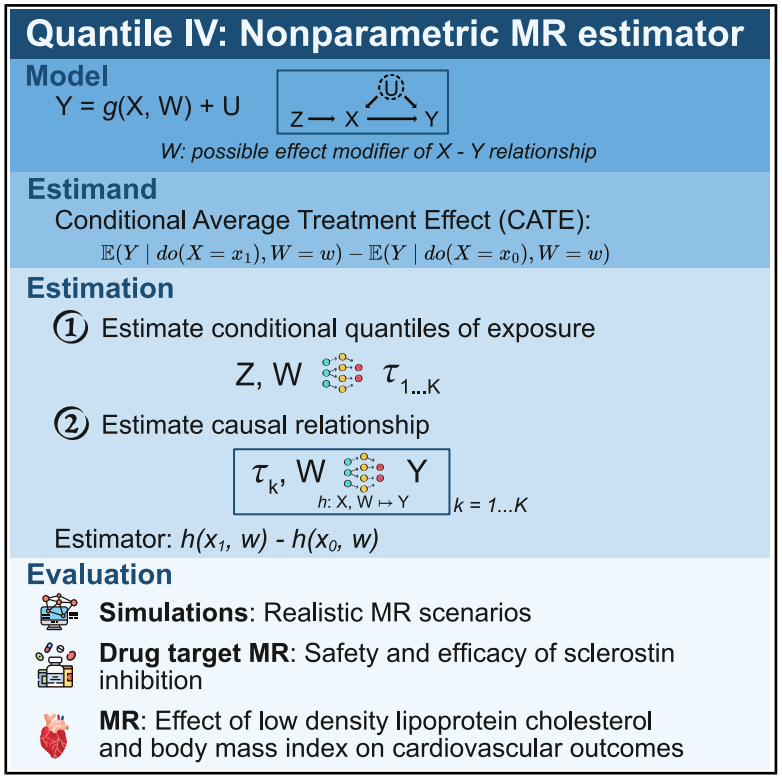
A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition

Graphical abstract



Authors

Marc-André Legault, Jason Hartford, Benoît J. Arsenault, Archer Y. Yang, Joelle Pineau

Correspondence

marc-andre.legault.1@umontreal.ca

Mendelian randomization (MR) enables the estimation of causal effects while controlling for unmeasured confounding factors, but MR estimators often rely on strong parametric assumptions. We propose an MR estimator named quantile instrumental variable (Quantile IV) that makes few statistical assumptions and performs well in simulations and real data studies.

Legault et al., 2025, The American Journal of Human Genetics *112*, 1–19
June 5, 2025
https://doi.org/10.1016/j.ajhg.2025.04.010

CellPress

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

ARTICLE

A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition

Marc-André Legault,[1,2,3,4,*] Jason Hartford,[5] Benoît J. Arsenault,[6,7] Archer Y. Yang,[2,8] and Joelle Pineau[1,2]

Summary

Mendelian randomization (MR) enables the estimation of causal effects while controlling for unmeasured confounding factors. However, traditional MR's reliance on strong parametric assumptions can introduce bias if these are violated. We describe a machine learning MR estimator named quantile instrumental variable (Quantile IV) that achieves a low estimation error in a wide range of plausible MR scenarios. Quantile IV is distinctive in its ability to estimate nonlinear and heterogeneous causal effects and offers a flexible approach for subgroup analysis. Applying quantile IV, we investigate the impact of circulating sclerostin levels on heel bone mineral density, osteoporosis, and cardiovascular outcomes. Employing various MR estimators and colocalization techniques, our analysis reveals that a genetically predicted reduction in sclerostin levels significantly increases heel bone mineral density and reduces the risk of osteoporosis while showing no discernible effect on ischemic cardiovascular diseases. As a second application, we estimated the effect of increases in low-density lipoprotein cholesterol and waist-to-hip ratio on ischemic cardiovascular diseases using this well-known association as a positive control analysis. Quantile IV contributes to the advancement of MR methodology, and the selected applications demonstrate the applicability of our estimator in various MR contexts.

Introduction

Instrumental variable (IV) estimation is a technique used to estimate the causal effect of an exposure on an outcome of interest from observational data. While IV estimation relies on the strong (and untestable) assumption of access to a valid IV—some variable that is assumed to only affect the outcome of interest via the exposure—it is a powerful technique because, unlike most other causal inference strategies, it allows estimation even in the presence of unobserved confounders of the exposure-outcome relationship. Because IV estimation relies on different assumptions than other study designs, it can be used when other designs are not applicable or susceptible to bias. Studies relying on IVs are increasingly being used to provide robust evidence when combined with designs that rely on different assumptions to triangulate a causal effect from multiple sources.[1]

Genetic variants can be used as IVs to infer causal effects in Mendelian randomization (MR) studies. MR leverages the fact that genetic variants are fixed throughout life and unaffected by environmental factors that may have confounding effects. The use of genetic variants as IVs has enabled the estimation of the causal effect of lipoprotein fractions[2,3] to predict the safety and efficacy of modulating drug targets[4,5] or estimate the causal effect of circulating proteins on diseases.[6,7] Despite these successes, MR studies may be biased either by failures of the untestable "exclusion restriction" assumption due to horizontal pleiotropy—when the genetic variant affects the outcome both via the exposure *and* some other pathway—or due to inappropriate assumptions on the functional form of the exposure-outcome relationship. Most recent work on MR has focused on the former, for example, by allowing a fraction of the IVs to be invalid (e.g., Verbanck et al.,[8] Mounier and Kutalik,[9] Burgess et al.,[10] and Bowden et al.[11]). However, few studies addressed the validity of the parametric assumptions made by MR estimators.

Most of the current MR estimators assume that both the genetic effect on the exposure and the causal effect of the exposure on the outcome are linear. Recent efforts have substantially relaxed these linearity assumptions by considering polynomial functional forms,[12] locally linear effects,[13,14] or semi-parametric models.[15] These innovations are important because they allow non-constant treatment effects to be estimated. They do not assume that the effect of a unit increase in the exposure is necessarily constant over the full range of the exposure, allowing for more complex but plausible dynamics such as threshold effects, diminishing returns, or exponential effects. Despite the progress made in the field of nonlinear MR, current models still rely on strong assumptions, including the assumption of a constant genetic effect among levels of exposure and covariates.[16–18]

[1]Department of Computer Science, McGill University, Montreal, QC, Canada; [2]Mila, Montreal, QC, Canada; [3]Faculté de pharmacie, Université de Montréal, Montreal, QC, Canada; [4]Centre de recherche Azrieli du CHU Sainte-Justine, Montreal, QC, Canada; [5]Valence Labs, Montreal, QC, Canada; [6]Centre de recherche de l'Institut universitaire de cardiologie et de pneumologie de Québec, Québec, QC, Canada; [7]Department of Medicine, Faculty of Medicine, Université Laval, Quebec, QC, Canada; [8]Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada
*Correspondence: marc-andre.legault.1@umontreal.ca
https://doi.org/10.1016/j.ajhg.2025.04.010.

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

Nonparametric IV estimators[19,20] do not require restricting the functional form relating the exposure and the outcome beyond an additive assumption on the confounding, and as a result, they allow for effect heterogeneity between the IV and exposure and between the exposure and outcome. Modeling effect heterogeneity allows for the estimation of causal effects for specified subgroups of individuals from a single model fit by conditioning the estimate on the levels of other covariates. Estimation of conditional treatment effects is a powerful tool to anticipate effect heterogeneity, for example, by investigating sex differences, age differences, or other clinically meaningful subgroup effects.

The development of nonparametric IV estimators has evolved independently of MR in the econometrics, statistics, and machine learning literature, and there have been limited efforts to bridge these worlds. Here, we harness recent developments in machine learning and nonparametric IV estimation to propose an estimator named Quantile IV that performs well in the context of MR.[19,21] Quantile IV, drawing from Hartford et al.'s DeepIV model,[21] incorporates a crucial simplification for enhanced performance in MR contexts without any added statistical assumption. Specifically, DeepIV is a two-stage procedure that first models the conditional distribution of the exposure given the IVs and covariates. This involves fitting a probabilistic model, often parametrized by a neural network, and the random sampling of values during training. In Quantile IV, we replace this step with neural network quantile regression and use simple averaging of the predicted conditional quantiles, avoiding the computational cost of sampling and the challenging task of explicitly modeling the probability density. In a realistic MR simulation setup, we show that our method outperforms the DeepIV estimator and other instantiations of this approach.[15] Using simulated data, we show that our estimator achieves low error in all of the considered MR scenarios, and we quantify the coverage, type 1 error rate, and type 2 error rate of confidence intervals (CIs) obtained by bootstrap aggregation (bagging). To evaluate our MR estimator in a real-world scenario, we first evaluated the causal effect of a decrease in circulating sclerostin on bone and cardiovascular diseases using a two-sample approach within the UK Biobank. Sclerostin is the drug target of romosozumab, an anti-sclerostin monoclonal antibody used to prevent fractures in individuals with osteoporosis.[22,23] Our investigation of possible cardiovascular adverse events aims to clarify safety concerns related to sclerostin inhibition stemming from the observation of a higher number of adjudicated serious cardiovascular events in the treatment arm of clinical trials of romosozumab.[22–25] To demonstrate the applicability of our method beyond the drug target MR context, we also estimated the effect of low-density lipoprotein cholesterol (LDL-c) and the waist-to-hip ratio (WHR) on the considered cardiovascular outcomes. Because the WHR and LDL-c are well-established causal cardiovascular disease risk factors,[26–28] this analysis serves as a positive control, and Quantile IV was able to detect the accrued risk caused by increases in the WHR and LDL-c. To demonstrate the advantage of our method, we investigated whether MR causal effects were predicted to vary with respect to sex or statin use without pre-specifying interactions. Our results suggest that, for the same genetically predicted increase in LDL-c levels, the increase in cardiovascular outcome risk is smaller in statin users vs. non-users. We attribute this effect to statins partially offsetting the atherogenic effect of genetically increased LDL-c.

Methods

Study population

The UK Biobank is a densely phenotyped population cohort of 500,000 participants that have been genotyped and imputed.[29] At the recruitment visit, UK Biobank participants undergo a thorough assessment with health questionnaires (touchscreen and verbal interview), blood and urine biomarker panels, and physical measurements, including ultrasound bone densitometry. Linkage to national health system hospitalization and death records further enables the algorithmic definition of many diseases, including acute cardiovascular events (supplemental methods; Table S1). In the current study, we used the linked medical records to define myocardial infarction (MI), acute coronary artery disease (CAD), ischemic stroke, and percutaneous coronary intervention/coronary artery bypass graft (PCI/CABG), which are surgical revascularization procedures. A subset of 46,673 randomly selected participants enrolled in the UK Biobank Pharma Proteomics Project have high-throughput proteomics data measuring around 3,000 circulating proteins using the Olink platform. Inclusion criteria and quality control are described in the supplemental methods. The UK Biobank participants provided informed consent, and the data used in this study were accessed under the UK Biobank application #20168. Research tissue bank (RTB) approval for the UK Biobank was provided by the North West Multi-centre Research Ethics Committee and the current study operates under this RTB approval.

Genetic association analyses

To identify genetic variants associated with circulating sclerostin levels (protein quantitative trait loci [pQTLs]), we conducted a genetic association analysis of 1,449 common (minor-allele frequency [MAF] $\geq$ 1%) bi-allelic genetic variants in the *SOST* (MIM: 605740) gene region in 42,830 UK Biobank participants with available sclerostin measurements. The circulating sclerostin measurements are taken from high-throughput proteomics measurements of circulating proteins (supplemental methods).[6] We defined the *SOST* locus using the gene boundaries and including 400 kb padding upstream and 200 kb padding downstream. The final coordinates of the locus on the GRCh37 reference build are chr17:41631099–42236156. We used Plink v.2.00a2LM AVX2 Intel (October 25, 2019) using the generalized linear model (–glm) option, implementing linear and logistic regression, for association testing. The association statistics (i.e., estimated coefficients and standard errors) were subsequently used for MR estimation using parametric models. We used the same procedure to estimate the effect of genetic variants at the *SOST* locus on the outcomes considered in the MR study. We used linear regression for heel bone mineral density (BMD) and logistic regression for osteoporosis, PCI/CABG, MI, acute CAD, and ischemic stroke. All the genetic association models were

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

adjusted for age at baseline, sex, and the first 5 principal components to adjust for residual population structure.

Summary statistics for two-sample MR

Two-sample MR relies on estimates of the effect of the genetic IVs on the exposure and the outcome from independent individuals, allowing for the use of published summary statistics for MR. This approach is often used to maximize statistical power and to mitigate the risk of false positives due to weak instrument bias.[30] We used data from the CARDIoGRAMplusC4D consortium when considering MI or CAD as the outcome in our two-sample MR analyses. We used summary statistics from a genome-wide association study (GWAS) meta-analysis considering genetic variants imputed using 1000 Genomes Project data and including 60,801 cases of CAD.[31] We also used data from the updated meta-analysis from CARDIoGRAMplusC4D that includes 10,801 additional CAD cases from the UK Biobank.[32]

Our sclerostin MR study uses circulating sclerostin levels measured in the UK Biobank as the exposure. We investigated whether the pQTLs were also *SOST* expression QTLs (eQTLs) in aorta and tibial artery tissues in GTEx v.8 using colocalization.[33] We selected these tissues because they are the ones with the most *SOST* expression, and they are plausible candidates to explain the cardiovascular impact of sclerostin. We note that bone tissue is not included in GTEx, hampering our ability to identify bone eQTLs of sclerostin.

For the MR study of the WHR, we used summary association statistics from the Genetics Investigation of Anthropometric Traits (GIANT) consortium.[34] The summary statistics are based on a GWAS meta-analysis of BMI-adjusted WHRs in 694,649 individuals of European ancestry, including 484,563 participants from the UK Biobank. For the MR study of LDL-c, we used summary association statistics from the Global Lipids Genetics Consortium (GLGC), including up to 842,660 individuals of European ancestry and were not participants in the UK Biobank.[35]

Fine-mapping and colocalization analyses

Fine-mapping is used to infer credible sets of genetic variants that, under the model assumptions, will include the true causal variant at a specified probability level. We used the "sum of single effects" (SuSiE) statistical model, which considers the sum of regression models with a single non-zero effect for fine-mapping. This approach can accommodate multiple causal variants within a region and allows the computation of variant posterior inclusion probabilities (PIPs) for every variant.[36] We used the implementation from the "susieR" R package (v.0.12.35).

Colocalization analysis compares the estimated association between genetic variants and a pair of traits to infer the presence or absence of shared causal variants. The "coloc" model estimates the posterior probability (PP) of five mutually exclusive hypotheses. H_0 is the hypothesis that there are no causal variants, H_1 that there is only a causal variant for trait 1, H_2 that there is only a causal variant for trait 2, H_3 that there are causal variants for both traits and that they are distinct, and H_4 that there is a shared causal variant for both traits. The original publication only accounted for a single causal genetic variant per association signal,[37] but this assumption was subsequently relaxed.[38] To account for multiple causal variants, fine-mapping is first used to derive credible sets, and pairwise colocalization between credible sets for the two traits is tested. We use this approach in our study in instances where we were able to infer credible sets with coverage ≥85%. When unable to infer credible sets, we assumed a maximum of a single causal variant per trait. We used the coloc R package (v.5.2.2) to conduct all of the colocalization analyses ("coloc.abf" and "coloc.susie" functions) with the default prior values and linkage disequilibrium (LD) matrices computed in the subset of UK Biobank participants that passed our genetic quality control. We followed the recommendations from the authors of the susieR package and verified that the λ statistic had low values and that the kriging plot did not have outlier variants to ensure adequate matching between the LD matrix and summary statistics and to detect allele flips.

Causal assumptions

IV estimation relies on three main assumptions (Figure S1).[39] We denote the exposure of interest as X, the outcome as Y, the observed covariables as W, and the IVs as Z. The first assumption (IV1), relevance, states that the IV is not independent of the exposure ($Z \not\perp\!\!\!\perp X|W$). The second assumption (IV2) assumes the unconfoundedness of the IV, meaning that the IV is independent of unobservable confounders of the exposure-outcome relationship ($Z \perp\!\!\!\perp U|W$). The third IV assumption is the exclusion restriction (IV3), also commonly known as the "no horizontal pleiotropy" assumption in the MR literature. This assumption requires that the effect the IV exerts on the outcome is exclusively through the modulation of the exposure ($Z \perp\!\!\!\perp Y|U, X, W$). Common violations of this assumption include direct pleiotropic effects of the IV on the outcome and, more perniciously, effects due to LD, with a variant influencing the outcome independently of the exposure. Many attempts to relax these assumptions rely on the identification of a subset of IVs with homogeneous effects, the inference of statistical structure in the causal effects, or the estimation of the mode of the causal effect (e.g., Verbanck et al.,[8] Mounier and Kutalik,[11] and Qi and Chatterjee[40]). When appropriate, we used estimators that relax the IV3 assumption in different ways (detailed in supplemental methods). However, these approaches are not suitable for use in *cis*-MR because a single set of correlated candidate IVs is used, hampering the estimation of modes, the inference of latent statistical structure, or the detection of outliers.

In our MR study investigating the effect of circulating sclerostin levels on bone and cardiovascular health, we were able to confirm the relevance assumption by observing a strong association between our IVs, rs6416905 and rs66838809, and circulating sclerostin levels ($p = 1.49 \times 10^{-18}$ and $p = 1.82 \times 10^{-16}$, respectively). The F statistic for these two IVs was 60. The most plausible violation of the unconfoundedness assumption is population structure, and we mitigated this risk by using a genetically homogeneous subset of participants within the UK Biobank and further adjusting all the MR estimates for the first 5 genetic principal components. Finally, since we only considered genetic variants associated with circulating sclerostin levels at the *SOST* locus, it is plausible that the observed effects are due to the modulation of sclerostin levels and not via other pathways. The risk of violations of the exclusion restriction assumption mostly arises from bias due to LD with other variants that may influence the considered outcomes. We use colocalization as an analytical approach to confirm that the causal variants underpinning genetic associations with the exposure and outcome are shared. We also conducted sensitivity analyses adjusting for genetic variants that are likely to have direct effects on the outcome when there is evidence for direct effects.

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

MR analyses

cis-*MR of sclerostin*

MR studies of exposures corresponding to specific proteins or genes are termed *cis*-MR.[41] This approach is useful for drug target validation or to investigate molecular traits instrumented by regulatory or functional variants. We used this approach to estimate the causal effect of circulating sclerostin levels on heel BMD, osteoporosis, MI, acute CAD, PCI/CABG, and ischemic stroke using genetic variants near the gene encoding sclerostin (*SOST*). We used the inverse variance weighted (IVW) and PC-GMM estimators for these *cis*-MR analyses. The IVW is a weighted average of the ratio estimates of the IVs where the weights are proportional to the precision of the ratio estimates.[42] The PC-GMM estimator was designed for the *cis*-MR setting and uses a principal-component analysis of a weighted LD matrix accounting for instrument strength and the precision of the effect estimates on the outcome.[43] The principal components from this decomposition are then used as IVs using a generalized method of moments for estimation.[44] We used the robust standard errors accounting for overdispersion in the current study.

We used a two-sample design within the UK Biobank (split sample). A subset of 42,830 participants with available proteomic measurements was used to estimate genetic effects on sclerostin levels, and the non-overlapping subset of 370,218 participants was used to estimate the genetic effects on our outcomes of interest. Descriptive statistics for both datasets are shown in Table S2.

MR of WHR and LDL-c

We first used a two-sample MR approach to estimate the causal effect of the WHR and LDL-c on cardiovascular outcomes. We selected variants associated with LDL-c levels at the genome-wide significance threshold ($p \leq 5 \times 10^{-8}$) in the GLGC summary statistics data. We clumped variants with an $r^2 \geq 0.01$ together and only considered variants with a MAF above 5% in the 1000 Genomes European individuals reference panel. The variant selection was done using *grstools* (v.0.4.0, https://github.com/legaultmarc/grstools). The same strategy was used to select genetic instruments for the WHR using summary statistics from the GIANT consortium. This selection strategy identified 539 IVs for LDL-c levels and 547 IVs for WHR levels (Tables S3 and S4).

We considered MR estimators with different statistical properties and that make different assumptions on the validity of the IVs. Namely, we used the IVW, weighted median, and MR-Egger estimators (described in the supplemental methods).[11,45] In addition, we used the contamination mixture method and the MR LASSO method, which provide variant-level statistical evidence of exclusion restriction violations.[46,47]

Quantile IV is not a robust method: it is biased if invalid IVs are included. To account for this limitation, we used the subset of variants predicted to be valid IVs by the contamination mixture and MR LASSO methods (across all of the considered outcomes). This IV selection procedure assumes that the largest set of variants that estimate a homogeneous causal effect is the set of valid IVs. These sets contained 311 variants for LDL-c and 298 variants for the WHR (Tables S3 and S4). Quantile IV requires individual-level data, and we used a one-sample MR approach in the UK Biobank (Table S5). In the one-sample MR context, weak instrument bias may bias the estimate toward the observational association.[30] However, we expect this bias to be small in our analyses because of the strength of our instruments (F statistics of 13,310 for the LDL-c instrument and 3,509 for the WHR instrument). Furthermore, the variants were selected based on external summary statistics, alleviating bias that could be due to Winner's curse. We used genetic scores (i.e., polygenic risk scores) as IVs following findings from our simulation study where we observed that Quantile IV does not perform optimally when hundreds of IVs are used. The externally estimated effects on LDL-c or the WHR served as weights in the scores. As a second model allowing for nonlinear MR estimates, we used the doubly ranked stratification method (R package DRMR v.0.1.0, https://github.com/HDTian/DRMR).[14] We used the implementation from the R "MendelianRandomization" package (v.0.9.0) for parametric MR.

Nonparametric MR estimation

We first consider the model introduced by Hartford et al. when describing the DeepIV estimator[21]:

$$Y = g(X, W) + U. \qquad \text{(Equation 1)}$$

In our context, Y represents the outcome, X denotes the exposure, and W encompasses observed confounders. The latent variables U account for the unobservable factors that may affect Y, X, and W. It enters the model additively. Here, $g(\cdot)$ is some unknown and potentially nonlinear function of both X and W. We further introduce an IV (Z) that satisfies the IV assumptions (IV1–3, Figure S1).

The goal is to estimate the conditional average treatment effect (CATE) $\eta(w, x_0, x_1)$, defined as

$$\eta(w, x_0, x_1) = \mathbb{E}(Y|do(X = x_1), W = w) - \mathbb{E}(Y|do(X = x_0), W = w).$$

Under the *do* operator and the assumption of model 1 (Equation 1), we have

$$\begin{aligned}\mathbb{E}(Y|do(X = x), W = w) &= \mathbb{E}(g(X, W)|do(X = x), W) \\ &\quad + \mathbb{E}(U|do(X = x), W) \\ &= g(X, W) + \mathbb{E}(U|W).\end{aligned} \qquad \text{(Equation 2)}$$

We can estimate the CATE by estimating the $h(\cdot)$ function, defined as

$$h(X, W) := g(X, W) + \mathbb{E}(U|W),$$

since the conditional expectation of the confounder given the covariates will not influence the estimation of contrasts such as the CATE: $\eta(w, x_0, x_1) = g(x_1, w) - g(x_0, w) = h(x_1, w) - h(x_0, w)$. Note that if we additionally assume that $\mathbb{E}(U|W) = 0$, then $h(X, W)$ directly characterizes the conditional effect of an intervention of X on Y. We emphasize that if the (conditional) causal relationship between the exposure and outcome is nonlinear, then the CATE will vary with respect to the choice of x_0 and x_1. These reference points can be selected based on the research question at hand, for example, with respect to the effect of existing interventions influencing the exposure. Here, for the MR of sclerostin levels, we standardized the exposure to have a mean of 0 and a variance of 1 and reported effects for a 1 or 2 SD reduction about the mean, effectively setting $x_0 = 0$ and $x_1 = -1$ or -2. We decided to use SDs because the Olink proteomic measurements do not have interpretable units. For the LDL-c and WHR MR, we set x_0 to the mean value from our sample and use x_1 to represent fixed increases about the mean. For LDL-c, we report effects for a 1 mmol/L increase and, for the WHR, a 0.1 increase in the ratio. In graphical representations, we report ATEs (average treatment effects) by varying x_1 while fixing x_0 to the reference level.

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

To estimate $h(\cdot)$, DeepIV uses the following result:

$$\begin{aligned}\mathbb{E}(Y|W,Z) &= \mathbb{E}(g(X,W)+U|W,Z)\\ &= \mathbb{E}(g(X,W)|W,Z)+\mathbb{E}(U|W)\\ &= \int (g(X,W)+\mathbb{E}(U|W))\, dF(X|W,Z)\\ &= \int h(X,W)dF(X|W,Z).\end{aligned}$$

Based on this result, Hartford et al.[21] suggest estimating $h(\cdot)$ by solving the following optimization problem:

$$\arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \int \widehat{h}(x,w_i)d\widehat{F}(x|w_i,z_i)\right)^2, \qquad \text{(Equation 3)}$$

which is found using a two-stage procedure. The first stage uses a treatment network to estimate the conditional cumulative distribution function $\widehat{F}(x|w,z)$ using any statistical or machine learning model, such as the mixture density network, which was initially suggested. The second stage then samples conditional exposure values from the first-stage network and relates these samples to the observed outcome to estimate $h(x,w)$.

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \int \widehat{h}(x,w_i)d\widehat{F}(x|w_i,z_i)\right)^2 \approx \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{M}\sum_{j=1}^{m}\widehat{h}(x_j,w_i)\right)^2,$$

where x_j is the samples from the estimated cumulative density function $x_j \sim \widehat{F}(x|w_i, z_i)$. This step is akin to any supervised learning task and can be done using a feedforward neural network. We note that this procedure is analogous to the two-stage least squares (2SLS) procedure that is a conventional estimator in IV analysis.

This procedure relaxes the parametric assumptions of conventional MR methods and only assumes that the confounder enters additively in the model. However, the sampling step needed to train the second stage introduces additional stochasticity in the training process and may limit performance in MR.[15]

Quantile IV algorithm

The current method, Quantile IV, proposes replacing the density estimation in the first stage of the DeepIV procedure with a quantile regression, which eliminates the need for sampling and simplifies the optimization. Quantile IV is a specific instantiation of DeepIV and an equally valid estimator while performing advantageously in realistic settings.

We now describe the estimation strategy for Quantile IV. In the first stage, we estimate K evenly spaced conditional quantiles of the exposure given the instruments using a neural network trained using the quantile loss. Specifically, we wish to estimate a fixed number of evenly spaced conditional τ-th quantiles of X given W and Z with $0<\tau<1$.

$$q_\tau(W,Z) = \inf\{x : F_{X|W,Z}(x) \geq \tau\},$$

where $F_{X|W,Z}(x) = P(X \leq x|W,Z)$ is the conditional distribution function of X. We know that the expectation of any function $f(X)$ of some random variable X can be related to the quantile using the following relation: $\mathbb{E}(f(X)) = \int_0^1 f(q_\tau)d\tau$, where q_τ represents the τ-th quantile of X. In our context, this provides a method to approximate the integral $\int h(X,W)dF(X|W,Z)$ using K conditional quantiles $q_{\tau_1}(W, Z)$, …, $q_{\tau_K}(W, Z)$. We partitioned (0,1) using $K+1$ evenly spaced points $0 = a_0 < a_1 < \cdots < a_K = 1$, where $a_k = k/K$, and chose quantiles $\tau_k = \frac{a_{k-1}+a_k}{2}$ for $k = 1, \ldots, K$. We have

$$\begin{aligned}\int h(X,W)dF(X|W,Z) &= \int_0^1 h\big(q_\tau(W,Z)\big)d\tau\\ &\approx \sum_{k=1}^{K} h\Big(q_{\tau_k}(W,Z)\Big)(a_k - a_{k-1})\\ &= \frac{1}{K}\sum_{k=1}^{K} h\Big(q_{\tau_k}(W,Z)\Big).\end{aligned}$$

To learn quantiles $q_{\tau_k}(w,z)$ for $k = 1, \ldots, K$, we use a neural network parametrized by a set of weights and biases denoted as ϕ and with a K-dimensional output layer $\mathbf{f}(w, z\,;\phi) : W \times Z \to X^K$. Let $\mathbf{f} = (\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(K)})$, where $\mathbf{f}^{(k)}$ is the k^{th} element of $\mathbf{f}$. The quantile loss estimates conditional quantiles[48] and can be expressed as

$$\mathcal{L}(w,z,x,\tau,\phi) : = \sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}\Big(x_i - \mathbf{f}^{(k)}(w_i,z_i\,;\phi)\Big), \qquad \text{(Equation 4)}$$

where $\rho_\tau(u) = (\tau - \mathbb{I}[u \leq 0])u$. Hence, K conditional quantiles $(q_{\tau_1}(w,z), \ldots, q_{\tau_K}(w,z))$ can be simultaneously estimated by $\mathbf{f}(w, z\,;\phi)$. In the first stage of Quantile IV, we train a neural network to minimize this quantile loss by solving

$$\widehat{\phi} : = \arg\min_{\phi}\mathcal{L}(w,z,x,\tau,\phi). \qquad \text{(Equation 5)}$$

The key insight of the method is that these K quantiles divide the conditional distribution of the exposure into equally probable regions and allow us to replace the sampling step in DeepIV using a simple average over these conditional quantiles as the input to the second-stage regression. Under this first-stage model and using a second neural network $h : X \times W \to Y$ to estimate the IV regression function, Equation 3 becomes

$$\widehat{h}(x,w\,;\theta) : = \arg\min_{h(x,w\,;\theta)}\sum_{i=1}^{n}\left(y_i - \frac{1}{K}\sum_{k=1}^{K}h\Big(\mathbf{f}^{(k)}(w_i,z_i\,;\widehat{\phi}),w_i;\theta\Big)\right)^2. \qquad \text{(Equation 6)}$$

The optimization of the weights and biases (denoted by θ) of this neural network can be achieved using conventional gradient-based optimizers (e.g., Adam[49]).

Neural network quantile regression

In our formulation of Quantile IV, we use a standard feedforward neural network trained with the quantile loss to estimate the conditional quantiles of the exposure given the IVs. This approach may pose a problem because it does not strictly constrain the learned conditional quantiles to be monotonically increasing, even though the quantile loss encourages the estimation of quantiles, satisfying this property. For example, there is no constraint in the model, forcing the 90% quantile to be smaller than the 95% quantile, which is required by the definition of quantiles. This problem is called the "quantile crossing" problem, and Moon et al. have proposed a neural network model and training procedure called noncrossing multiple quantile regression with neural networks (NMQN) to address it.[50] We implemented NMQN and tested the impact of its use on Quantile IV estimates. We observed that the estimated conditional quantiles are slightly more spread out when comparing NMQN to the naive implementation, but there was an increase in the average mean-squared error in the preliminary analysis of the simulation study, prompting us to rely on the naive implementation. The option to use NMQN remains available in the implementation.

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

Table 1. Values taken by the different parameters in the simulation study

Simulation parameter	Simulated values
Structural equation	$0.4x$, $0.2(x-1)^2$,[a] $\max(x-2,0)$
Sample size	10,000, 50,000,[a] 100,000
Proportion of the variance in the exposure explained by the IV	0.05, 0.1,[a] 0.5
Error correlation (confounding strength)	−0.6, 0.3,[a] 0.6
Number of independent instrumental variables	2, 10,[a] 100

We repeated every simulation 200 times.
[a]The reference values for the parameters. The reference value represents the fixed value of a simulation parameter used when other simulation parameters are varied.

Quantile IV implementation and estimation

We implemented the Quantile IV estimator using the Python programming language and the PyTorch framework (v.1.13.0, https://pytorch.org/). Our implementation is publicly available online as part of our *ml-mr* Python package (https://github.com/legaultmarc/ml-mr). The Quantile IV estimator is effectively composed of two neural networks, and our implementation allows setting the corresponding hyperparameters, including the number of layers, number of hidden units, and learning rate. To avoid overfitting, we use a sample splitting strategy, which randomly selects 20% (by default) of the samples to be used as a validation dataset. Training is based strictly on the training dataset, but the validation dataset is used to stop training when there are no further improvements on the validation loss, a strategy known as early stopping. Models can be trained using either the CPU or specialized hardware such as graphical processing units (GPUs). In practice, because the neural networks used in Quantile IV are relatively small, we have not observed significant performance improvements when training on GPUs. Reasonable default values (Table S6) for all of the hyperparameters were selected based on our experimentation on both real and simulated data during model development. We used these default values from our software implementation unless otherwise specified.

Quantile IV is a computationally intensive algorithm, as it requires the training of two neural networks. It is, however, possible to fit Quantile IV with over 300,000 individuals in under 1 h with one CPU and under 1 GB of RAM. We show the distribution of runtimes for a hyperparameter sweep and the relationship between the outcome learning rate and runtime in Figure S2. To alleviate this computational complexity, other flexible regression models (e.g., random forests) could be used to implement the Quantile IV estimator. The impact of alternative formulations on the computational burden and performance of the estimator could be considered in future work.

MR simulation study

We conducted a simulation study to assess the performance of nonparametric IV estimators in the context of MR. To achieve this, we selected simulation parameters covering a range of plausible MR settings. We vary the functional form of the causal effect between the exposure and outcome using linear, J-shaped (quadratic) and threshold relationships. These forms of nonlinearity are the most relevant to medical applications and are commonly studied in epidemiology. We use the specific parametrization from previous simulations of nonlinear MR as summarized in Table 1.[13] We vary the sample size between 10,000 and 100,000 individuals, corresponding to plausible modern sample sizes for small and large genetic studies. We use the proportion of the variance in the exposure explained by the IVs (h_x^2) to vary the strength of the genetic IVs. We simulate traits with $h_x^2 = 0.05$ corresponding to traits with a small fraction of the variance explained by genetic factors, traits with moderate variance explained ($h_x^2 = 0.1$), and traits with high variance explained ($h_x^2 = 0.5$). For reference and context, using the phenome-wide heritability browser based on the UK Biobank data (https://nealelab.github.io/UKBB_ldsc/h2_browser.html), we observed that the BMI had an estimated $h^2 = 0.25$, systolic blood pressure had $h^2 = 0.15$, heel BMD (right) had $h^2 = 0.33$, and forced expiratory volume in 1 s had $h^2 = 0.43$. We emphasize that in our study, we only consider the fraction of the variance explained by the IV, which is lower than the genome-wide SNP heritability.

To simulate genetic variants to be used as IVs, we follow the procedure described in Sulc et al.[12] Briefly, we sample allele frequencies $p_i \sim \text{beta}(1,3)$ and draw the number of alternative alleles following a binomial distribution with two draws. We then standardized the genotypes and assigned the effects using the baseline LDAK heritability model as $\beta_i \sim \mathcal{N}(0, p_i(1-p_i)^{-0.25})$ and rescaling to reach the desired heritability. The exposure and outcomes are then simulated as

$$X = \sum_{i=1}^{k} \beta_i G_i + \varepsilon_x \text{ and} \qquad \text{(Equation 7)}$$

$$Y = f(X) + \varepsilon_y, \qquad \text{(Equation 8)}$$

where G_i is the simulated standardized genotypes, $f(X)$ is the simulated structural relationship between the exposure and the outcome, and with

$$\begin{bmatrix} \varepsilon_x \\ \varepsilon_y \end{bmatrix} \sim \mathcal{N}_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right). \qquad \text{(Equation 9)}$$

The formulation for the errors implicitly models the effect of unmeasured confounding and follows the simulation model from He et al.[15] The advantage of this formulation is that it provides a convenient way of varying the strength of the latent confounding variable using a single simulation parameter (ρ). When varying the simulation parameters, we hold the others fixed at a reference value indicated in Table 1 (values marked with a footnote).

Estimating treatment effects and their CIs

The Quantile IV algorithm estimates the IV regression function $\hat{h}(x, w; \theta)$. From this fitted model, an estimator for the CATE is given by

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

$$\mathrm{CATE}(x_0, x_1, w) = \widehat{h}(x_1, w) - \widehat{h}(x_0, w),$$

where we dropped parameters for notational convenience. Similarly, an estimate of the ATE is obtained by averaging the CATE over the empirical data distribution.

$$\mathrm{ATE}(x_0, x_1) \approx \frac{1}{n} \sum_{i=1}^{n} (\widehat{h}(x_1, w_i) - \widehat{h}(x_0, w_i))$$

To obtain CIs, we rely on bootstrapping.[51] We resample the dataset with replacement and refit Quantile IV for every one of the B bootstrap resamples. This yields a bag of IV regression functions $\{\widehat{h}(x, w, \theta_b)\}_{b=1}^{B}$. Bootstrapping in this way first allows us to ensemble the predictions to obtain a bagging estimator for the CATE (and consequentially the ATE):

$$\mathrm{CATE}_{\mathrm{bs}} = \frac{1}{B} \sum_{b=1}^{B} (\widehat{h}(x_1, w; \theta_b) - \widehat{h}(x_0, w; \theta_b)).$$

A CI at the $1 - \alpha$ coverage level is obtained by selecting the $\alpha/2$ and $1 - \alpha/2$ quantiles of the bootstrap estimates of the CATE. To test the hypothesis that the CATE is null, we derive p values from these CIs by assuming asymptotic normality to avoid the high computational cost of computing bootstrap p values. Interaction p values are computed using one-way ANOVA of the bootstrap estimates of the CATEs at varying levels of the conditioning variable.

Adjustment for selection bias

Non-random sampling of individuals can induce bias in MR studies.[52,53] For example, if the exposure or outcome considered in the MR study influences the probability of being included in the analysis, then MR estimates may be biased. In some instances, it is possible to account for the sampling mechanism by comparing the observed distribution of variables in a sample to an unselected population (i.e., census data). A machine learning model trained to predict study inclusion in the UK Biobank from variables found in both UK Biobank and census data has been developed.[54] This model relies on variables related to health, lifestyle, education, and demographics to derive sampling probabilities, which can be used within an inverse probability weighting strategy. To account for sampling bias due to mechanisms captured by these variables, we allow the bagging resampling to be weighted according to these inverse probability weights. We used the inverse probability of sampling weights to correct Quantile IV estimates in our MR of the WHR and LDL-c because genetic associations with LDL-c and anthropomorphic traits were shown to be biased due to sampling in the UK Biobank.[54]

Evaluation of the quantile IV estimator

In simulation scenarios, we had access to the true causal relationship between the exposure and outcome $g(X,W)$. In the simulation, the conditional expectation of the confounder given the covariates is 0, meaning that the estimated IV function is an estimate of the interventional effect under our model (i.e., $\mathbb{E}(U|W) = 0$ in Equation 2 and $\mathbb{E}(Y|\mathrm{do}(X), W) = g(X, W) = h(X, W)$). Leveraging this relationship and following previous work,[21] we assessed model performance by directly comparing the IV regression function (h) to the known interventional value of the outcome ($\mathbb{E}(Y|\mathrm{do}(X))$). Specifically, we report the root of the mean-squared differences between these two functions over a grid spanning the range of the exposure.

Results

Evaluation of nonparametric IV estimators in realistic MR simulation scenarios

To evaluate the use of neural network-based nonparametric IV estimators in MR, we evaluated the performance of two recently proposed estimators (DeepIV[21] and DeLIVR[15]) and our proposed method (Quantile IV); we also compared all three with the traditional linear 2SLS estimator. There is no single parameter describing the shape of the causal relationship between the exposure and outcome in nonlinear settings. We use the square root of the mean-squared error (RMSE) between the true causal relationship and the IV regression function to evaluate the performance of the different estimators. The RMSE is taken over evenly spaced points spanning the range of the exposure. We computed this metric for 200 simulation replicates, and we consider scenarios varying the causal relationship shape, the sample size, the variance explained in the exposure by the IVs, the strength of confounding, and the number of IVs (Table 1). We report the results over the range of the exposure encompassing 95% of the distribution (Figure 1) but observed similar results over the full range (Figure S3). Quantile IV was competitive and achieved low RMSEs across all of the simulation scenarios. The linear estimate (2SLS) is provided as a baseline comparison that is not expected to outperform the nonparametric estimators except in the linear model. Furthermore, we observed that the DeepIV estimator had high error and variability across the considered parameters, prompting us to focus on DeLIVR and Quantile IV for quantitative comparisons. We used t tests paired by simulation replicates to compare DeLIVR and Quantile IV (Tables 2 and S7). In most of our simulation scenarios, Quantile IV significantly outperformed DeLIVR at the nominal p value threshold of 0.01 (12/15 scenarios when considering 95% of the exposure range and 9/15 scenarios when considering the full range). DeLIVR significantly outperformed Quantile IV when the number of instruments was set to 100, the largest number of IVs considered in our study. In all our simulations, there was a linear and homogeneous effect of the IVs on the exposure, which is an assumption made by DeLIVR, favoring this model.

To model binary outcomes, we use a logistic model implemented by treating the linear output of the Quantile IV estimator as the log odds of the outcome. We investigated the performance of Quantile IV with simulated binary outcomes with 2,500 (prevalence 5%) and 15,000 (prevalence 30%) cases, and we compared the estimate to DeLIVR and the two-stage control function estimator (supplemental methods D.2).[15,55] In the first scenario with 2,500 cases, Quantile IV crudely estimated the causal log odds and was outperformed by DeLIVR and a linear estimator (Figures S4 and S5). In the second scenario with a high number of cases, Quantile IV outperformed both of the other models (Figures S4 and S5). We note

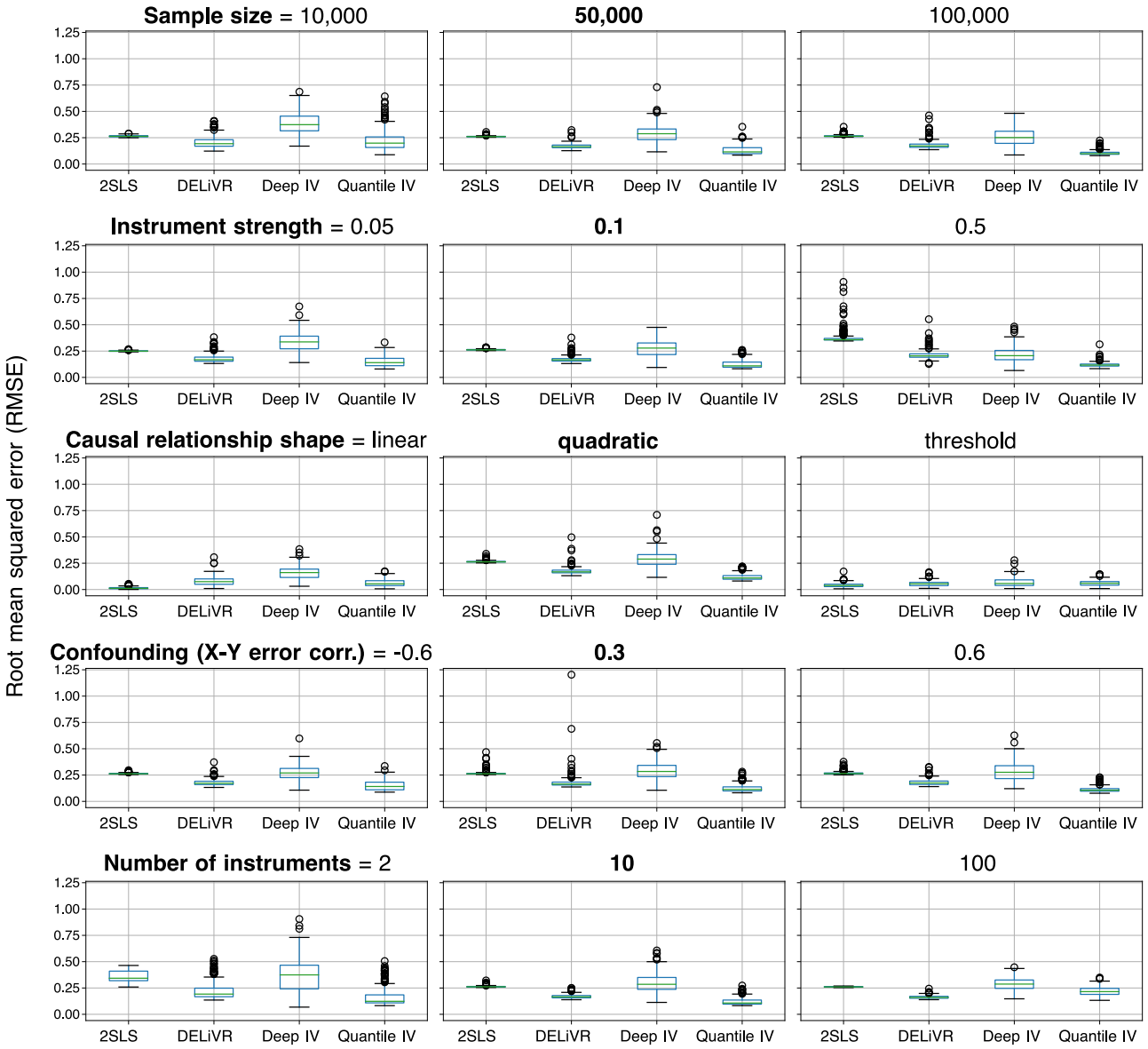


Figure 1. Root-mean-square error between the estimated IV regression and the true causal function over a grid spanning 95% of the empirical range of the exposure

The boxplot for every estimator represents variability over 200 simulation replicates. The simulation parameter values in bold correspond to the reference values. 2SLS (two-stage least squares), DeLIVR,[15] DeepIV,[21] and Quantile IV (proposed method).

that in the real data application considered in this paper, the binary outcomes have between 8,535 cases (ischemic stroke) and 33,516 cases (acute CAD; Table S2), and we expect Quantile IV to perform well given the large number of cases.

A drawback of many nonparametric IV methods is that they do not easily allow for the computation of CIs or provide means to quantify uncertainty. We evaluated the use of bagging to construct CIs focusing on the simulation scenario that consisted of the baseline value for all of the simulation parameters (Table 1). We observed that coverage of the ATE for a unit increment in the exposure was close to or upwards of the nominal value across most of the exposure range (Figure S6). We recommend reporting effects within the central range of the exposure, where coverage was more consistent (e.g., between the 2.5th and 97.5th percentiles of the empirical distribution of the exposure). We also observed a surprising drop in coverage below the nominal level for exposure values near 0 despite a low absolute error in the estimation of the ATE (Figure S6). We attribute this to overconfidence of the bagging CI in this region, which could be due to factors related to the selection of the neural network parameter initialization or architecture. We used a similar strategy to evaluate the false positive rate and observed that our estimator did not exceed the nominal level in the simulation scenario where all parameters are fixed at their reference value (i.e., the values marked with a footnote in

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

Table 2. Comparison between the mean square root of the mean-squared error between DeLIVR and quantile IV across the different Mendelian randomization simulation scenarios and over the central 95% of the empirical exposure distribution

Sim. value	DeLIVR mean RMSE (SD)	Quantile IV mean RMSE (SD)	*p* value
Sample size			
10,000	0.21 (0.06)	0.22 (0.10)	0.094
50,000	0.17 (0.03)	0.13 (0.04)[a]	2.81×10^{-22}
100,000	0.18 (0.04)	0.11 (0.02)[a]	1.34×10^{-57}
Proportion of the variance in the exposure explained by the IV			
0.05	0.18 (0.04)	0.15 (0.05)[a]	1.50×10^{-10}
0.1	0.17 (0.03)	0.13 (0.04)[a]	2.74×10^{-24}
0.5	0.21 (0.04)	0.12 (0.02)[a]	3.01×10^{-71}
Causal relationship shape			
Linear	0.08 (0.04)	0.06 (0.03)[a]	1.79×10^{-5}
Quadratic	0.18 (0.04)	0.12 (0.03)[a]	7.89×10^{-38}
Threshold	0.06 (0.03)	0.06 (0.03)	0.061
Confounding			
−0.6	0.18 (0.03)	0.15 (0.05)[a]	8.53×10^{-12}
0.3	0.18 (0.09)	0.13 (0.04)[a]	4.07×10^{-15}
0.6	0.18 (0.03)	0.11 (0.03)[a]	9.62×10^{-58}
No. of instruments			
2	0.23 (0.09)	0.17 (0.09)[a]	3.38×10^{-20}
10	0.17 (0.02)	0.12 (0.04)[a]	5.58×10^{-40}
100	0.16 (0.01)[a]	0.22 (0.04)	2.35×10^{-41}

The *p* value is from a paired t test by simulation replicate.
[a]The smallest error when the difference is statistically significant.

Table 1; Figure S7). We also estimated the type 2 (false negative) error rate for a unit increase from different reference exposure values (Figure S8). Because the simulated relationship is nonlinear, the ATE for a unit increase in the exposure can be taken at various reference points with different expected causal effects. Our model had the power to detect non-null ATEs in the regions where the exposure had a larger effect on the outcome but not when the ATE was closer to the null. Parametric models will generally have greater statistical power than nonparametric methods and will detect causal effects with high sensitivity. However, the estimated effects could be biased under violations of their modeling assumptions.

These MR simulations characterized the performance of Quantile IV in realistic settings, but they do not demonstrate the added benefits of fully nonparametric IV estimation. We considered another simulation scenario with complex confounding and causal effect heterogeneity to demonstrate the advantages of our method more explicitly (supplemental methods). In this scenario, Quantile IV estimates heterogeneous causal effects with low error, whereas DeLIVR exhibits substantial bias, as exhibited by marked deviations from the true CATE line in Figure S9.

Genetic association and colocalization analysis of circulating sclerostin and outcomes in the UK Biobank

To identify *cis*-pQTLs associated with circulating sclerostin levels, we conducted a genetic association analysis of 1,449 common variants at the *SOST* locus in the UK Biobank (Figure S10). Our goal was to select strong genetic predictors of sclerostin levels for subsequent MR analyses by using fine-mapping to select IVs. A SuSiE analysis[36] revealed two 95% credible sets with $\log_{10}$ of the Bayes factor of 10.3 and 5.0. The first set contained 17 variants, and the variant with the largest PIP was rs6416905 (PIP = 0.078). The second set contained 6 variants, and rs66838809 had the largest PIP (PIP = 0.48). Variants within the two credible sets were in high LD, with a mean $r^2 = 1$ for the first set and a mean $r^2 = 0.89$ for the second set. We selected the two variants with the highest PIP for subsequent MR analyses, as they were the best representatives of the two independent signals identified by SuSiE. As a sensitivity analysis, we also used a forward stepwise regression procedure and obtained concordant results (supplemental note). The estimated effects of the two selected IVs on circulating sclerostin levels are presented in Table S8.

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

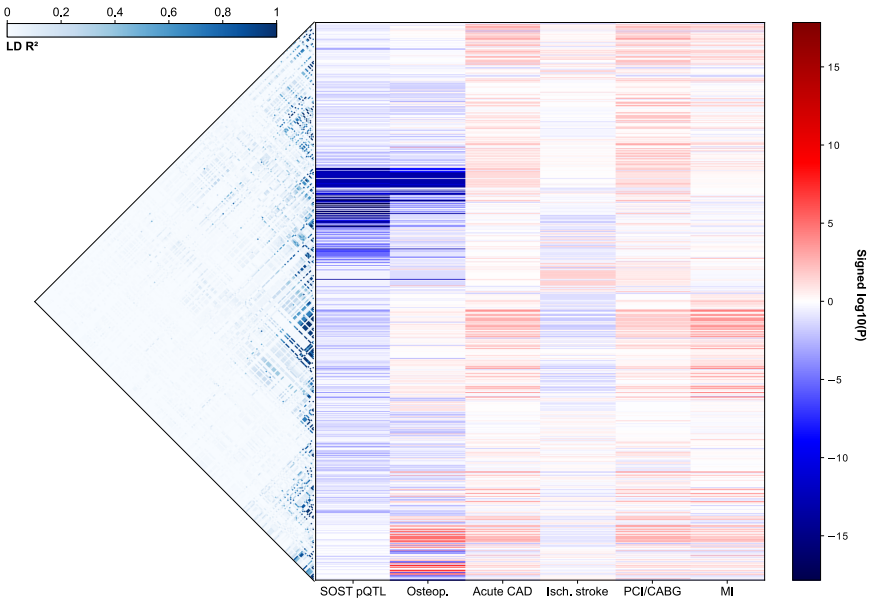


Figure 2. Association between common genetic variants at the *SOST* locus (chr17:41631099–42236156) and the exposure and outcomes considered in our MR study in the UK Biobank

The variants are ordered with respect to their genomic coordinate, and the leftmost part of the plot shows the observed LD between the variants. The color lines represent association *p* values for every genetic variant and phenotype colored with respect to the sign of the regression coefficient (red for trait increasing and blue for trait decreasing). The sclerostin-decreasing allele is the coded allele throughout.

We used the same genetic association and fine-mapping procedure for the outcomes of interest, namely heel BMD, osteoporosis, and cardiovascular outcomes (PCI/CABG, MI, ischemic stroke, and acute CAD). The association *p* values expressed with respect to the sclerostin-decreasing allele are shown in Figure 2, along with the measured LD in the region. We identified significant associations with heel BMD, and the top association was with rs66838809 (chr17:41798621G/A), whose A allele had $\widehat{\beta} = 0.072$, 95% CI (0.061, 0.083), and $p = 5.8 \times 10^{-38}$ (Table 3). This variant is also a lead variant for one of the sclerostin pQTL credible sets (Table S8). We identified an association between genetic variants at the *SOST* locus and osteoporosis, and the lead variant was again rs66838809 (chr17:41798621G/A) whose "A" allele had an odds ratio (OR) = 0.83, 95% CI (0.80, 0.87), and $p = 3.8 \times 10^{-17}$. To formally test whether the observed associations were due to shared causal variants with the sclerostin levels, we conducted a colocalization analysis.[38]

The colocalization algorithm estimates the PP of four scenarios denoted by H_1 to H_4, including the existence of shared causal variants underlying the association signal in both of the considered traits (H_4). It is possible to test for these hypotheses assuming a single causal variant or to use fine-mapping and test pairs of credible sets to allow for multiple causal variants per region.[38] Colocalization analysis revealed strong evidence of a shared causal variant between heel BMD and circulating sclerostin levels. The two sclerostin pQTL credible sets colocalized with the heel BMD credible set with $PPH_4 = 0.964$ and 0.997. There was also strong evidence for colocalization between osteoporosis and one of the pQTL credible sets with PPs of a shared causal variant of 0.996. The second sclerostin pQTL credible set (represented by rs6416905) had no evidence of colocalization with osteoporosis ($PPH_4 = 0.013$). Overall, there was robust evidence of colocalization between genetic associations with circulating sclerostin levels, heel BMD, and osteoporosis, recapitulating the known protective effect of pharmacological sclerostin inhibition in osteoporosis.[22,23]

There was evidence of genetic associations at the *SOST* locus with some of the considered cardiovascular outcomes (Figure 2). The outcome with the most significant genetic association was MI, and the T allele of the lead variant, rs75086002 (chr17:42021918C/T), had an OR of 0.92 95% CI (0.89, 0.96) and $p = 5.9 \times 10^{-6}$. This association does not reach the genome-wide significance threshold, but it did cross the conservative Bonferroni threshold considering multiple testing of 1,449 variants ($3.5 \times 10^{-5} = 0.05/1{,}449$). Using SuSiE, we inferred a 90% credible set for this association signal, but it did not colocalize with the sclerostin pQTL credible sets (PPH_4 with the rs6416905 credible set = 0.005 and PPH_4 with the rs66838809 credible set = 0.004). Visual inspection of the association signals and inferred credible sets is also consistent with the hypothesis of distinct association signals for MI and sclerostin levels (Figure S11). The most probable hypothesis, according to colocalization analysis, is that the associations are due to distinct genetic variants ($PPH_3 = 0.78$). Because the inference of credible sets can be statistically challenging, we also tested colocalization assuming the existence of a unique causal variant in the region. The results were concordant with the absence of colocalization between sclerostin pQTLs and cardiovascular diseases (largest $PPH_4 = 0.01$ with ischemic stroke). We retrieved the summary association statistics between our two selected sclerostin pQTLs and the considered outcomes in the UK Biobank and in data from the CARDIoGRAMplusC4D consortium, which includes over 70,000 cases of CAD (Table 3). In our UK Biobank analysis, there was a nominally significant association of rs66838809 with acute CAD ($p = 0.02$) and PCI/CABG ($p = 0.003$) but no evidence of association in data from the CARDIoGRAMplusC4D consortium (*p* values for the

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

Table 3. Association statistics between the two lead variants for the sclerostin pQTL credible sets identified by SuSiE and quantitative traits and diseases

	Coef./OR (95% CI)		Coef./OR (95% CI)	
Dataset	**rs6416905**	***p* value**	**rs66838809**	***p* value**
UK Biobank quantitative traits				
pQTL	−0.059 (−0.072, −0.046)	1.50×10^{-18}	−0.097 (−0.120, −0.074)	1.80×10^{-16}
Heel bone mineral density	0.032 (0.026, 0.039)	3.30×10^{-25}	0.072 (0.061, 0.083)	5.80×10^{-38}
UK Biobank diseases (OR scale)				
Osteoporosis	0.954 (0.934, 0.976)	3.10×10^{-5}	0.834 (0.800, 0.870)	3.80×10^{-17}
Myocardial infarction	0.988 (0.967, 1.010)	0.28	1.017 (0.979, 1.057)	0.39
Acute CAD	1.000 (0.981, 1.019)	0.99	1.042 (1.008, 1.078)	0.02
Ischemic stroke	0.986 (0.955, 1.018)	0.38	1.002 (0.946, 1.061)	0.95
PCI/CABG	1.007 (0.984, 1.030)	0.58	1.064 (1.022, 1.108)	0.003
Summary statistics (consortia) quantitative traits				
GTEx v.8 eQTL (artery tibial)	−0.253 (−0.300, −0.206)	6.00×10^{-23}	−0.311 (−0.410, −0.211)	2.40×10^{-9}
GTEx v.8 eQTL (artery aorta)	−0.247 (−0.319, −0.176)	7.00×10^{-11}	−0.401 (−0.542, −0.261)	5.00×10^{-8}
Summary statistics (consortia) diseases (OR scale)				
CARDIoGRAM MI	1.002 (0.981, 1.024)	0.83	1.043 (0.998, 1.091)	0.06
CARDIoGRAM CAD	0.991 (0.973, 1.010)	0.37	0.997 (0.958, 1.038)	0.90
CARDIoGRAM and UKB CAD	0.994 (0.978, 1.011)	0.48	0.999 (0.966, 1.033)	0.96

All of the effects are presented with respect to the sclerostin-reducing allele. The UK Biobank associations were estimated in the current study, and we present associations published by the GTEx (sclerostin expression) and CARDIoGRAMplusC4D (cardiovascular diseases) consortia. Coef., coefficient; UKB, UK Biobank.

largest CAD study were 0.48 for rs6416905 and 0.96 for rs66838809).

Previous studies have considered eQTLs of sclerostin levels as instruments in MR studies.[24,56] We sought to evaluate whether there was a shared genetic basis for the regulation of sclerostin eQTLs and circulating sclerostin levels. The selected sclerostin pQTLs were robustly associated with *SOST* expression in aorta and tibial arteries ($p \leq 5 \times 10^{-8}$; Table 3). We observed that the eQTLs and pQTLs were located near each other on the chromosome and were significantly associated with both sclerostin gene expression and circulating levels (Figures S12–S14). Statistical colocalization analysis, however, suggests that the underlying causal variants are different (max $PPH_4 = 0.12$ for aorta and 0.004 for tibial artery; Figure S15). This finding could be due to the presence of different genetic regulatory mechanisms behind sclerostin gene expression in comparison to circulating protein levels, but it could also be due to limited statistical power in GTEx or a mismatch between the GTEx and UK Biobank populations, hampering statistical fine-mapping analyses.

We conclude that association analysis identified strong *cis*-pQTLs of circulating sclerostin levels that colocalize with genetic associations with heel BMD and osteoporosis. Despite some evidence of genetic associations with cardiovascular outcomes at the *SOST* locus, evidence from large consortia of CAD and colocalization analyses suggest that they may be unrelated to genetic variants, influencing the regulation of circulating sclerostin levels.

MR of the effect of circulating sclerostin on bone and cardiovascular diseases

To estimate the causal effect of a genetically predicted reduction in circulating sclerostin levels on bone and cardiovascular traits and outcomes, we considered three complementary approaches suitable for the *cis*-MR context.[41] Using the fine-mapped sclerostin pQTLs as IVs, we used the IVW estimator accounting for LD and our Quantile IV nonparametric estimator. As a complementary approach, we used PC-GMM, an estimator that can leverage all of the variants in the region.[44] All of the methods estimated that reducing circulating sclerostin would result in a statistically significant increase in heel BMD and a reduction in the risk of osteoporosis (Table S9). However, the magnitude of the estimates was different across estimators. We report the estimated effect for both a 1 SD reduction in sclerostin levels about the mean and a 2 SD reduction about the mean (Table S9). These two contrasts correspond, assuming normality in the distribution of circulating sclerostin levels, to a reduction from the mean to the level of the bottom 16% or the bottom 2% of the distribution of the exposure. The IVW estimator suggests that a 2 SD reduction in sclerostin levels about the mean reduces the odds of osteoporosis by a surprising 92% (OR = 0.085). For the same change

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

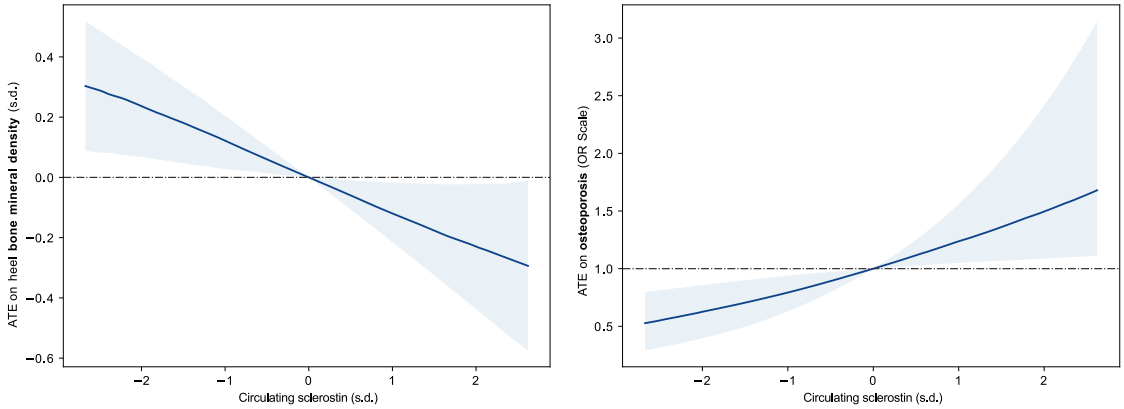


Figure 3. Quantile IV estimate of the average effect varying the levels of circulating sclerostin about the mean on heel bone mineral density and osteoporosis in the UK Biobank
The shaded region corresponds to 90% bootstrap confidence intervals. The plots cover the central 99% of the exposure range.

in circulating sclerostin levels, Quantile IV estimates an OR of 0.626. Upon visual inspection of the Quantile IV estimate, there was no evidence of pronounced nonlinearity (Figure 3). When considering heel BMD and osteoporosis as outcomes, Quantile IV estimated smaller, albeit significant, causal effects than conventional linear methods.

Nonparametric IV estimation enables the estimation of CATEs without explicitly specifying an interaction model. Such effects represent the average effect of a treatment in a specific subset of individuals and can be used to assess treatment response heterogeneity. To evaluate if the causal effect of varying circulating sclerostin levels differs across levels of covariables, we estimated CATEs in males, females, and individuals with different values of age at baseline. Age did not modify the effect of sclerostin inhibition on heel BMD or osteoporosis ($p = 0.43$ and $p = 0.35$, respectively). However, we observed that sex heterogeneity as the CATE for a 1 SD decrease in circulating sclerostin levels was 0.15 for females vs. 0.09 for males (interaction $p = 2.3 \times 10^{-22}$; Table S10; Figure S16). We observed directionally concordant sex differences on the effect of sclerostin inhibition on osteoporosis with interaction $p = 0.04$, but the effect difference was small.

When estimating the causal effect of a reduction of sclerostin levels on cardiovascular outcomes, there was disagreement between the results obtained by different estimators. The IVW and PC-GMM estimated a nominally significant increase in the risk of PCI/CABG ($p_{IVW} = 0.026$, $p_{PC-GMM} = 0.017$; Figures S17 and S18), and PC-GMM estimated a nominally significant increase in the risk of acute CAD ($p = 0.021$; Figure S18). Quantile IV, on the other hand, estimated a nominally significant protective effect on PCI/CABG (OR = 0.74, 95% CI (0.52, 0.97); Figures S19 and S20). However, we interpret these results with care, as estimates from the three estimators were heterogeneous, and there was no supporting colocalization evidence for shared causal variants between sclerostin pQTLs and cardiovascular diseases in our previous analysis. We attribute these results to bias due to LD with other genetic variants that have effects on the outcome independently of sclerostin levels and investigate this in the next section.

MR analyses of sclerostin accounting for pleiotropic effects

In the MR analyses adjusted for age, sex, and ancestry principal components, we observed conflicting effects for PCI/CABG despite limited evidence of genetic associations between *SOST* variants and this outcome. The top association was with rs370088062, chr17:41657403 C to CT insertion (dbSNP rs36035748, GenBank: NC_000017.10, g.41657419dup), with $p = 2.9 \times 10^{-4}$. This variant showed no association with circulating levels of sclerostin ($p = 0.40$), indicating that the observed effects in MR might be biased due to LD. More precisely, if the IVs are in LD with other genetic variants that have effects on the outcome independently of sclerostin levels, the exclusion restriction assumption will be violated, biasing the MR estimates. This problem is particularly challenging in the *cis*-MR context because of LD and the limited number of candidate IVs. To identify the variants most at risk of violating the MR assumptions, we repeated the association analysis with PCI/CABG, adjusting for all of the pQTLs identified in the stepwise conditional analysis (supplemental note). We then compared the association p values before and after adjustment for the pQTL associations under the premise that variants whose association is unattenuated may influence PCI/CABG risk through pathways unrelated to sclerostin (Figure S21). This analysis revealed a group of correlated variants, including rs113533733, that had low-association p values with PCI/CABG after adjusting for the sclerostin pQTL variants.

We sought to confirm that rs113533733 may have effects on other genes by consulting the Open Targets Genetics platform.[57] This online resource includes a variant-to-gene prioritization module based, in part, on the distance to transcription start sites and pQTL, splice QTL (sQTL), and eQTL data. On this platform, the most likely

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

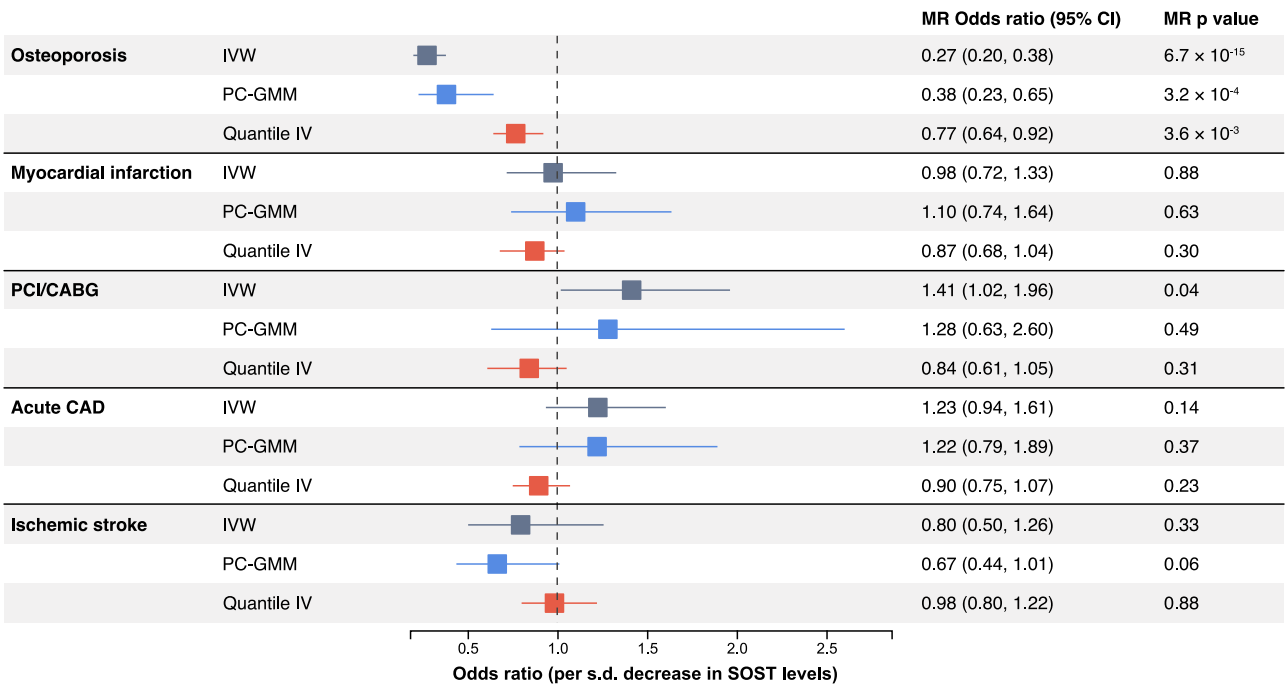


Figure 4. Mendelian randomization estimates of a 1 SD decrease in circulating sclerostin levels on osteoporosis and cardiovascular diseases accounting for direct effects by rs113533733 in the UK Biobank

gene assigned to rs113533733 is *MPP3* (MIM: 601114) (score = 0.31), with support from sQTL and eQTL data. The other prioritized genes are *DUSP3* (MIM: 600183) (score = 0.18), *CFAP97D1* (MIM: 619866) (nearest gene, score = 0.15), and *MPP2* (MIM: 600723) (score = 0.13). There is no evidence linking *SOST* to rs113533733 except for its distance to the transcription start site of 41 kb, and the assigned score is 0.06. In GTEx v.8, the strongest eQTL for this variant was with *MPP3* in tissue from the left ventricle of the heart ($p = 2.1 \times 10^{-5}$). Considering this external evidence and our association analysis conditional on sclerostin pQTLs, we concluded that rs113533733 may induce bias in the MR analysis and repeated our MR estimation, adjusting for this variant.

In the MR analysis adjusted for rs113533733, the effect of a 1 SD reduction in circulating sclerostin levels on heel BMD and osteoporosis remained significant, with no attenuation in the *p* value for all methods (Figure 4; Table S11). After adjustment for rs113533733, the estimated causal effects of sclerostin levels on the considered cardiovascular diseases were null for PC-GMM and Quantile IV (Figures 4 and S22). The IVW had an inconsistent estimate for the effect of sclerostin reduction on PCI/CABG (OR = 1.41, 95% CI (1.02, 1.96), $p = 0.04$), but the large CI, discordance with the other estimators, absence of effect with related traits, and lack of support from colocalization analyses suggest the true effect is likely null.

Post hoc analysis of linear effect underestimation by Quantile IV

In MR analyses of the effect of circulating sclerostin on heel BMD and osteoporosis, the effect estimated by Quantile IV is substantially smaller than the effect estimated by linear models. For example, the estimated effects for a 1 SD reduction in circulating sclerostin on heel BMD are 0.122 for Quantile IV vs. 0.526 for PC-GMM. The contrast between the estimates is surprising, and our simulation study did not include a comparable case with two IVs with modest effect sizes. To address this, we conducted a post hoc semi-synthetic simulation study aimed at reproducing the results from the real data application in terms of sample and effect sizes (supplemental methods). When we simulated a linear causal effect of −0.526 (equivalent to the PC-GMM estimate in our study), the mean linearized Quantile IV underestimated the effect by 22% (Figure S23). This difference remains far from the effect observed in the MR of sclerostin on heel BMD, where the effect was 77% smaller than the PC-GMM estimate. This result suggests that the underestimation of the causal effect by Quantile IV alone is unlikely to explain the discrepancy between Quantile IV and the linear estimators.

MR estimation of the effect of LDL-c and WHR on ischemic cardiovascular diseases

To estimate the causal effect of an increase in LDL-c and the WHR on cardiovascular outcomes, we first used a two-sample MR. All of the estimators found that increasing LDL-c increased the risk of MI and CAD (Table S12). The estimated ORs per 1 mmol/L increase in LDL-c across estimators and outcomes ranged from 1.65 to 2.10 (assuming an SD of 0.87 mmol/L; Table S5), which is concordant with prior studies.[58] A genetically predicted 1 SD increase in the WHR was consistently predicted to

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

Table 4. MR estimate of the effect of a 1 mmol/L increase in LDL-c or a 0.10 unit increase in WHR on cardiovascular outcomes in the UK Biobank using quantile IV

	CATE, OR scale (95% CI)			CATE, OR scale (95% CI)		
	Female	Male	*p* value (itx.)	Statin non-user	Statin user	*p* value (itx.)
Exposure: 1 mmol/L increase in LDL-c						
MI	1.63 (1.26, 1.97)	1.63 (1.26, 1.98)	0.743	1.38 (1.25, 1.57)	1.33 (1.18, 1.54)	1.78×10^{-18}
PCI/CABG	2.02 (1.53, 2.44)	2.02 (1.54, 2.43)	0.955	1.76 (1.55, 2.04)	1.65 (1.40, 2.01)	9.24×10^{-27}
Acute CAD	1.63 (1.35, 2.00)	1.65 (1.34, 2.00)	0.127	1.40 (1.28, 1.54)	1.36 (1.20, 1.51)	3.99×10^{-15}
Ischemic stroke	1.19 (1.00, 1.35)	1.21 (1.00, 1.37)	0.033	1.07 (0.94, 1.24)	1.07 (0.92, 1.23)	0.415
Exposure: 0.10 unit increase in the WHR						
MI	1.45 (1.29, 1.62)	1.44 (1.26, 1.69)	0.813	1.46 (1.31, 1.66)	1.45 (1.29, 1.74)	0.432
PCI/CABG	1.59 (1.42, 1.90)	1.58 (1.36, 1.93)	0.172	1.56 (1.41, 1.86)	1.54 (1.36, 1.88)	0.252
Acute CAD	1.42 (1.29, 1.60)	1.41 (1.26, 1.61)	0.054	1.40 (1.31, 1.57)	1.38 (1.26, 1.57)	0.015
Ischemic stroke	1.41 (1.25, 1.69)	1.43 (1.25, 1.71)	0.033	1.45 (1.27, 1.76)	1.42 (1.22, 1.78)	0.070

The estimates are adjusted for age, sex, the first 5 ancestry principal components (PCs), and statin use. Selection weights are used in the resampling procedure to lessen the possible impact of selection bias. The reported *p* values represent interaction (itx.). *p* values computed using one-way ANOVA of the CATE estimates from the compared groups across bootstrap replicates.

increase the odds of cardiovascular outcomes with ORs ranging from 1.29 to 1.83 (Table S13). Overall, these analyses represent positive controls recapitulating the well-established effects of the WHR and LDL-c as causal risk factors for ischemic cardiovascular diseases.

We repeated the parametric MR analyses using individual-level data from the UK Biobank and reported the MR effects per 1 mmol/L increase in LDL-c or per 0.1 increase in the WHR (Tables S14 and S15). Using MI as a point of comparison between the summary statistics and one-sample analyses of the causal effect of a 1 mmol/L increase in LDL-c, the IVW MR ORs were 1.88 for the one-sample analysis vs. 1.39 for the summary statistics analysis. This difference could be attributed to differences in the handling of the LDL-c measurements, as the GLGC re-scaled observed LDL-c measurements by a factor of 0.7× in individuals who were users of an LDL-c-lowering drug.[35] Then, to evaluate the possibility of nonlinear or heterogeneous causal effects, we used Quantile IV. There was some qualitative evidence of nonlinearity of the effect of LDL-c on cardiovascular outcomes on the OR scale (Figure S24). Increasing LDL-c was predicted to increase the odds for all of the considered cardiovascular outcomes, albeit marginally for ischemic stroke, for which the analysis was likely underpowered due to the lower number of cases in the UK Biobank. Using MI as a point of comparison across MR estimates, Quantile IV estimated an OR of 1.65 (1.36, 1.93), which is consistent with the parametric estimates. The Quantile IV estimates for the effect of the WHR showed a significant increase in cardiovascular risk, albeit with a smaller effect magnitude when compared to the parametric estimates (Figure S25).

To highlight the advantage of Quantile IV over alternative methods, we estimated *conditional* treatment effects. We used the fitted Quantile IV model to estimate the CATEs for males, females, statin users, and statin non-users and used one-way ANOVA to test for differences in the conditional estimates between groups (Table 4). We found evidence of effect heterogeneity in the MR effect of increasing LDL-c on MI, PCI/CABG, and acute CAD, where the increase in cardiovascular risk was attenuated in statin users compared to statin non-users. This observed gene-by-environment interaction suggests that the increased cardiovascular disease risk associated with LDL-c can be partially offset by the use of statins, which pharmacologically lower LDL-c. This result must be interpreted with care, as adjustment for statin use could induce bias in certain scenarios (supplemental note). Additionally, there was some evidence for sex differences in the effect of LDL-c and the WHR on ischemic stroke. For the same increase in LDL-c or the WHR, males are predicted to have a larger increase in the odds of stroke (Table 4).

We repeated the analysis with the doubly ranked nonlinear MR (DRMR) estimator as a secondary nonlinear MR method (Figures S26 and S27). For the effect of LDL-c on MI, PCI/CABG, and acute CAD, the DRMR estimate suggested a nonlinear relationship, with larger MR effects at lower values of the LDL-c distribution. However, the DRMR estimates had wide CIs, suggesting a high degree of uncertainty. The nonlinear relationship estimated by DRMR was not corroborated by Quantile IV, which suggested a linear effect close to the IVW estimate.

Discussion

This study introduces Quantile IV, a nonparametric MR estimator that offers computational stability while introducing few statistical assumptions. This innovation is important given the limited exploration of nonparametric IV estimators in MR contexts. We evaluated Quantile IV across many realistic MR scenarios and applied it to study

Please cite this article in press as: Legault et al., A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition, The American Journal of Human Genetics (2025), https://doi.org/10.1016/j.ajhg.2025.04.010

the causal effect of circulating sclerostin inhibition on heel BMD, osteoporosis, and cardiovascular diseases. As a second application, we estimated the causal effect of LDL-c and the WHR on cardiovascular outcomes. Compared to DeepIV, another nonparametric IV estimator, Quantile IV, consistently showed lower error and greater stability in all simulations. These scenarios probed various elements, such as the impact of differing sample sizes, the strength of the IVs, the shape of the causal effect, the degree of confounding influences, and the total number of IVs employed. Quantile IV's performance varied more in simulations with the smallest sample size (n = 10,000). The only scenario where DeLIVR, a semi-parametric relaxation of the DeepIV estimator, significantly outperformed Quantile IV is when the number of IVs was set to the largest number (100 independent IVs), but we reiterate that the simulations favored DeLIVR as the linearity and constant variance assumptions held. A limitation of Quantile IV is its lack of direct methods for uncertainty quantification. We used bagging to construct CIs and compute p values for the hypothesis that the ATE (or CATE) is null. The bagging CIs had good coverage of the true value on average, but we did notice some localized miscoverage despite low estimation error. We attribute this finding to the neural network regularization and weight initialization, which may favor null effects, resulting in conservative estimates when the ATE is close to zero. The false positive rate was well controlled in our simulations, never exceeding the nominal level.

The causal effect of inhibiting sclerostin on bone and cardiovascular health has been predicted using MR in the past.[24,56,59] However, the results from previous MR studies are conflicting, and we opted to revisit the question using updated MR estimators, support from fine-mapping and colocalization analyses, and circulating sclerostin measurements from the UK Biobank Pharma Proteomics Project. To briefly summarize previous work, Bovijn et al. used IVs at the *SOST* locus that were ascertained based on their effect on BMD.[24] Using MR, they recapitulated the protective effect of sclerostin inhibition on osteoporosis and fracture risk and estimated an increase of 18% in the odds of MI per 0.09 g/cm^2 of BMD (p = 0.003).[24] The second MR study by Holdsworth et al. selected genetic variants at the *SOST* locus that were eQTLs of *SOST* in arterial or heart tissue and associated with BMD.[56] These variants were reported to not be associated with cardiovascular outcomes, including MI and CAD, in CARDIoGRAMplusC4D. Another study by Zheng et al. conducted a GWAS meta-analysis of circulating sclerostin, including 33,961 individuals from 9 cohorts.[59] A total of 18 conditionally independent variants were associated with circulating sclerostin, and the authors have conducted MR based on all of the variants (including *trans* effects) and a subset of *cis*-acting variants. The *cis*-MR analysis from Zheng et al. suggests an increased risk of MI with an OR of 1.35 (p = 0.04) per 1 SD lowering of sclerostin levels.[59] Faced with these contradictory results, we first investigated the possibility of bias due to LD with variants that could influence cardiovascular disease risk independently of sclerostin. Using fine-mapping, we were able to infer two credible sets explaining the *cis*-regulatory signal of circulating sclerostin levels. The variants in the circulating sclerostin credible sets colocalized with the osteoporosis and heel BMD credible sets but not with MI or other cardiovascular diseases. This result suggests that previous MR estimates may have been biased due to the presence of LD with cardiovascular risk variants that act independently of the modulation of circulating sclerostin levels.

We then conducted *cis*-MR analyses to estimate the effect of a reduction in circulating sclerostin on heel BMD, osteoporosis, and cardiovascular outcomes. In accordance with the well-known clinical effect of pharmacological sclerostin inhibition,[22,23] our MR estimates showed that a genetically predicted reduction in circulating sclerostin levels leads to an increase in heel BMD and a decrease in the risk of osteoporosis. There was evidence of sex differences with a larger increase in BMD for the same reduction in sclerostin levels in females compared to males (0.15 vs. 0.09). This is concordant with the difference observed comparing the ARCH trial of postmenopausal women to the BRIDGE trial of men, where monthly romosozumab injections increased lumbar spine BMD by 13.7% vs. 12.1% and total hip BMD by 6.2% vs. 2.5%.[22,23] These results illustrate how conditional MR effect estimation may detect effect heterogeneity. In our study, the magnitude of the Quantile IV estimate for the effect of sclerostin inhibition on heel BMD and osteoporosis was 4–5 times smaller than the parametric estimates from IVW and PC-GMM. We validated that this underestimation was unlikely to be attributable to a downward bias of the Quantile IV estimator using a realistic post hoc simulation model based on the observed values of the instrument's strength and causal effects. The larger estimates from parametric models could be due to the linear extrapolation of small genetic effects (e.g., allelic effects of <0.1 SD) to predict the effect of a comparatively larger 1 SD decrease in circulating sclerostin. MR estimates are often larger than effects estimated in randomized controlled trials, and the difference is typically attributed to the comparison of lifelong vs. short-term effects.[60] Whether violations of parametric assumptions contribute to the inflation of effect estimates in real-world settings is unclear. Finally, our MR estimates accounting for possible bias due to LD were concordant with colocalization analyses and results from large CAD genetics consortia and found a null relationship between genetically predicted circulating sclerostin levels and ischemic cardiovascular diseases. Because of the limited number of observed cardiovascular outcomes in our study, it is possible that a small causal effect would remain undetected due to low statistical power. However, a false negative finding is unlikely because the lead sclerostin pQTLs were not significantly associated with MI or

CAD in the largest available dataset of GWAS summary statistics from the CARDIoGRAMplusC4D consortium. The increase in cardiovascular events observed in individuals treated with romosozumab in clinical trials could be explained by off-target effects or effects that are tissue specific and not captured by our genetic model of sclerostin inhibition. Furthermore, violations of the MR assumptions remain possible, as the genetic instruments used in our study could have pleiotropic effects (directly or through variants in LD).

As a second application of our MR estimator, we considered the effect of varying the levels of LDL-c and the WHR on atherosclerotic cardiovascular outcomes. Overall, the Quantile IV estimates were concordant with other methods, replicating the positive causal effect of the WHR and LDL-c on ischemic cardiovascular outcomes. A significant advantage of Quantile IV over alternative approaches is its ability to estimate causal effects corresponding to different levels of covariates included in the model. In the current study, a constant, genetically predicted increase in LDL-c was associated with a smaller increase in atherosclerotic cardiovascular disease risk in statin users compared to non-users. This observation is likely due to statins partially offsetting the increased cardiovascular risk of LDL-c-increasing genetic variants. Quantile IV, in contrast with the DRMR method, did not predict the nonlinear effects of LDL-c on coronary outcomes. Nonlinear causal inference of the effect of LDL-c on coronary outcomes remains a challenging problem. DRMR is less biased than other localized average causal effect estimators,[14] but there have been documented instances of false positive nonlinear relationships inferred by DRMR using negative control outcomes.[61] Whether these false positives are due to biases inherent to the estimation procedure or data-related factors, such as selection bias, remains uncertain.

In this study, we proposed an MR estimator, Quantile IV, and demonstrated its performance in real data and simulation models. Our estimator makes few modeling assumptions when compared to traditional methods, and it allows for nonlinearity and effect heterogeneity. Unlike other MR estimators, Quantile IV allows the estimation of CATE without specifying an interaction model, which is an important tool for assessing treatment response heterogeneity. However, this increased model flexibility is reliant on the use of individual-level data. Despite good overall performance, Quantile IV is a computationally intensive method, especially if CIs need to be estimated via bootstrapping. This could be alleviated by using more computationally efficient model formulations or by developing semi-parametric inference for deep IV models, which will be considered in future work.

Data and code availability

The Quantile IV algorithm is implemented in *ml-mr*: https://github.com/pgx-ml-lab/ml-mr.

Acknowledgments

M.-A.L. received a fellowship from the Canadian Institutes of Health Research (CIHR) that supported this work. M.-A.L. receives funding from IVADO. B.J.A. holds a Senior Scholar Award from the Fonds de Recherche du Québec - Santé.

Author contributions

M.-A.L., B.J.A., and J.P. contributed to the study design and application of the method to UK Biobank data. M.-A.L., J.H., and A.Y. Y. contributed to the methodological development of the quantile IV algorithm and designed the MR simulation study. All of the authors contributed to writing the manuscript and approved the final submitted version.

Declaration of interests

J.H. was an employee of Recursion during the course of this work and has received optional ownership interest in Recursion. B.J.A. is a consultant for Eli Lilly, Silence Therapeutics, Editas Medicine, and Novartis and has received research contracts from Pfizer, Ionis Pharmaceuticals, Eli Lilly, and Silence Therapeutics.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2025.04.010.

Web resources

dbSNP, https://www.ncbi.nlm.nih.gov/snp/
OMIM, https://www.omim.org

Received: February 26, 2024
Accepted: April 22, 2025
Published: May 15, 2025

References

1. Lawlor, D.A., Tilling, K., and Davey Smith, G. (2016). Triangulation in Aetiological Epidemiology. Int. J. Epidemiol. *45*, 1866–1886. https://doi.org/10.1093/ije/dyw314.
2. Voight, B.F., Peloso, G.M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M.K., Hindy, G., Hólm, H., Ding, E.L., Johnson, T., et al. (2012). Plasma HDL Cholesterol and Risk of Myocardial Infarction: A Mendelian Randomisation Study. Lancet *380*, 572–580. https://doi.org/10.1016/S0140-6736(12)60312-2.
3. Burgess, S., and Thompson, S.G. (2015). Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects. Am. J. Epidemiol. *181*, 251–260. https://doi.org/10.1093/aje/kwu283.
4. Triozzi, J.L., Hsi, R.S., Wang, G., Akwo, E.A., Wheless, L., Chen, H.-C., Tao, R., Ikizler, T.A., Robinson-Cohen, C., Hung, A.M.; and VA Million Veteran Program (2023). Mendelian Randomization Analysis of Genetic Proxies of Thiazide Diuretics and the Reduction of Kidney Stone Risk. JAMA Netw. Open *6*, e2343290. https://doi.org/10.1001/jamanetworkopen.2023.43290.
5. Ference, B.A., Robinson, J.G., Brook, R.D., Catapano, A.L., Chapman, M.J., Neff, D.R., Voros, S., Giugliano, R.P., Davey

Smith, G., Fazio, S., and Sabatine, M.S. (2016). Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. N. Engl. J. Med. *375*, 2144–2153. https://doi.org/10.1056/NEJMoa1604304.

6. Sun, B.B., Chiou, J., Traylor, M., Benner, C., Hsu, Y.-H., Richardson, T.G., Surendran, P., Mahajan, A., Robins, C., Vasquez-Grinnell, S.G., et al. (2023). Plasma Proteomic Associations with Genetics and Health in the UK Biobank. Nature *622*, 329–338. https://doi.org/10.1038/s41586-023-06592-6.
7. Henry, A., Gordillo-Marañón, M., Finan, C., Schmidt, A.F., Ferreira, J.P., Karra, R., Sundström, J., Lind, L., Ärnlöv, J., Zannad, F., et al. (2022). Therapeutic Targets for Heart Failure Identified Using Proteomics and Mendelian Randomization. Circulation *145*, 1205–1217. https://doi.org/10.1161/CIRCULATIONAHA.121.056663.
8. Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization between Complex Traits and Diseases. Nat. Genet. *50*, 693–698. https://doi.org/10.1038/s41588-018-0099-7.
9. Mounier, N., and Kutalik, Z. (2023). Bias Correction for Inverse Variance Weighting Mendelian Randomization. Genet. Epidemiol. *47*, 314–331. https://doi.org/10.1002/gepi.22522.
10. Burgess, S., Bowden, J., Dudbridge, F., and Thompson, S.G. (2018). Robust Instrumental Variable Methods Using Multiple Candidate Instruments with Application to Mendelian Randomization. Aug *30*, 2018. https://doi.org/10.48550/arXiv.1606.03729.
11. Bowden, J., Davey Smith, G., Haycock, P.C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet. Epidemiol. *40*, 304–314. https://doi.org/10.1002/gepi.21965.
12. Sulc, J., Sjaarda, J., and Kutalik, Z. (2022). Polynomial Mendelian Randomization Reveals Non-Linear Causal Effects for Obesity-Related Traits. HGG Adv. *3*, 100124. https://doi.org/10.1016/j.xhgg.2022.100124.
13. Burgess, S., Davies, N.M., Thompson, S.G.; and EPIC-InterAct Consortium (2014). Instrumental Variable Analysis with a Nonlinear Exposure-Outcome Relationship. Epidemiology *25*, 877–885. https://doi.org/10.1097/EDE.0000000000000161.
14. Tian, H., Mason, A.M., Liu, C., and Burgess, S. (2023). Relaxing Parametric Assumptions for Non-Linear Mendelian Randomization Using a Doubly-Ranked Stratification Method. PLoS Genet. *19*, e1010823. https://doi.org/10.1371/journal.pgen.1010823.
15. He, R., Liu, M., Lin, Z., Zhuang, Z., Shen, X., and Pan, W. (2024). DeLIVR: A Deep Learning Approach to IV Regression for Testing Nonlinear Causal Effects in Transcriptome-Wide Association Studies. Biostatistics *25*, 468–485. https://doi.org/10.1093/biostatistics/kxac051.
16. Wade, K.H., Hamilton, F.W., Carslake, D., Sattar, N., Davey Smith, G., and Timpson, N.J. (2023). Challenges in Undertaking Nonlinear Mendelian Randomization. Obesity *31*, 2887–2890. https://doi.org/10.1002/oby.23927.
17. Small, D.S. (2014). Commentary: Interpretation and Sensitivity Analysis for the Localized Average Causal Effect Curve. Epidemiology *25*, 886–888. https://doi.org/10.1097/EDE.0000000000000187.
18. Burgess, S. (2023). Violation of the Constant Genetic Effect Assumption Can Result in Biased Estimates for Non-Linear Mendelian Randomization. Hum. Hered. *88*, 79–90. https://doi.org/10.1159/000531659.
19. Newey, W.K., and Powell, J.L. (2003). Instrumental Variable Estimation of Nonparametric Models. Econometrica *71*, 1565–1578.
20. Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. (2011). Nonparametric instrumental regression. Econometrica *79*, 1541–1565.
21. Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A Flexible Approach for Counterfactual Prediction. In . Proceedings of the 34th International Conference on Machine Learning, *70*, D. Precup and Y.W. Teh, eds. *(PMLR), pp. 1414–1423*.
22. Saag, K.G., Petersen, J., Brandi, M.L., Karaplis, A.C., Lorentzon, M., Thomas, T., Maddox, J., Fan, M., Meisner, P.D., and Grauer, A. (2017). Romosozumab or Alendronate for Fracture Prevention in Women with Osteoporosis. N. Engl. J. Med. *377*, 1417–1427. https://doi.org/10.1056/NEJMoa1708322.
23. Lewiecki, E.M., Blicharski, T., Goemaere, S., Lippuner, K., Meisner, P.D., Miller, P.D., Miyauchi, A., Maddox, J., Chen, L., and Horlait, S. (2018). A Phase III Randomized Placebo-Controlled Trial to Evaluate Efficacy and Safety of Romosozumab in Men With Osteoporosis. J. Clin. Endocrinol. Metab. *103*, 3183–3193. https://doi.org/10.1210/jc.2017-02163.
24. Bovijn, J., Krebs, K., Chen, C.-Y., Boxall, R., Censin, J.C., Ferreira, T., Pulit, S.L., Glastonbury, C.A., Laber, S., Millwood, I.Y., et al. (2020). Evaluating the Cardiovascular Safety of Sclerostin Inhibition Using Evidence from Meta-Analysis of Clinical Trials and Human Genetics. Sci. Transl. Med. *12*, eaay6570. https://doi.org/10.1126/scitranslmed.aay6570.
25. Tobias, J.H. (2023). Sclerostin and Cardiovascular Disease. Curr. Osteoporos. Rep. *21*, 519–526. https://doi.org/10.1007/s11914-023-00810-w.
26. Emdin, C.A., Khera, A.V., Natarajan, P., Klarin, D., Zekavat, S.M., Hsiao, A.J., and Kathiresan, S. (2017). Genetic Association of Waist-to-Hip Ratio With Cardiometabolic Traits, Type 2 Diabetes, and Coronary Heart Disease. JAMA *317*, 626–634. https://doi.org/10.1001/jama.2016.21042.
27. Baigent, C., Blackwell, L., Emberson, J., Holland, L.E., Reith, C., Bhala, N., Peto, R., Barnes, E.H., Keech, A., et al.; Cholesterol Treatment Trialists' CTT Collaboration (2010). Efficacy and Safety of More Intensive Lowering of LDL Cholesterol: A Meta-Analysis of Data from 170 000 Participants in 26 Randomised Trials. Lancet *376*, 1670–1681. https://doi.org/10.1016/S0140-6736(10)61350-5.
28. de Koning, L., Merchant, A.T., Pogue, J., and Anand, S.S. (2007). Waist Circumference and Waist-to-Hip Ratio as Predictors of Cardiovascular Events: Meta-Regression Analysis of Prospective Studies. Eur. Heart J. *28*, 850–856. https://doi.org/10.1093/eurheartj/ehm026.
29. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank Resource with Deep Phenotyping and Genomic Data. Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.
30. Burgess, S., Davies, N.M., and Thompson, S.G. (2016). Bias Due to Participant Overlap in Two-Sample Mendelian Randomization. Genet. Epidemiol. *40*, 597–608. https://doi.org/10.1002/gepi.21998.
31. Nikpay, M., Goel, A., Won, H.-H., Hall, L.M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C.P.,

Hopewell, J.C., et al. (2015). A Comprehensive 1000 Genomes–Based Genome-Wide Association Meta-Analysis of Coronary Artery Disease. Nat. Genet. *47*, 1121–1130. https://doi.org/10.1038/ng.3396.

32. Nelson, C.P., Goel, A., Butterworth, A.S., Kanoni, S., Webb, T. R., Marouli, E., Zeng, L., Ntalla, I., Lai, F.Y., Hopewell, J.C., et al. (2017). Association Analyses Based on False Discovery Rate Implicate New Loci for Coronary Artery Disease. Nat. Genet. *49*, 1385–1391. https://doi.org/10.1038/ng.3913.
33. GTEx Consortium (2020). The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.
34. Pulit, S.L., Stoneman, C., Morris, A.P., Wood, A.R., Glastonbury, C.A., Tyrrell, J., Yengo, L., Ferreira, T., Marouli, E., Ji, Y., et al. (2019). Meta-Analysis of Genome-Wide Association Studies for Body Fat Distribution in 694 649 Individuals of European Ancestry. Hum. Mol. Genet. *28*, 166–174. https://doi.org/10.1093/hmg/ddy327.
35. Graham, S.E., Clarke, S.L., Wu, K.-H.H., Kanoni, S., Zajac, G.J. M., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T.W., et al. (2021). The Power of Genetic Diversity in Genome-Wide Association Studies of Lipids. Nature *600*, 675–679. https://doi.org/10.1038/s41586-021-04064-3.
36. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. J. R. Stat. Soc. Series B Stat. Methodol. *82*, 1273–1300. https://doi.org/10.1111/rssb.12388.
37. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. *10*, e1004383. https://doi.org/10.1371/journal.pgen.1004383.
38. Wallace, C. (2021). A More Accurate Method for Colocalisation Analysis Allowing for Multiple Causal Variants. PLoS Genet. *17*, e1009440. https://doi.org/10.1371/journal.pgen.1009440.
39. Didelez, V., and Sheehan, N. (2007). Mendelian Randomization as an Instrumental Variable Approach to Causal Inference. Stat. Methods Med. Res. *16*, 309–330. https://doi.org/10.1177/0962280206077743.
40. Qi, G., and Chatterjee, N. (2019). Mendelian Randomization Analysis Using Mixture Models for Robust and Efficient Estimation of Causal Effects. Nat. Commun. *10*, 1941. https://doi.org/10.1038/s41467-019-09432-2.
41. Schmidt, A.F., Finan, C., Gordillo-Marañón, M., Asselbergs, F. W., Freitag, D.F., Patel, R.S., Tyl, B., Chopade, S., Faraway, R., Zwierzyna, M., and Hingorani, A.D. (2020). Genetic Drug Target Validation Using Mendelian Randomisation. Nat. Commun. *11*, 3255. https://doi.org/10.1038/s41467-020-16969-0.
42. Burgess, S., and Thompson, S.G. (2021). Mendelian Randomization: Methods for Causal Inference Using Genetic Variants (CRC Press).
43. Batool, F., Patel, A., Gill, D., and Burgess, S. (2022). Disentangling the Effects of Traits with Shared Clustered Genetic Predictors Using Multivariable Mendelian Randomization. Genet. Epidemiol. *46*, 415–429. https://doi.org/10.1002/gepi.22462.
44. Patel, A., Gill, D., Shungin, D., Mantzoros, C.S., Knudsen, L. B., Bowden, J., and Burgess, S. (2024). Robust Use of Phenotypic Heterogeneity at Drug Target Genes for Mechanistic Insights: Application of Cis-Multivariable Mendelian Randomization to GLP1R Gene Region. Genet. Epidemiol. *48*, 151–163. https://doi.org/10.1002/gepi.22551.
45. Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. (2017). A Framework for the Investigation of Pleiotropy in Two-Sample Summary Data Mendelian Randomization. Stat. Med. *36*, 1783–1802.
46. Rees, J.M.B., Wood, A.M., Dudbridge, F., and Burgess, S. (2019). Robust Methods in Mendelian Randomization via Penalization of Heterogeneous Causal Estimates. PLoS One *14*, e0222362. https://doi.org/10.1371/journal.pone.0222362.
47. Burgess, S., Foley, C.N., Allara, E., Staley, J.R., and Howson, J. M.M. (2020). A Robust and Efficient Method for Mendelian Randomization with Hundreds of Genetic Variants. Nat. Commun. *11*, 376. https://doi.org/10.1038/s41467-019-14156-4.
48. Angrist, J.D., and Pischke, J.-S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion (Princeton: Princeton University Press). https://doi.org/10.1515/9781400829828.
49. Kingma, D.P., and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In 3rd International Conference for Learning Representations. https://doi.org/10.48550/arXiv.1412.6980.
50. Moon, S.J., Jeon, J.-J., Lee, J.S.H., and Kim, Y. (2021). Learning Multiple Quantiles With Neural Networks. J. Comput. Graph Stat. *30*, 1238–1248. https://doi.org/10.1080/10618600.2021.1909601.
51. Davison, A.C., and Hinkley, D.V. (1997). Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge: Cambridge University Press). https://doi.org/10.1017/CBO9780511802843.
52. Hughes, R.A., Davies, N.M., Davey Smith, G., and Tilling, K. (2019). Selection Bias When Estimating Average Treatment Effects Using One-sample Instrumental Variable Analysis. Epidemiology *30*, 350–357. https://doi.org/10.1097/EDE.0000000000000972.
53. Gkatzionis, A., and Burgess, S. (2019). Contextualizing Selection Bias in Mendelian Randomization: How Bad Is It Likely to Be? Int. J. Epidemiol. *48*, 691–701. https://doi.org/10.1093/ije/dyy202.
54. Schoeler, T., Speed, D., Porcu, E., Pirastu, N., Pingault, J.-B., and Kutalik, Z. (2023). Participation Bias in the UK Biobank Distorts Genetic Associations and Downstream Analyses. Nat. Hum. Behav. *7*, 1216–1227. https://doi.org/10.1038/s41562-023-01579-9.
55. Palmer, T.M., Thompson, J.R., Tobin, M.D., Sheehan, N.A., and Burton, P.R. (2008). Adjusting for Bias and Unmeasured Confounding in Mendelian Randomization Studies with Binary Responses. Int. J. Epidemiol. *37*, 1161–1168. https://doi.org/10.1093/ije/dyn080.
56. Holdsworth, G., Staley, J.R., Hall, P., van Koeverden, I., Vangjeli, C., Okoye, R., Boyce, R.W., Turk, J.R., Armstrong, M., Wolfreys, A., and Pasterkamp, G. (2021). Sclerostin Downregulation Globally by Naturally Occurring Genetic Variants, or Locally in Atherosclerotic Plaques, Does Not Associate With Cardiovascular Events in Humans. J. Bone Miner. Res. *36*, 1326–1339. https://doi.org/10.1002/jbmr.4287.
57. Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E.M., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., et al. (2021). Open Targets

Genetics: Systematic Identification of Trait-Associated Genes Using Large-Scale Genetics and Functional Genomics. Nucleic Acids Res. *49*, D1311–D1320. https://doi.org/10.1093/nar/gkaa840.

58. Holmes, M.V., Asselbergs, F.W., Palmer, T.M., Drenos, F., Lanktree, M.B., Nelson, C.P., Dale, C.E., Padmanabhan, S., Finan, C., Swerdlow, D.I., et al. (2015). Mendelian Randomization of Blood Lipids for Coronary Heart Disease. Eur. Heart J. *36*, 539–550.
59. Zheng, J., Wheeler, E., Pietzner, M., Andlauer, T.F.M., Yau, M. S., Hartley, A.E., Brumpton, B.M., Rasheed, H., Kemp, J.P., Frysz, M., et al. (2023). Lowering of Circulating Sclerostin May Increase Risk of Atherosclerosis and Its Risk Factors: Evidence From a Genome-Wide Association Meta-Analysis Followed by Mendelian Randomization. Arthritis Rheumatol. *75*, 1781–1792. https://doi.org/10.1002/art.42538.
60. Ference, B.A. (2018). How to Use Mendelian Randomization to Anticipate the Results of Randomized Trials. Eur. Heart J. *39*, 360–362. https://doi.org/10.1093/eurheartj/ehx462.
61. Hamilton, F.W., Hughes, D.A., Spiller, W., Tilling, K., and Davey Smith, G. (2024). Non-Linear Mendelian Randomization: Detection of Biases Using Negative Controls with a Focus on BMI, Vitamin D and LDL Cholesterol. Eur. J. Epidemiol. *39*, 451–465. https://doi.org/10.1007/s10654-024-01113-9.

The American Journal of Human Genetics, Volume *112*

Supplemental information

A flexible machine learning Mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition

Marc-André Legault, Jason Hartford, Benoît J. Arsenault, Archer Y. Yang, and Joelle Pineau

A Supplemental Figures

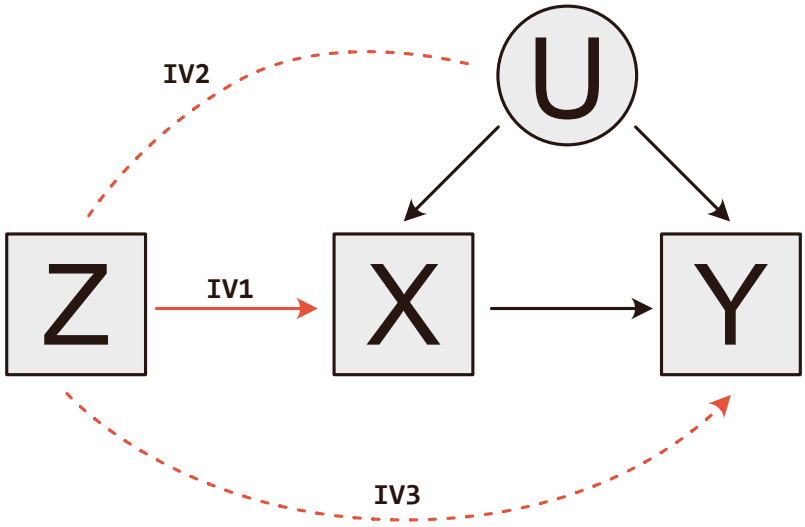


Figure S1: Graph illustrating the instrumental variable assumptions. Dashed lines represent effects that are assumed absent. Solid lines denote effects that are assumed present. Squares denote observed variables and circles denote unobserved variables. When there are no arrow heads, the relationship is assumed to be bi-directional.

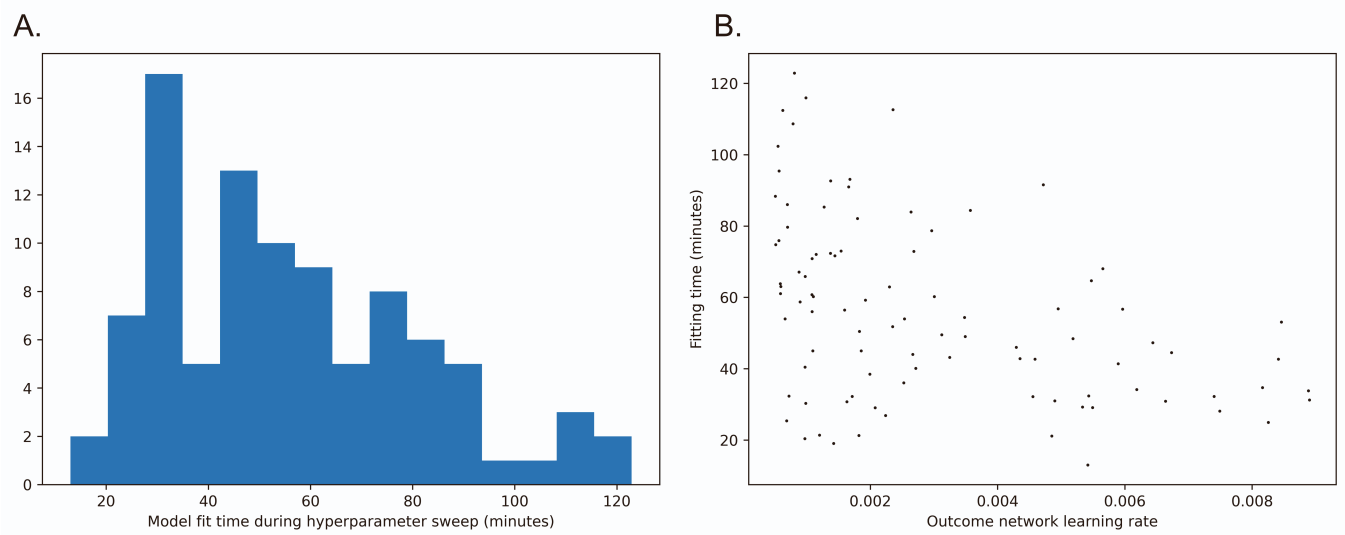


Figure S2: Runtime of Quantile IV in a real data application in the UK Biobank. The runtimes are for a 100 fit hyperparameter sweep for the effect of LDL-c on MI. The histogram of runtimes in minutes is presented (A.) and the relationship between the selected outcome neural network learning rate and runtime is shown in (B.). Smaller learning rates lead to considerable longer fitting times. The CPUs used to fit the model were AMD Rome 7532 or AMD Rome 7502.

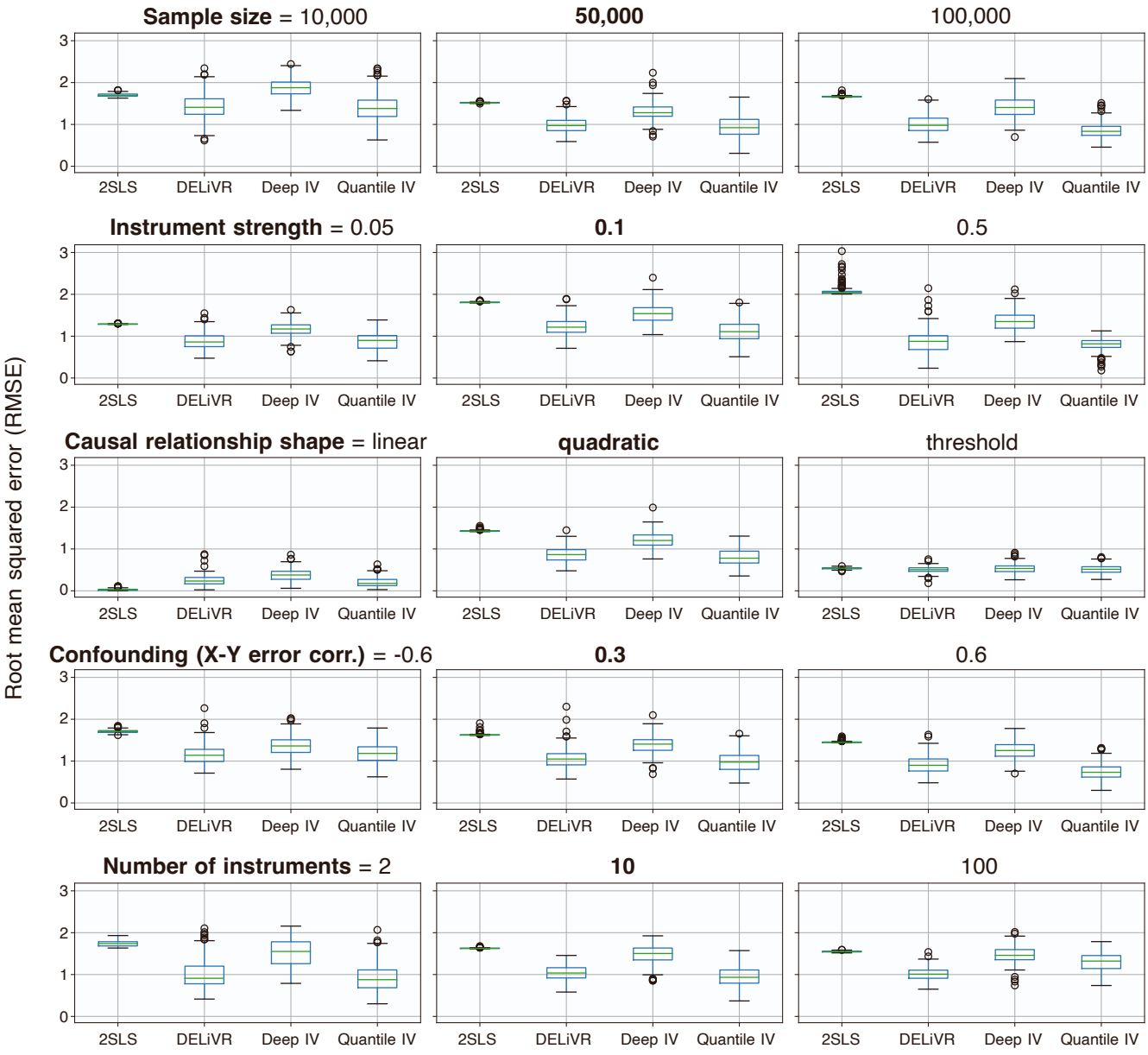


Figure S3: Root mean squared error between the estimated IV regression and the true causal function over a grid spanning the full range of the exposure. The boxplot for every estimator represents variability over 200 simulation replicates. The simulation parameter values in bold correspond to the reference values. 2SLS: Two-stage least squares, DeLIVR [14], DeepIV [20], Quantile IV: proposed method.

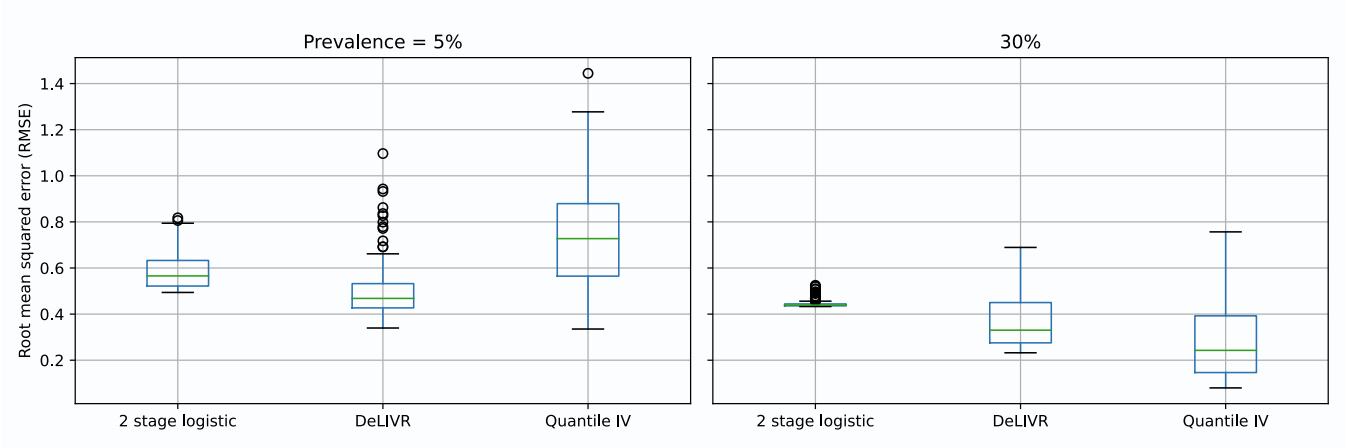


Figure S4: Root mean squared error between the estimated IV regression and the true causal log odds over a grid spanning 95% of the range of the exposure. The boxplot for every estimator represents variability over 200 simulation replicates. 2 stage logistic: Two-stage residual inclusion (control function) estimator [55], DeLIVR [14], Quantile IV: proposed method.

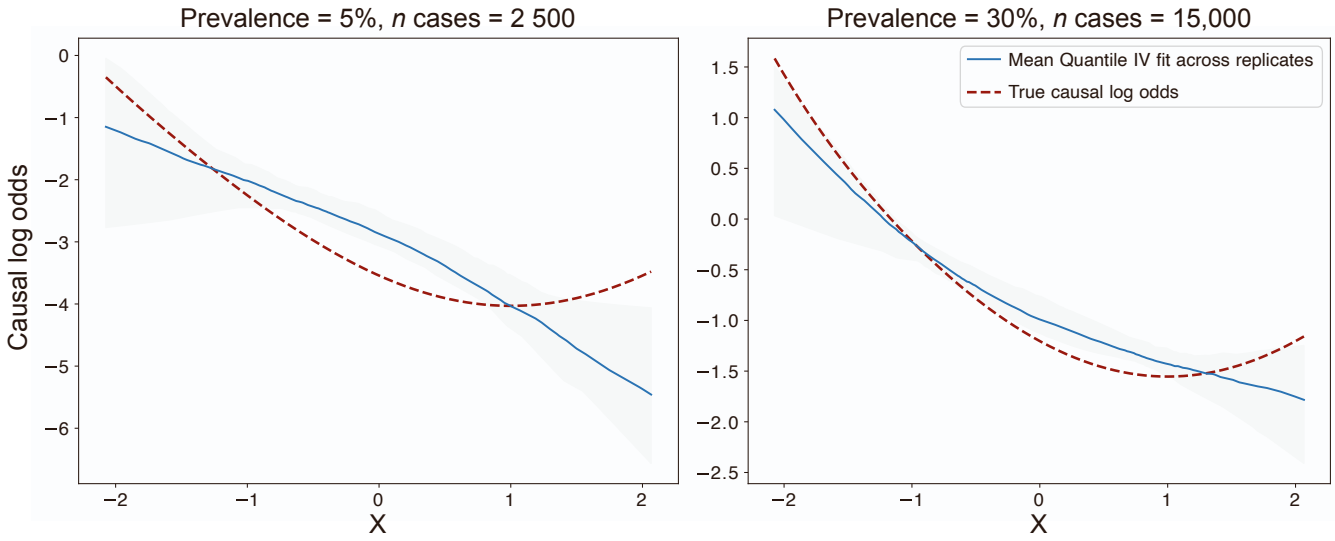


Figure S5: Mean estimated IV regression function targeting the causal log odds for simulated binary outcomes. The mean estimate over 200 simulation replicates is shown along with the theoretical function according to the simulation model. The shaded area represents the 5^{th} and 95^{th} percentiles of the estimated regression function over simulation replicates.

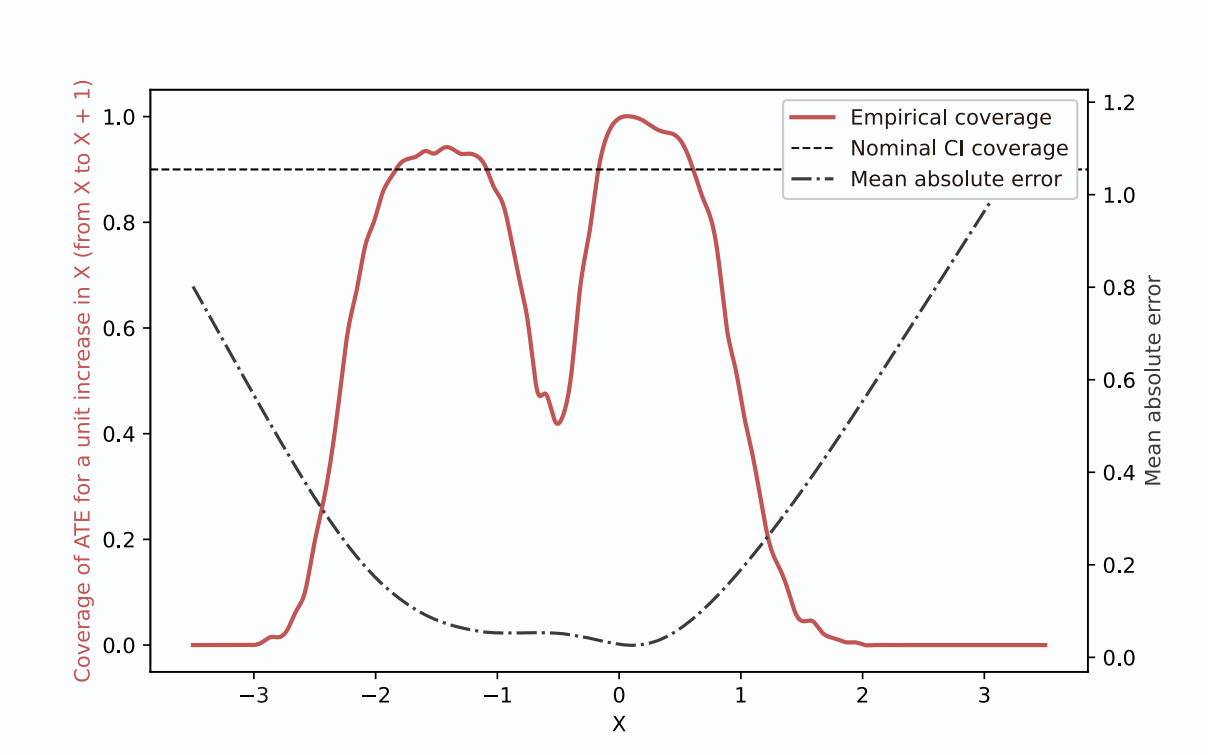


Figure S6: Coverage and mean absolute estimation error of the ATE for a unit increment in the exposure. The coverage is estimated using 200 simulation replicates and 50 bootstrap iterations are used to derive the confidence intervals. The simulation scenario used for this analysis uses the baseline parameter values for every simulation parameter.

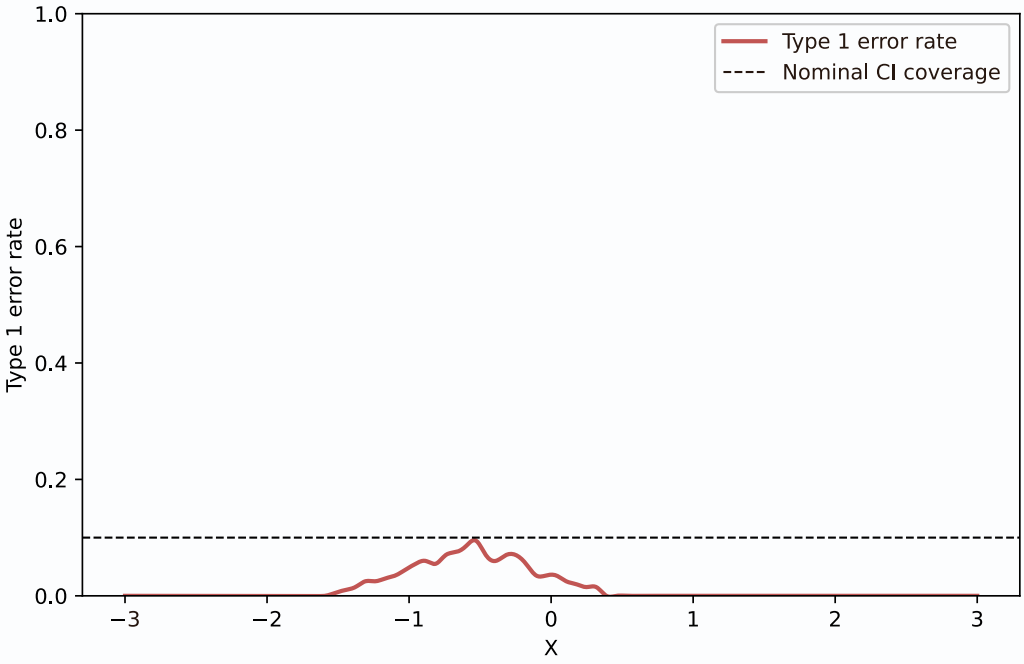


Figure S7: Type 1 error rate when estimating the ATE for a unit increment in the exposure across different values of the exposure. The type 1 error rate is estimated using 200 simulation replicates and 50 bootstrap iterations are used to derive the confidence intervals. The simulation scenario used for this analysis uses the baseline parameter values for every simulation parameter except for the structural relationship where $f(X) = 0$.

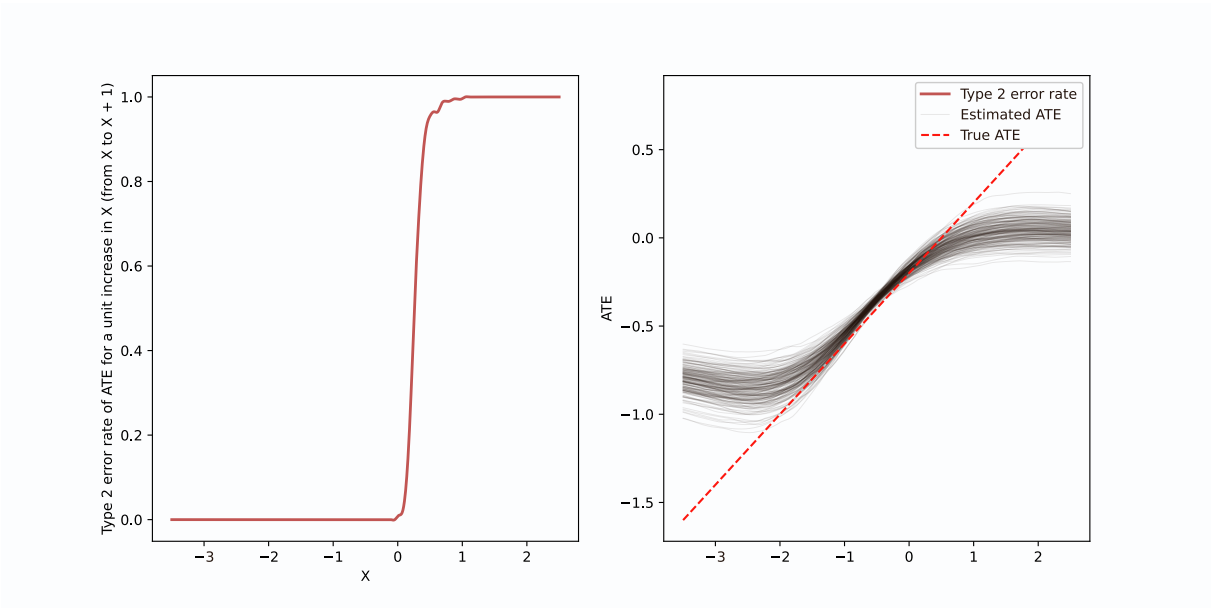


Figure S8: Type 2 error rate and estimated ATEs for a unit increment in the exposure across different values of the exposure. The type 2 error rate is estimated using 200 simulation replicates and 50 bootstrap iterations are used to derive the confidence intervals. The simulation scenario used for this analysis uses the baseline parameter values for every simulation parameter. The plot on the left shows the type 2 error rate, and the plot on the right shows the true and estimated ATEs across the simulation replicates. Note that for some included values of the exposure, the true ATE is null (near 0).

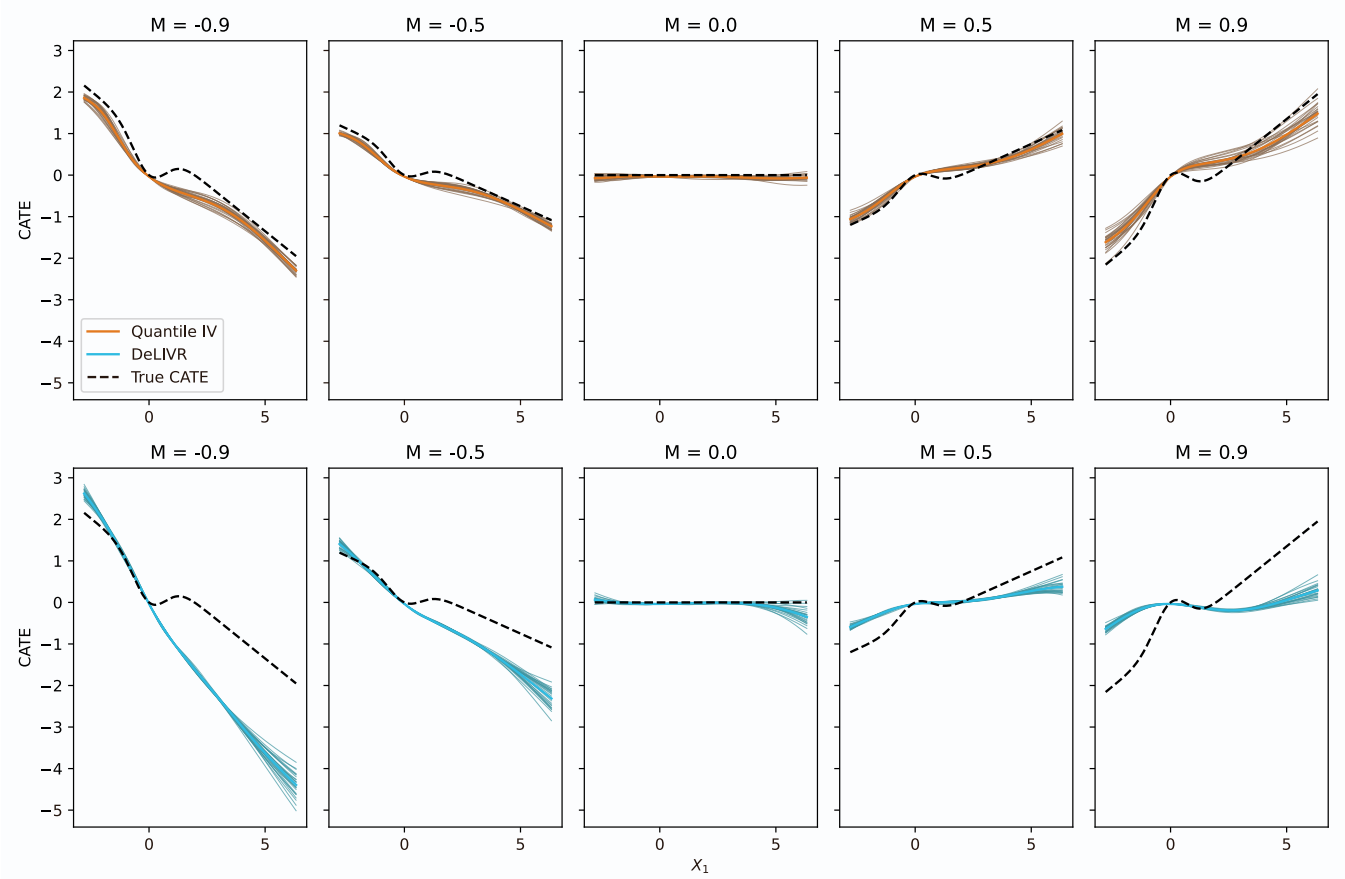


Figure S9: Estimated CATE by Quantile IV (top) and DeLIVR (bottom) at different values of the effect modifier variable (M) and for varying interventions about the mean of the exposure. Fits for 25 bootstrap iterations are shown for both estimators, and the colored line represents the mean value (bagging estimate). The true CATE is shown in the dashed black line.

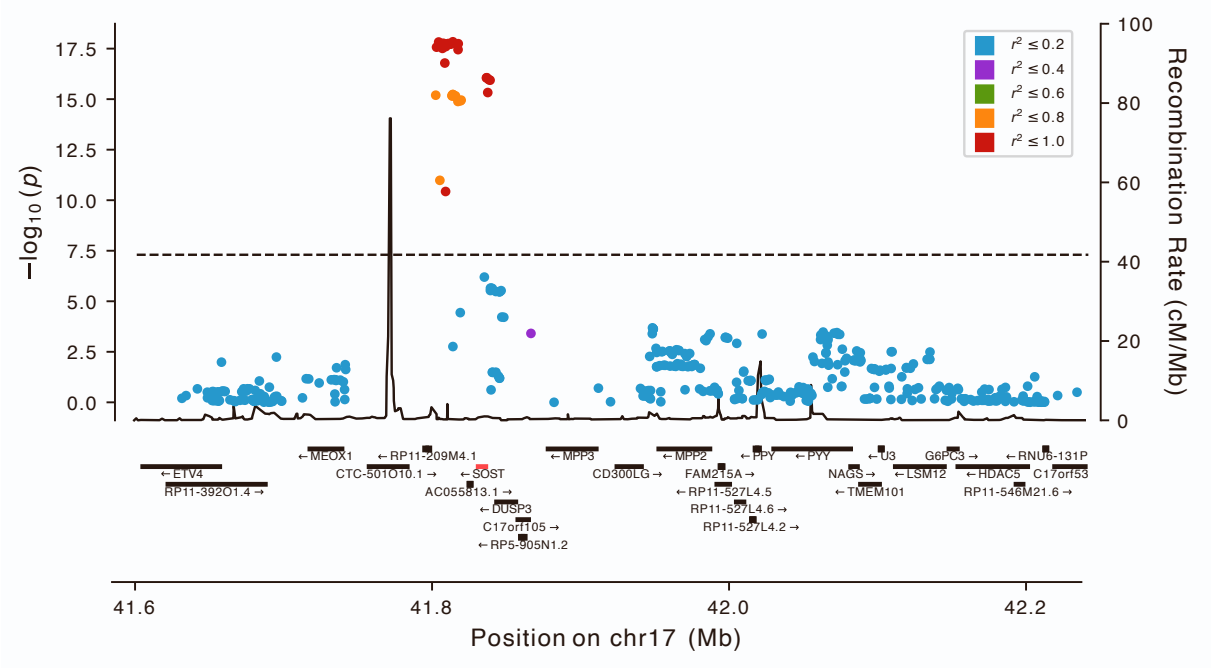


Figure S10: Genetic association between variants at the *SOST* locus and circulating sclerostin protein levels as measured using the Olink platform in the UK Biobank. Genes in the region are shown below the locus plot and the *SOST* gene is highlighted in red.

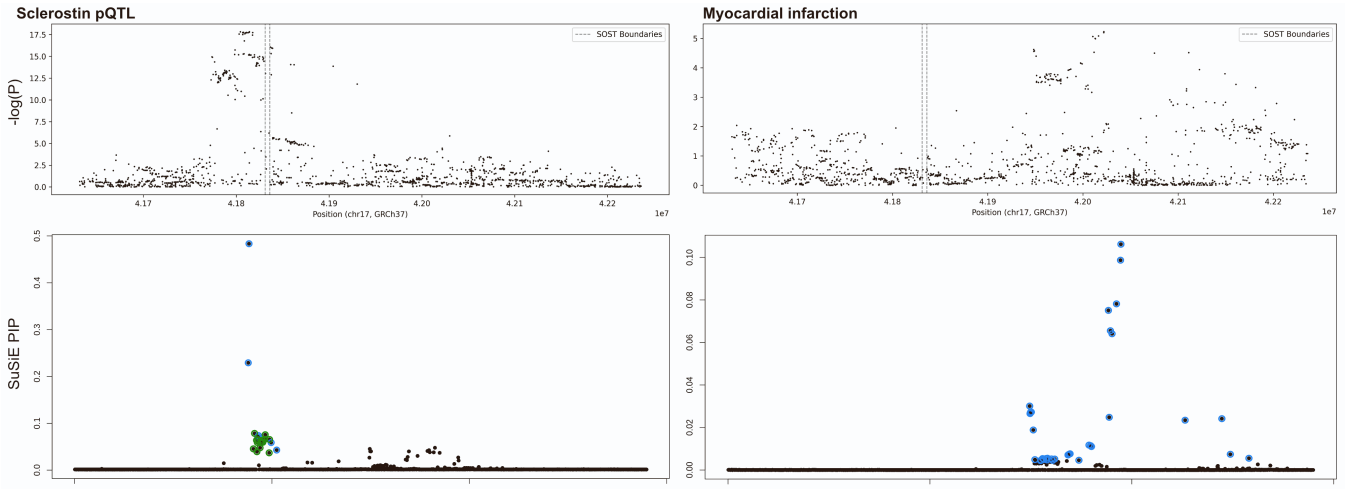


Figure S11: Genetic associations at the SOST locus with circulating sclerostin levels (left) and myocardial infarction (right) and variant posterior inclusion probabilities in finemapping credible sets as inferred by SuSiE. The boundaries of the *SOST* gene are denoted by dashed lines.

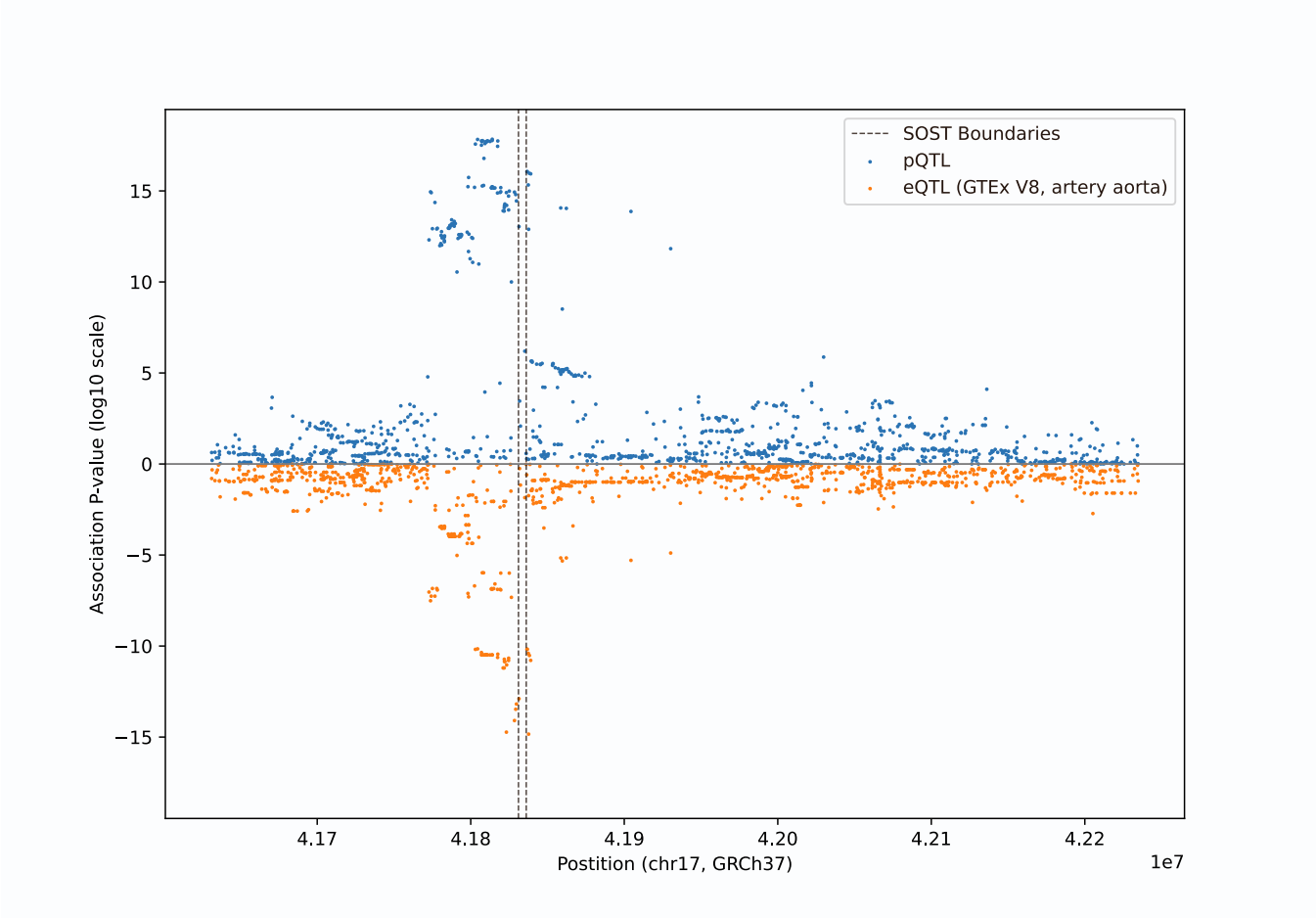


Figure S12: Association p values (log scale) between genetic variants at the *SOST* locus and *SOST* expression in the aorta in GTEx V8 and circulating sclerostin levels in the UK Biobank.

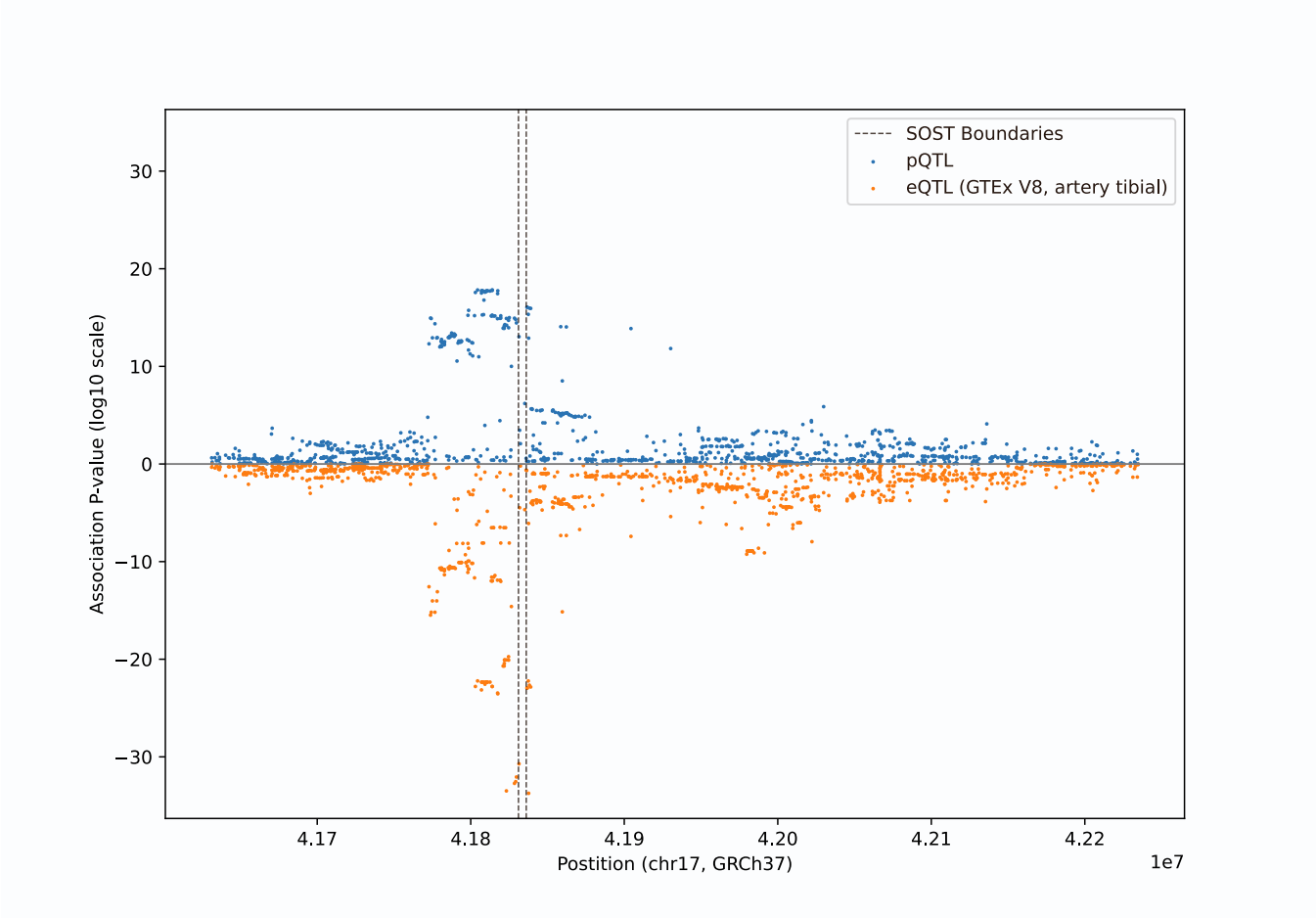


Figure S13: Association p values (log scale) between genetic variants at the *SOST* locus and *SOST* expression in the tibial artery in GTEx V8 and circulating sclerostin levels in the UK Biobank.

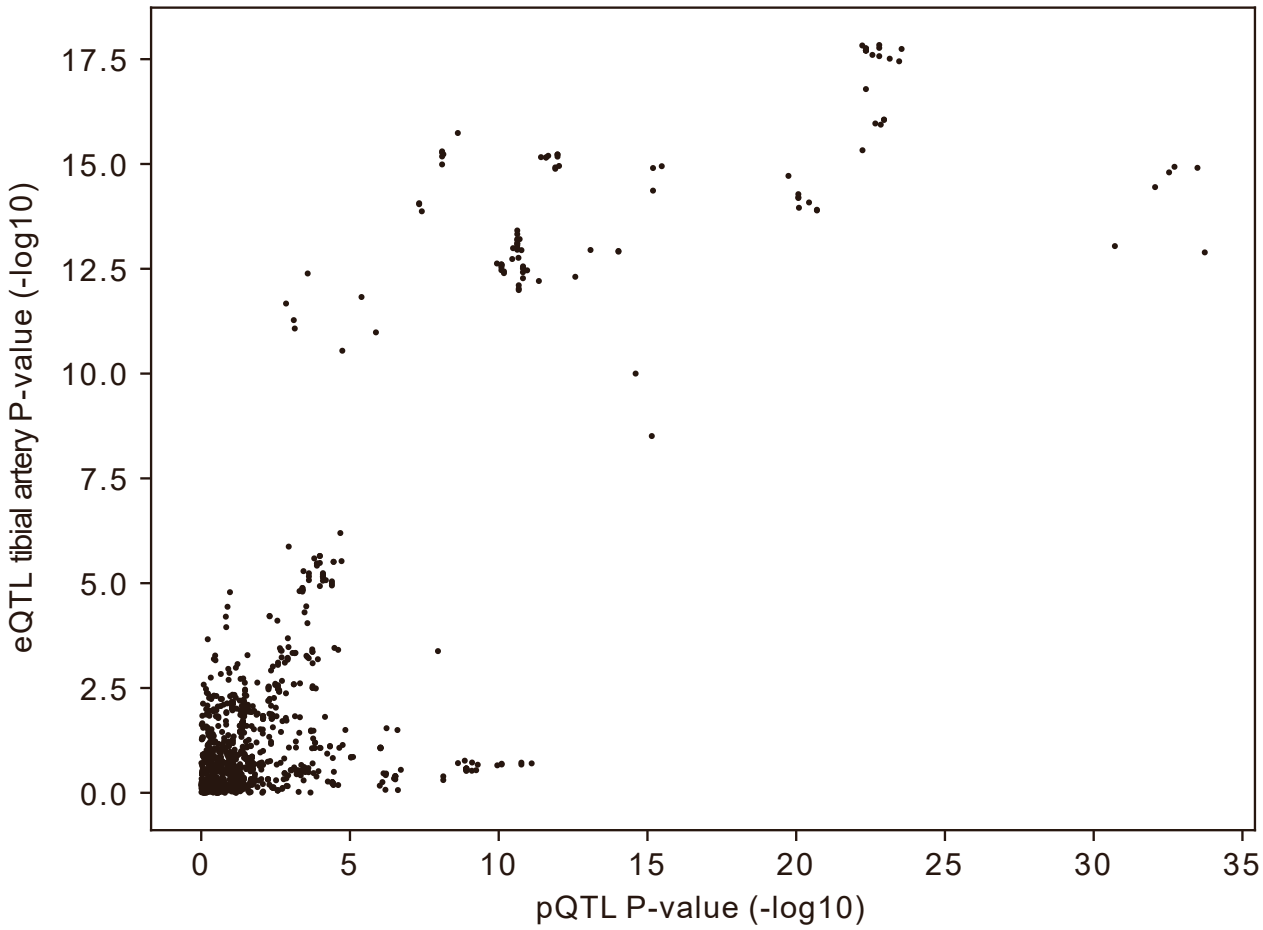


Figure S14: Comparison of sclerostin tibial artery eQTL and pQTL p values (log scale).

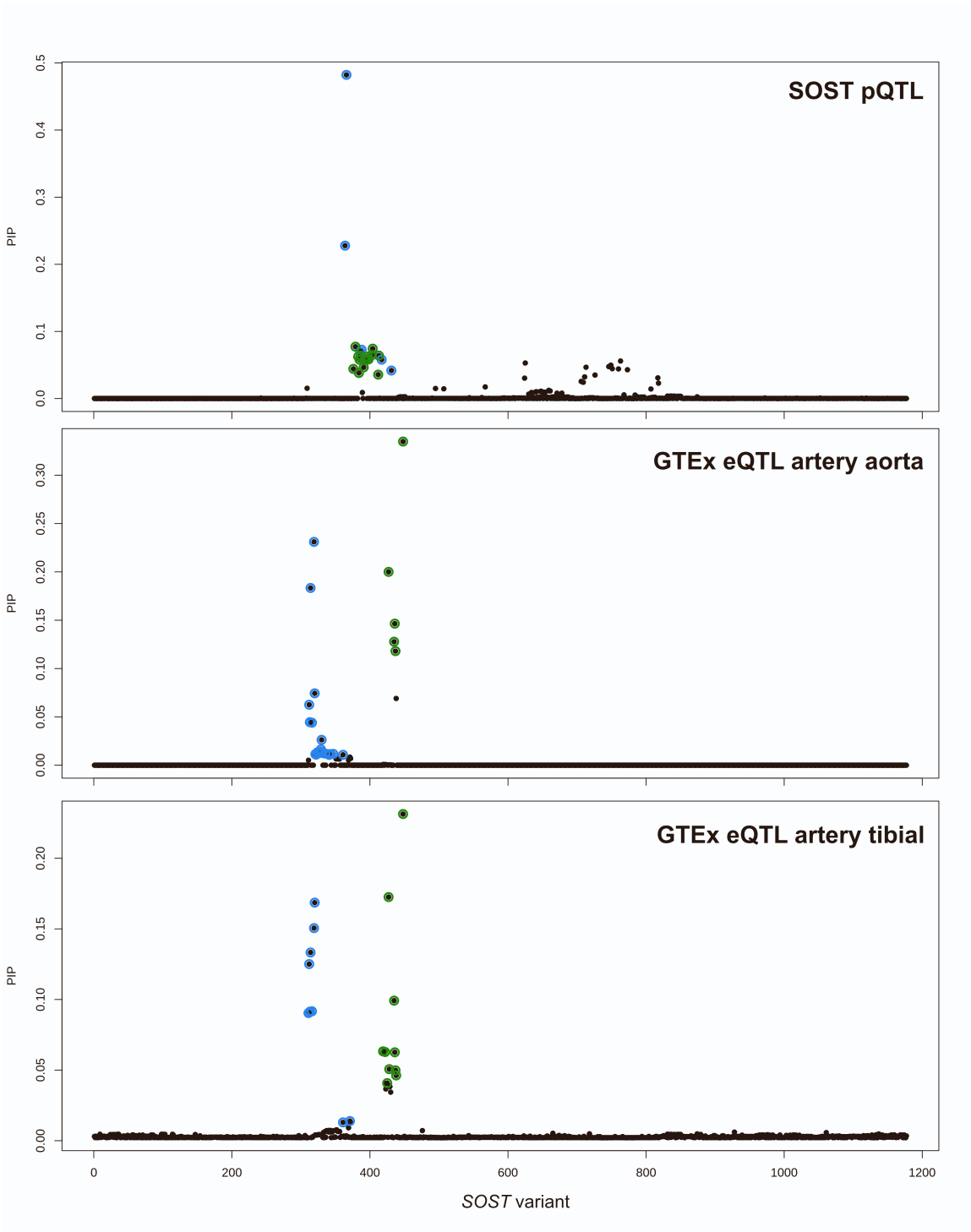


Figure S15: Posterior inclusions probabilities of variants at the *SOST* locus in finemapping credible sets inferred by SuSiE for association with circulating protein levels (UK Biobank) and sclerostin expression (aorta and tibial artery, GTEx).

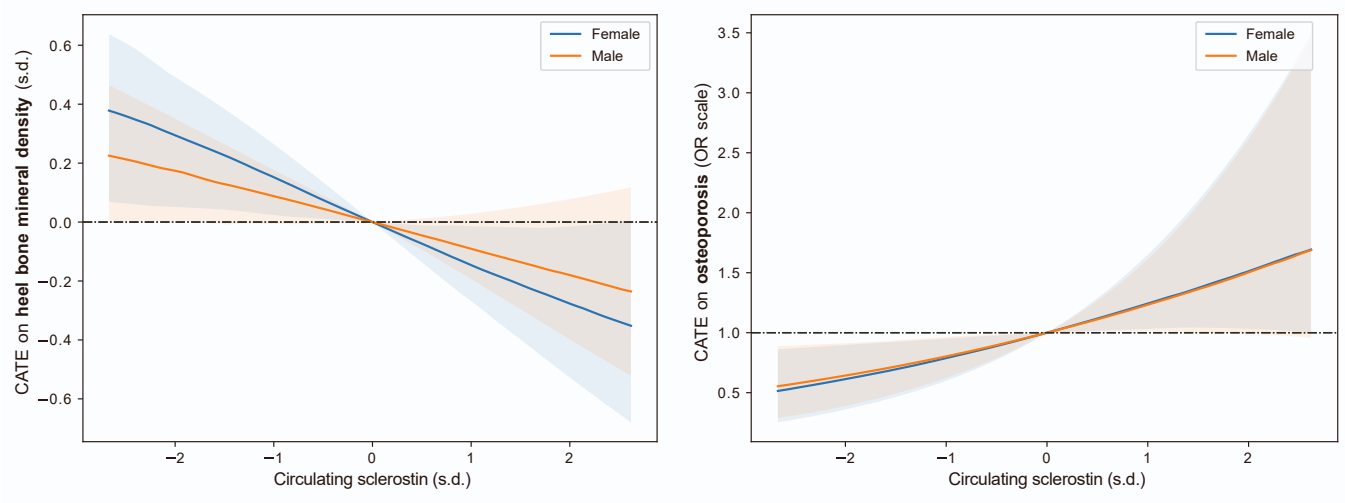


Figure S16: Conditional average treatment effect of circulating sclerostin levels on heel bone mineral density and osteoporosis in males and females estimated using Quantile IV in the UK Biobank. The shaded region corresponds to 90% bootstrap confidence intervals. The plots cover the central 99% of the exposure range.

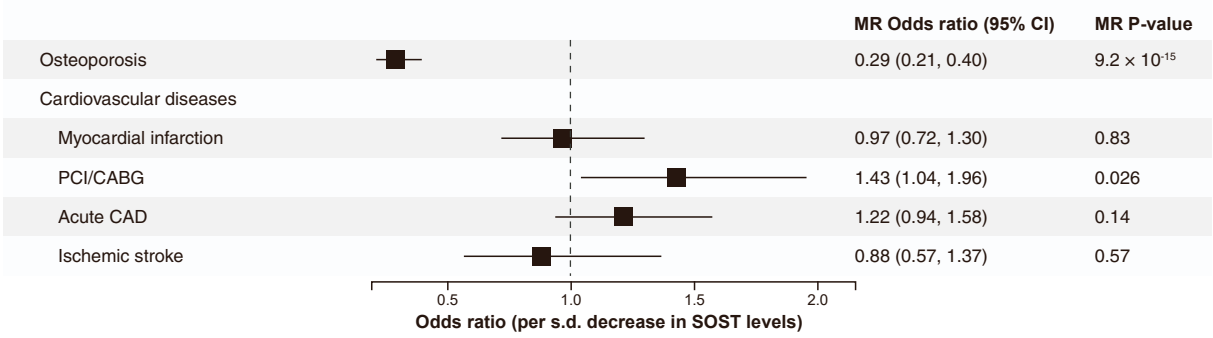


Figure S17: Mendelian randomization of the effect of a 1 s.d. reduction in circulating sclerostin using the Inverse Variance Weighted estimator based on the finemapped sclerostin pQTLs in the UK Biobank

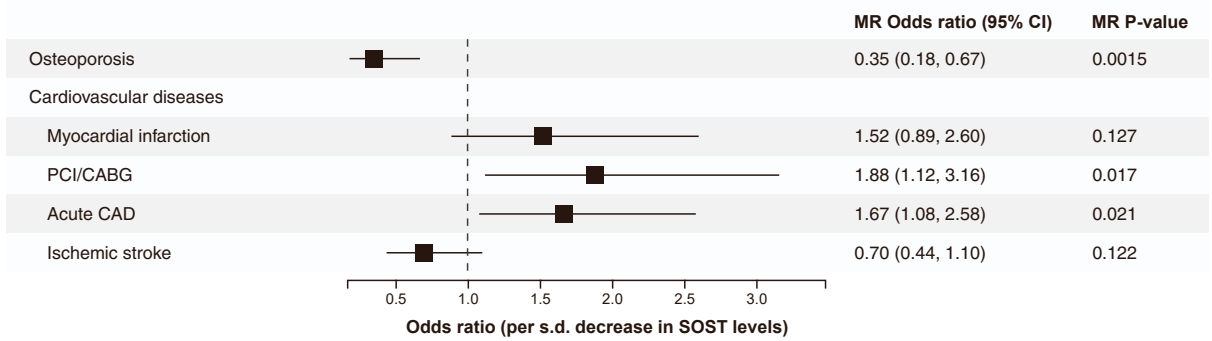


Figure S18: Mendelian randomization of the effect of a 1 s.d. reduction in circulating sclerostin using the PC-GMM estimator based on LD pruned variants at the *SOST* locus.

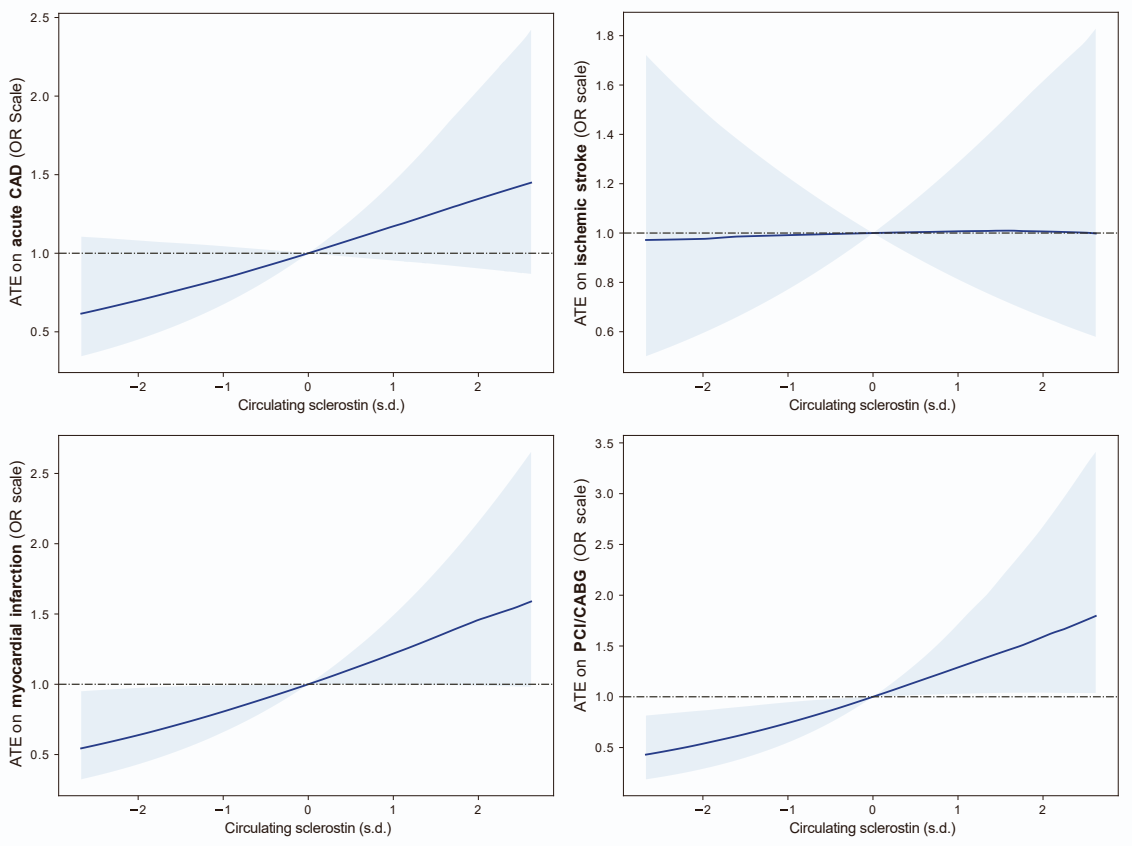


Figure S19: Quantile IV estimate of the average effect of varying the levels of circulating sclerostin about the mean on cardiovascular outcomes in the UK Biobank. The shaded region corresponds to 90% bootstrap confidence intervals. The plots cover the central 99% of the exposure range.

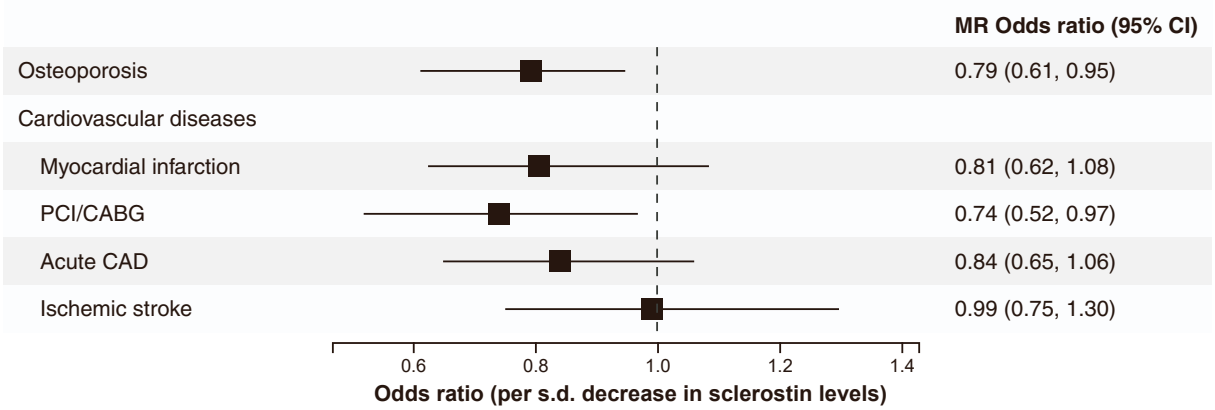


Figure S20: Mendelian randomization of the effect of a 1 s.d. reduction in circulating sclerostin using the Quantile IV estimator in the UK Biobank.

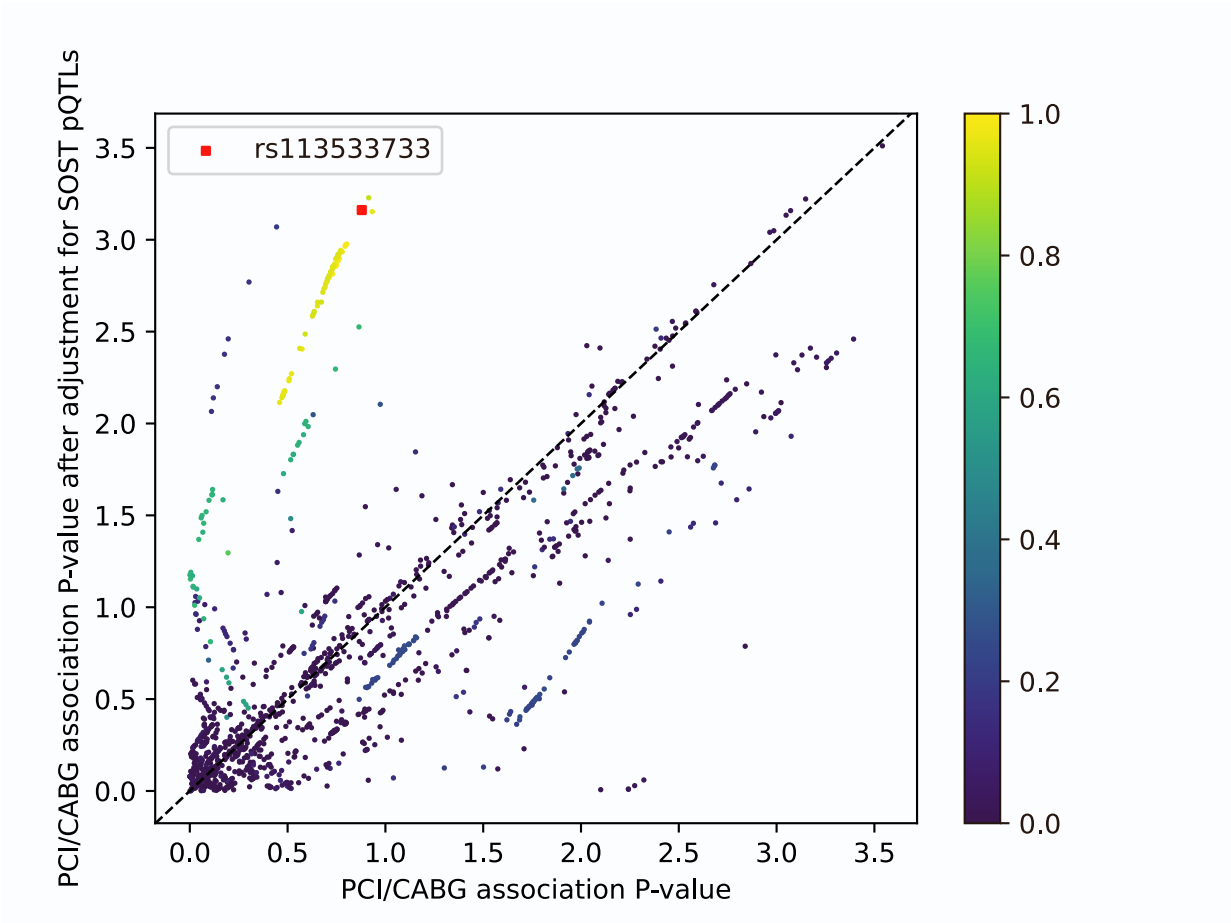


Figure S21: Comparison of the genetic association p values before and after further adjustment for the sclerostin pQTLs. The square on the scatter plot denotes the variant rs113533733 which we identified as having possible direct effects do to its high association p value despite robust adjustment for variants associated with circulating sclerostin levels. The color represents LD (r^2) with this variant.

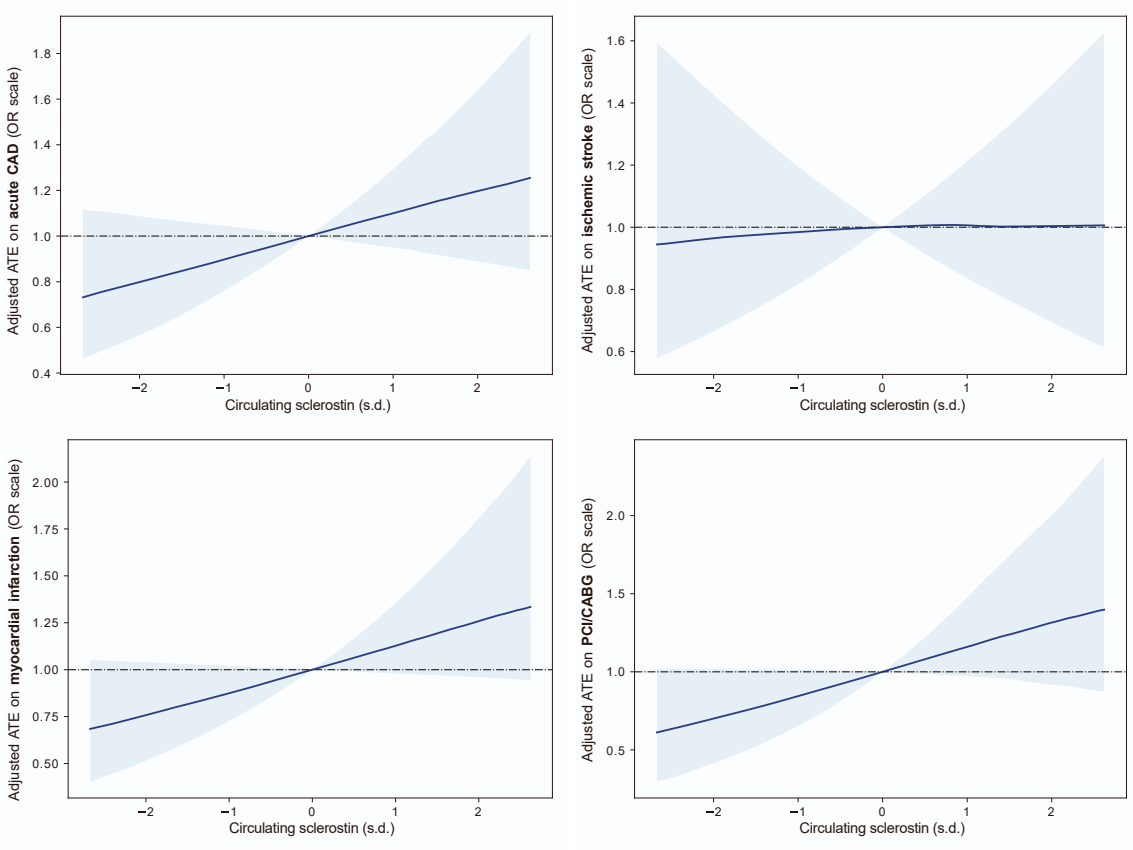


Figure S22: Adjusted Quantile IV estimate of the average effect of varying the levels of circulating sclerostin about the mean on cardiovascular outcomes accounting for direct effects by rs113533733 in the UK Biobank. The shaded region corresponds to 90% bootstrap confidence intervals. The plots cover the central 99% of the exposure range.

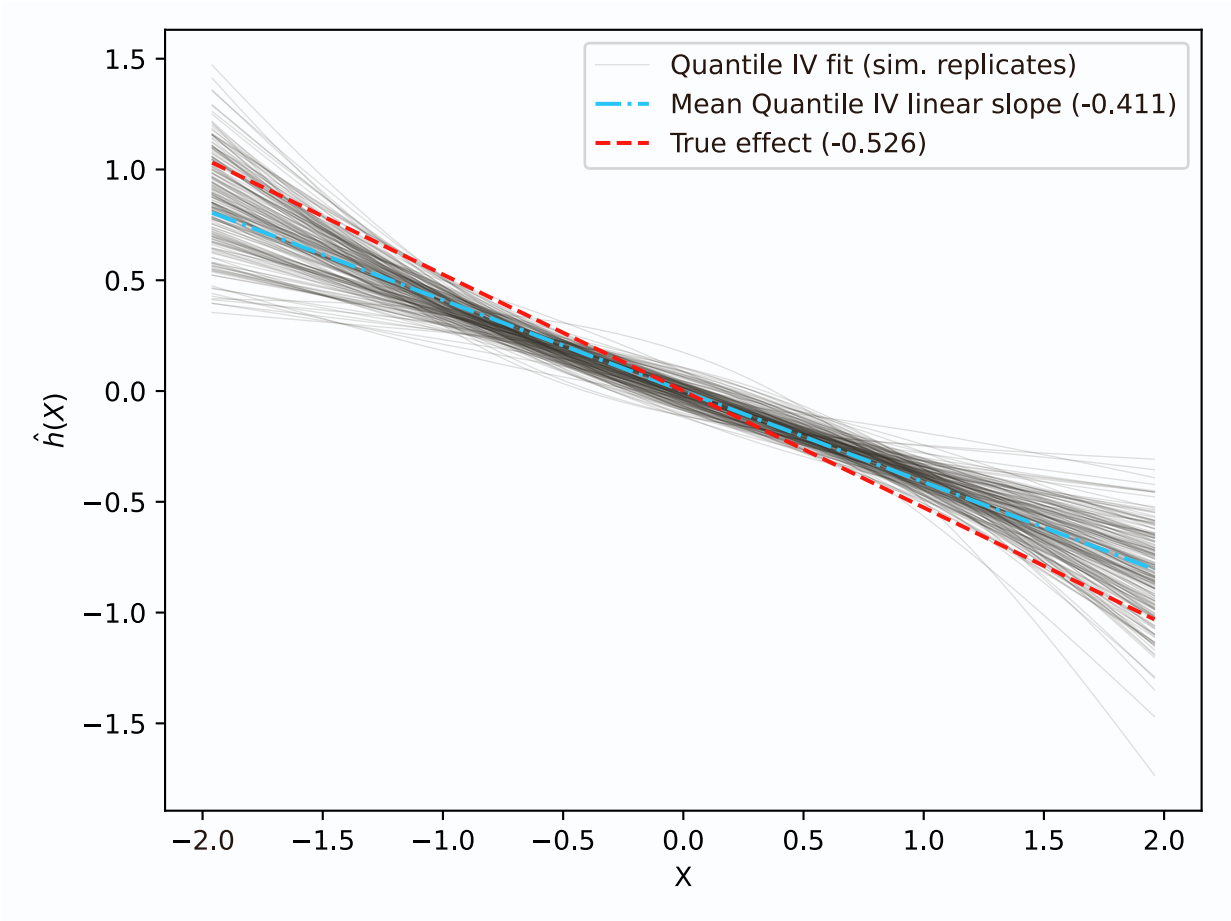


Figure S23: Estimated causal relationship by Quantile IV for a *post hoc* simulation analysis replicating the real data analysis of the effect of sclerostin on heel bone mineral density. The simulated relationship (in red) is linear with a slope of -0.526 as was estimated by PC-GMM in the real data analysis. The pale black lines represent the Quantile IV fits for 200 simulation replicates. The blue line is a mean linearized estimate of the slope taken over the simulation replicates.

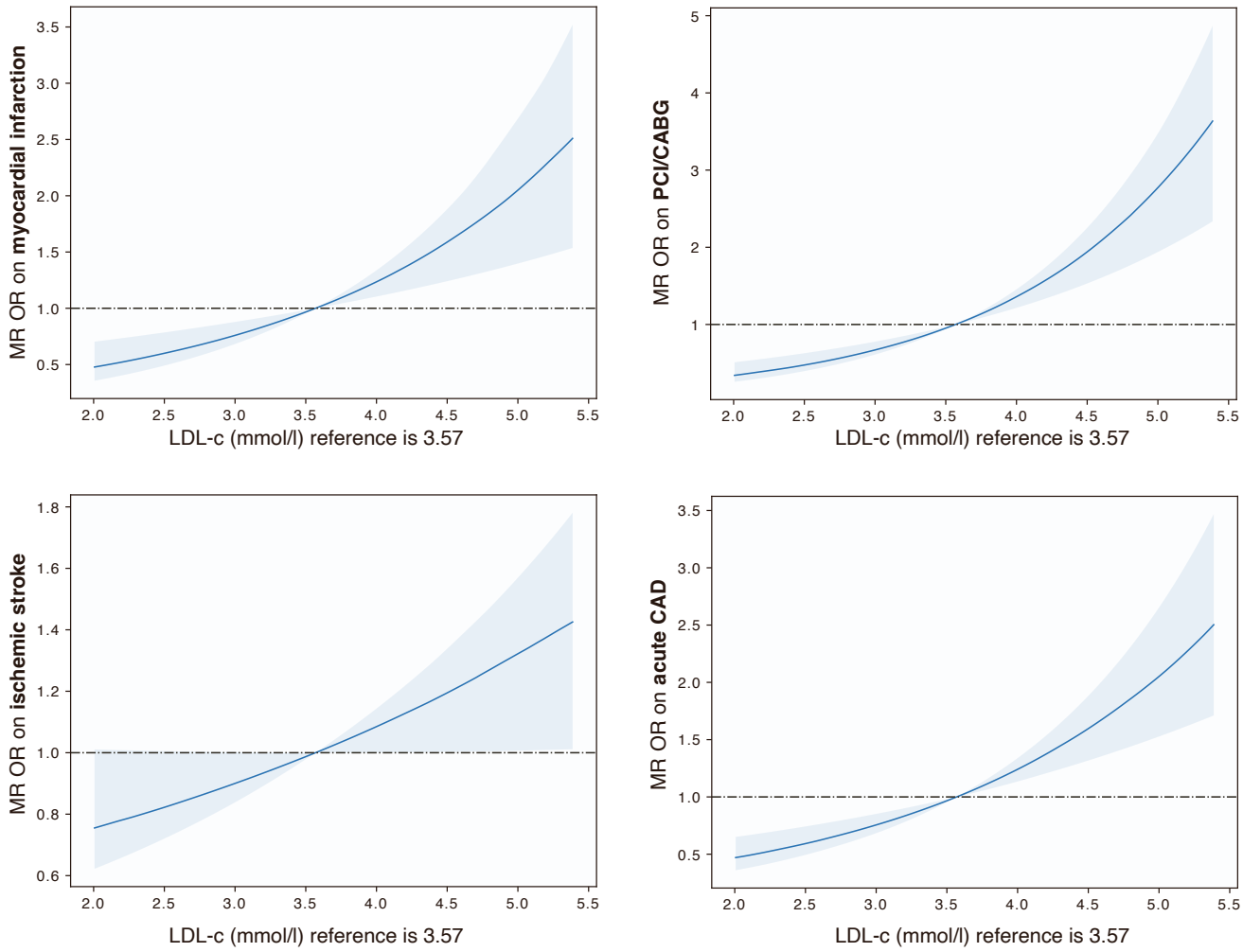


Figure S24: Quantile IV estimates of the effect of varying LDL-c levels about the mean on the odds of ischemic cardiovascular outcomes in the UK Biobank. The plot covers the central 95% of the exposure range.

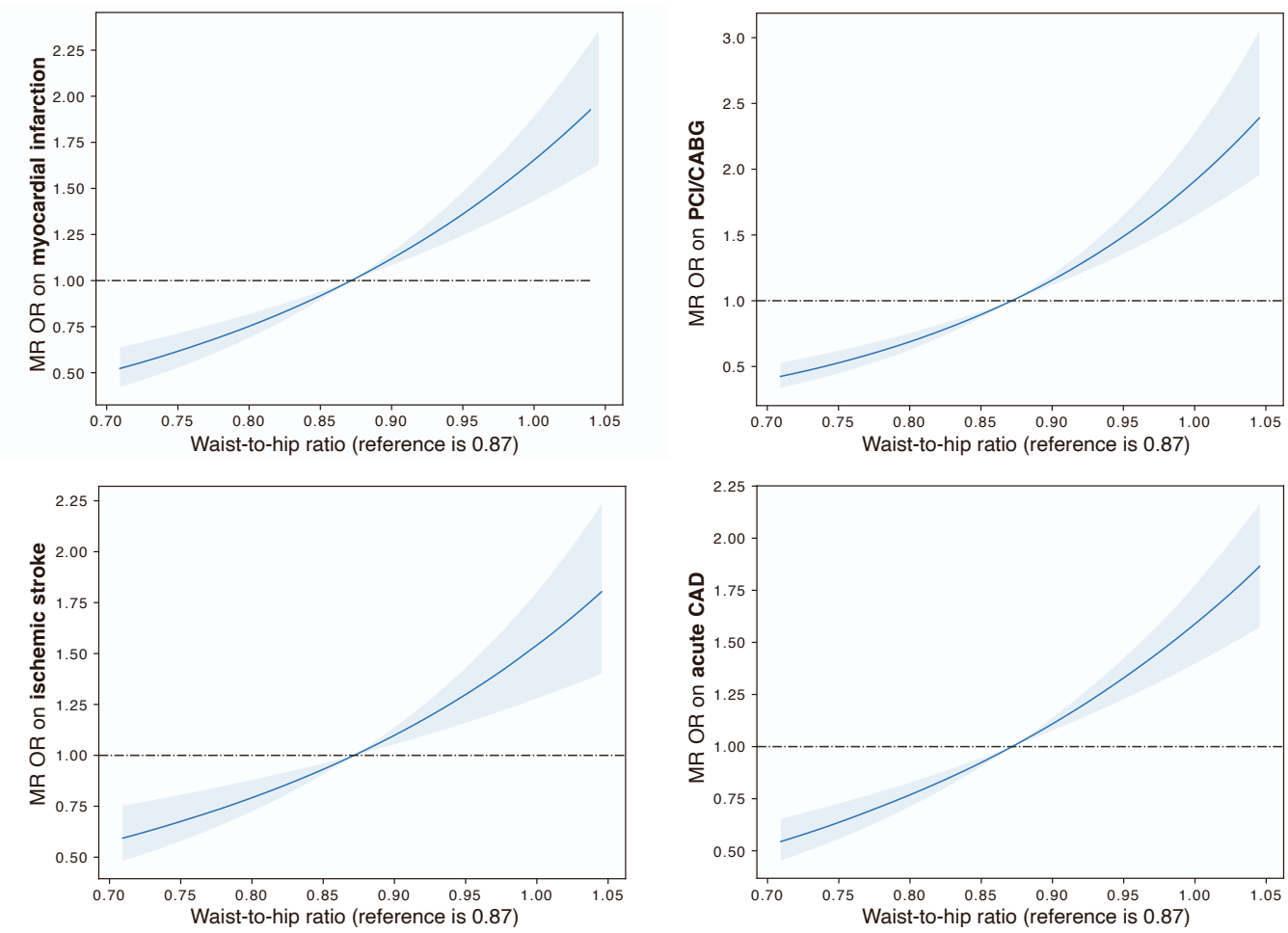


Figure S25: Quantile IV estimates of the effect of varying waist-to-hip ratio levels about the mean on the odds of ischemic cardiovascular outcomes in the UK Biobank. The plot covers the central 95% of the exposure range.

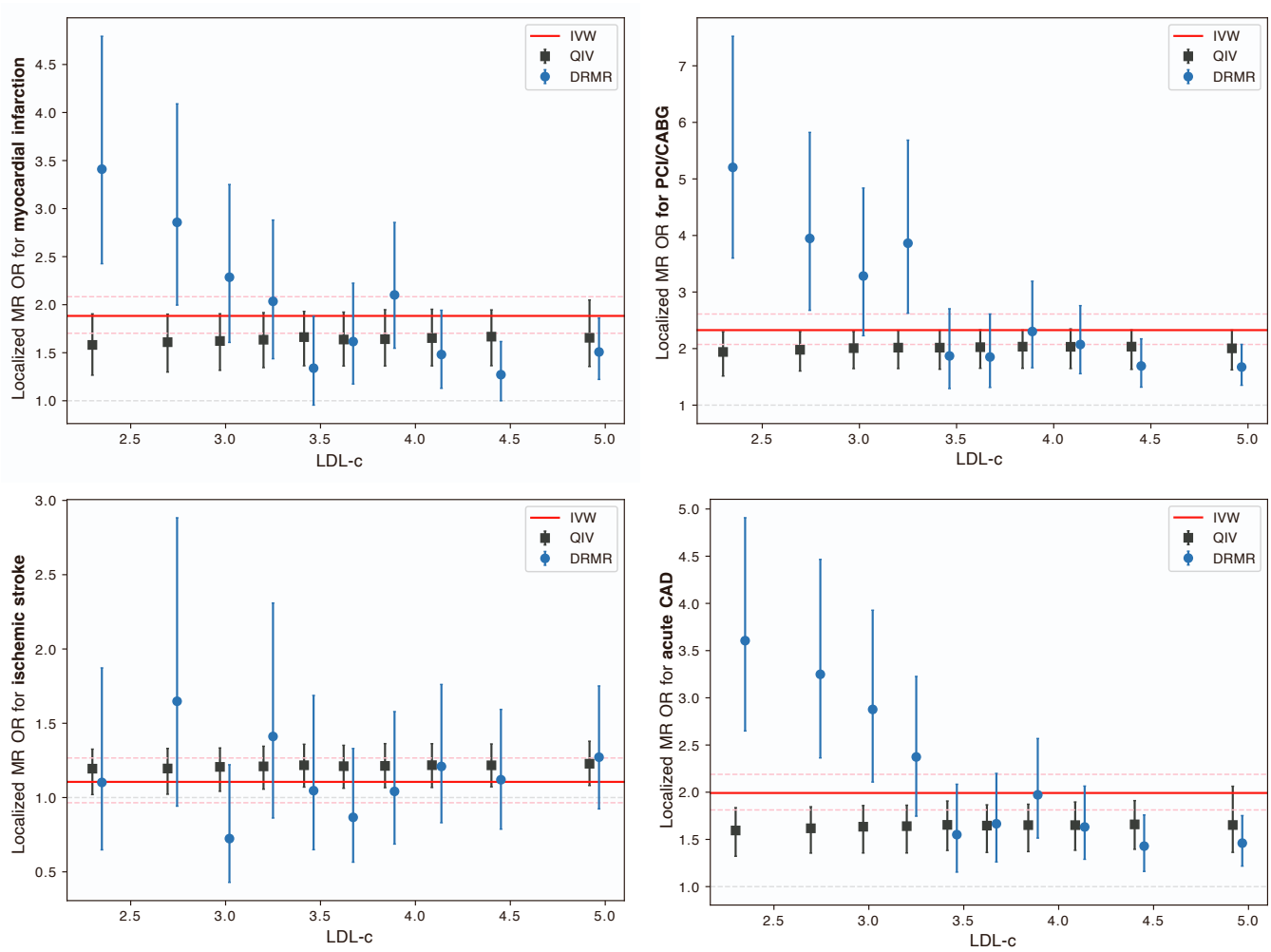


Figure S26: Predicted effect of a 1 mmol/l increase in LDL-c on cardiovascular outcomes at different values of LDL-c corresponding to the localized average causal effects estimated by the doubly-ranked and Quantile IV MR methods. The IVW estimate is provided to facilitate comparisons.

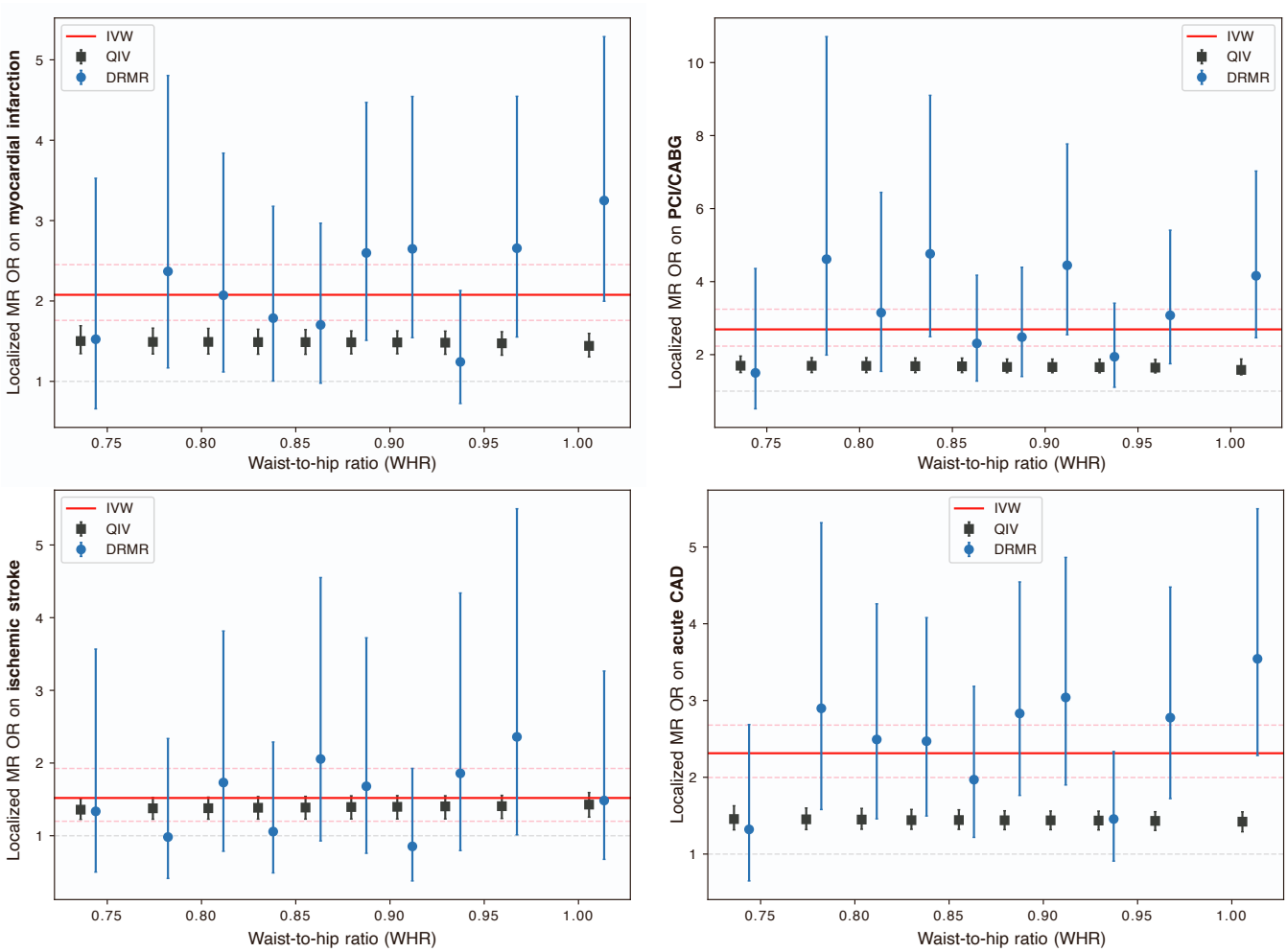


Figure S27: Predicted effect of a 0.1 unit increase in the waist-to-hip ratio on cardiovascular outcomes at different values of the waist-to-hip ratio corresponding to the localized average causal effects estimated by the doubly-ranked and Quantile IV MR methods. The IVW estimate is provided to facilitate comparisons.

B Supplemental Tables

Outcome	Code voc.	Code	Label
Myocardial infarction (MI)	ICD9	410	Acute myocardial infarction
		412	Old myocardial infarction
		411.0	Postmyocardial infarction syndrome
		429.7	Certain sequelae of myocardial infarction, not elsewhere classified
	ICD10	I21	Acute myocardial infarction
		I22	Subsequent myocardial infarction
		I23	Certain current complications following acute myocardial infarction
		I25.2	Old myocardial infarction
Percutaneous Coronary Intervention and Coronary Artery Bypass Graft (PCI/CABG)	OPCS	K40	Saphenous vein graft replacement of coronary artery
		K41	Other autograft replacement of coronary artery
		K42	Allograft replacement of coronary artery
		K43	Prosthetic replacement of coronary artery
		K44	Other replacement of coronary artery
		K45	Connection of thoracic artery to coronary artery
		K46	Other bypass of coronary artery
		K49	Transluminal balloon angioplasty of coronary artery
		K50	Other therapeutic transluminal operations on coronary artery
		K75	Percutaneous transluminal balloon angioplasty and insertion of stent into coronary artery
Acute Coronary Artery Disease (Acute CAD)	Prev. defined outcome	MI	
		PCI/CABG	
	ICD10	I20.0	Unstable angina (for this particular instance, we only consider primary cause for hospitalization or death)
Ischemic stroke	ICD9	434	Occlusion of cerebral arteries
		434.0	Cerebral thrombosis
		434.1	Cerebral embolism
		434.9	Cerebral artery occlusion, unspecified
		436	Acute, but ill-defined, cerebrovascular disease
	ICD10	I63	Cerebral infarction
		I64	Stroke, not specified as haemorrhage or infarction

Table S1: Algorithmic definition of the cardiovascular outcomes considered for the Mendelian randomization study. Unless otherwise specified, codes are considered present if they are found as the primary cause of death or as the primary or secondary cause of hospitalization in the electronic records.

Exposure dataset (circulating sclerostin)	
n	42,830
Genetic Female - n (%)	22,981 (53.7%)
Age - Mean (s.d.)	57.2 (8.1)
Outcome dataset	
n	370,218
Genetic Female - n (%)	199,656 (53.9%)
Age at baseline - Mean (s.d.)	56.8 (8.0)
Heel bone mineral density - n	211,692
Mean in g/cm^2 (s.d.)	0.54 (0.14)
Osteoporosis - n cases / n controls (%)	18,937 / 351,281 (5.4%)
Myocardial infarction - n cases / n controls (%)	19,925 / 348,722 (5.7%)
PCI/CABG - n cases / n controls (%)	17,261 / 352,957 (4.9%)
Ischemic stroke - n cases / n controls (%)	8 535 / 361,651 (2.4%)
Acute CAD - n cases / n controls (%)	33,516 / 335,167 (10.0%)

Table S2: Descriptive statistics of the datasets used for the Mendelian randomization study evaluating the effect of a reduction in circulating sclerosting levels on outcomes related to bone and cardiovascular health in the UK Biobank.

Table S3: Genetic variants used as instrumental variables in the MR study of the effect of LDL-c on cardiovascular outcomes. The subset of variants predicted to be valid by MR-LASSO and the contamination mixture method and used for one-sample analyses within the UK Biobank are labeled. The reported effect estimates are from the Global Lipid Genetics Consortium. Position are on the GRCh37 build.

See Excel file `ldl_variants_table.xlsx`

Table S4: Genetic variants used as instrumental variables in the MR study of the effect of the wais-to-hip ratio on cardiovascular outcomes. The subset of variants predicted to be valid by MR-LASSO and the contamination mixture method and used for one-sample analyses within the UK Biobank are labeled. The reported effect estimates are from the GIANT consortium. Position are on the GRCh37 build.

See Excel file `whr_variants_table.xlsx`

Exposures	
LDL-c - n	392,333
Mean in mmol/l (s.d.)	3.57 (0.87)
Waist-to-hip ratio - n	412,222
Mean (s.d.)	0.87 (0.09)
Covariates	
Age - Mean (s.d.)	56.8 (7.96)
Genetic Female - n (%)	222,187 (53.9%)
Outcomes	
Myocardial infarction - n cases / n controls (%)	22,437 / 389,785 (5.4%)
PCI/CABG - n cases / n controls (%)	19,337 / 392,885 (4.7%)
Ischemic stroke - n cases / n controls (%)	9,625 / 402,597 (2.3%)
Acute CAD - n cases / n controls (%)	30,056 / 382,166 (7.3%)

Table S5: Descriptive statistics of the datasets used for the Mendelian randomization study evaluating the effect of LDL-c and waist-to-hip ratio on outcomes related to cardiovascular health in the UK Biobank.

Parameter	Default value
n quantiles	5
Exposure network units per layer	128, 64
Outcome network units per layer	64, 32
Activation function	Gaussian Error Linear Units (GELU)
Learning rate	5×10^{-4}
Minibatch size	10,000
Maximum number of epochs	1000
Weight decay (l_2 penalty)	10^{-4}

Table S6: Default values for tunable hyperparameters of the Quantile IV estimator. When the neural network is not specified, the default value applies for both the exposure and the outcome neural network.

Simulation parameter	Sim. value	DeLIVR Mean RMSE (s.d.)	Quantile IV Mean RMSE (s.d.)	p value
Sample size	10,000	1.44 (0.30)	1.39 (0.33)	0.196
	50,000	0.98 (0.18)	0.94 (0.23)	0.037
	100,000	1.01 (0.23)	*0.86 (0.20)*	1.1×10^{-11}
Instrument strength	0.05	0.88 (0.20)	0.87 (0.22)	0.67
	0.1	1.23 (0.21)	*1.13 (0.26)*	1.7×10^{-5}
	0.5	0.86 (0.28)	*0.80 (0.15)*	8.1×10^{-4}
Causal relationship shape	Linear	0.25 (0.13)	*0.20 (0.11)*	4.4×10^{-5}
	Quadratic	0.87 (0.17)	*0.80 (0.20)*	2.9×10^{-4}
	Threshold	0.50 (0.07)	0.52 (0.10)	0.076
Confounding	-0.6	1.15 (0.22)	1.19 (0.24)	0.127
	0.3	1.06 (0.23)	*0.98 (0.23)*	2.1×10^{-4}
	0.6	0.91 (0.21)	*0.75 (0.19)*	1.3×10^{-15}
Number of instruments	2	1.03 (0.36)	*0.96 (0.35)*	3.0×10^{-3}
	10	1.04 (0.18)	*0.96 (0.23)*	5.1×10^{-5}
	100	*1.01 (0.16)*	1.30 (0.21)	2.8×10^{-35}

Table S7: Comparison between the mean root mean squared error between DeLIVR and Quantile IV across the different Mendelian randomization simulation scenarios and over the full range of the exposure distribution. The p value is from a paired t-test by simulation replicate. The italicized values highlight the smallest error when the difference is statistically significant.

ID	Pos. (Chr. 17)	Alleles (ref./coded)	Effect on sclerostin levels (95% CI)	p value
rs6416905	41,804,464	A/G	-0.059 (-0.046, -0.072)	1.49×10^{-18}
rs66838809	41,798,621	G/A	-0.097 (-0.074, -0.120)	1.82×10^{-16}

Table S8: Estimated genetic associations with circulating sclerostin levels for the top variants assigned to the two credible sets identified in the SuSiE finemapping analysis.

Phenotype	Estimator	ATE 1 s.d. reduction (95% CI)	ATE 2 s.d. reduction (95% CI)
Heel BMD	IVW	0.639 (0.554, 0.723)	1.277 (1.108, 1.446)
	PC-GMM	0.526 (0.162, 0.889)	1.051 (0.324, 1.778)
	Quantile IV	0.122 (0.117, 0.125)	0.236 (0.230, 0.243)
Osteoporosis (OR scale)	IVW	0.291 (0.213, 0.398)	0.085 (0.045, 0.158)
	PC-GMM	0.349 (0.182, 0.670)	0.122 (0.033, 0.449)
	Quantile IV	0.794 (0.612, 0.949)	0.626 (0.373, 0.857)

Table S9: Mendelian randomization estimate of a 1 s.d. or 2 s.d. reduction in sclerostin levels about the mean on heel bone mineral density and osteoporosis. Three different MR estimators are considered, and the estimates are from a two-sample MR within the UK Biobank.

Outcome	Conditioning	CATE* (95% CI)	Interaction p value
Osteoporosis	Female	0.79 (0.60, 0.95)	0.04
	Male	0.80 (0.62, 0.96)	
	Age = 48.8 (mean − 1 s.d.)	0.79 (0.61, 0.94)	0.35
	Age = 56.8 (mean)	0.79 (0.61, 0.95)	
	Age = 64.8 (mean + 1 s.d.)	0.81 (0.61, 0.98)	
Heel bone mineral density	Female	0.15 (0.02, 0.26)	2.3×10^{-22}
	Male	0.09 (-0.01, 0.18)	
	Age = 48.8 (mean − 1 s.d.)	0.12 (0.03, 0.20)	0.43
	Age = 56.8 (mean)	0.13 (0.02, 0.23)	
	Age = 64.8 (mean + 1 s.d.)	0.12 (-0.02, 0.24)	

* The CATE is reported on the odds ratio scale for osteoporosis and in standardized units (*i.e.* z-scores) for heel bone mineral density.

Table S10: Conditional average treatment effects estimated using Quantile IV for a 1 s.d. reduction in circulating sclerostin levels on osteoporosis and heel bone mineral density. The estimates are conditioned on different values of the covariates, specifically by conditioning on the genetic male vs female variable or varying the age at the mean (56.8 years), 1 s.d. below the mean (48.8 years) and 1 s.d. above the mean (64.8 years).

Estimator	ATE on heel BMD (95% CI)	p value
IVW	0.66 (0.57, 0.75)	1.7×10^{-49}
PC-GMM	0.52 (0.18, 0.86)	2.7×10^{-3}
Quantile IV	0.11 (0.03, 0.21)	6.1×10^{-3}

Table S11: Mendelian randomization estimates of the effect of a 1 s.d. reduction in circulating sclerostin levels on heel bone mineral density adjusting for possible direct effects by rs113533733. The average treatment effect on heel bone mineral density for a 1 s.d. reduction in circulating sclerostin is presented.

	All IVs ($n = 519$) MR OR (95% CI)	p value	Restricted subset MR OR (95% CI)	p value
Myocardial infarction ($n_v = 383$)				
IVW	1.60 (1.49, 1.73)	6.04×10^{-36}	1.66 (1.58, 1.75)	4.98×10^{-80}
Weighted median	1.63 (1.48, 1.79)	7.93×10^{-23}	1.63 (1.48, 1.80)	1.61×10^{-22}
MR-Egger	1.66 (1.47, 1.87)	$\leq 10^{-8}$	1.56 (1.44, 1.70)	$\leq 10^{-8}$
MR-Egger intercept		0.48		0.07
Contamination mixture (385 valids)			1.66 (1.57, 1.78)	1.67×10^{-30}
MR LASSO (422 valids)			1.62 (1.54, 1.71)	7.77×10^{-75}
Coronary artery disease ($n_v = 393$)				
IVW	1.66 (1.55, 1.78)	2.23×10^{-46}	1.81 (1.72, 1.90)	2.42E-131
Weighted median	1.59 (1.46, 1.74)	1.26×10^{-25}	1.68 (1.55, 1.83)	1.30×10^{-35}
MR-Egger	1.75 (1.57, 1.96)	$\leq 10^{-8}$	1.68 (1.56, 1.82)	$\leq 10^{-8}$
MR-Egger intercept		0.24		0.02
Contamination mixture (395 valids)			1.82 (1.71, 1.97)	3.69×10^{-41}
MR LASSO (443 valids)			1.73 (1.65, 1.81)	3.34×10^{-119}
Coronary artery disease (incl. UKB, $n_v = 363$)				
IVW	1.64 (1.54, 1.75)	3.02×10^{-50}	1.81 (1.73, 1.91)	5.58×10^{-127}
Weighted median	1.54 (1.43, 1.66)	2.21×10^{-29}	1.83 (1.68, 1.98)	1.85×10^{-46}
MR-Egger	1.71 (1.54, 1.90)	$\leq 10^{-8}$	1.61 (1.48, 1.76)	$\leq 10^{-8}$
MR-Egger intercept		0.28		1.83×10^{-3}
Contamination mixture (377 valids)			1.91 (1.78, 2.03)	6.15×10^{-50}
MR LASSO (401 valids)			1.68 (1.61, 1.75)	1.40×10^{-121}

Table S12: Two-sample MR estimates of the effect of a 1 s.d. increase in LDL-c levels on the odds of cardiovascular outcomes. The estimates are based on summary statistics from the GLGC and CARDIoGRAMplusC4D consortia. The estimates are provided for the full set of IVs and for a restricted subset containing the variants with no evidence of direct effects according to the contamination mixture and MR LASSO estimators. The number of variants used for the restricted subset for every outcome is identified as n_v.

	All IVs (n=545) MR OR (95% CI)	p value	Restricted subset MR OR (95% CI)	p value
Myocardial infarction ($n_v = 389$)				
IVW	1.33 (1.23, 1.43)	6.82×10^{-13}	1.70 (1.58, 1.83)	2.13×10^{-44}
Weighted median	1.35 (1.22, 1.51)	1.54×10^{-8}	1.63 (1.46, 1.82)	1.11×10^{-17}
MR-Egger	1.45 (1.17, 1.79)	6.62×10^{-4}	1.53 (1.23, 1.91)	1.58×10^{-4}
MR-Egger intercept		0.39		0.32
Contamination mixture (394 valids)			1.75 (1.46, 1.93)	8.47×10^{-11}
MR LASSO (484 valids)			1.34 (1.26, 1.43)	2.62×10^{-18}
Coronary artery disease ($n_v = 352$)				
IVW	1.31 (1.22, 1.42)	1.80×10^{-12}	1.70 (1.58, 1.83)	3.48×10^{-48}
Weighted median	1.36 (1.24, 1.50)	3.87×10^{-10}	1.68 (1.51, 1.87)	3.24×10^{-22}
MR-Egger	1.53 (1.24, 1.88)	7.92×10^{-5}	1.69 (1.37, 2.08)	1.07×10^{-6}
MR-Egger intercept		0.13		0.94
Contamination mixture (370 valids)			1.83 (1.66, 2.02)	2.11×10^{-12}
MR LASSO (458 valids)			1.29 (1.22, 1.38)	1.61×10^{-16}
Coronary artery disease (incl. UKB, $n_v = 376$)				
IVW	1.31 (1.22, 1.40)	1.20×10^{-14}	1.57 (1.48, 1.66)	3.08×10^{-48}
Weighted median	1.33 (1.22, 1.44)	7.75×10^{-11}	1.52 (1.38, 1.66)	2.05×10^{-18}
MR-Egger	1.47 (1.22, 1.77)	6.00×10^{-5}	1.50 (1.25, 1.80)	8.82×10^{-6}
MR-Egger intercept		0.20		0.61
Contamination mixture (381 valids)			1.60 (1.48, 1.79)	3.68×10^{-12}
MR LASSO (458 valids)			1.30 (1.23, 1.37)	1.39×10^{-21}

Table S13: Two-sample MR estimates of the effect of BMI adjusted WHR on the odds of cardiovascular outcomes. The estimates are based on summary statistics from the GIANT and CARDIoGRAMplusC4D consortia. The estimates are provided for the full set of IVs and for a restricted subset containing the variants with no evidence of direct effects according to the contamination mixture and MR LASSO estimators. The number of variants used for the restricted subset for every outcome is identified as n_v.

	MR OR per mmol/l (95% CI)	p value
Myocardial infarction		
IVW	1.88 (1.70, 2.08)	8.85×10^{-35}
Weighted median	1.83 (1.58, 2.12)	5.98×10^{-16}
MR-Egger	1.63 (1.39, 1.91)	1.86×10^{-9}
MR-Egger intercept		0.022
Contamination mixture (299 valids)	1.99 (1.84, 2.18)	6.54×10^{-56}
MR LASSO (273 valids)	1.87 (1.71, 2.04)	3.32×10^{-44}
PCI/CABG		
IVW	2.33 (2.07, 2.61)	1.12×10^{-46}
Weighted median	2.19 (1.87, 2.57)	5.49×10^{-22}
MR-Egger	1.91 (1.59, 2.29)	3.03×10^{-12}
MR-Egger intercept (283 valids)		6.15×10^{-3}
Contamination mixture (269 valids)	2.25 (2.00, 2.51)	7.36×10^{-53}
MR LASSO	2.41 (2.19, 2.65)	8.67×10^{-72}
Acute CAD		
IVW	1.99 (1.81, 2.19)	2.40×10^{-46}
Weighted median	1.95 (1.71, 2.22)	1.22×10^{-23}
MR-Egger	1.64 (1.41, 1.90)	6.73×10^{-11}
MR-Egger intercept		7.92×10^{-4}
Contamination mixture (292 valids)	2.10 (1.90, 2.28)	6.11×10^{-66}
MR LASSO (273 valids)	1.94 (1.79, 2.10)	3.71×10^{-62}
Ischemic stroke		
IVW	1.11 (0.97, 1.27)	0.147
Weighted median	1.16 (0.94, 1.44)	0.166
MR-Egger	1.09 (0.88, 1.35)	0.446
MR-Egger intercept		0.850
Contamination mixture (290 valids)	1.12 (1.00, 1.27)	0.072
MR LASSO (289 valids)	1.15 (1.02, 1.30)	0.024

Table S14: One-sample parametric MR estimates of the effect of a 1 mmol/l increase in LDL-c levels on the odds of cardiovascular outcomes in the UK Biobank. The MR estimates used the restricted subset of instrumental variables containing the variants with no evidence of direct effects according to the contamination mixture and MR LASSO models in the summary statistics based analyses.

	MR OR per 0.1 unit increase in WHR (95% CI)	p value
Myocardial infarction		
IVW	2.08 (1.76, 2.45)	6.84×10^{-18}
Weighted median	2.12 (1.68, 2.68)	2.45×10^{-10}
MR-Egger	2.57 (1.68, 3.92)	1.32×10^{-5}
MR-Egger intercept		0.288
Contamination mixture (243 valids)	2.76 (2.01, 3.45)	1.43×10^{-13}
MR LASSO (286 valids)	2.16 (1.85, 2.51)	6.87×10^{-23}
PCI/CABG		
IVW	2.69 (2.23, 3.24)	2.59×10^{-25}
Weighted median	2.87 (2.22, 3.71)	8.62×10^{-16}
MR-Egger	2.34 (1.46, 3.78)	4.59×10^{-4}
MR-Egger intercept		0.537
Contamination mixture (227 valids)	4.35 (3.18, 5.39)	6.27×10^{-21}
MR LASSO (280 valids)	2.95 (2.50, 3.49)	6.86×10^{-37}
Acute CAD		
IVW	2.31 (2.00, 2.68)	4.72×10^{-29}
Weighted median	2.40 (1.95, 2.95)	1.48×10^{-16}
MR-Egger	2.91 (2.00, 4.23)	2.33×10^{-8}
MR-Egger intercept		0.194
Contamination mixture (238 valids)	3.25 (2.67, 3.88)	2.96×10^{-25}
MR LASSO (285 valids)	2.43 (2.13, 2.79)	1.01×10^{-37}
Ischemic stroke		
IVW	1.52 (1.20, 1.93)	5.35×10^{-4}
Weighted median	1.58 (1.10, 2.25)	0.012
MR-Egger	1.62 (0.88, 2.96)	0.119
MR-Egger intercept		0.824
Contamination mixture (243 valids)	1.65 (1.18, 2.84)	5.43×10^{-3}
MR LASSO (286 valids)	1.55 (1.23, 1.95)	1.91×10^{-4}

Table S15: One-sample parametric MR estimates of the effect of a 0.10 unit increase in WHR on the odds of cardiovascular outcomes in the UK Biobank. The MR estimates used the restricted subset of instrumental variables containing the variants with no evidence of direct effects according to the contamination mixture and MR LASSO models in the summary statistics based analyses.

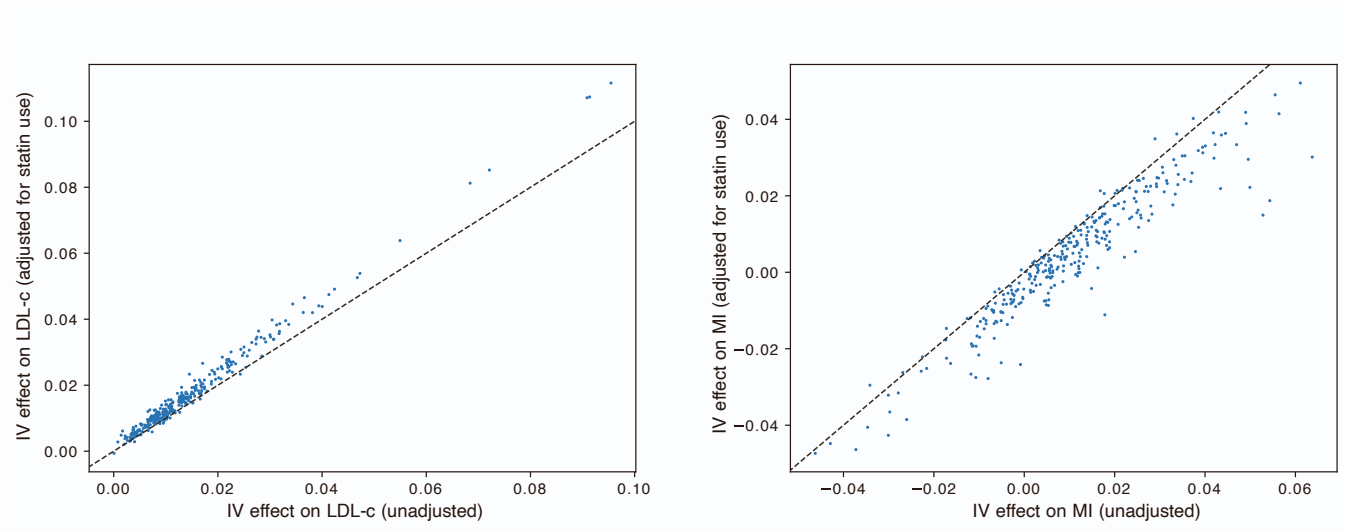
C Supplemental Note

C.1 Forward stepwise regression analysis of sclerostin pQTLs

In addition to the finemapping analysis, we also conducted a forward stepwise regression analysis where we iteratively condition on the top variant and repeat the association study. The initial association scan had identified rs9303537 as the lead variant ("T" allele $\hat{\beta} = 0.059$ 95% CI (0.045, 0.071), $p = 1.5 \times 10^{-18}$). This variant is almost perfectly correlated with the lead variant from the first credible set ($r^2 = 0.99$ in 1,000 genomes European sample). Conditioning on this top variant in the forward stepwise analysis, the variant rs66838809 was identified as the second stage top variant. The conditional association coefficient for the rs66838809 "A" allele is $\hat{\beta} = -0.078$, 95% CI (-0.091, -0.064) $p = 1.2 \times 10^{-10}$. After conditioning on these two variants, no residual signal crossed the genome-wide significance threshold. The top variant (rs75508812) had a conditional association $p = 3.3 \times 10^{-5}$. This variant passes the conservative Bonferroni correction at $\alpha = 5\%$ considering the 1,449 variants we included and can be considered significant.

C.2 Adjustment for statin use in MR of the effect of LDL-c on cardiovascular outcomes

In the MR study of the effect of LDL-c on cardiovascular, we considered effect heterogeneity between statin users and non-users. We noticed that the IVW estimate varied substantially with the addition of statin use as a covariate. For example, the IVW MR OR for a 1 mmol/l increase in LDL-c on MI was 1.88 95% CI (1.70, 2.08) in the model unadjusted for statin use and 1.23, (1.13, 1.34) in the model adjusted for statin use. This difference suggests that the IV assumptions are violated in one scenario or the other. To investigate this effect, we plotted the effect estimate of the IVs on LDL-c and MI in both models (see below).



Estimated effect of genetic variants on LDL-c levels and MI in models adjusted and unadjusted for statin use in the UK Biobank. Every point represents a genetic variant used as an IV in the MR study of LDL-c levels. The effects are all expressed with respect to the LDL-c increasing allele. The diagonal line represents the identity.

We observe that adjustment for statin use increases the effect estimates for LDL-c and reduces the MI effect estimates on average resulting in deflated ratio estimates in the model adjusted for statin use. We could hypothesize that this effect is due to a statistical interaction where statin use modifies the effect of genetic variants on LDL-c. For instance, that the true causal model is:

$$\text{LDL-c} = \beta G + \beta_i \cdot G \cdot \text{statin} + U + \epsilon_x$$

$$Y = \theta \cdot \text{LDL-c} + U + \epsilon_y$$

Then, the regressions of LDL-c on the genetic variant ($\beta_{\text{LDL-c}|G}$) and of the outcome on the genetic variant ($\beta_{Y|G}$) would be:

$$\beta_{\text{LDL-c}|G} = \beta + \beta_i \cdot p$$

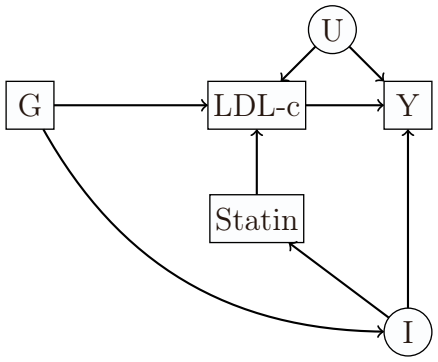
$$\beta_{Y|G} = \theta \cdot (\beta + \beta_i \cdot p)$$

$$\implies \beta_{Y|G} / \beta_{\text{LDL-c}|G} = \theta$$

With $p = P(\text{statin} = 1)$ the prevalence of statin use. This shows that the Wald estimator (and,

by extension, the IVW estimator) is unbiased in the presence of unaccounted for genetic effect heterogeneity when the true causal effect is linear.

We hypothesize that the difference between MR estimates of the effect of LDL-c on MI with and without adjustment for statin use are due to a more complex causal structure. For example, a genetic association with propensity for statin use (*indication bias*, I), which can be independently associated with cardiovascular outcomes, could partially explain the observed effects (see tentative causal graph below):



Directed acyclic graph representing a causal structure where adjustment for statin use could result in changes in the genetic effect estimates on the exposure and outcome. The variable labeled "I" represents latent indication bias. It can be interpreted as an unmeasured clinical variable physicians may use to decide to initiate statin pharmacotherapy and that may also influence the outcome. Under that DAG, adjusting for statin use reduces bias in the genetic effect estimates by blocking the indirect path mediated by the latent indication.

D Supplemental Methods

D.1 Parametric MR estimators

The inverse variance weighted MR estimator is a weighted average of the ratio (or Wald) estimators where the weights are the precision of the estimates. This method assumes that all variants are valid IVs. The MR-Egger estimator relaxes the exclusion restriction assumption and relies on the Instrument Strength Independent of Direct Effects (InSIDE) assumption instead. This assumption states that the direct (or pleiotropic) effects should be independent from the IV–exposure effects. The median estimator uses the median of the ratio estimators as the causal effect estimate and is a

consistent estimator of the causal effect if at least 50% of the IVs are valid. The robust counterpart of these estimates use strategies from robust regression to reduce the impact of extreme data points on the estimate (see [62]). The penalized counterparts add a penalty based on the Cochran's Q statistic as a measure of heterogeneity in the causal estimates. All of these estimators are implemented in the MendelianRandomization R package and are documented therein.

In addition to those estimators, we used the contamination mixture and the MR LASSO methods that provide variant level statistical evidence of violations of the IV assumptions. The contamination mixture algorithm is a two-component mixture of gaussians with a component centered at the causal effect and a second component centered at zero accounting for the variance from invalid IVs [47]. The MR LASSO introduces a LASSO (ℓ_1 norm) penalty on the intercept term representing variant direct effects. Because of the sparsity inducing property of the LASSO penalty, the variants with non-zero intercepts are considered to have pleiotropic effects and the others to be valid IVs [46].

D.2 Evaluation of Quantile IV with simulated binary outcomes

To evaluate the performance of Quantile IV with binary outcomes, we discretized the continous outcome corresponding to the simulation scenario with $n = 50,000$, 10% of the variance in the exposure explained by the IV, a quadratic relationship, a correlation of 0.3 between the errors of the exposure and outcome and 10 simulated IVs. The binary variable was defined using an indicator variable to achieve a prevalence of either 5% of 30%. Specifically, we computed the 70^{th} and 95^{th} percentiles of the continuous outcome and set individuals as "cases" if their realized value of the outcome was greater than the threshold. Because the total sample size for the simulation model is 50,000, the simulated number of cases was either 2500 of 15,000. We repeated the discretization procedure for the 200 simulation replicated and fit Quantile IV once for each replicate. By comparison, the number of cases included in our real data analysis ranged between 5 535 for ischemic stroke and 33 516 for acute CAD (Table S2).

We now describe the causal log odds used as the target in the calculation of the root mean

squared error. We denote the latent continuous outcome as Y^* and the discretized outcome $Y := \mathbb{I}(Y^* \geq \tau_{1-p})$ where p is the simulated prevalence and τ represents quantiles of Y^*. From Equation (8), we have:

$$\begin{aligned} P(Y = 1 \mid \text{do(X=x)}) &= P(Y^* \geq \tau_{1-p} \mid \text{do}(X = x)) \\ &= P(f(X) + \epsilon_y \geq \tau_{1-p} \mid \text{do}(X = x)) \\ &= P(\epsilon_y \geq \tau_{1-p} - f(x)) \\ &= 1 - \Phi(\tau_{1-p} - f(x)) \end{aligned}$$

Where Φ is the CDF of the standard normal distribution, the marginal distribution of ϵ_y in our simulation model. The causal log odds is

$$\ln\left(\frac{P(Y = 1 \mid \text{do(X=x)})}{1 - P(Y = 1 \mid \text{do(X=x)})}\right)$$

Using this simulation setup, we compare the RMSE between the estimated causal log odds and the true causal log odds for the compared models. For the binary trait simulation analysis, we compare Quantile IV to the 2-stage control function estimator which is less biased than naive two-stage procedure when estimating the log odds ratio [55]. We also compare to DeLIVR [14] using the same strategy as for Quantile IV of modeling the outcome linearly on the log odds scale. We excluded DeepIV from this comparison because the optimization of the estimator relies on a modified mean squared error loss that is incompatible with binary outcomes [20]. This custom loss was designed to avoid bias in the gradient computation.

D.3 Simulation scenario with causal effect heterogeneity and complex confounding

The main simulation study aimed to assess the applicability of Quantile IV in realistic MR scenarios. For this reason, it was designed to be conservative in the complexity of the data generating process. Here, we consider a different structure aimed at illustrating the advantages of using nonparametric

IV estimators. For this example, we compare DeLIVR and Quantile IV because the strong simulated causal effect heterogeneity warrants estimation of the CATE (i.e. methods estimating the ATE would not be competitive).

We consider a scenario with a unique continuous instrumental variable Z (e.g. a genetic risk score), a normally distributed latent confounder of the exposure–outcome relationship (U) and a uniformly distributed effect modifier (M) of the causal relationship between the exposure and outcome taking values between -1 and 1. The exogenous random errors for the exposure (ϵ_x) are sampled from the exponential distribution.

Drawing inspiration from [16], we introduce additional model complexity by inducing a correlation between the instrumental variable effect on the exposure (α) and the errors of the outcome (ϵ_y). This implies that there is a latent variable with an effect on the outcome that also influences the effect of the IV on the exposure. Specifically, we set

$$\begin{bmatrix} \alpha \\ \epsilon_y \end{bmatrix} \sim \mathcal{N}_2 \left(\begin{bmatrix} 1.25 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.04 & 0.1 \\ 0.1 & 1 \end{bmatrix} \right) \tag{10}$$

Where α is the coefficient for the effect of the genetic IV on the exposure.

We introduce additional complexity in the form of effect modification by the observed variable M and interactions with the latent confounder in the exposure model:

$$X = \alpha Z - 1.25 \cdot Z \cdot M + Z \cdot U + U + \epsilon_x$$

The outcome is defined as

$$Y = f(X, M) - 0.8U + 0.1U^2 + \epsilon_y \tag{11}$$

$$f(X, M) := M \left(\exp(-x^2) + \frac{x - m}{2} \right) \tag{12}$$

The causal function f was selected because the change in the CATE with respect to X is

globally decreasing at low values of the effect modifier (M), CATE $= 0$ when $M = 0$ and it becomes globally increasing at high values of M while being nonlinear (Figure S9).

D.4 Genetic quality control

We excluded variants or individuals with a missing rate $> 2\%$. The genetic and self-reported sex variables were compared to ensure concordance between genetic and self-reported sex and individuals with discrepancies or sex chromosome aneuploidies were excluded from the analysis dataset. To mitigate bias due to population stratification, we restricted our analysis to individuals of European ancestry because they represent the largest genetically homogeneous subgroup in the UK Biobank and excluded participants that fell outside of a manually defined region on the principal component analysis plot. Related individuals were excluded based on a kinship coefficient cutoff of 0.0884, corresponding to a relationship no closer than the 3rd degree. A total of 413,138 individuals remained after this quality control process.

D.5 Mendelian randomization exposure and outcome definitions

The exposure for the MR study is circulating sclerostin levels as measured using the Olink Explore high throughput proteomics platform. This platform uses the proximity extension assay technology where antibodies bind the target protein and allow hybridization of complementary affixed oligonucleotides which are amplified and used to quantify protein levels. Sclerostin is assayed on this platform (protein #2527) and we extracted measurements in the normalized protein expression (NPX) format at the baseline instance. The NPX values are provided by the UK Biobank following the Olink QC and quantification process and have an approximately logarithmic interpretation (see UK Biobank resources #4654 to #4658 for additional information on the quality control and quantification) [6]. We visually confirmed that the extracted sclerostin levels had an approximately normal distribution with no outliers. A total of 42,830 individuals had available sclerostin measurements at the baseline visit and were included in our dataset accounting for our genetic quality control (Table S2).

For the MR study, we considered heel bone mineral density, osteoporosis, MI, PCI/CABG,

acute CAD and ischemic stroke as outcomes. We used data from ultrasound bone densitometry which provides an estimate of heel bone mineral density in g/cm^2. We extracted measurements from the initial assessment visit and used the average measurement when multiple values were present for a same individual. We standardized the measurements to have a mean of 0 and standard deviation (s.d.) of 1. The reported effects in the s.d. scale can be converted back to g/cm^2 by multiplying by 0.14 and represent changes about the mean which is of 0.54 g/cm^2.

We defined the osteoporosis variable as self-reported osteoporosis (UK Biobank variable #20002) or from ICD10 codes M80 (osteoporosis with pathological fracture), M81 (osteoporosis without pathological fracture), M84.4 (pathological fracture, not elsewhere classified) or M85.9 (disorder of bone density and structure, unspecified) in the hospitalization records. We relied on hospitalization and death records to capture acute cardiovascular events using the definitions we previously developed in collaboration with physician researchers at the Montreal Heart Institute (Table S1). Our previous definitions did not include ischemic stroke, so we relied on the definition from the UK Biobank Outcome Adjudication Group based on self-reported ischemic stroke and the codes reported in Table S1). In addition to the coding of cases based on the diagnostic codes, we excluded individuals with self-reported MI from the controls of the MI and acute CAD variable and the individuals with self-reported ischemic stroke from the controls of the ischemic stroke variable.

D.6 *Post hoc* simulation analysis of linear effect underestimation by Quantile IV

The causal effect estimate of a 1 s.d. reduction in sclerostin levels on heel BMD by Quantile IV is much smaller than the estimate from linear models. To evaluate whether Quantile IV systematically underestimates linear effects under our conditions, we conducted a simulation based evaluation using a semi-synthetic data approach.

For every simulation replicate, we sampled with replacement 413,048 real genotypes for the two instruments (rs6416905 and rs66838809) from the UK Biobank genetic dataset. This ensured that the distribution of the genetic variants was preserved in our simulation in terms of frequency

and LD. We then simulated an exposure variable with unit variance while fixing the genetic effects to the observed allelic effects for the two variants. We used the point estimates from the joint linear regression model adjusted for age, sex and ancestry principal components (coefficient of 0.049 for rs6416905 "A" allele and coefficient of -0.076 for rs66838809 "A" allele). We simulated a standard normal outcome while fixing the linear causal effect of the exposure to have a coefficient of -0.526 corresponding to the point estimate from PC-GMM in our real data analysis. The latent confounder was modeled using the same correlated errors strategy as for the original MR simulation study with a correlation between the errors of -0.4. To mimic our study design, we partitioned our dataset into a stage 1 sample ($n_1 = 42,830$) used to train the model relating the IV to the exposure and a stage 2 sample ($n_2 = 370,218$) used to train the model relating the IV to the outcome. We repeated the simulation procedure for 200 replicates.

After fitting Quantile IV, we obtained a linear approximation of the IV regression function by using linear regression to predict the estimator's output. This allowed us to report a mean slope estimate that can be easily compared to the true slope parameter (-0.526). We additionally report the raw nonlinear estimates to show that Quantile IV did not estimate functions that substantially deviated from a linear effect.