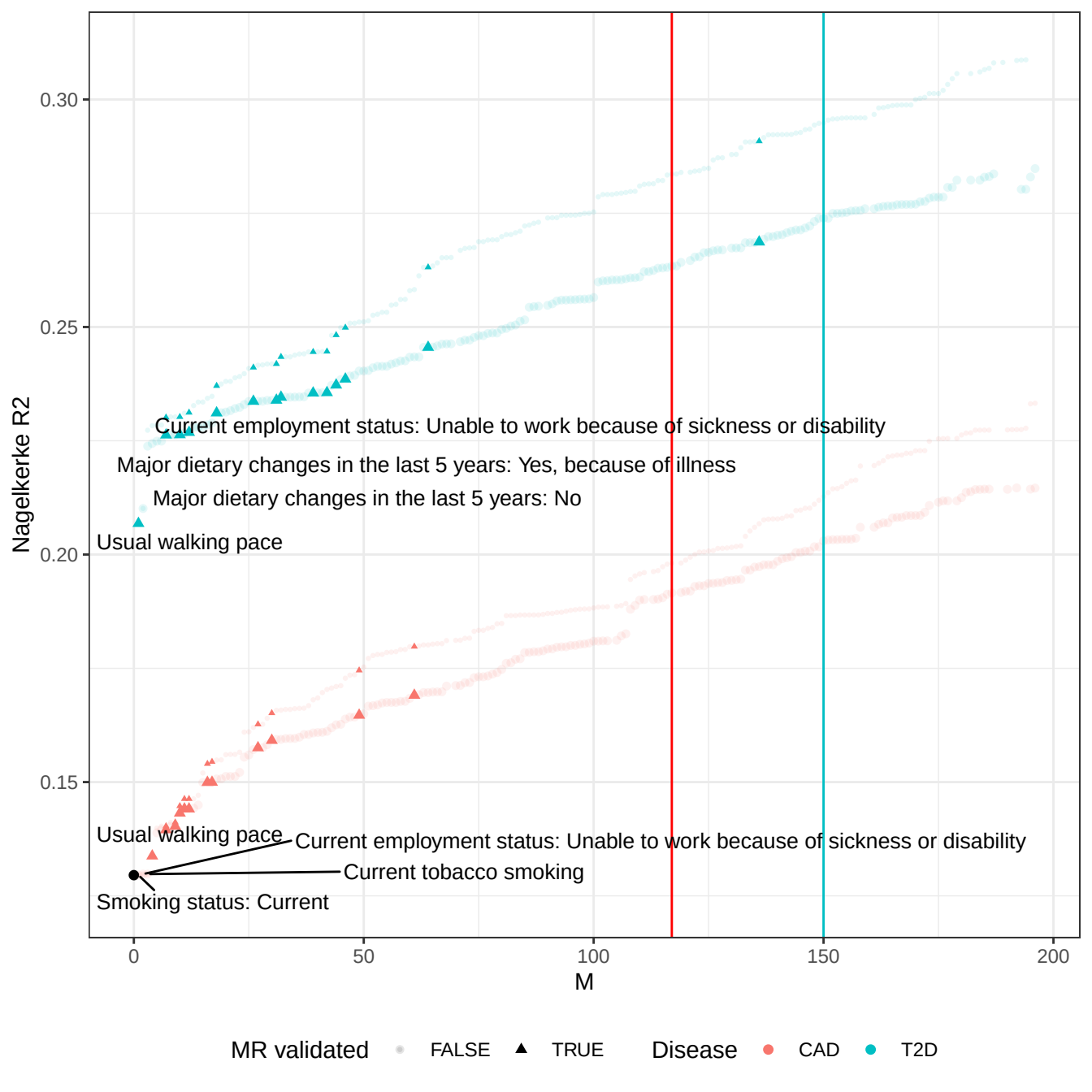

a
Current tobacco smoking
Usual walking pace
Age first had sexual intercourse
Smoking status: Never
Alcohol intake frequency
75
50
25
0
$-\log_{10}$(FDR)
0.50
0.75
1.00
1.25
1.50
Hazard Ratio
Significance
FDR < 0.05
FDR > 0.05


b
CAD
Binary
CAD
Numeric
T2D
Binary
T2D
Numeric
1.00
0.75
0.50
0.25
0.00
F(Hazard Ratio)
1.2
1.6
2.0
Hazard Ratio
Significance
ExWAS FDR < 0.05
ExWAS FDR > 0.05


c
CAD
Binary
CAD
Numeric
T2D
Binary
T2D
Numeric
0.4
0.3
0.2
0.1
0.0
density
0
10
20
30
40
$-\log_{10}$(p − value)


d
Major dietary changes in the last 5 years: Yes, because of illness
Gap coding: Unable to work due to sickness or disability
Current employment status: Unable to work because of sickness or disability
Types of physical activity in last 4 weeks: None of the above
Reason for reducing amount of alcohol drunk: Illness or ill health
Own or rent accommodation lived in: Rent – from local authority, local council, housing association
Smoking status: Current
Major dietary changes in the last 5 years: No
Transport type for commuting to job workplace: Cycle
Types of physical activity in last 4 weeks: Strenuous sports
3
2
1
0
T2D HR
0
1
2
3
CAD HR


Age first had sexual intercourse
Age completed full time education
Qualifications: College or University degree
Qualifications: A levels/AS levels or equivalent
No major dietary changes in the last 5 years
MR OR
ExWAS OR
2.5
2.0
1.5
1.0
0.5
0.0
0.8
1.2
1.6
Significant MR p−value
no
yes
Possible Reverse Causality
no
yes


0.30
0.25
0.20
0.15
Nagelkerke R2
Current employment status: Unable to work because of sickness or disability
Major dietary changes in the last 5 years: Yes, because of illness
Major dietary changes in the last 5 years: No
Usual walking pace
Usual walking pace
Current employment status: Unable to work because of sickness or disability
Current tobacco smoking
Smoking status: Current
0
50
100
150
200
M
MR validated
FALSE
TRUE
Disease
CAD
T2D